Appendix 1

Latency

ADTRAN Submission to NTIA/RUS 4-13-09

Introduction

In its recent Notice of Inquiry (FCC 09-31) [1], the FCC seeks comment "to inform the development of a national broadband plan for our country." One of the primary topics for which comment is solicited is "Defining Broadband Capability," for example, whether broadband should be defined "in terms of bandwidth and latency" or by other metrics related to user experience.

Both bandwidth and latency significantly affect the user's experience. Bandwidth has been addressed in a separate ADTRAN submission analyzing the relationship between different definitions of "speed" and user experience [2]. Speed, however, is only one factor affecting the quality of the user experience. The latency associated with the network connection is equally important, and in many cases more so, for interactive applications requiring response times that should be perceived by the user as instantaneous. Even for non-real time applications such as web browsing, small additions to network latency can have a multiplicative effect that results in latency, and not speed, frequently being the dominant factor in web page download times.

This appendix summarizes the requirements that have been specified for delay (or latency) by a number of standards development organizations. The requirements are shown both from the application perspective and from the network perspective. In addition, the appendix includes a brief discussion of the disproportionate effect of network latency on web browsing.

1 User requirements

The perceived quality of a user's experience related to response time has been classified into several perceptual regions by Nielson (bullets reprinted from [3]):

- **0.1 second** is about the limit for having the user feel that the system is reacting instantaneously, meaning that no special feedback is necessary except to display the result.
- **1 second** is about the limit for the user's flow of thought to stay uninterrupted, even though the user will notice the delay. Normally, no special feedback is necessary during delays of more than 0.1 but less than 1.0 second, but the user does lose the feeling of operating directly on the data.
- **10 seconds** is about the limit for keeping the user's attention focused on the dialogue. For longer delays, users will want to perform other tasks while waiting for the computer to finish, so they should be given feedback indicating when the computer expects to be done. Feedback during the delay is especially important if

the response time is likely to be highly variable, since users will then not know what to expect.

Other sources, including Miller [4] and Cheshire [5], have also identified 100 msec as a threshold response time, below which a user will perceive that the response is virtually instantaneous. The International Telecommunications Union (ITU) has categorized a set of QoS categories along lines similar to [3] in Recommendation G.1010 [6]. Figure 2/G.1010 is reprinted here as Figure 1.

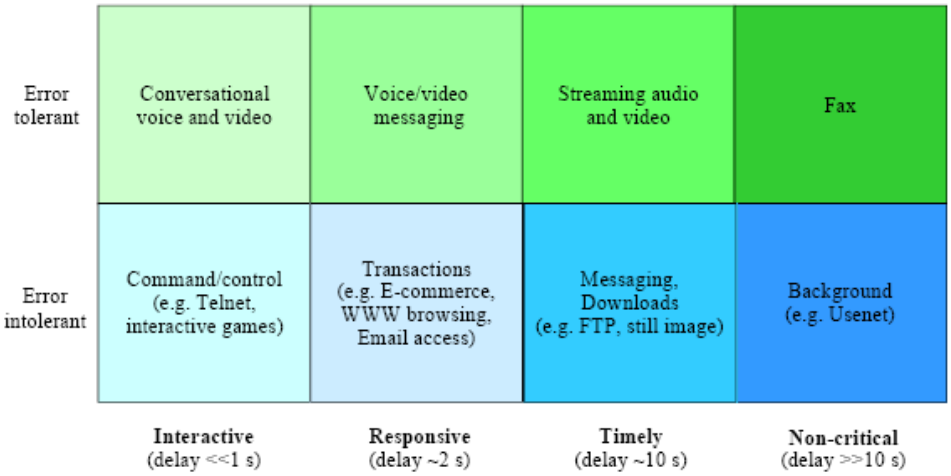

Figure 1 – Model for user-centric QoS categories (reprinted from [6])

Since the interactive applications shown at the left end of Figure 1 represent the most challenging requirements for latency in High Speed Internet Access (HSIA) deployments, this paper focuses on them (with the exception of telemetry and Telnet, which are not considered residential applications). In addition, we discuss web browsing, since (as will be shown) small increases in network latency can have disproportionately large effects on download times for many web pages.

One way delay (or latency) requirements as defined by the ITU, the 3rd Generation Partnership Project (3GPP), and the Broadband Forum, are consolidated in Table 1 for the relevant applications in Figure 1.

