Availability and Latency of World Wide Web Information Servers

Charles L. Viles and James C. French
University of Virginia

ABSTRACT: During a 90-day period in 1994, we measured the availability and connection latency of HTTP (hypertext transfer protocol) information servers. These measurements were made from a site in the Eastern United States. The list of servers included 189 servers from Europe and 324 servers from North America. Our measurements indicate that on average, 5.0 percent of North American servers and 5.4 percent of European servers were unavailable from the measurement site on any given day. As seen from the measurement site, the day-to-day variation in availability was much greater for the European servers than for the North American servers. The measurements also show a wide variation in availability for individual information servers. For example, more than 80 percent of all North American servers were available at least 95 percent of the time, but 5 percent of the servers were available less than 80 percent of the time. The pattern of unavailability suggests a strong correlation between unavailability and geographic location. Median connection latency from the measurement site was in the 0.2–0.5 s range to other North American sites and in the 0.4–2.5 s range to European sites, depending upon the day of the week. Latencies were much more variable to Europe than to North America. The magnitude of the latencies suggest the addition of an

MGET method to HTTP to help alleviate large TCP set-up times associated with the retrieval of web pages with embedded images. The data show that 97 percent and 99 percent of all successful connections from the measurement site to Europe and North America respectively were made within the first 10 s. This suggests the establishment of client-side time-out intervals much shorter than those used for normal TCP connection establishment.

1. Introduction

The World Wide Web (WWW or W3 or Web) [Berners-Lee et al. 1994, Berners-Lee 1994] hypertext paradigm, combined with the availability of good public-domain server and browsing software, has enabled a true explosion of information resources. By many accounts, both anecdotal and objective, the size of the Web, in terms of number of servers [Beebee 1994, Gray 1994], number of resources [Fletcher 1994, McBryan 1994], and network traffic [Merit 1994] has increased exponentially since the Web's conception at CERN in the early 1990's.

Synonymous with the Web is an information transfer protocol (Hypertext Transfer Protocol or HTTP [Berners-Lee 1993.3]), a mark-up language with which to compose documents (HyperText Markup Language or HTML [Berners-Lee and Connolly 1993]), and a method to address information resources (Universal Resource Identifiers or URIs [Berners-Lee 1993.5]). Complementary technology to allow extensible typing of information resources (MIME types [Borenstein 1992]) has also been instrumental in the growth and popularity of the Web. It seems apparent that the World Wide Web and its technology is more than a passing fancy and represents a fundamental change in the way information can be provided and used on the Internet.

If we imagine ideal performance on the Web, two measures of interest are availability and latency. Ideally, we would like every site to be 100 percent available and the latency between the selection of a hyperlink and the appearance of the information that link represents to be undetectable. For exposition's sake we can talk of the *100–100 Web*: 100 percent availability for all servers and 100 millisecond latency to every server. Anything less than 100 msec is perceived as instantaneous by most humans and is a design criteria in the current development of HTTP [Berners-Lee 1993.3]. In Section 3, we define exactly what we mean by latency and availability.

While 100 percent availability of individual servers is a realistic goal, 100 percent availability of *all* servers is not. Failures happen. 100 millisecond user-level latencies are unlikely in the general case, given current physical networks and protocols. In this study, we attempt to identify how far the actual Web is from the ideal 100–100 Web. We characterize the latency and availability of a large group (> 500) of Web servers distributed throughout the world, but concentrated in Europe and North America, *as measured from an Eastern United States site*[1] with typical Internet connectivity. The important contributions of this paper include

1. We emphasize this phrase to underscore the fact that some of the results reported here must be interpreted with respect to the measurement site.

- A characterization of the typical connection latencies to European and North American Web servers from a North American site.
- A characterization of the availability of 'The Web' over an extended (90-day) time period, as well as the availability of particular servers over the same period.
- The observation that relatively short (10–20 s) client-side time-out intervals would significantly improve worst-case response times on expansion of a hyperlink *without* perceptibly affecting availability.
- Presentation of convincing empirical evidence that the 95–500 Web is a more realistic goal than the 100–100 Web.
- A call for the addition of a new method definition for HTTP analogous to the MGET available in some FTP client implementations. This method would be very useful for efficient retrieval of Web pages that currently require multiple GETs to assemble in their entirety.

In addition, we feel the information presented here will be useful in the development of new server and client software and in the design of distributed information retrieval systems, particularly those that require exhaustive search of all participating sites.

The organization of the paper is as follows. We describe some of the other Internet monitoring activities we are aware of in Section 2, paying particular attention to those involving the Web. In Section 3, we describe our experimental methodology. In Section 4 we present our results, followed by a discussion in Section 5. We conclude with a short synopsis of our contributions in Section 6.

2. *Related Work*

2.1. *Measurement Activities on the World Wide Web*

We know of only a few measurement activities on the Web as of the writing of this paper. Most involve the use of automated programs that crawl over the Web in search of new documents or servers.[2] Some of these activities have as a stated

2. These programs are often called web robots or spiders. A description of currently known robots is maintained by Martijn Koster at http://www.nexor.co.uk/mak/doc/robots/robots.html. A voluntary standard for the use and construction of web robots is also available at the same site. The standard attempts to ensure that robots 'do no harm' to individual servers or the Web itself.

goal the estimation of the size of the Web either in number of documents or number of servers [Beebee 1994, Gray 1994], but most are used to collect documents for indexing and searching purposes [Eichmann 1994, Fletcher 1994, Mauldin and Leavitt 1994, McBryan 1994]. Web size estimates are a side-effect of the collection effort. Other robots are used to aid in the maintenance of hypertext infostructures [Fielding 1994]. Among other things, the WWWMM robot [Tronche 1994.32] was built to estimate the latency of single and multi-link paths in the Web. Latency was defined as the time it took from the document request until the receipt of the first byte of the document [Tronche 1994.33]. Our definition includes only the time to set up the connection and send the request. No data or summary of results from WWWMM was available as of the writing of this paper.

Padmanabhan and Mogul [1994] propose modifications to the evolving Hypertext Transfer Protocol to improve user-level latency. Braun and Claffy [1994] characterized HTTP traffic at a popular server and examined how caching heavily accessed documents at sites closer to the requestor would improve bandwidth at the main server.

2.2. Other Internet Measurement Activities

A variety of performance data about the NSFNET backbone is collected each month, summaries of which are made publicly available at ftp://nic.merit.edu/nsfnet/statistics. This data includes one-way delays between backbone nodes and traffic breakdowns by port, country, network and day. Using this data Claffy et al. [1993] described general trends in the NSFNETs T1 backbone up until its retirement in late 1992.

The Internet Domain Survey [Lottor 1992, Lottor 1994], is a long-running activity to estimate the size of the Internet by counting the number of hosts and networks in the Domain Name System.

Each month, a flow analysis of traffic on Usenet news groups is published [DEC Network Systems Laboratory 1994]. This information includes statistics on the size of articles, size of news sites, traffic distribution by newsgroup, and the top news sites by traffic volume.

Using packet traces, Caceres et al. [1991] characterized the attributes of both interactive and bulk-transfer applications in wide-area TCP/IP conversations. No latency or failure measurements were made. Danzig et al. [1992] found that UDP-based DNS traffic consumed considerably more network bandwidth than was strictly necessary, and attributed the excess traffic to buggy implementations of name servers and resolvers. In public FTP archives, Maffeis [1993] found that access to files generally exhibited high locality, with a few files being accessed most

of the time. Most files (99 percent) were less than 1MB in size and transfers of files less than 100K made up 80 percent of total file retrievals.

Much has been done in examining end-to-end delays on the Internet [Agrawala and Sanghi 1992, Bolot 1993, Mills 1983, Sanghi et al. 1993] but most of this work has been at the packet level, not at the connection level.

3. Description of Monitoring Activities

In this section, we first supply some general background information on the Web. We leave the details to the cited works. We then define the measurements of interest, followed by the description of our experiments.

3.1. Background

3.1.1. World Wide Web (WWW or W3 or Web)

The World Wide Web provides a hypertext environment where users are able to share information regardless of its physical location. Links in a hypertext document may lead to many types of information resources located throughout the Internet. The Web concept was developed at CERN, the European Particle Physics Laboratory, but was quickly embraced globally. More detailed descriptions of the Web, its design goals, and its current and potential capabilities are available elsewhere [Berners-Lee et al. 1994, Berners-Lee 1994].

