Analysis of HTTP Performance

Joe Touch, John Heidemann, and Katia Obraczka

USC/Information Sciences Institute

June 24, 1996

Initial Release, V1.1

Abstract:

We discuss the performance effects of using per-transaction TCP connections for HTTP access, and the proposed optimizations of avoiding per-transaction re-connection and TCP slow-start restart overheads. We analyzed the performance penalties of the interaction of HTTP and TCP. Our observations indicate that the proposed optimizations do not affect Web access for the vast majority of users. Most users see end-to-end latencies of about 100-250 ms and use modem lines, resulting in only 1-2 packets in transit. At these rates, the optimizations reduce the overall transaction time by 13%. Rates over 240 Kbps are required in order to achieve user-noticeable performance enhancement, reducing the time per transaction by half.

Table of contents:

Introduction

There have been several recent discussions about the performance problems of HTTP over TCP. This Web page continues that discussion with a description of the performance benefits of the proposed approaches. We discuss the evolution of the HTTP protocol, and the potential for inefficiencies. We then present analysis of these inefficiencies in current Web systems. We have found that, except for users with Ethernet-speed end-to-end links between the client and server, the performance enhancements proposed in [web-analysis] and [p-http] have limited benefit for current Web access. The proposed application-layer persistent connection mechanisms achieve only a 20% improvement in response time in these cases, whereas we consider a 50% improvement the minimum noticeable (i.e., factor of 2 speedup). We conclude that such mechanisms are of limited benefit, and observe that they may interfere with emerging Internet services.

Why HTTP?

The HTTP protocol was originally developed to reduce the inefficiencies of the FTP protocol [www], [ftp]. The goal was fast request-response interaction without requiring state at the server. To see the performance advantage of HTTP over FTP, we can compare the process of file retrieval transactions in each protocol. Both protocols use a reliable connection-oriented transport protocol, TCP [tcp].

In FTP, a client opens a TCP connection with the server for control (Figure 1). Once that connection is established, a request for a file is sent on that channel. The server then opens a separate TCP connection for the file transfer, and returns the file in that other connection. Each connection requires one round-trip time (RTT) to open. The request takes 1/2 a RTT to get to the server, and the response takes another 1/2 RTT to return, in addition to the transmission time of the file. The overall time required for an FTP transaction is:

```
1   RTT control-channel OPEN
0.5 RTT send request on control-channel
1   RTT file-channel OPEN
0.5 RTT file starts to arrive on file-channel
Ftrans  time to transmit the file
--------
3 RTT + Ftrans = time to get the first file in FTP
```

This is shown in Figure 1, below. The control channel interaction is shown in red, and the file channel is shown in blue.

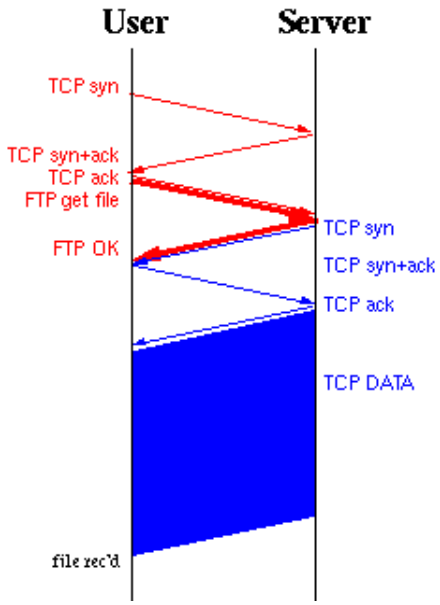

Figure 1: FTP File Transfer (first file)

Subsequent transactions to the same server may take less time, because the control channel is already open. A new TCP connection is required for each transaction in conventional use (block and compressed mode do not require this, but are not commonly used). The interaction is shown in Figure 2.

```
0.5 RTT send request on control-channel
1   RTT file-channel OPEN
0.5 RTT file starts to arrive on file-channel
Ftrans  time to transmit the file
--------
2 RTT + Ftrans = time to get subsequent files in FTP
```

