An ADTRAN White Paper

Defining Broadband: Network Latency and Application Performance

Defining Broadband: Network Latency and Application Performance

Executive Summary

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.

Adtran has provided comments [2] noting that both bandwidth and latency (or delay) significantly affect the user's Quality of Experience (QoE). Bandwidth has been addressed in detail in separate submissions. This paper addresses delay, and provides the following information:

- Multiple Standards Development Organizations (SDOs) provide guidance regarding delay limits for real time, near-real time, and other network applications. The applicable requirements, and the Quality of Service (QoS) classes that have been defined to support them, are summarized.
- Delays in the access and core segments of the network have different sources. Delay in the core is due to speed-of-light propagation and to switching/routing delays. Delay in the access network is due to factors related to speed and to the technologies and protocols used to bridge the "last mile" to the subscriber. Each segment must be considered separately to understand its contribution to end-to-end delay.
- Despite being a non-real time application and having requirements measured in seconds rather than milliseconds, web browsing performance suffers relative to its SDO requirements more than any other application. The primary reason is network delay, which has a multiplicative effect on response times due to the many round trips required to load the average web page. Without improvements in the underlying causes for multiple round trips (protocols and web site design practices), web browsing performance will not improve significantly no matter how much access speeds continue to improve.
- With recommended delay limits of 35 ms upstream and 15 ms downstream in the access network, QoS-enabled services such as "home phone" services can provide delay and jitter performance meeting SDO preferred requirements in continental connections and limit requirements in intercontinental connections.
- Assuming typical performance comparable to the limits recommended above for QoS-enabled services, real time applications (VoIP, conversational video, gaming) running over best effort Internet access services should typically perform close to or at the levels specified in the SDO requirements. Best effort services cannot provide

guaranteed delay, however, and performance will be strongly dependent on other traffic in the network.

1 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.

Bandwidth and latency (or delay) both significantly affect the user's Quality of Experience (QoE). Bandwidth has been addressed in a separate ADTRAN white paper analyzing the relationship between different definitions of "speed" and user experience [3]. This paper addresses network delay, beginning with a summary of the requirements that have been specified for delay by a number of standards development organizations. It then breaks down those requirements by network segment and by application.

In the case of web browsing, the multiplicative effects of delay on response time are examined and it is shown that delay, rather than speed, is frequently the dominant factor in web page response times. Real-time applications such as VoIP are then analyzed to allocate delay between applications, core network delay, and access network delay. The results are then summarized to generate recommendations for access network delay for broadband access. We find that allocating 35 ms to upstream traffic and 15 ms to downstream traffic in the access network supports VoIP when offered as a managed service. Other interactive applications (which are generally provided over best effort services) will also generally work well with the above allocations, but those delays are highly dependent on congestion and cannot be guaranteed. The same values also support web browsing response time limits as long as other factors such as local caching and/or web page optimizations are applied.

2 Terminology

The terms **latency** and **delay** tend to be used interchangeably in the industry, and the two terms are equated in ITU Recommendation G.114 [4]. In this document, both terms refer to the one way delay between two points, which may or may not be within the IP transmission path (for example, a talker's mouth to a listener's ear). ITU Recommendation Y.1540 [5] introduces the term **IP packet Transfer Delay (IPTD)** as the time between packet ingress to and packet egress from an IP network or a section thereof. When the applicable measurement points are connected to an IP network, latency, delay, and IPTD are all interchangeable.

Unless otherwise noted, all references to IPTD in this document refer to the arithmetic mean value for the parameter. ITU Recommendation Y.1541 [6], section 5.3.2 discusses measurement intervals and recommends intervals on the order of one minute for telephony applications.

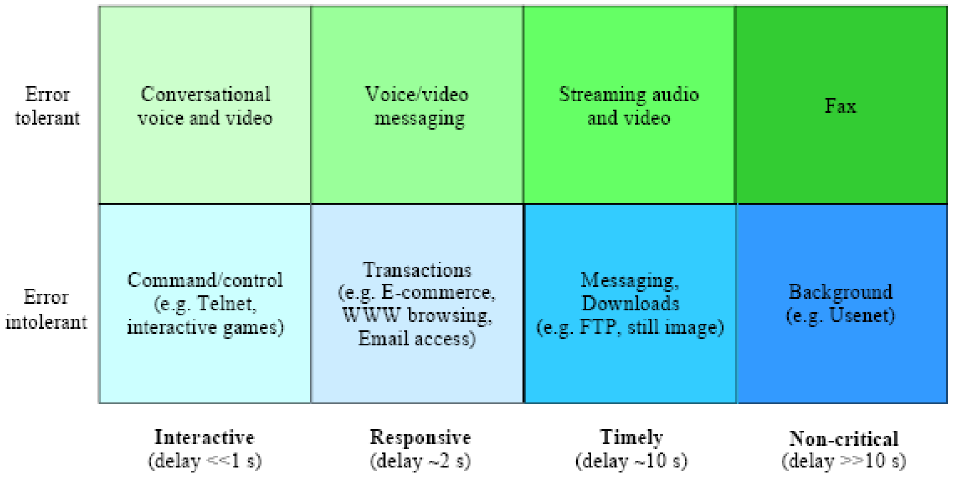
Jitter (or **IP packet Delay Variation [IPDV]** in the ITU Y series of recommendations), is the peak-to-peak variation in IPTD. Y.1541 specifies the quantiles at which IPDV is defined for different "QoS classes."