3.1.2. HTTP

The Hypertext Transfer Protocol (HTTP) is an evolving protocol for the exchange of hypertext information over wide area networks [Berners-Lee 1993.3] and is the native client/server protocol for the World Wide Web. HTTP is an application level protocol that runs on top of the layer four protocol (nominally, TCP). HTTP treats documents as objects and defines a set of methods that can be invoked on the objects. These methods support search and retrieval and are designed to be extensible to encompass other functionality, including update and annotation. HTTP is stateless and is designed to be as lightweight as possible in order to support short response times. There are 14 methods in the current proposed standard but only a subset of these are implemented in most servers: GET, POST, PUT, DELETE, and HEAD. Because HTTP is stateless, each method request is handled as a separate transaction. The server terminates the conversation with the client after performing each method. The result is a separate TCP connection for each method request.

3.2. Measurement Definitions

3.2.1. Connection

In our experiments, we attempted to contact a large set of HTTP servers. Each contact attempt was a TCP stream socket connection to the port where the server was listening (generally, but not always, port 80). The main loop of the measurement program is depicted in pseudo code in Figure 1.

The `Resolve_Address` routine includes resolution of the hostname through DNS, and the `TCP_Connect` routine includes the building of the TCP stream socket. To minimize server-side system delays, a nonsense method request was sent to the server instead of the normal HTTP "GET" [Berners-Lee 1993.3]. A known HTTP method would often require the server to go to disk to resolve the request, thus introducing additional system delays into the latency measurement. The method request that was sent was

```
''TESTCOMMAND ForInfo-->
                http://uvacs.cs.virginia.edu/~clv2m/webtest.html''.
```

The argument to TESTCOMMAND was the URI for a document describing the purpose of the experiment. We found that this forestalled a flurry of e-mail from curious server administrators.

For each contact to an HTTP server, two measurements were taken: the resolution of the connection attempt (successful or unsuccessful), and the time to either establish the connection or get a failure. We were able to distinguish DNS failures from those due to other problems, but we were unable to further distinguish those failures in the "other" category. From this data, we present two kinds of performance metrics, *availability* and *latency*.

```
While more hosts
begin
        read host and port;
        start_timer;
        address = Resolve_Address(host);
        successful = TCP_Connect (address, port);
        send_Nonsense_Request();
        stop_timer;
        record success and timer value;
        clean-up connection;
end loop
```

Figure 1. Pseudo-code for the main loop in the measurement program.

3.2.2. Availability

We define two types of availability, *WAvail* and *SAvail*.

$$WAvail(t) = \frac{\text{Number Successful Contacts at } t}{\text{Total Contacts Attempted } t}$$

WAvail is a measure of the proportion of Web servers that are active and serving information at any particular time. Ideally, measurements of *WAvail* should be made by contacting all servers simultaneously. In practice, such a method would be ill-advised, since the measurement activity would likely bias the measurement itself. Sequential contacts, as were made in this study, allow only one pending contact and thus do not cause congestion at the measurement site. In this work, all measurements of *WAvail* were calculated by contacting the target set of HTTP servers over a short time period (about 30 minutes) rather than instantaneously.

We define server availability in terms of a particular server and a time period rather than a particular time.

$$SAvail(s, t_1, t_2) = \frac{\text{Number of successful contacts to server } s \text{ between } t_1 \text{ and } t_2}{\text{Total number of contacts to server } s \text{ between } t_1 \text{ and } t_2}$$

In this paper, we present a single estimate of *SAvail* for each server, with t_1 and t_2 set to be the start and stop days of the long-term measurement period.

3.2.3. Latency

Latency is the time it takes to resolve a logical name, set up a connection, and transmit a request to an information server, given the server's logical name. We consider this latency to be a lower bound on the wait that interactive users experience when they click on a hyperlink and wait for the first byte of a document to appear. This is because our latency does not include the time spent waiting for a server to fetch and return the requested document.

3.3. The Experiments

3.3.1. The List of HTTP Servers

In our experiments, we repeatedly contacted a set of 542 HTTP servers. This set of servers was obtained from a published list of World Wide Web servers available on the Web. This list was generated by a Web-walking automaton called the World Wide Web Wanderer [Gray 1994]. The automaton starts with a collection of known documents and conducts a depth-first search of the Web. The stated goal of the automaton is to estimate the total size of the Web in terms of the number of servers. It only expands on HTTP links, ignoring links like ftp, wais, telnet, and

others. In January of 1994, this list contained 623 sites (in October 1994, 4600 sites). We culled the January list down to 542 sites, eliminating all sites with corrupted domain names or that were otherwise inaccessible. When the experiment started, all 542 sites were up and available. For data analysis, we split the list of sites into three groups, Europe (EU), North America (NA), and "Other." The European group (189 servers) contained all servers with European country codes in their domain names. This list was verified using packet traces [Sun 1993.31]. The North American list (324 servers) comprised all .edu, .mil, .org, and .com sites plus all Canadian sites. The "Other" group (29 servers) contained a miscellaneous group of servers from the Far East, Australia, and Central and South America.

3.3.2. Measurement Period

Measurements took place over two time periods. We measured *WAvail* and latency every 2 hours for the first 7 days and every 4 hours for the next 5 days over the period from February 14, 1994 to February 25, 1994. For the 90-day period from March 1, 1994 to May 29, 1994, we measured *WAvail* and latency twice a day at approximately 11 AM and 11 PM Eastern Time. For both time periods, *WAvail* was measured for the set of servers and latency was measured for each server. At the conclusion of the longer time period, we also were able to measure *SAvail* for the entire 90-day period for each server. Due to local problems at the measurement site, on two days over the longer time period, only a single measurement was made.

At any particular measurement time, a *run* though all servers consisted of contacting each server on the list in sequence. A run normally took about 30 minutes. The list of servers was randomized before each run to ensure that the connection attempt to any particular server occurred at a slightly different time on each run. This was to avoid regularly scheduled activities on the server side that might bias the measurements for that server. For example, we wanted to avoid repeatedly contacting a machine in the middle of running its daily backups.

3.3.3. Measurement Environment

For all of the 12-day period and the first 15 days of the 90-day period, measurements were made from a SPARCstation IPC running SunOS 4.3.1 with 32 MB of memory and an attached disk. For the last 75 days of the 90-day period, measurements were made from a different but similarly configured SPARCstation IPC. Network connections within the University of Virginia are fiber-optic based, with a T1 (1.544 mBit/sec) link to the wider Internet. Non-local traffic generally takes three T1 hops and about 15 msec to get to the NSFNET T3 backbone. The University of Virginia runs three DNS servers whose caches are purged each night.

4. Results

In the results that follow, we present the time of day in terms of the measurement site, not the destination server. Thus, a 'midnight' measurement for a European server is actually four to six hours later in terms of the server's local time. Because of the small number of servers in the Other group, results are presented only for the Europe and North America groups.

It is very important to realize that all results presented below are from measurements made at an Eastern United States site. Any interpretation of the results must be made accordingly. To avoid verbosity and misinterpretation, we will use the following conventions. When speaking of measurements made of European servers, we will use "NA-to-EU". Similarly, for North American server measurements, we will use "NA-to-NA".

4.1. Availability

4.1.1. First Measurement Period: 12-day Intensive

In Figure 2 we present measurements of WAvail for the 12-day period from February 14 to February 25. NA-to-NA is at the top, and NA-to-EU is at the bottom. Both NA-to-EU and NA-to-NA measurements show daily minima and maxima in availability, occurring more or less at the same local server time. Minimal availability normally occurred in the early to mid-morning for North America and around midnight for Europe. Availability was highest in the evening in North America and in the early afternoon in Europe. This behavior is not surprising, since servers going down over night might not get re-booted until their administrators arrive for work the following day. Morning is also the time when many administrators bring servers down for configuration changes and other maintenance tasks. Web availability was roughly 95 percent over the life of the experiment.

Figure 2 also shows a slight weekend drop in availability for both groups of servers. It is difficult to say whether this is a consistent phenomenon, since only one weekend period is shown.