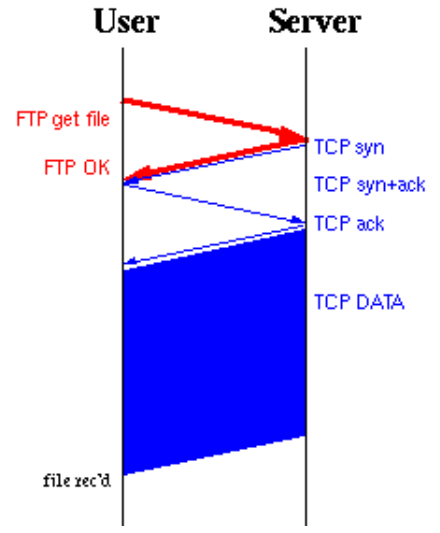

Figure 2: FTP File Transfer (subsequent files)

HTTP uses a single TCP connection for the entire transaction, achieving FTP's best response time, even for the first file requested. Further, HTTP doesn't require the control-channel to be maintained at the server or client, so is stateless and simpler to implement. The transaction is also shown in Figure 3.

```
1   RTT channel OPEN
0.5 RTT send request
0.5 RTT file starts to arrive
Ftrans  time to transmit the file
--------
2 RTT + Ftrans = time to get a file in HTTP
```

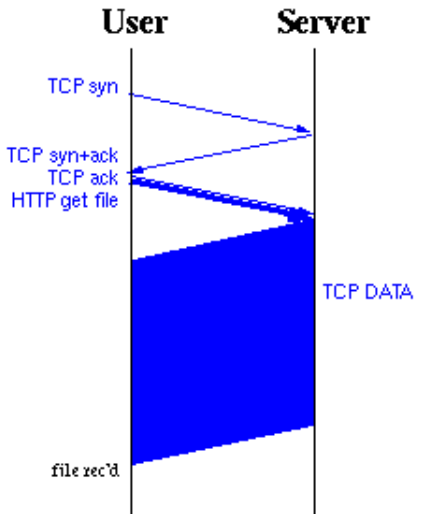

Figure 3: HTTP File Transfer

Potential Protocol Inefficiencies

There are inefficiencies in using HTTP over TCP. TCP establishes a connection prior to transferring any data, namely the request. TCP also includes a congestion avoidance mechanism [tcp-ss]. In both cases, these mechanisms are restarted for each file request, possibly resulting in excessive overheads.

Connection Establishment

A minimal reliable transfer could occur with as little as one round-trip of overhead, plus the file transmission time, as shown in Figure 4.

```
1   RTT channel OPEN and send request, file starts to arrive
Ftrans  time to transmit the file
--------
1 RTT + Ftrans = time to get a file optimally
```

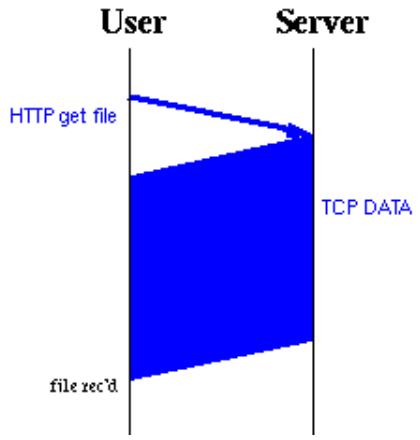

Figure 4: Optimal Transaction

TCP does not support the minimal transaction because the initial request cannot be delivered to the server until the connection has been established, which takes 1.5 RTTs, from the server's perspective [tcp]. This is true even if the request is enclosed with the initial "SYN" OPEN packet; delivery of the request at the server is stalled until the third packet of the exchange arrives.

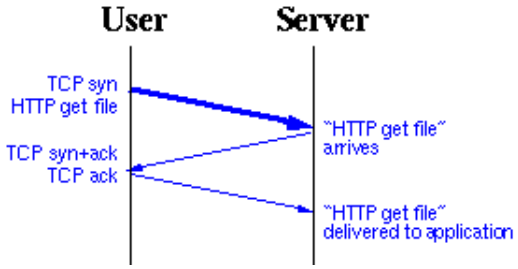

Figure 5: Delayed request delivery

Slow-Start (congestion avoidance)

TCP also includes a congestion avoidance mechanism called "slow-start" [tcp-ss]. When a connection opens, only one packet is sent until an ACK is received. For each ACK received, the number of packets that can be sent is increased by one. For each round-trip, the number of outstanding packets doubles, until a set of thresholds have been reached.