3 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 [7]):

- **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 [8] and Cheshire [9], 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 [7] in Recommendation G.1010 [10]. Figure 2/G.1010 is reprinted here as Figure 1.



Error tolerant	Conversational voice and video	Voice/video messaging	Streaming audio and video	Fax
Error intolerant	Command/control (e.g. Telnet, interactive games)	Transactions (e.g. E-commerce, WWW browsing, Email access)	Messaging, Downloads (e.g. FTP, still image)	Background (e.g. Usenet)
	Interactive (delay <<1 s)	**Responsive** (delay ~2 s)	**Timely** (delay ~10 s)	**Non-critical** (delay >>10 s)

Figure 1 – Model for user-centric QoS categories (reprinted from [10])

Since the interactive applications shown at the left end of Figure 1 represent the most challenging requirements for delay 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 delay can have disproportionately large effects on download times for many web pages.

One way delay 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 [10], TS 22.105 [11]
	< 150 ms	TR-126 [12]
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
	< 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

4 Network delay requirements

ITU Recommendation Y.1541 [6] 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 IPDT and IPDV, as well as objectives for lost and errored packets.

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 [6])

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 $1\cdot10^{-3}$ quantile of IPTD minus the minimum IPTD	50 ms	50 ms	U	U	U	U

*U = Unspecified

4.1 IPTD and IPDV

The notes accompanying Table 1/Y.1541 state that “For planning purposes, the bound on the mean IPTD may be taken as an upper bound on the minimum IPTD and, therefore, the bound on the $1\cdot10^{-3}$ quantile may be obtained by adding the mean IPTD and the IPDV value (e.g., 150 ms in Class 0).” This implies that the minimum and mean IPTD values should be close to each other and should not be affected significantly by IPDV. For networks that are not suffering from excessive congestion, this tends to be the case. Figure 2 shows most of the response times from 1000 pings to google.com (three additional times, at 1109, 625, and 1375 ms, are literally off the chart).

We can estimate minimum and mean one way delays at half the round trip times, or approximately 27 and 31 ms respectively. The outlier delay values, however, are primarily due to congestion in one direction. The best estimate of the maximum one way delay from this sample is the maximum RTT minus the mean one way delay, or approximately 1344 ms, making the IPDV 1344-27=1317 ms. Using the Y.1541 guidance, the bound on the $1\cdot10^{-3}$ quantile is 1317+31=1348 ms, very close to the sample estimate.

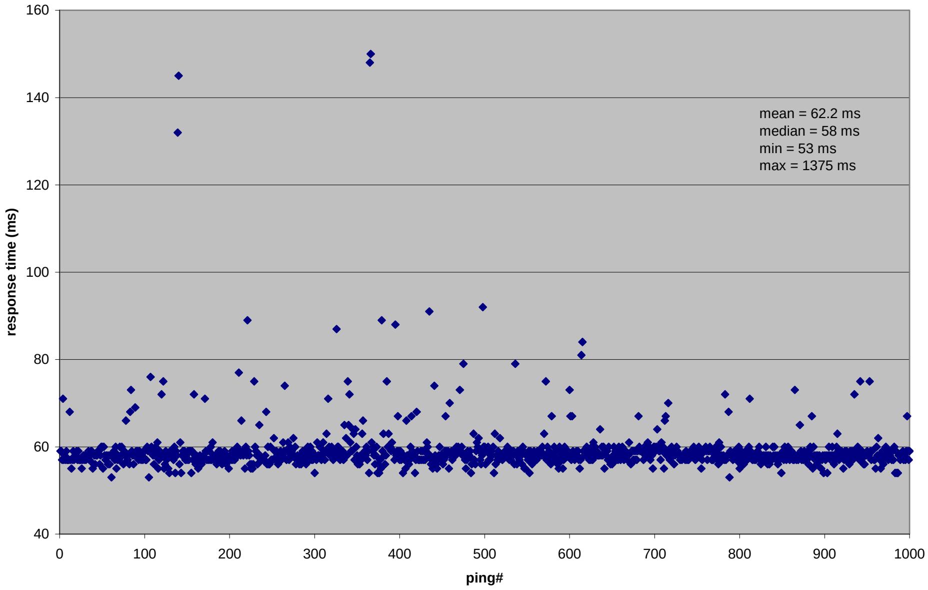


Figure 2 – Ping response times

5 Network delay

Delay in the network can be broken down into a number of components, depending on the type of application. For the purposes of this paper, the network itself is broken into two components. The first component is the access network, through which the user connects to the Internet. The second component is the core network, or the Internet backbone providing transport between the user's access network and the destination (or source, depending on direction) of the traffic.