Table 1 – Response time requirements

Application	One way delay	Sources
Conversational voice	< 150 ms preferred < 400 ms limit	G.1010 [6], TS 22.105 [7]
	< 150 ms	TR-126 [8]
Videophone	< 150 ms preferred < 400 ms limit	G.1010, TS 22.105
Interactive games	< 200 ms	G.1010, TR-126
	< 75 ms preferred	TS 22.105

Application	One way delay	Sources
	< 50 ms (objective)	TR-126
Web browsing	< 2 s/page preferred < 4 s/page acceptable	G.1010, TR-126
	< 4 s/page	TS 22.105

2 Network latency requirements

ITU Recommendation Y.1541 [10] defines performance objectives for the network that complement the user-driven performance requirements defined in G.1010. Y.1541 defines a total of eight "QoS class definitions" (two of which are provisional) which define performance objectives for IP packet Transfer Delay (IPDT), IP packet Delay Variation (IPDV), IP packet Loss Ratio (IPLR), and IP packet Error Ratio (IPER).

Table 2/Y.1541 provides guidance linking the QoS classes with applications and routing distances. In that table, QoS classes 0 and 1 are recommended for "real time, jitter sensitive, high interaction" applications such as conversational voice, videophone, and interactive games. Within that application set, QoS class 0 is recommended for networks with "constrained routing and distance," and QoS class 1 is recommended for networks with "less constrained routing and distances." QoS class 5 is recommended for "traditional applications of default IP networks" such as web browsing. Other QoS classes are recommended for applications such as signaling and video streaming.

The performance objectives for the non-provisional QoS classes are defined in Table 1/Y.1541. The specific objectives for IPTD and IPDV are reproduced here as Table 2.

Table 2 – QoS class performance objectives (from [10])

Network performance parameter	Nature of network performance objective	QoS Classes					
		Class 0	Class 1	Class 2	Class 3	Class 4	Class 5 Unspecified
IPTD	Upper bound on the mean IPTD	100 ms	400 ms	100 ms	400 ms	1 s	U*
IPDV	Upper bound on the $1 \cdot 10^{-3}$ quantile of IPTD minus the minimum IPTD	50 ms	50 ms	U	U	U	U

*U = Unspecified

2.1 Web browsing

Compared to the interactive requirements in Table 1, the requirements for web browsing would seem significantly less stringent. Experience, however, shows that many web pages take longer than 4 seconds to download, even over high speed access links. The reason for this has as much or more to do with network latency as with speed.

Most web pages are composed of a number of objects, including text, graphics, and applets. When a web page is accessed, the first object requested is the base file for the page. That file provides directions for accessing other objects. Some of those objects may point to yet other objects. Each object must be requested with a separate HTTP "Get" command and retrieved via a TCP connection. There are limits in most consumer operating systems on how many concurrent TCP connections may be opened, so only so many objects can be downloaded in parallel.

Each HTTP command, and each TCP connection, generates at least one sequence of messages between the client and server that requires receipt of the previous message before the response can be transmitted. Each of these sequences requires a round trip through the network, or a "turn," to complete.[1] The number of turns required to download a web page was incorporated into a formula for download time originally developed by Sevcik and Bartlett [11] and simplified in an article by Savoia [12]. The simplified version is shown below:

$$R \approx \frac{Size}{Bandwidth} + Turns \cdot RTT + Cs + Cc \ . \qquad (1)$$

Where: R = the approximate response time, or total time to download the web page,
$Size$ = the total amount of data to be transferred,
$Bandwidth$ = the effective speed of the connection between client and server,
$Turns$ = the effective number of turns required to download the page,
RTT = the round trip time of the connection (upload plus download latency),
Cs = the server processing time, and
Cc = the client processing time.

An example web page download sequence is shown in Figure 2. This figure is based on an actual download of a simple web page containing the base file and a single JPEG image [13]. The figure, simplified from the actual download sequence, shows the following turns:

1. The turn required to establish the initial TCP connection with the web site server.
2. The turn associated with the HTTP Get command requesting the base file.
3. An additional turn, associated with TCP "slow start," required before download of the base file is complete.
4. The turn associated with the HTTP Get command requesting the JPEG image.

Additional turns (and partial turns) associated with TCP "slow start" occur during download of the JPEG image, but are not shown in the figure for the sake of simplicity. In all, between 9 and 10 turns (at RTT of approximately 70 ms) were required to load this page and the resulting accumulated latency took 656 ms out of a total 736 ms for the download. In this simple example, almost 90% of the total response time was due to network latency.