The packet size is negotiated, and commonly the largest end-to-end link packet size supported. The default is 536 bytes in TCP, although most implementations round this down to 512. Hosts on Ethernets typically use 1460, but only for local connections.

Slow-start occurs when a connection is initialized, when a packet is lost, or may occur when there is a significant gap in the packet transmission, when the packet burst of resuming the transmission encourages packet loss.

Alternative Protocol Mechanisms

There have been several proposals to address the potential inefficiencies of using HTTP over TCP. These include persistent-connection HTTP, Transaction TCP, and (recently) sharing TCP control blocks. These proposals address either connection or slow-start overheads, and in some cases, both issues.

Persistent HTTP

Persistent HTTP addresses both connection and slow-start overheads. There are several distinct proposals for p-HTTP, including [http], [http-ng], and [p-http]. For the purposes of this discussion, we treat them as sufficiently similar.

P-HTTP attempts to achieve optimal transaction time for sequences of transactions to the same server. The initial transaction occurs as in HTTP, but the connection is not closed. Subsequent requests occur without needing to re-open the connection.

In addition, p-HTTP attempts to avoid slow-start restart for each new transaction, again by using a single connection for a sequence of transactions. Unfortunately, sufficiently large gaps in the arrival of requests may cause a restart of slow-start anyway, due packet loss during the resulting packet burst when the transmission resumes. The p-HTTP method is useful primarily for multiple adjacent requests, as would occur on pages with embedded GIFs, for example.

P-HTTP achieves this efficiency at the expense of application-layer complexity. Re-using a single connection requires application-layer multiplexing or can stall concurrent requests arbitrarily. Consider retrieving a large PostScript file, and issuing a small HTML file request during the transfer. The HTML response will either be stalled until the end of the PostScript file transmission, or the PostScript file will be segmented. The server cannot know whether this segmentation is required or not when it started to send the PostScript file. MIME-style headers are in-line and not encoded via "escapes"; only the specified length is used to determine when to parse the next header. As a result, the inefficiency of application-layer segmentation and reassembly occurs for every transaction. In addition, the multiplexing algorithm in the application interferes with emerging Integrated Services multiplexing in the kernel, for Type-of-Service and Quality-of-Service mechanisms [int-svc].

Transaction TCP

Transaction TCP (T/TCP) provides transaction-oriented service over TCP, via extensions to the TCP protocol [ttcp-c], [ttcp-f]. T/TCP uses cached per-host state to avoid the delayed delivery of data carried with an OPEN, as discussed earlier. TCP delays that data to avoid delivery to the wrong connection, especially in cases of aborted connections. T/TCP uses cached values of extended state to avoid such errors, and permits early delivery before the third packet in the exchange.

In addition, T/TCP caches other TCP protocol control block parameters, such as round-trip time measures, to avoid inefficiencies with reconnecting to the same host. Reusing slow-start information, which would avoid slow-start restart, is discussed briefly in the T/TCP specification.

Shared TCP Control Blocks

Shared TCBs (s-TCB) augment the TCB state-sharing mechanism of T/TCP and show how to aggregate parameters such as window size across sets of concurrent connections [tcb]. T/TCP state caching is aimed predominantly at serial connection state reuse, whereas s-TCBs address both serial and concurrent shared state reuse.

S-TCBs optimize only the inefficiency of the slow-start restart component of HTTP over TCP. Also described in the s-TCB memo are the effects of application-layer multiplexing, and ways in which kernel-based multilevel feedback queuing, as in Integrated Services, would be adversely affected.

When s-TCB and T/TCP are coupled, they provide similar efficiency to p-HTTP, but at the kernel-level rather than requiring application-layer multiplexing.

Prior Analyses

Earlier analyses have claimed significant performance problems with HTTP over TCP [web-analysis], [web-why-slow], [http-lat]. Both Spero's and Mogul's analyses focused on client/server interactions in well-connected (1 Mbps) hosts [web-analysis], [http-lat]. Their conclusions do not apply to the vast majority of web accesses, which are for small files over modem and ISDN links.