Depending on the specific application, traffic may transit either one or two access networks in its end-to-end path. For most client-server applications, where the server is a commercial host connected to the Internet via a low-delay path, the server can be considered to be connected directly to the core network and only one access network transit applies. For point-to-point applications such as video telephony, traffic will typically transit access networks at each user location, as well as the core. The data path is discussed in more detail in the individual application sections below.

5.1 Core network delay

Several network providers [13, 14] provide online public access to round trip delay statistics on their networks. At the time of writing, the average round trip delay on continental US routes for the AT&T network is about 37 ms, and the average round trip delay on North American routes (which may include non-US destinations) for Global Crossing's network is about 43 ms. The maximum round trip delay on AT&T's network performance site (which includes statistics between specific US cities) is 84 ms, between Orlando and Seattle. This is roughly twice the theoretical speed-of-light minimum, which seems reasonable given non-straight fiber routing and delays through network devices.

With the above numbers in hand, and adding a few ms for margin, we can reasonably state that worst case one way (not round trip) core network delay can be estimated at 50 ms for routes within the continental United States. Taking the mean round trip values provided by the sources and dividing by two, we can estimate the average one way core network delay at approximately 20 ms.

Core network latencies for intercontinental routes can be significantly longer than for intra-continental routes. At the time of writing, the reported round trip delay from New York to London on AT&T’s network was 77 ms, on a par with New York to San Diego. The reported round trip delay from San Francisco to Hong Kong was 159 ms. Routes from the East coast to Asia, or from the West coast to Europe, would of course be even longer.

5.2 Access network delay

Most delay on access networks is not caused by speed-of-light propagation over very long distances.[1] Instead, it is caused by a number of factors that tend to be unique to the access infrastructure.

- The “last mile” of the access network (that is, the connection closest to the user) frequently operates at speeds significantly slower than the core network. Serialization delay, or the time it takes to transmit all bits of a packet from first to last, can become a significant component of IPTD at access network speeds. For a constant connection speed, the serialization delay of a packet is proportional to its size, so if the traffic contains variable packet sizes serialization delay will also contribute to IPDV.

- Some access technologies (for example, wireless access networks) can rapidly adapt to changing transmission path loss and noise conditions by adapting the transmission rate and/or changing the allocated part of the bandwidth used by a particular subscriber. These changes cause variation in serialization delay and contribute to IPDV.

- Channel sharing protocols such as Time Division Multiple Access (TDMA) and Time Division Duplexing (TDD) add to both IPTD and IPDV as transceivers wait their turn to send data.

- Many access technologies use some form of error correction to reduce the error rate on noisy channels. In addition, interleaving of data may be enabled to maximize the effectiveness of the error correction. Both of these techniques introduce IPTD.

- Some access technologies make use of local retransmission to reduce errors on noisy channels. Packets for which data must be retransmitted are delayed more than other packets, adding to IPDV.

[1] One notable exception is access via geostationary satellite networks, for which latency is dominated by speed-of-light propagation over the 44,000 mile round trip between earth and the satellite.

- Queuing delays add to both IPTD and IPDV in both the access and core networks. Due to the lower data rates involved, queues in the "last mile" contribute a larger share to the overall network delay than do queues in the core for similar levels of congestion. For instance, the servicing time required for a single 1500-byte packet is 12 ms at 1 Mbps, but only 12 µs at 1 Gbps.

6 Network delay by application

In this section we develop recommendations for delay in the access network by starting at the application-level delay requirements (for applications delivered over services using QoS) and subtracting out the delay of other known components. We then test the applicable values from those recommendations against applications that are delivered over the best effort HSIA service. The applications we consider here are: web browsing, conversational voice (both QoS-based and best effort), conversational video, and interactive games.

Any recommendations for IPTD and IPDV in the access network must be applied to specific services, and not to the access channel as a whole. For instance, IPTD and IPDV can be managed by the access provider in a QoS-based voice service providing a "home phone" application, separate from best effort HSIA. In this case, the access provider marks VoIP traffic with a higher Class of Service (CoS), and that traffic is given priority over best effort traffic across the access network. Non-local VoIP traffic is then routed across a network designed for voice, rather than competing with best effort traffic on the Internet.

Consider the second variety of VoIP, which might be described as "Internet VoIP." In this case, the VoIP traffic accesses the Internet as best effort data using the HSIA service. The access provider is usually not aware that the traffic is associated with VoIP, and provides no priority for it. This "Internet VoIP" traffic can be delayed by other best effort traffic generated by the same user – for instance, web browsing or large file transfers. It can also be delayed by best effort traffic from other users in the access network. In this case – which applies to all applications delivered over the best effort HSIA service – IPTD and IPDV cannot be constrained with any guarantee.[2] The access and core networks may perform well enough most of the time to support delay sensitive applications but, as shown in Figure 2, there will be moments when best effort packets suffer from excessive delay.