[1] Additional turns may be required even before the HTTP transaction begins – for instance, if the client must request the IP address for the page from a DNS server.

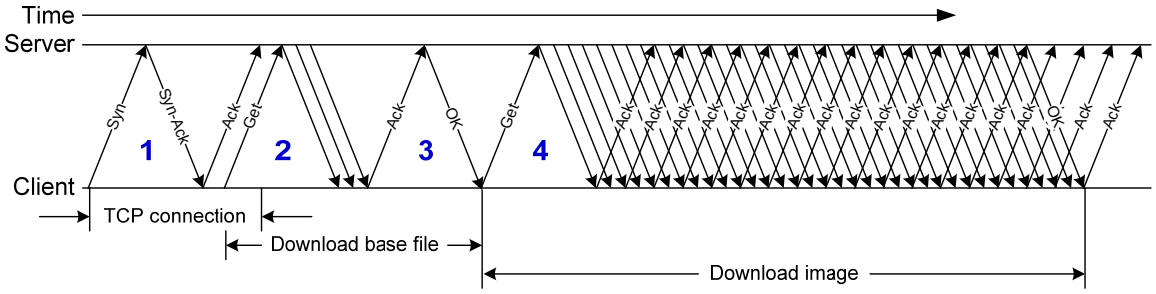

Figure 2 – Simple web page download sequence

In [11], the authors note that a typical Keynote Business-40 web site[2] requires 40 turns. With a RTT of 100 ms, such a site requires a minimum of 4 seconds to download, even with infinite bandwidth and zero processing time at client and server.

There are techniques for optimizing the design of networks and web sites, including caching content closer to the users and optimizations in HTTP and other protocols, which mitigate the latency issue. Caching of content closer to the user reduces, but doesn't eliminate, network latency. Opening multiple concurrent TCP connections reduces the effective number of turns required, as does intelligent web page design. Given the overwhelming variety of web pages on the Internet, however, the importance of network latency to web page response time is likely to remain disproportionately high for the foreseeable future.

3 Summary

Performance objectives for both application response time and network latency, as defined by several standards development organizations, are summarized in this appendix. The list of applications includes those which are likely to be accessed via residential broadband, and for which latency is a significant factor. The applications listed include conversational voice, conversational video, and gaming, highly interactive applications for which the user's quality of experience demands a perception of near instantaneous response time. Perhaps surprisingly, the applications also include web browsing, which is not considered a real time or interactive application. It is shown, however, that the disproportionate effect of small increases in network latency on web browsing response times justifies the application's inclusion in this list.

4 References

[1] GN Docket No. 09-51 [FCC 09-31], released April 8, 2009

[2] Adtran, "Defining Broadband Speeds: an Analysis of Peak vs. Sustained rates in Network Access Architectures," April 2009.

[3] Nielsen, J., "Response Times: The Three Important Limits," excerpt from *Usability Engineering*, Morgan Kaufmann, San Francisco, 1994.
Excerpt available at http://www.useit.com/papers/responsetime.html.

[2] The Keynote Business-40 is an index which tracks the performance of 40 leading business web sites.

[4] Miller, R. B., "Response time in man-computer conversational transactions," Proc. AFIPS Fall Joint Computer Conference**,** Vol. 33, pp 267-277.

[5] Cheshire, S., "Latency and the Quest for Interactivity," November 1996, available at http://www.stuartcheshire.org/papers/LatencyQuest.html.

[6] ITU Recommendation G.1010, "End-user multimedia QoS categories," November 2001.

[7] 3GPP TS 22.105 V9.0.0, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Service aspects; Services and service capabilities (Release 9)," December 2008.

[8] Broadband Forum, Technical Report TR-126, "Triple-play Services Quality of Experience (QoE) Requirements," 13 December 2006.

[9] ITU Recommendation G.114, "One-way transmission time," May 2003.

[10] ITU Recommendation Y.1541, "Network performance objectives for IP-based services," February 2006.

[11] Sevcik, P., and Bartlett, J., "Understanding Web Performance," NetForecast Report 5055, available at http://www.netforecast.com/Reports/NFR%205055%20Understanding%20Web%20Performance.pdf

[12] Savoia, A., "Web Page Response Time 101," STQE Magazine, Vol. 3, Issue 4, July/August 2001, pp. 48-53, available at http://www.webperformancematters.com/papers-and-talks/performance-management/Web%20page%20response%20time%20101%20Savoia%20STQE%202001%20.pdf

[13] www.chineseviking.com