Moskowitz described performance problems at the server, where buffering limitations in the operating systems affected transaction performance. The claim is that the server runs out of buffers to create new TCP connections; the purported evidence is the "Host Connected, Waiting Reply" message. This message is emitted **after** the TCP connection is established, which is in turn **after** the TCB control block is allocated. This message is possibly evidence of processing bottlenecks at the server after the connection is established, although it is counter-proof of the claimed lack of buffers.

We assume server processing time is zero, to maximize the potential benefit of the proposed optimizations. When processing is longer, the benefits are reduced.

We are currently looking at HTTP performance over several transport protocols, including TCP, p-HTTP, T/TCP, and UDP-based RPC protocols, over a wider variety of network conditions [web-transp].

Environment Characteristics

The degree of inefficiency of HTTP over TCP depends on the environment in which the Web operates. Here we describe some characteristics of that environment. These factors will be used in the evaluation of the performance inefficiencies of HTTP over TCP later.

Networks

Current network environments can be characterized by a small set of classes: remote (satellite), modem, ISDN, leased-line direct (T-1), and high-performance (ATM, fast). These classes are named for dominant factor in the path between client and server, as follows:

```
        Net     BW      MSS             Latency (ms)
```

	(bps)	(bytes)	LAN/MAN	WAN
sat.	9 K	512	250	500
modem	29 K	512	150	250
ISDN	112 K	1460/512	30	130
direct	1 M	1460/512	2	100
fast	155 M	8192/512	2	100

The default TCP MSS is 536 bytes for data, although most current implementations round this down to 512 bytes. Fast links support larger MSSs, but TCP often ignores them and uses the default for connections where MTU discovery is not implemented.

The Web

Current Web use can also be characterized by classes of file types accessed. Several studies have shown that the vast majority of Web accesses retrieve small files, on the order of 6 KB. We describe these classes as:

Web	File Size (bytes)	File Type
HTML	6 K	ASCII text
"Web page"	6+2+2 K	HTML and links to 2 icons
text	60 K	ASCII text
icon	2 K	small GIF (icons)
image	20 K	large GIF (clickable map)
photo	200 K	very large GIF (photo)

Although this describes the characteristics of web pages in general, the majority of accesses are to 6 KB files, and that is the focus of the discussion. Other analysis at ISI shows that the "Web page" case has a greater potential for optimization than the HTML case, because it is composed of multiple files [web-transp]. In particular, the aggregate of files denoted by a Web page can be retrieved in a single connection with a single aggregate ’GET-ALL’ request, rather than even using persistent connections [web-lat].

The extent of the optimization depends on the network properties; for modem links it is 10%, for ISDN it is also low. The potential for optimization increases significantly for faster connections or for higher latency paths [web-transp], which is similar to the results in [web-analysis] and [p-http].

Evaluation

We want to determine the potential for inefficiency in HTTP over TCP. For this purpose, we analyze the time required for an HTTP interaction, considering both per-transaction connection establishment and potential slow-start overheads, and compare that to the optimal time for transfer.

These analysis consider optimal performance of the system. We assume that the hosts are limited only by the network bandwidth, that server processing time is negligible, and that disk I/O and other bottlenecks are minimal. For this analysis, we consider the worst-case performance of HTTP over TCP, assuming no packet loss. This maximizes the benefit of persistent connections. Even so, these optimizations benefit most Web users only minimally. When other factors, such as server processing,

packet loss, etc., are included, the optimizations are even less noticeable.

The following section presents analysis for HTTP over TCP. A more complete analysis of HTTP over several transport protocols and p-HTTP is currently underway [web-transp]. The formulae below are simplified versions of those developed there.