6.1 Web browsing

Starting an examination of application latency with web browsing may seem unusual given that it isn't normally considered a real-time application. However, it is one of the most common residential applications accessed over HSIA and the user's QoE frequently fails to meet the requirements in Table 1. Experience shows that many web pages take

[2] At least one access provider has attempted to address this issue by providing a control panel to allow users to customize their broadband connection by application [15]. There is also at least one software product providing the same type of customization for individual PCs [16]. It isn't clear that either of these products provides QoS across the network.

longer than 4 seconds to download, even over high speed access links. The reason for this has more to do with network delay than with speed.

6.1.1 Response time factors

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.[3] The number of turns required to download a web page was incorporated into a formula for download time originally developed by Sevcik and Bartlett [18] and simplified in an article by Savoia [19]. 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[4],
$Turns$ = the effective number of turns required to download the page,
RTT = the round trip time of the connection (upload plus download delay),
Cs = the server processing time, and
Cc = the client processing time.

An example web page download sequence is shown in Figure 3. This figure is based on an actual download of a simple web page containing the base file and a single JPEG image [20]. 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.

[3] 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.

[4] Since TCP transfer rates can be affected by the network's bandwidth-delay product, the RTT can have a secondary effect in Equation 1 by limiting the effective value of *Bandwidth*.

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 delay 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 delay.

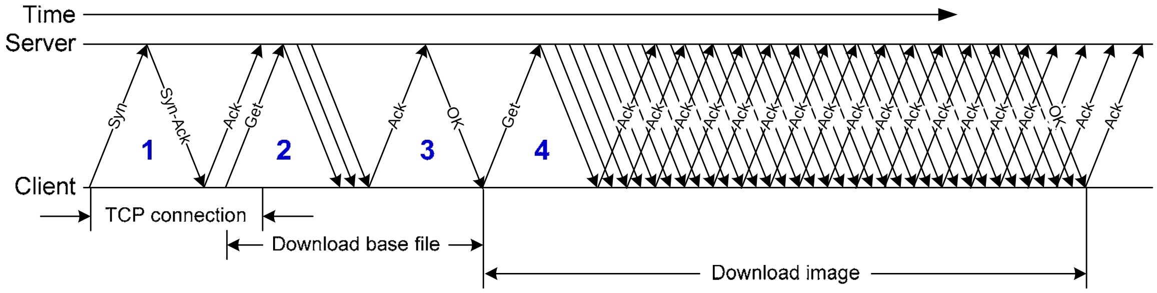


Figure 3 – Simple web page download sequence

6.1.2 Web page statistics

In [18] (written in 2001), the authors note that a typical Keynote Business-40 web site[5] requires 40 turns and contains approximately 100 kbytes of data. In an attempt to both modernize the data and apply it to sites of interest to consumers, transactions were recorded while browsing to the 25 most popular web sites for March 2009 as reported by ranking.com [21] (see the Appendix for details). The resulting parameters are shown in Table 3.

Table 3 – Parameters for ranking.com top 25 web sites

Grouping	Average size (kbytes)	Average effective turns	Average client processing time* (s)
Search engines	71.4	19.5	1.17
Portals	481	58.4	1.55
All	339	44.8	1.42

*Client processing time is client hardware- and software-dependent.

The measured parameters placed the sites into two distinct groupings. The search engines are distinguished by efficiency and fast load times, with relatively few objects per page. The portal pages are much larger both in terms of raw size and number of objects, with visual interest taking precedence over response time.

6.1.3 Response time analysis

Using the above data and Equation 1, graphs for web page response times are shown in Figure 4 plotted against rate for different values of RTT. The following parameters were used:

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

- Cs = 100 ms, Cc = 1 s
- Size = 340 kbytes
- Turns = 45

Equation 1 was modified slightly to account explicitly in the graphs for serialization delays, which change with download and upload rates (elsewhere in the paper, serialization delays are included in the overall delay figures for the network segments). The modified equation is

$$R \approx \frac{Size}{R_d} + Turns \cdot \left(RTT' + \frac{P_d}{R_d} + \frac{P_u}{R_u} \right) + Cs + Cc \,. \tag{2}$$

Where: P_d = 1500 bytes = the download packet size,

P_u = 128 bytes = the upload packet size,

R_d = the download rate, and

R_u = the upload rate.

The figure also shows the lower limits on response time performance due to increasing delay. For instance, if the RTT exceeds 20 ms, no finite rate allows the response time to meet the preferred value of 2 s. In general, download speeds beyond about 10 Mbps make little difference in response time for average web pages, but reducing delay has a significant effect on response time for all rates over a few hundreds of kbps.