4.1.2. Second Measurement Period: 90-day Long Term

In Figures 3 and 4 we present the results of every single connection attempt made to every server in North America and Europe respectively in the second measurement period, March 1 to May 29, 1994. Each row in the images represents the results of the 178 connection attempts made to a single server. The attempts are presented in chronological order from left to right. There is

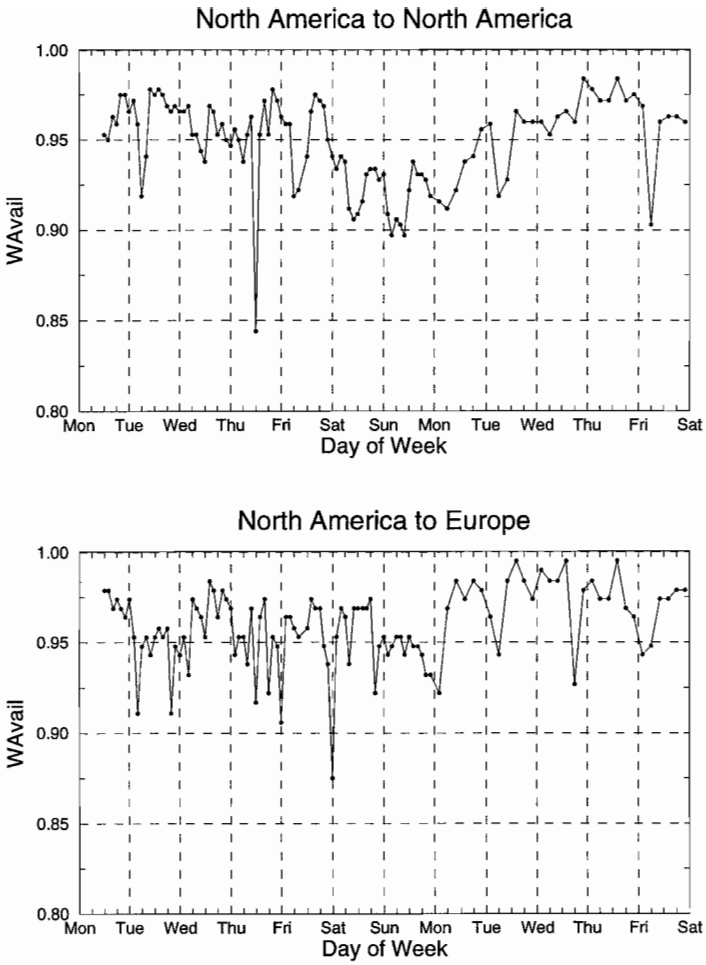

Figure 2. Web availability from the measurement site over 12 days. Vertical grid lines represent the start of the day.

no vertical relationship in Figures 3 and 4, as the servers are presented in alphabetical order of their domain names. White areas in the figures represent times that a server was down and black areas mean a server was up. A white line indicates that a server was down for consecutive connection attempts. The length of the line then represents the number of consecutive attempts a server was down.

The total area in black(white) for either of the images represents the overall availability (down-time) of the group of servers over the life of the experiment. For NA-to-NA, this turns out to be 95.0 percent (5.0 percent) and for NA-to-EU, 94.6 percent (5.4 percent). Another interesting observation is in the pattern

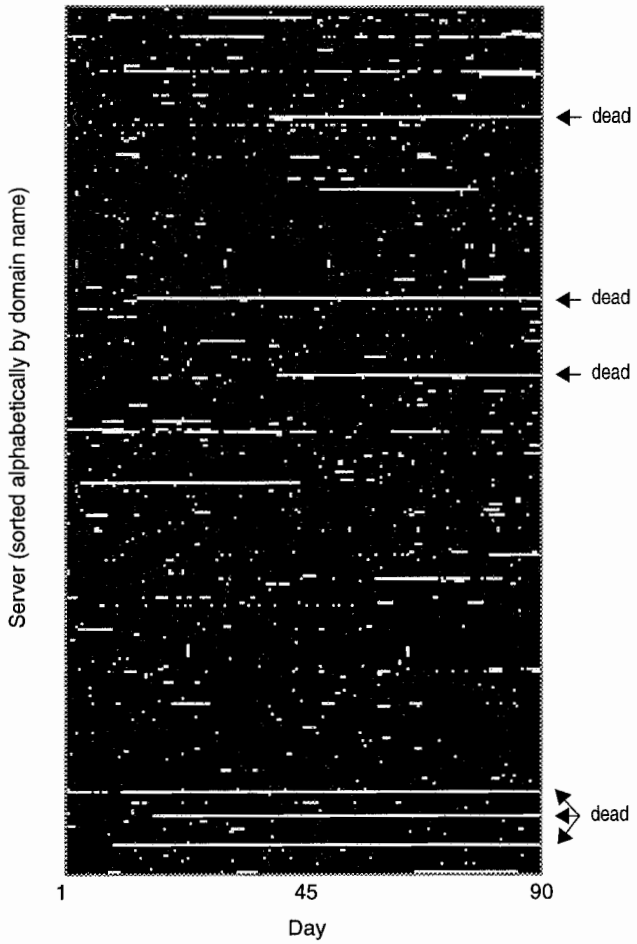

Figure 3. Summary of North American HTTP server availability over 90 days.

of down times at a particular server. In some cases, servers are down intermittently. This is shown by those rows that have an occasional single white point, but are otherwise black. In other cases, servers have long periods of down-time but eventually come back up. We see this in rows that are broken into one or more long white lines. By inspection, the North American group seems to show more of this behavior than the European group, but we have not attempted to quantify this observation. One possible explanation for long down-times is negligence or apathy on the part of the server's administrator. The server itself may have been improperly configured so that it does not survive re-booting. Hardware problems at the server or in the network close to the server are another possible cause.

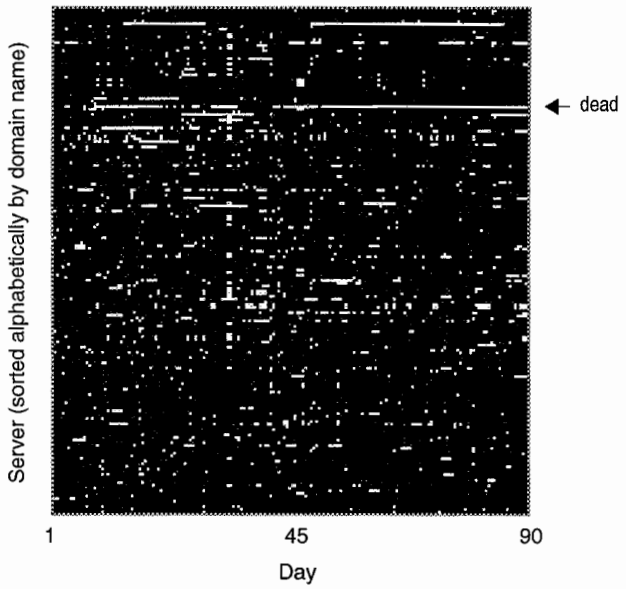

Figure 4. Summary of European HTTP server availability over 90 days.

We identified six North American servers and one European server that were down for 40 or more consecutive days and were down at the end of the measurement period. These servers are marked along the right margin of Figures 3 and 4. We believe these servers are dead or have moved to a different address.

In Table 1, we summarize some of the failure information from Figures 3 and 4. The most interesting additional information to be gleaned from this table is that the DNS failure rate was about five times higher when resolving European domain names than when resolving North American domain names. When DNS failures are removed, the failure rates (and thus overall availability) for both continents are very similar.

To see if there were some geographic relationship to server down-time, we clustered all of the European servers hierarchically by the network path to the server. This path was determined using packet traces generated by *traceroute* [Sun 1993.31]. In general, the clustering consisted of the first two or three European networks traversed to the destination server. Within a "network" cluster, we sorted hierarchically using domain names. The effect of this clustering is a rough geographic sorting, where servers are grouped first by the major network to which they are attached and second by the organization to which they belong. The result, depicted in Figure 5, shows some clear vertical relationships between server down times. The most obvious example of this was on the ja.net path at day 34 of

Table 1. Summary of Failures.