Analysis

The following notation is used in the analysis:

```
R   = RTT
bw  = bandwidth
MSS = maximum segment size (packet size)

K = number of packets in the file
  = filesize / MSS

L = round trip time in packets, i.e., length of the pipe
  = number of packets to fill the pipe
  = bw * R / MSS

M = max useful window size (lower bound)
  = min(floor(L), floor(K))

S = round trips stalled in slow-start, assuming no loss (upper bound)
    (window starts at 2, see [web-transp])
  = max(0, ceil(log2( M ))-1)

W = amount of wasted time
  = (upper-bound on waste -- not all slow-start is wasted, though)
  = slow-start + connection-setup
  = R * S + R

F = min. file transmission time
  = filesize / bw

P = server processing time

T = transaction time
  = F + P + R
  = F (minimally)
```

Things start to matter when the wasted time is the same as the minimum transaction time, or larger. At that point, including re-connection and slow-start restart avoidance will halve the time of access. This assumes no network loss; otherwise, the wasted time may not be avoidable.

```
F + P <= W

(under best conditions, assume P goes to zero)

F <= W
```

So we can plot the ratio of time wasted to file transmission, as an upper-bound on the optimization possible. This ignores processing time and other impediments at the server, as mentioned earlier. We also count the entire round-trip of each slow-start exchange as wasted, which is not strictly true. Up to one RTT-worth of data is sent during this exchange, at most; by ignoring this, we achieve a further

upper-bound.

```
 W
--- = wasted time, relative to file transaction time
 F

      100 * F
100 - ------- = percent reduction in overall transaction time
      (W + F)
```

Some common cases

We plotted the ratio of waste to useful time on a contour plot. For a given network RTT, we want to see what bandwidth is required for the proposed optimizations to reduce the overall transaction by half, i.e., where the waste is the same as the useful time.

The graph is fixed for a constant filesize of 6 KB. We consider bandwidths from 10 Kbps - 1 Mbps and latencies from .01-1 seconds. Typical latencies are 70 ms for end-to-end latency for average Web surfing in the USA, with 30-150 ms of additional latency for modem or ISDN links. We therefore consider 250 ms total latency for modem links and 100 ms total latency for other types of directly-connected networks. Satellite network latencies are higher, but not considered below.

Shown in Figure 6 are contour lines where the waste to useful time is .25, .50, 1, and 10. I.e., for .50, removing the overhead halves the effective transaction time. The shaded area shows where this 2x speedup (or greater) applies. For this graph, we consider Internet interactions, so that the MSS is 512 bytes.

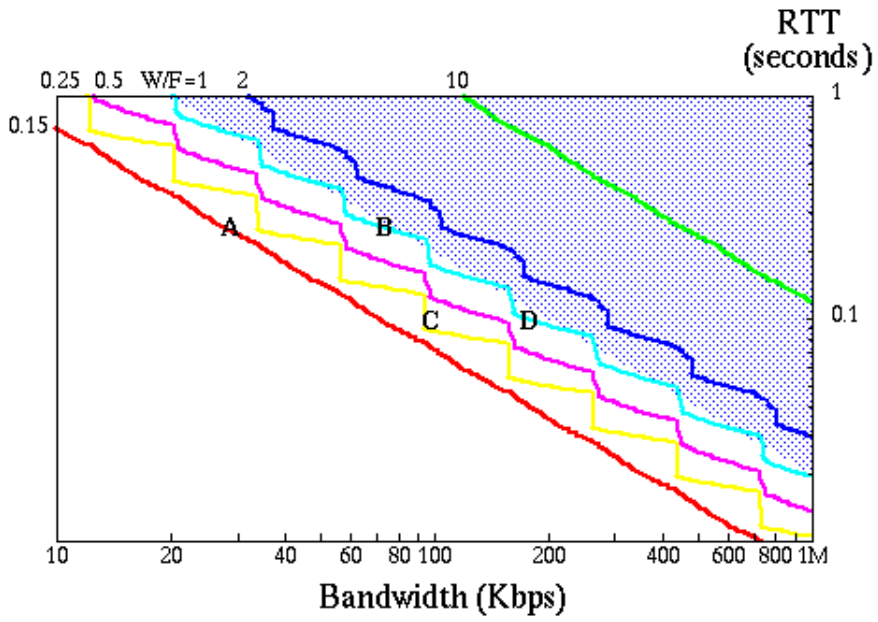

Figure 6: Effect of optimizations (for Internet MSS)

Contour plot of (wasted time)/(useful time)

Current modem links, at 28.8 Kbps and 250 ms total round-trip latency, have a waste ratio of only .15 (A). In this case, the overhead is 13% of the time of a 6 Kbyte file transfer. A waste ratio of 1.0 is achieved around 75 Kbps at 250 ms latency (B).