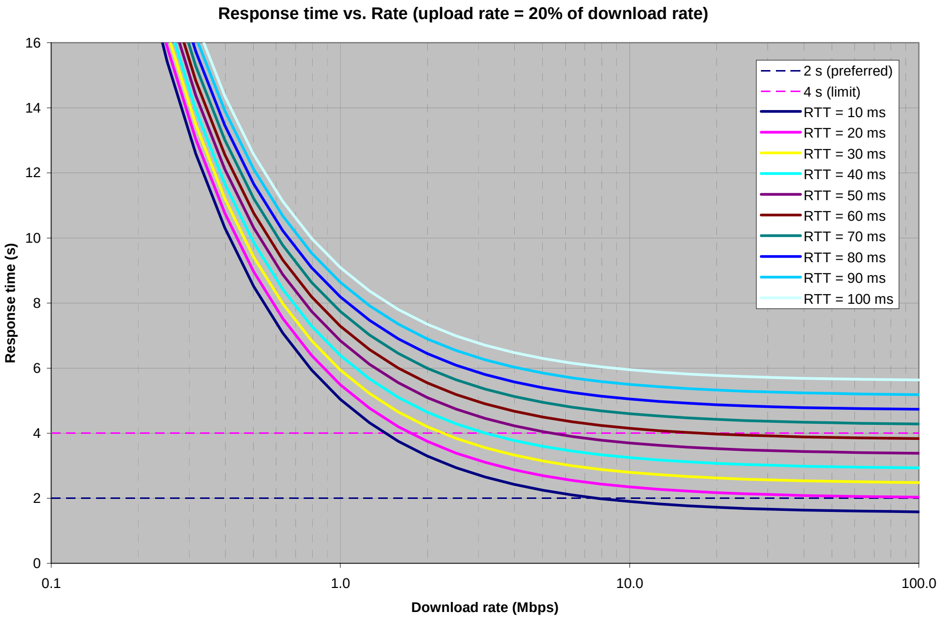


Figure 4 – Response times for average web pages

It may be helpful to generalize these results in order to develop rules of thumb or intuition that survive particular choices in the *Size* and *Turns* parameters. To do this, we need to simplify the equation used to generate Figure 4, repeated below:

$$R \approx \frac{Size}{R_d} + Turns \cdot \left(RTT' + \frac{P_d}{R_d} + \frac{P_u}{R_u} \right) + Cs + Cc \,. \tag{2}$$

If $RTT' >> \frac{P_d}{R_d} + \frac{P_u}{R_u}$ which occurs if $R_d >> \frac{P_d}{RTT'}$ and $R_u >> \frac{P_u}{RTT'}$, we can simplify our equation. With P_d = 1500 bytes = 12000 bits, P_u = 128 bytes = 1024 bits, and RTT_{min} = 20 ms, R_d should be much greater than 12000*50 = 600 kbps and R_u should be much greater than 1024*50 = 51.2 kbps, which we achieve with minimum rates of 6 Mbps down and 512 kbps up.

This reduces our equation to

$$R \approx \frac{Size}{R_d} + Turns \cdot RTT + Cs + Cc \,, \tag{3}$$

which uses the original *RTT* (inclusive of the serialization delays).

Looking at this equation, the two things we can affect with the access network are the download speed (R_d) and *RTT*, which is a function of the transport delay. By holding *R* constant at the 2 and 4 second levels from Table 1, we can see how R_d and *RTT* interact.

Because the interaction changes with the size of the web page and the number of turns required, we try to normalize our factors and look at the effect of relative changes in the two terms. We define a normalized download data rate R_N

$$R_N = \frac{R_d\,(bps)}{Size_{bits}} \tag{4}$$

and total network delay (which is proportional to *RTT*)

$$D_N = Turns \cdot RTT \,. \tag{5}$$

Then we can write the expression

$$D_N = R_{\max} - \frac{1}{R_N} - Cs - Cc \,, \tag{6}$$

where R_{max} is the maximum allowed response time (2 or 4 seconds).

Plotting this equation in Figure 5 with R_{max} = 2 and R_{max} = 4 and using Cs + Cc = 1.1 seconds, we can see that there are three basic regions, based on the normalized rate R_N.

Region 1: Once D_N is sufficiently low, additional decreases in network delay have little effect on response time. In this region, the response time contribution from D_N is under 200 ms and the value of R_N dominates the ability to meet the R_{max} requirement.

Region 2: When R_N is above a threshold R_H = 5, the value of D_N dominates the ability to meet the R_{max} requirement and R_N has little effect. In this region, the response time contribution from R_N is less than 200 ms.

Region 3: When neither of the above conditions is true, rate and delay both make significant contributions to response time. The values can be traded off against each other to meet a response time specification.