	Europe	North America	Both
Runs	178	178	178
Sites	189	324	513
Connections	33642	57672	91314
DNS Failures	218	60	278
Total Failures	1801	2868	4669
DNS Fail Rate (%)	0.6	0.1	0.3
Other Fail Rate (%)	4.7	4.9	4.8
Total Fail Rate (%)	5.4	5.0	5.1

the experiment, where all servers (exclusively UK based in this experiment) connected via this network were unavailable for two consecutive runs. This behavior points to a network failure or network congestion that effectively partitioned the UK servers from the measurement site. Other examples of this geographic locality

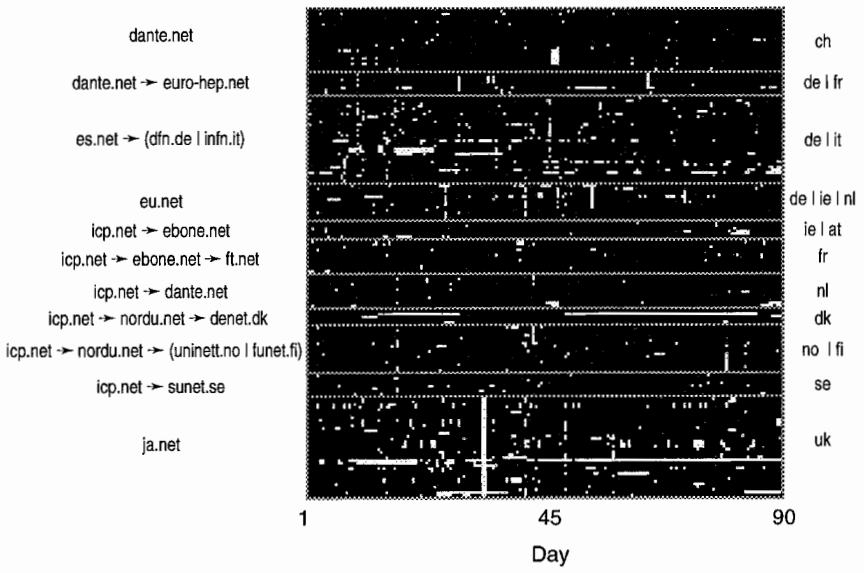

Figure 5. Summary of European HTTP server availability with rows clustered by physical networks traversed. Servers within a cluster were sorted hierarchically by domain name. The path for each cluster is shown to the left, and the countries represented in each cluster are depicted to the right.

appear on the dante.net path (3 runs around day 46); the dante.net/euro-hep.net path (1 run at days 30 and 64); the eu.net path (1 run each at several days); and the icp.net/nordu.net/uninett.no path (one run at days 43 and 80). For geographically close servers, down periods do not appear to be independent events. We did not attempt to cluster the North American servers in this fashion.

In Figure 6 (left), we track *WAvail* over the life of the experiment for NA-to-NA (top), NA-to-EU (center) and NA-to-All (bottom) servers respectively. *WAvail* for NA-to-NA was around 95 percent for the entire experiment, with only two dips below 91 percent and a single peak above 98 percent. This steady behavior is represented in the histogram for *WAvail* (Figure 6 top, right), which shows a very tight central tendency around 95 percent and very few outliers. For NA-to-EU, *WAvail* was also around 95 percent, but was more variable, with several dips below 90 percent and two adjacent dips below 80 percent. The difference between the NA-to-NA and NA-to-EU availability measurements was small, and the two measures were found to be statistically indistinguishable using the Wilcoxon rank sum test. Variability was especially high between days 20 and 60. European servers were also 100 percent available for one run. Some, but not all, of these dips correspond to the geographically correlated down periods depicted in Figure 5. For example, the two dips below 80 percent are in fact the two measurement periods the UK was unavailable via the ja.net path from the measurement site. The wider variability in *WAvail* is reflected in the histograms for NA-to-EU (Figure 6 middle, right), which show a wider distribution than NA-to-NA as well as more outliers.

Figure 7 shows histograms and cumulative histograms of *SAvail* for North American servers (top), European servers (middle), and all servers (bottom). For both NA-to-NA and NA-to-EU, a significant portion of the servers show very good availability: 80 percent of North American servers and 70 percent of the European servers were available from North America 95 percent of the time or better. However, at least 5 percent of both North American and European servers were available from North America less than 80 percent of the time. Even accounting for the seven servers (six NA and one EU) that died or moved, there is a small but significant group of servers with poor availability. This data suggests that this small group might be responsible for a large portion of the Web downtime presented earlier. In Table 2, we show that overall availability and down-time would be better if we remove some of this small group of poor performing servers. For example, if we remove the bottom 5 percent (as measured by *SAvail*) of the EU and NA servers, overall down time on the Web drops from 5.0 percent to 2.5 percent in North America and from 5.4 percent to 4.1 percent in Europe, as measured from North America.

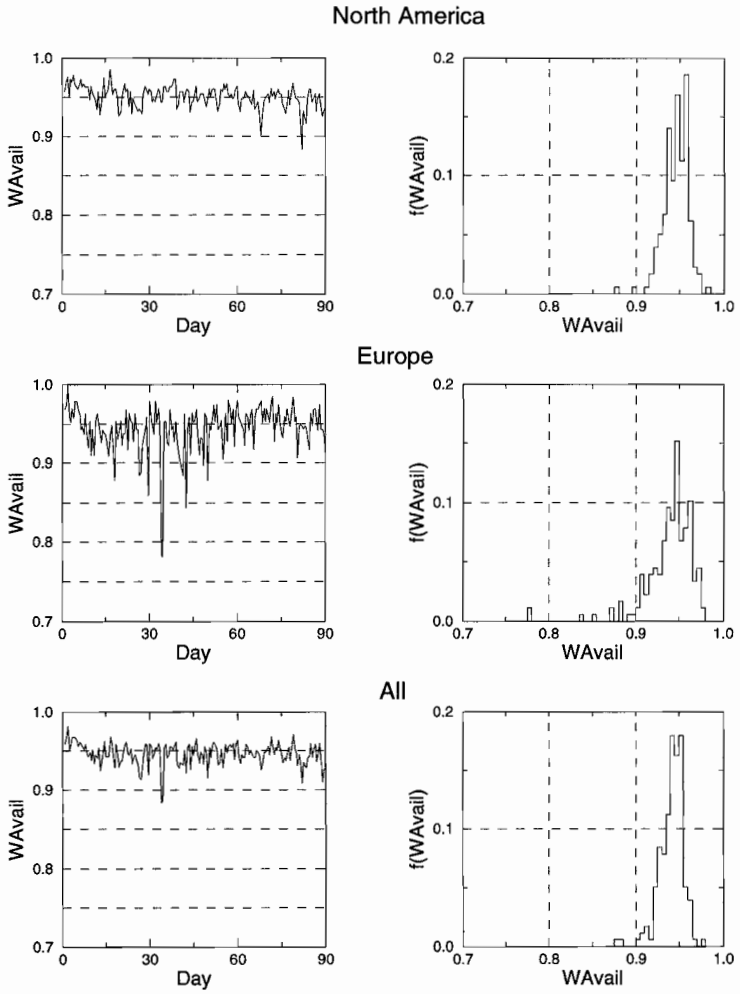

Figure 6. Web availability (WAvail) from the measurement site over 90 days. Longitudinal tracks are on the left and frequency distributions are on the right.

4.2. Latency

The distribution of latencies for successful connections in a typical run is presented in Figure 8. In the case of both NA-to-NA and NA-to-EU, the distributions presented are remarkably typical, e.g. one NA-to-NA run is very similar to any other NA-to-NA run. The spread of the distributions changed with time of day, but the long-tailed shape was characteristic of all runs.

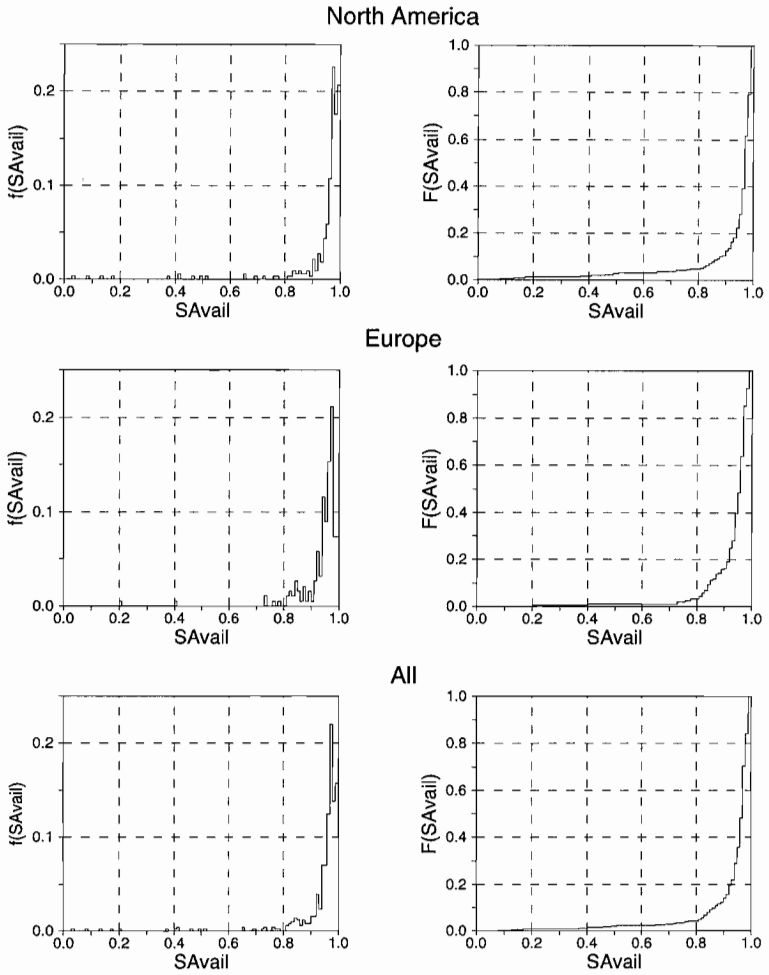

Figure 7. HTTP server availability (SAvail) from the measurement site over 90 days. Frequency distributions are on the left, and cumulative frequency distributions are on the right.