ISDN (112 Kbps) at 100 ms latency has a waste ratio of .24 (C). At 100 ms latency, 180 Kbps end-to-end links are the minimum required to approach the waste ratio of 1.0, i.e., the 2x speedup sought (D).

```
    MODEM (Internet MSS):
        R   = .250 s
        bw  = 28,800 bps
        MSS = 512 B = 4096 bits

        K   = 12 packets
        L   = 2
        M   = min(2, 12) = 2 packets
        S   = 0 rtts
        W   = .250 s
        F   = 1.71 s

        W/F = wasted time is 15% of optimal file transaction time

        100- 100*F/(W+F) = 13% benefit

    ISDN (Internet MSS):
        R   = .100 s
        bw  = 112,000 bps
        MSS = 512 B = 4096 bits

        K   = 12 packets
        L   = 3 packets
        L   = min(3, 12) = 3 packets
        S   = 1 rtt
        W   = .100 s
        F   = .42 s

        W/F = wasted time is 24% of optimal file transaction time

        100- 100*F/(W+F) = 19% benefit
```

We re-evaluated the graph for the case where MTU discovery is implemented, and packets contain the Ethernet-MSS (1460 bytes). In this case, the results of the optimizations are even less impressive. End-to-end rates of 150 Kbps are required at 250 ms latency (E), and 240 Kbps is required for 100 ms latency (F). ISDN at 100 ms gains only 19% from the optimizations. Going to higher MSSs, such as ATM (9 Kbytes) will further reduce the gains.

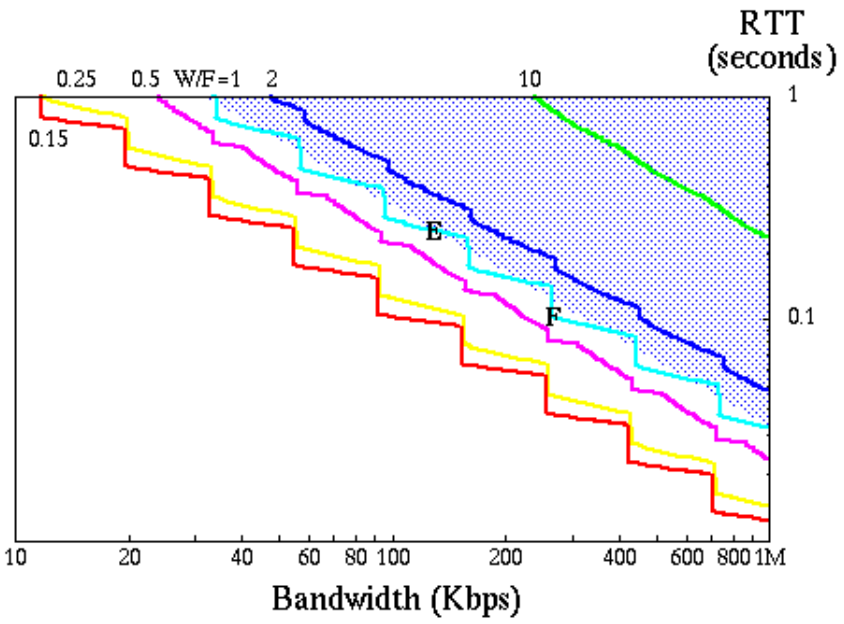

Figure 6: Effect of optimizations (for Ethernet MSS)

Contour plot of (wasted time)/(useful time)

```
ISDN (Ethernet to ISDN):
    R   = .100 s
    bw  = 112,000 bps
    MSS = 1460 B = 11680 bits

    K   = 5 packets
    L   = 1
    L   = min(.96, 4.11) = .96 = 1 packet
    S   = 1 rtt
    W   = .100 s
    F   = .42 s

    W/F =  wasted time is 24% of optimal file transaction time

    100- 100*F/(W+F) = 19% benefit
```

These observations indicate that the proposed optimizations do not affect current Web access for the vast majority of users. Most users see end-to-end latencies of about 250 ms and use modem lines. At these rates, the optimizations reduce the overall transaction time by 19%. Rates over 240 Kbps are required to provide user-noticeable performance.