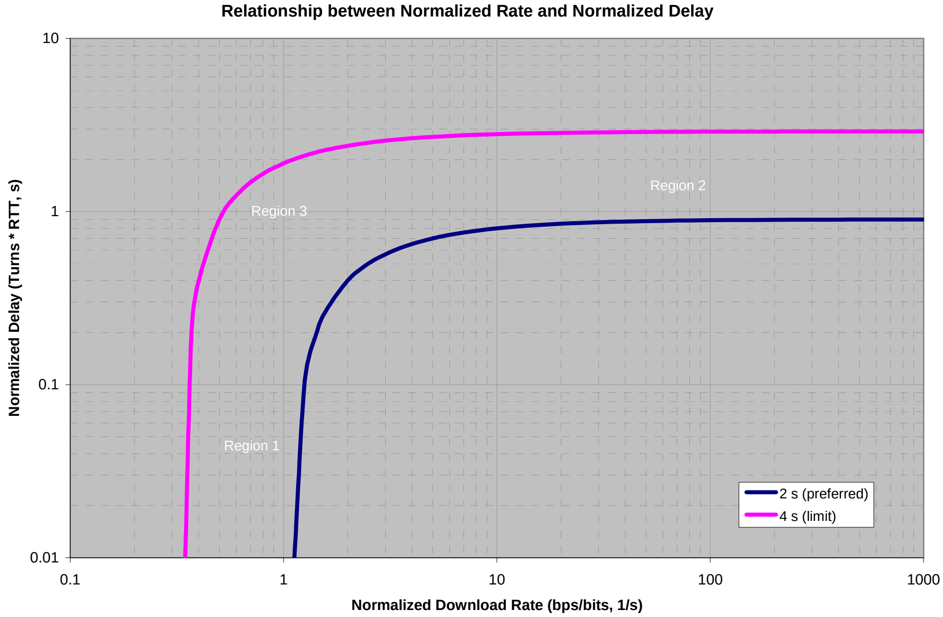


Figure 5 – Normalized rate vs. delay

De-normalizing the above parameters, the response time becomes relatively insensitive to download rate once $R_d > 5 \cdot Size_{bits}$. For the parameters in Table 3, the threshold occurs at about 14 Mbps. Response time becomes insensitive to RTT once $RTT < 0.2/Turns$. Turning once again to the averages in Table 3, the low delay threshold is met at RTT = 4.5 ms.

While download rates may or may not exceed 14 Mbps depending on the access network and other factors, the round trip time will virtually never be less than 4.5 ms due to speed of light limitations. So, network delay will be a significant factor in web response times as long as web page design and protocols continue to require multiple turns.

6.1.4 Average response time

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 delay issue. Caching of content closer to the user reduces, but doesn't eliminate, network delay. Opening multiple concurrent TCP connections reduces the effective number of turns required, as does thoughtful web page design. The overwhelming variety of web pages on the Internet, and the wide variety of options for designing both those pages and the networks that access them, make it impractical to associate a specific threshold for network delay with a corresponding response time requirement. However, we can check performance using the following assumptions and delay values borrowed from Table 3:

- Cs = 100 ms, Cc = 1 s

- Size = 340 kbytes
- Turns = 45
- Caching of content near the access network (no core network delay)
- Access network round trip delay = 50 ms (including serialization delay)
- Download rate = 10 Mbps

Plugging the above values into Equation 1 provides a response time of 3.62 seconds. While this doesn't meet the preferred value of 2 seconds, it is within the acceptable limit value of 4 seconds.

Applying these average values to the normalized analysis in section 6.1.3, we get the following contributions to response time:

- Normalized Rate = 10 Mbps /(340*8 kbits) = 3.67 1/s
- Normalized Delay = 45 * 50 ms = 2.250 s

This point is in Region 3 of Figure 5, close to the boundary where region 3 meets Region 2. Note that at the normalized delay value, the 2 second preferred response time cannot be met at any download rate.

6.1.5 Future trends

In section 6.1.2, we noted that the average business web page required about 40 turns and contained about 100 kbytes of data in 2001, and that the average popular consumer site requires about 45 turns and contains about 340 kbytes of data in 2009. While business and consumer web sites may have different statistics, comparing the above values provides little confidence that either web page size or effective number of turns is on a decreasing trend. Unless widespread adoption of both protocol and web site design improvements reduces the turns requirement, latency will continue to dominate web browsing response times.

6.2 QoS-based VoIP

Many access providers offer VoIP as a "home phone" service apart from the HSIA service. This service is frequently marketed without any specific reference to VoIP, and customers' expectations for performance are comparable to those for traditional phone service. In this case, the VoIP traffic is provided with a different Class of Service (CoS) and given priority over best effort traffic across both the access and core networks.

6.2.1 Network segments

VoIP calls can be either VoIP-to-VoIP or VoIP-to-PSTN. A VoIP-to-VoIP call is shown in Figure 6. In this case, the VoIP traffic transits two access networks – one at each end of the connection – in addition to the core network. A VoIP-to-PSTN call, in which the traffic transits only a single access network, is shown in Figure 7. Other than one vs. two access networks, there are no significant differences between Figure 6 and Figure 7 with regard to delay. Both routes require the same encoding, decoding and buffering, although in Figure 7 some of those functions take place in the VoIP-to-PSTN gateway rather than a second VoIP terminal. For similar sources and destinations, the sum of the latencies in the Core IP network and the PSTN in Figure 7 will approximately equal the Core IP network delay in Figure 6.

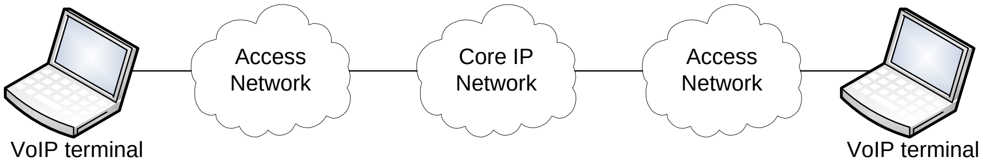


Figure 6 – VoIP-to-VoIP call

Since the PSTN is still the prevalent network for voice traffic, most VoIP calls are routed to the PSTN at one end as in Figure 7. However, since the VoIP-to-VoIP case presents the worst case with regard to the number of network segments, it will be considered for "home phone" services.