With such long-tailed behavior, the sample median and sample mean were far apart. As an example, the mean and median for the NA-to-NA run depicted in Figure 8 (top) were 1.19 s and 0.27 s respectively, and 3.65 s and 1.48 s respectively for the NA-to-EU run. In fact, in all cases the sample mean was considerably larger than the sample median. We felt that the sample median was a better indicator of typical connection latency than the sample mean, since the vast majority of connections showed latencies below the sample mean. For this reason, we use the median as our presentation statistic.

Table 2. Overall Availability with Poor Performing Servers Removed.

	Europe percent available (percent down)	North America percent available (percent down)
All Servers	94.6 (5.4)	95.0 (5.0)
Dead Servers Removed	95.0 (5.0)	96.4 (3.6)
Bottom 5% Removed	95.9 (4.1)	97.5 (2.5)
Bottom 10% Removed	96.5 (3.5)	98.1 (1.9)

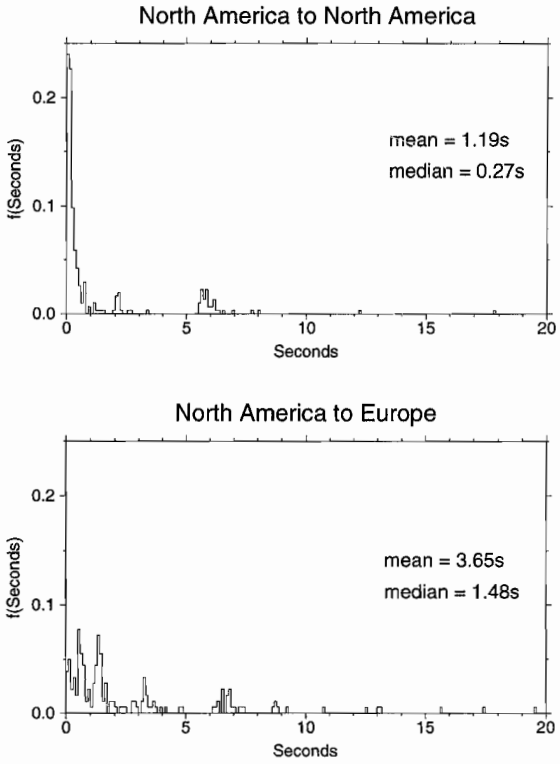

Figure 8. Typical daily latency distributions to North America (top) and Europe (bottom) from the measurement site. This is day 31 of the long-term experiment.

4.2.1. *First Measurement Period: 12-day Intensive*

Figure 9 shows how connection latency varied by time of day. For both North American and European servers, periodic behavior was observed. For NA-to-NA connections, latencies were shortest in the early morning, between 2 and 6 AM, and longest in the mid-afternoon, between 2 and 6 PM. For NA-to-EU connections, latencies were shortest in the late evening to early morning, and longest in the late morning to early afternoon. Given the time difference between the two continents, it appears that latency to a particular site is correlated with local server time. The periodic behavior was particularly evident on the weekdays and less so on the weekend. NA-to-NA latencies were considerably less than NA-to-EU latencies. This is not surprising given that the measurement site was in North America.

4.2.2. *Second Measurement Period: 90-day Long Term*

In Figure 10, we present the median with 25 percent and 75 percent quartiles for both groups of servers and both times of day. European and North American servers both show higher variability in the 75 percent quartile than the median or 25 percent quartile. This underscores the typical long-tailed latency distributions we mentioned earlier. If we compare the first 45 days to the last 45 days, then it appears for both groups and both times that latencies were higher and more variable in the first half of the experiment than the last half. We were unable to trace this phenomenon to any local event (e.g. a change in hardware).

In Figure 11, we make several comparisons of median latencies over the 90-day period. On the left, we compare NA-to-EU and NA-to-NA at 11 AM (top) and 11 PM (bottom) Eastern time. For both times, NA-to-NA latencies are markedly lower than NA-to-EU latencies. For both times of day, NA-to-NA latencies were found to be significantly lower at the 0.01 level than NA-to-EU latencies using a Wilcoxon rank sum test. This is to be expected, not only because distances from North America to Europe are generally greater than distances within North America, but because inter-continent network bandwidth is generally less than intra-continent bandwidth. The NA-to-EU latencies also show a marked increase between Days 15 and 40 (March 15–April 8) of the experiment. We are unsure as to what to attribute this activity other than some temporary network phenomenon that may have reduced bandwidth and connectivity for a time. These increases in latency also corresponded with a slight decrease in NA-to-EU availability (Figure 6 center-left).

Figure 11 also shows how latency varies with the day of the week. At top-left of Figure 11, this is clearly illustrated by the periodic behavior of the NA-to-EU latencies, with dips occurring on the weekends and peaks occurring midweek. We see the same behavior for NA-to-NA, just on a different scale.

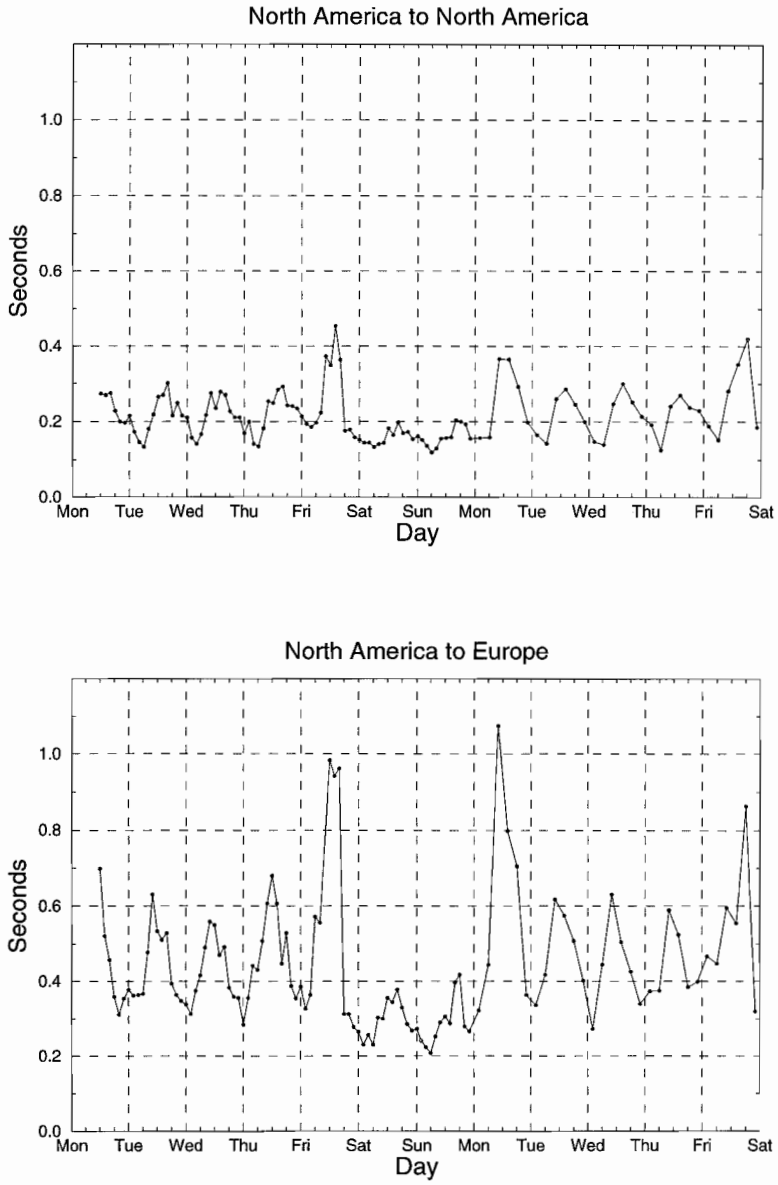

Figure 9. Median connection latency to North America (top) and Europe (bottom) over 12 days. Vertical grid lines represent the start of the day (00:00 AM eastern standard time).