Conclusions

These observations indicate that the persistent connection optimizations do not substantially affect Web access for the vast majority of users. Most users see end-to-end latencies of about 250 ms and use modem lines. At these rates, the optimizations reduce the overall transaction time by 15%. Bandwidths over 240 Kbps are required to provide user-noticeable performance improvements.

The optimizations require network and file characteristics that are not true for most users. Connection optimizations assume that the file is as small as the round-trip bandwidth-delay product, or smaller. Slow-start optimizations assume that there are a large number of packets in the round-trip, and that the overall number of packets is a small number of round-trips' worth. Neither of these assumptions hold for users over modem or ISDN lines accessing the vast majority of Web files.

In such cases, only 1-2 packets are typically in round-trip, negating the effects of slow-start optimizations. Typically, files are over 10x larger than the round-trip bandwidth-delay product, negating the effects of connection optimizations.

In the future, bandwidths are sure to increase. Packet sizes are also likely to increase, e.g., to 9 Kbytes for ATM, and MSS discovery should be more widely available. File sizes may increase as well. Given all three of these advances, it is not easy to predict the overall effect. This is discussed in further detail in ongoing work [web-trans].

Acknowledgements

We would like to thank the members of ISI's HPCC Division for their assistance with this page. Specifically, Ted Faber helped with the analysis section. This document was the result of discussions on the HTTP-NG and WEB-TALK maillists, and we also thank the members of those lists for their feedback.

References

[ftp]
Postel, J., and Reynolds, J., "File Transfer Protocol (FTP)," RFC-959 / STD-009, USC/ISI, October 1985.

[http]
Fielding, et al., "Hypertext Transfer Protocol - HTTP/1.1," (working draft), June 7, 1996.

[http-ng]
Spero, S., "Progress on HTTP-NG," (URL)

[p-http]
Mogul, J., "The Case for Persistent-Connection HTTP," ACM Sigcomm '95, August 1995, pp. 299-313. A longer, more comprehensive version of this paper is available on line at Digital Equipment Corporation Western Research Laboratory Research Report 95/4, May, 1995.

[http-lat]
Padmanabhan, V., Mogul, J., "Improving HTTP Latency," Proc. of the Second International WWW Conference, Oct. 1994.

[int-svc]
Clark, D., Shenker, S., and Zhang, L., "Supporting Real-Time Applications in an Integrated

Services Packet Network: Architecture and Mechanism," Sigcomm ’92, pp. 14-26.

[tcb]
Touch, J., "TCP Control Block Interdependence," (work in progress), USC/ISI, June 1996.

[ttcp-c]
Braden, R., "Extending TCP for Transactions -- Concepts," RFC-1379, USC/ISI, November 1992.

[ttcp-f]
Braden, R., "T/TCP -- TCP Extensions for Transactions: Functional Specification," RFC-1644, USC/ISI, July 1994.

[tcp]
Postel, J., "Transmission Control Protocol," RFC-793 / STD-007, USC/ISI, September 1981.

[tcp-ss]
Jacobson, V., "Congestion Avoidance and Control," ACM Sigcomm ’88, August 1988.

[web-analysis]
Spero, S., "Analysis of HTTP Performance Problems," (URL)

[web-slow]
Touch, J., Heidemann, J., and Obraczka, K., "Why the Web is Slow," (in progress).

[web-transp]
Heidemann, J., Obraczka, K., and Touch, J., "Analysis of HTTP Transport Protocols," (in progress).

[web-why-slow]
Moskowitz, R., "Why in the World is the Web So Slow?" Network Computing, March 15, 1996, pp. 22-24.

[www]
Berners-Lee, T.J, R. Cailliau and J.-F. Groff, The World-Wide Web, Computer Networks and ISDN Systems 25 (1992) 454-459. Noth-Holland.

Go back to the LSAM home page. / Go back to the ISI home page.

This page written and maintained by the LSAM Group.
Please mail any problems with or comments about this page.
Last modified June 24, 1996.