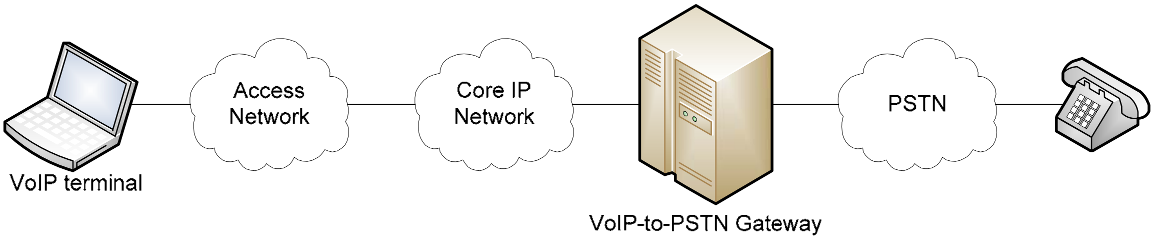


Figure 7 – VoIP-to-PSTN call

6.2.2 IPTD recommendations

The total one-way delays specified for conversational voice in Table 1 include all elements from ear to mouth, including encoding and buffering in the transmitter, network delay, and buffering and decoding in the receiver. To find the portion that can be allocated to network delay, we need more detail. As noted in section 6.2.1, the VoIP-to-VoIP case will be used to allocate allowable delay between network segments. The end-to-end delay calculation is shown in Equation 7.

$$L_{total} = \left(L_{enc} + L_{TB} + L_{dec}\right) + L_{A1} + L_N + L_{A2} + L_{RB} \qquad (7)$$

Where: L_{total} = the total mouth-to-ear one way delay,

L_{enc} = encoding delay at the transmitter,

L_{TB} = buffering delay in the transmitter,

L_{A1} = upstream access network delay (at sending end),

L_{A2} = downstream access network delay (at receiving end),

L_N = core network delay,

L_{RB} = buffering delay in the receiver, and

L_{dec} = decoder delay in the receiver.

Y.1541 associates interactive voice applications with "constrained routing and distance" with QoS class 0, which limits IPTD to 100 ms and IPDV to 50 ms. G.114 provides values for encoding, decoding, and buffering delays applicable for different voice codecs. Assume a "home phone" service with high quality encoding (such as G.711) and 10 ms packets. In this case, the voice frame size is one sample, or 0.125 ms, and 80 frames are grouped into each 10 ms packet. Per G.114 section A.2.4, the minimum encoding, decoding and transmit frame buffering delay (not counting serialization delay, which will be allocated to the access network separately) is given by

$$L_{enc} + L_{TB} + L_{dec} = (N+1) \times \mathit{framesize} + \mathit{lookahead} \,. \tag{8}$$

With N = 80, *framesize* = 0.125 ms, and *lookahead* = 0, Equation 8 becomes

$$L_{enc} + L_{TB} + L_{dec} = (80+1) \times 0.125 = 10.125 \,\text{ms.} \tag{9}$$

G.114 and Y.1541 also recommend that the delay added by the de-jitter buffer in the receiver be estimated for planning purposes at one half the peak delay. Using the 50 ms IPDV limit for QoS class 0 as the peak delay, the de-jitter buffer contribution is estimated at 25 ms, making the total encoding, decoding, and buffering delay in the transmitter and receiver approximately 35 ms. Subtracting that value from the 150 ms preferred value in Table 1, we get approximately 115 ms allocated to the network. Since that is slightly higher than the 100 ms allocated for IPTD by QoS class 0, we'll use the more stringent 100 ms value.

Meeting the "preferred" delay from Table 1: From section 5.1, worst case core network delay for the continental US is estimated at 50 ms. Allotting that value to the core network leaves 50 ms for the two access networks.

Most residential broadband access networks (including most DSL, DOCSIS, and wireless networks) provide higher average data rates in the downstream direction (going away from the core network) than in the upstream direction, so serialization delays are generally longer going upstream. Also, traffic aggregation in the upstream direction creates opportunities for delay as flows from different users are queued together. For these reasons, it makes sense to allocate the access network delay budget asymmetrically, with a higher limit in the upstream direction than downstream. We suggest allocating 35 ms to L_{A1} (upstream traffic) and 15 ms to L_{A2} (downstream traffic). These allocations meet both the QoS Class 0 requirement for IPTD and the preferred value of 150 ms for conversational voice from Table 1.