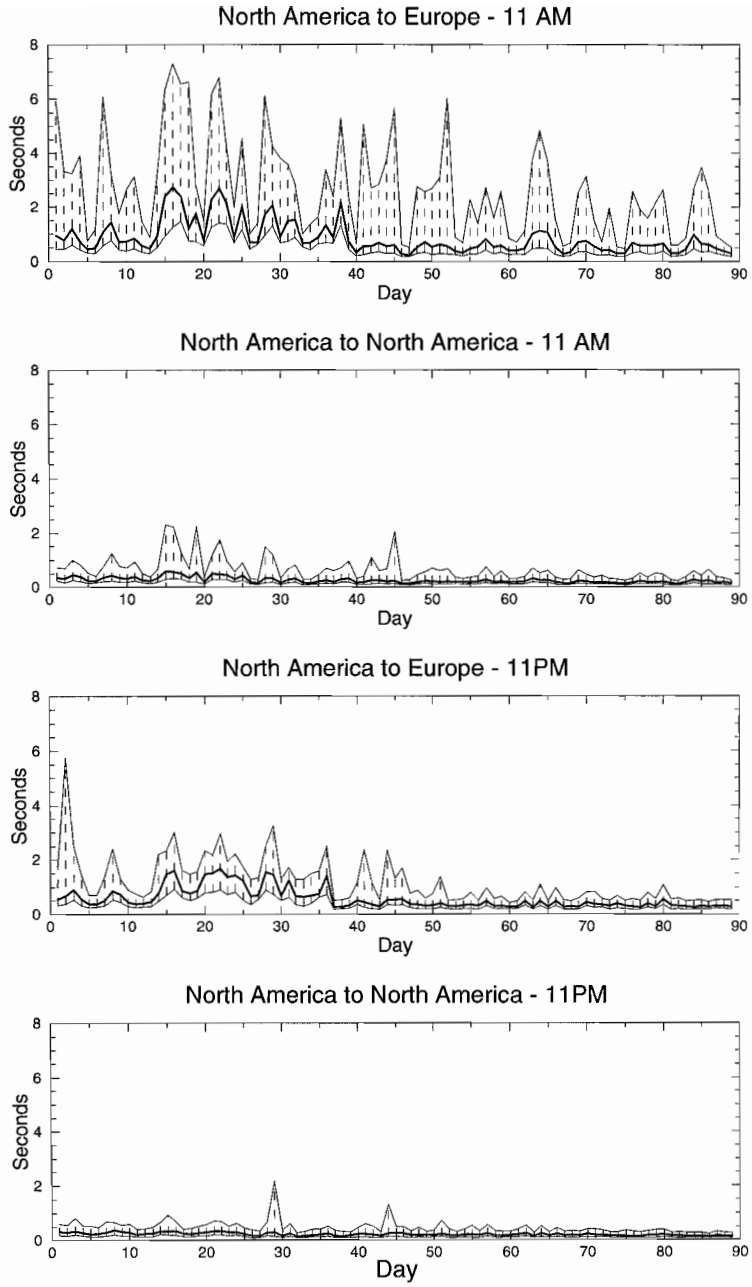

Figure 10. 25 percent (bottom line), median (middle line), and 75 percent (top line) quartiles for latency of successful connections. The top two figures show the latency at 11 AM to North America and Europe, and the bottom two figures show the latency at 11 PM (both times are eastern standard time).

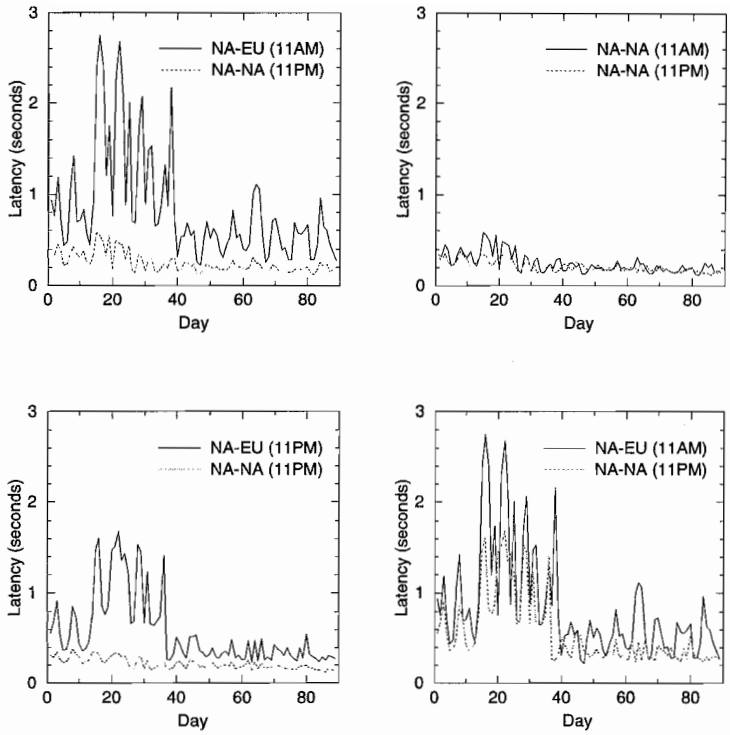

Figure 11. Comparison of median connection latencies by location of servers (left) and time of day (right).

On the right of Figure 11, we compare NA-to-NA (top) and NA-to-EU (bottom) latencies at 11 AM and 11 PM. NA-to-NA latencies are very similar, but NA-to-EU latencies differ for the two times. This is not surprising when we look back to the more intensive measurements depicted in Figure 9. There we can see that latencies at late morning and late evening are in fact very similar for NA-to-NA, but differ for NA-to-EU due to the time difference. In the long-term experiment, we did not sample the curve depicted in Figure 9 at a sufficiently high rate to pick up its inherent periodicity. In the NA-to-NA case, the two samples have occurred on the upslope and downslope of the curve, effectively eliminating the observation of any high frequency behavior. In the NA-to-EU case, where the curve is effectively phase-shifted by several hours from the NA-to-NA curve, the samples occur at different relative points along the curve, and so we see some of the periodic behavior of Figure 9 evidenced in the NA-to-EU plots of Figure 11.

We show the distribution of latencies for all successful and failed connections in Figure 12. NA-to-NA and NA-to-EU measurements are displayed separately. These plots represent all of the latency data, 178 runs to all sites, over 90 days.

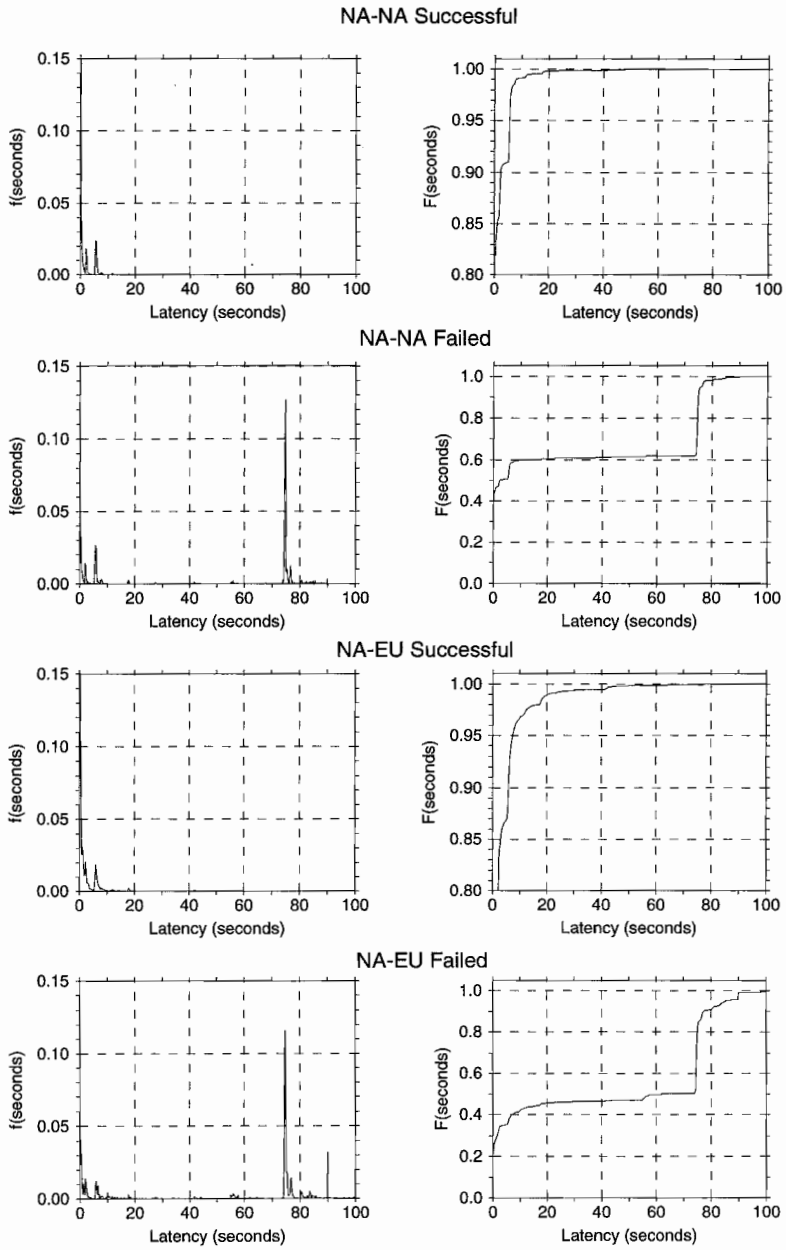

Figure 12. Frequency distributions (left) and cumulative frequency distributions (right) for all connection attempts over 90 days. The top two rows show successful and failed connections respectively for North American servers, and the bottom two rows show similar data for European servers.

If we examine the histogram plots (left), we can make several observations. First, the distribution of failed connections is bimodal, with a large group of failures occurring in the first few seconds and a second group failing at 75 or 90 seconds. A small portion of both of these groups (see Table 1) are DNS failures, but most are generated after name resolution. We believe failures that resolved quickly were those that were rejected by the destination site, either because the server was down or was heavily loaded, or the machine was down. Almost all of the failures clustered at 75 and 90 seconds are due to time-outs generated by the TCP service running on the client machine.

The second observation is that the vast majority of successful connections occur quickly. In fact, the data represented by the cumulative distributions of Figure 12 show that 99.1 percent of all NA-to-NA connections occur in the first 10 s, and 99.8 percent in 20 s. For NA-to- EU, the numbers are slightly lower, 96.7 percent and 98.9 percent for 10 s and 20 s, but still very high. This data suggests that client performance would be greatly improved by establishing a client-side time-out interval drastically lower than that often supplied by the TCP service. For example, a time-out period of 10s for NA-to-NA connection attempts and 20s for NA-to-EU attempts would have caused the loss of only about 1 percent of the successful connections, while saving tens of seconds in user wait-time for each connection attempt. The distribution of successful connections for NA-to-NA shows two additional peaks after the main body of the distribution, one at 2 s and another at 6 s. NA-to-EU also has an additional peak at 6 s. Though not shown by the plots, a small number of sites are responsible for these peaks. For example, the 6 s peak for NA-to-NA is almost entirely composed of connections to about 20 of the 324 North American servers. We attribute this locality to a gateway or other physical bottleneck between the measurement site and the destination servers.

5. *Discussion*

In this section we address in more detail some of the issues that our measurements have raised.

5.1. *HTTP and Latency*

One of the original justifications for HTTP was a protocol lightweight enough to ensure 100 millisecond response times [Berners-Lee 1993]. Our measurements clearly show that only a handful of servers showed median latency in the 100 msec range, and many servers showed median latencies an order of magnitude greater. This in itself is not an indictment of HTTP: these latencies are

a fact of life on the Internet. Median *one-way* message latency between nodes on the current NSFNET T3 backbone can be as great as 50 msec [Merit 1994]. Round-trip message latency from our site on the East Coast of North America to the West Coast of North America is consistently greater than 100 msec for ICMP ECHO_REQUEST messages generated by the *ping* program [Sun 1993.30]. These measurements are good estimates of the inherent network latency between two sites. Given that setting up a TCP connection takes a minimum of three messages [Postel 1981], with data usually not being transmitted until the fourth message, by the time a TCP connection is set up (and before any HTTP action occurs), latency from the user's perspective is often well above 100 msec. Since TCP is the de facto standard for reliable communication on the Internet, connection latency for Web servers is bounded by both the current physical infrastructure and the protocols used to deliver the information.

5.2. The Case for MGET

The stateless nature of HTTP has its drawbacks. Many Web pages integrate text and images in order to effectively present the information of interest. Since HTML includes images by reference, a client must create separate TCP connections for the text of the document and for each included image. For example, a page with three included images must pay the TCP connection latency price four times. As the measurements presented in this paper attest, this price is non-trivial. Ideally, the document text and images would all be transmitted during the same transaction. Unfortunately, this would require the server to parse the document in order to ascertain what if any included references there were. While this capability is present in some HTTP servers, there can be a considerable server-side performance cost associated with the parsing. A reasonable compromise solution would be to extend HTTP by creating a new method called MGET, analogous to the capability of the same name implemented in some FTP clients. The Web client would then issue at most two HTTP method requests: the first would be a GET for the document text, and the second request would be a MGET for all images included by reference in the first document. This solution obviates parsing at the server, limits the number of TCP connections for any Web page to two, and retains the stateless nature of HTTP. One possible disadvantage might arise on clients with low speed links, in that MGET would commit the client to fetch all included items. A client level user-abort option is appropriate for these situations.

Spero [1994] discusses other performance problems that arise due to interactions between HTTP and TCP. Padmanabhan and Mogul [1994] performed a more detailed study that quantified the performance cost of multiple GETs. They show marked reductions in latency for implementations of a GETALL (get all

images associated with this document) and GETLIST (get the following list of documents). GETLIST is semantically similar to our MGET.

5.3. Long-Term Availability

It does not appear that Web availability changed over the duration of the experiment. As mentioned previously, the slight downward trend in *WAvail* (Figure 6) for North American sites is mostly due to the death or migration of six servers. If we remove these servers from the measurements, then *WAvail* was stable. In fact, a simple linear regression of the *WAvail* data including the dead servers yielded a line whose slope translated to the loss of 5.9 servers over the 90-day period. The fit of this line was relatively poor ($r^2 = 0.15$), but this result at least lends some credence to our supposition about stable Web availability.

The cumulative histograms of *WAvail* (Figure 6) illustrate the good news and bad news about Web availability. The good news is that the Web is almost always at least 90 percent available and is usually in the 95 percent range. The bad news is that at no time was the Web completely available. While some servers showed 100 percent measured availability over the life of the experiment, the majority of servers were down at least some of the time.

The World Wide Web is a recent phenomenon and it is interesting to posit about the effect of its tender age on our measurements, particularly availability. In particular, we wonder how much effect the skill of server administrators has on availability. World Wide Web server software can be installed and maintained by almost anyone. We feel this is one reason we saw such a high variation in the *SAvail* measurements. Some servers are meticulously maintained, others are not. In addition, public domain server software underwent major changes over the life of the experiment. We have no way of measuring the effect of these changes on availability. An interesting question to examine would be to see if there were any difference in availability between "new" servers and "old" servers.

5.4. Client-Side Time-out Intervals

The distributions of successful and failed connections depicted in Figure 12 provide useful information to the client. It is clear that there is very little payoff to waiting longer than 20 seconds to establish a connection. If the entire cumulative distribution were available to the client, then users could set the time-out interval for their sessions based on their immediate information needs. Simple browsing and wandering might call for a short time-out interval, while exhaustive searching might require longer intervals to ensure that no servers that were up were missed. Users might use the cumulative distribution directly by specifying the interval they

were willing to wait, or they might use the inverse function to specify the coverage they want to achieve and then set the interval needed to achieve that coverage.

5.5. Naming

The observation that some servers have moved or died underscores a general resource and document naming problem that the Web community has experienced in its initial growing period and which will likely only get worse. In its initial formulation, and as of the writing of this paper, resources were expressed in terms of Uniform Resource Locators (URLs). The URL mechanism can be considered the instantiation of the URI [Berners-Lee 1993.5] concept in the existing naming schemes and protocols on the Internet. A URL is essentially an address for the resource, not its name. Whenever a resource moves, its URL becomes useless. The technical community is currently attempting to address this issue through the introduction of Uniform Resource Names (URNs) [Weider and Deutsch 1994], an attempt at location transparent names. At this point, the exact shape of the naming system and the migration from URLs to URNs is a subject of debate.

5.6. Study Methodology

As we mentioned at the beginning of Section 4, any interpretation of the latency and availability measurements needs to be made with the location of the measurement site in mind. The results reported here consistently show higher availability and lower latency for North America. If the measurement site were in Europe, then we would likely see the opposite behavior, with longer latencies and lower availability in North America.

We caution against too general an interpretation of the results here. The measurements were made from a single end point in the Web and would likely be different if made elsewhere. However, the measurement site has typical Internet connectivity for a US university. A site with poorer connectivity would almost assuredly show longer latencies, and perhaps lower apparent availability due to local network congestion and more time-outs. A site with better connectivity might take less time to get to a high bandwidth network (e.g. NSFNET or Dante), but would still be bound by the network close to the destination server.

Choosing a nonsense method to send to the HTTP servers eliminated some but not all of the arbitrary system delays on the server side. For example, some HTTP servers run out of inetd and some run as a stand-alone daemon. We did not attempt to contact server administrators to get this information. Paging and machine load are other activities that might introduce delays.

Latencies presented here are optimistic from the user's perspective, because they do not include the retrieval of any documents by the contacted server, only the time to set up the connection. If we view availability in terms of obtaining the requested document rather than simply servers being up, then the availability measurements are also optimistic, since they do not include the possibility of server time-outs after connection establishment, or of stale URLs.

We feel our splitting of the sites list into European and North American lists by domain name represents a fair split of servers by geographical area. There are likely some sites in the .org and .edu hierarchies that are not North American based, but the vast majority of these sites are located according to our assumptions. Examination of network connections using *traceroute* [Sun 1993.31] verified the European list.

6. Summary

Measurements we have made of a large number of World Wide Web servers indicate consistent availability of servers in the 95 percent range. We view this as an optimistic estimate of Web availability. We found no statistically significant differences in availability between North American and European servers. Examination of the European data showed a correlation between server availability and geographic location of the server (as expressed by the network path to the server).

Latency measurements show connection establishment taking in the 200-500 msec range for North America to North America connections, and in the 400-2500 msec range for North America to Europe. From the user perspective, these latency estimates are optimistic because they do not include document retrieval. Statistical differences were found between the median latencies to the two groups of servers. We attribute this difference to the location of the measurement site. Variability in these latencies was consistently higher in the first half of the experiment than in the last half, and it is unclear which half constitutes 'normal' behavior. The magnitude of connection latency is affected not only by the physical connection between client and server, but the long set-up time for TCP connections. We advocate the addition of an MGET method to HTTP to help overcome the large start-up costs associated with retrieval of "compound" documents.

Our data (end of Section 4.2.2 and Figure 12) indicate that setting of client-side time-out intervals would drastically improve worst-case connection attempts with only a very minor effect on availability. The particular interval may vary over time, but ongoing monitoring would allow up-to-date estimates of this interval. Monitoring would also allow server administrators to see availability and latencies of their servers as they are viewed from other points on the Internet.

Acknowledgments

This work was supported by the NASA Goddard Space Flight Center under NASA Graduate Student Research Program Fellowship NGT-51018 and by NASA/CESDIS Grant 5555-25. We thank Jorg Liebeherr, Bert Dempsey, and the anonymous reviewers for their helpful comments. A version of this paper appeared as University of Virginia, Department of Computer Science Technical Report # CS-94-36.

References

1. A. K. Agrawala, D. Sanghi. Network Dynamics: An Experimental Study of the Internet, *Proc. Globecom*, pp. 782–786, 1992.
2. P. Beebee, *The SG-Scout Home Page*, available at http://www-swiss.ai.mit.edu/ptbb/SG-Scout.html, 1994.
3. T. Berners-Lee, *Hypertext Transfer Protocol (HTTP)*, Working Draft of the Internet Engineering Task Force, 1993.
4. T. Berners-Lee, D. Connolly, *Hypertext Markup Language (HTML)*, Working Draft of the Internet Engineering Task Force, 1993.
5. T. Berners-Lee, *Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web*, Internet RFC 1630, 1993
6. T. Berners-Lee, R. Cailliau, A. Luotonen, H.F. Neilsen, A. Secret. The World Wide Web, *Communications of the ACM*. 37(8):76–82, 1994.
7. T. Berners-Lee, *The World Wide Web Initiative: The Project*, Available at http://info.cern.ch/hypertext/WWW/TheProject.html, 1994.
8. J. C. Bolot, End-to-End Packet Delay and Loss Behavior in the Internet, *ACM SIGCOMM*, San Francisco, California, pp. 289–298, 1993.
9. N. Borenstein, N. Freed. *MIME (Multipurpose Internet Mail Extensions): Mechanisms for Specifying and Describing the Format of Internet Message Bodies*, Internet RFC 1341, 1992.
10. H. W. Braun, K. Claffy, Web Traffic Characterization: An Assessment of the Impact of Caching Documents from NCSA's Web Server, *Second International Conf. on the World Wide Web*, Chicago, IL, October 20–23, 1994.
11. R. Caceres, P. B. Danzig, S. Jamin, D. J. Mitzel, Characteristics of Wide-Area TCP/IP Conversations, *ACM SIGCOMM*, Zurich, Switzerland, pp. 101–112, 1991.
12. K. C. Claffy, G. C. Polyzos, H. Braun, Traffic Characteristics of the T1 NSFNET Backbone, *Proceedings of IEEE INFOCOM*, pp. 885–892, 1993.
13. Peter B. Danzig, K. Obraczka, A. Kumar. An Analysis of Wide-Area Name Server Traffic, *ACM SIGCOMM*, Baltimore, Maryland, pp. 281–292, 1992.
14. DEC Network Systems Laboratory, *Usenet Flow Analysis*, available from Usenet newsgroup news.lists, 1994.
15. D. Eichmann, The RBSE Spider: Balancing Effective Search Against Web Load, *Proc. First Intl. Conf on the World Wide Web*, Geneva, Switzerland, May 25–27, 1994.
16. R. T. Fielding, Maintaining Distributed Hypertext Infostructures: Welcome to MOMspider's Web, *Proc. First Intl. Conf on the World Wide Web*, Geneva, Switzerland, May 25–27, 1994.
17. J. Fletcher, *The Jumpstation*, available at http://www.stir.ac.uk/jsbin/js, 1994.
18. M. T. Gray, *Growth of the World Wide Web*, Available at http://www.mit.edu:8001/afs/sipb/user/mkgray/ht/web-growth.html, 1994.

19. M. Lottor, *Internet Growth (1981-1991)*, Internet RFC 1296, 1992.
20. M. Lottor, *Internet Domain Survey*, available at http://www.nw.com/, 1994.
21. S. Maffeis, File Access Patterns in Public FTP Archives and an Index for Locality of Reference, *ACM SIGMETRICS Performance Evaluation Review*, 20(5):22–35, 1993.
22. M. Mauldin, J. R. R. Leavitt, *The Lycos Home Page: Hunting WWW Information*, available at http://fuzine.mt.cs.cmu.edu/mlm/lycos-home.html, 1994.
23. O. A. McBryan, GENVL and WWWW: Tools for Taming the Web, *Proc. First Intl. Conf. on the World Wide Web*, Geneva, Switzerland, May 25–27, 1994.
24. Merit Networks, Inc., *ANSNET T3 Delay Matrix Report*, available by anonymous ftp at ftp://nic.merit.edu/nsfnet/statistics, 1994.
25. D.L. Mills, *Internet Delay Experiments*, Internet RFC 889, 1983.
26. V. N. Padmanabhan, J. C. Mogul. Improving HTTP Latency, *Second International Conf on the World Wide Web*, Chicago, IL, October 20–23, 1994.
27. J. Postel, 1981. *Transmission Control Protocol*, Internet RFC 793, 1981.
28. D. Sanghi, A. K. Agrawala, O. Gudmundsson, B. N. Jain, Experimental Assessment of End-to-End Behavior on Internet, *Proc IEEE INFOCOM*, pp. 867–874, 1993.
29. S. Spero, *Analysis of HTTP Performance Problems*, available at http://elanor.oit.unc.edu/http-probs.html, 1994.
30. Sun Microsystems, Ping, *SunOS User's Manual*, 1993.
31. Sun Microsystems, Traceroute, *SunOS User's Manual*, 1993.
32. C. Tronche, *The WWWMM Robot*, available at http://www-ihm.lri.fr/~tronche/W3M2/, 1994.
33. C. Tronche, Personal Communication, 1994.
34. C. Weider, P. Deutsch, *Uniform Resource Names*, Working Draft of the Internet Engineering Task Force, 1994.