Meeting the "limit" delay from Table 1: Using the values in Table III.3/Y.1541, a worst case hypothetical reference path for intercontinental traffic incurs speed-of-light propagation delay of 138 ms (over 27,500 km) and an additional 44 ms IPDT through the distribution and core routers and internetworking gateways, for a total IPTD through the core network of 182 ms. Although the worst-case IPDV (including access networks) from Table III.3/Y.1541 is 86 ms, Y.1541 Appendix IV provides additional data showing that voice traffic assigned the appropriate CoS and given priority over best effort data should be able to maintain an IPDV within 50 ms even over this route. Allowing the above values for the access network and assuming de-jitter buffer delay of one-half the peak value, the total delay for this case is

$$L_{total} = 20.125 + 35 + 182 + 15 + 25 \approx 277 \,\text{ms,} \tag{10}$$

which is well below both the IPTD requirement for QoS class 1 and the ear-to-mouth limit value of 400 ms from Table 1.

6.2.3 IPDV recommendations

As noted above, Y.1541 Appendix IV addresses IPDV for delay sensitive traffic across the IP network. It shows (in section IV.3.1) that, for an IP core network with 12 routers at STM-1 rates, the total IPDV is conservatively estimated by the following components:

- Variation in routing lookup = 0.36 ms
- Queuing delay variation due to other priority traffic ≈ 1.36 ms (note that this assumes 1500 byte packets and 50% loading for priority traffic on the network, both conservative assumptions)
- Queuing delay variation due to lower priority packets already in service when the voice packets arrive = 0.72 ms.

Adding the components (rather than convolving) generates a conservatively calculated IPDV of 2.44 ms for the core network.

In the same appendix, section IV.3.3 shows that a single low speed access link can dominate delay variation on the network. If we assume two VoIP flows on a 1 Mbps access link (as could be the case for a “home phone” service with two lines), the maximum delay variation on the access queue is

$$DV_{\max} = \frac{(120+1500)*8}{10^6}*10^3 = 12.96\,\text{ms}, \tag{11}$$

where the maximum delay is encountered when a 1500-byte best effort packet has just begun being serviced and a 120-byte voice packet from the other flow arrives at the queue just in front of the delayed packet.

Adding twice the $DV_{\max}$ in equation 11 to the calculated IPDV for the core network yields a total IPDV of just under 30 ms for the example shown. The overall IPDV requirement for QoS classes 0 and 1 is 50 ms. Given the additional contributors to IPDV outlined in section 5.2, it seems prudent to allocate as much IPDV as possible to the access networks. As is the case for IPTD, there are more aggregation points at which delay variation can be introduced in the upstream than in the downstream direction. If IPDV is to be allocated between network segments, the suggested values are:

- Access network (upstream): 30 ms
- Access network (downstream): 15 ms
- Core network: 5 ms

These values are roughly proportional to the IPDV examples provided for core and access networks in Y.1541 Appendix IV. Following the lead of Y.1541, delays are broken out as if they were additive, which should generate conservative allocations.

6.3 Best effort real-time applications

The applications below are almost always accessed via a best effort HSIA service. As noted in the introduction to section 6, the access provider cannot guarantee IPTD or

IPDV values over the best effort service, and (as shown in Figure 2) IPDV in particular is unconstrained for best effort traffic.

Having said that, if a network is designed to perform at a given IPTD for one or more QoS-based applications, that same network may deliver best effort traffic with a similar IPTD *most of the time when the network is not suffering from congestion*. Even that heavily qualified statement must be considered further – for instance, serialization delay could have one value for the small packets sent for VoIP traffic, but could be substantially longer for the large data packets sent by a different application.

Considering all the above – in this section, we will discuss nominal performance of several delay-sensitive applications delivered over best effort service, on a network designed to provide QoS for “home phone” service. Any congestion on the network, particularly due to “home phone” traffic (which will be given priority over the best effort traffic) but also due to other best effort traffic, is likely to cause deterioration in the perceived performance of the applications.

6.3.1 “Internet VoIP” service

The second variety of VoIP might be described as “Internet VoIP.” In this case, the VoIP traffic accesses the Internet as best effort data using the HSIA service. The access provider is usually not aware that the traffic is associated with VoIP, and provides no priority for it. While the network performance for “Internet VoIP” cannot be managed as for “home phone” services, users accessing “Internet VoIP” generally do so with the awareness that it is not traditional phone service, and performance expectations are correspondingly lower.

Since “Internet VoIP” sends delay sensitive traffic across the access network (and the Internet) as best effort data, there is no way to enforce IPTD or IPDV requirements on it. We can estimate and roughly allocate IPTD, but (as shown by the ping response times in Figure 2) IPDV will be largely uncontrollable.

“Internet VoIP” applications usually use lower bit rate codecs than “home phone” services due to the varying bandwidth of the HSIA connections over which they run. A frequently used codec is CS-ACELP (G.729), which runs at 8 kbps with 10 ms frames. For this codec, $L_{enc} + L_{TB} + L_{dec} = 25$ ms. One slight advantage of this codec, in addition to the lower bandwidth requirement, is lower serialization delay since the transmitted packets are only 64 bytes long, rather than 120 bytes as in section 6.2. At 1 Mbps, however, the savings is less than 0.5 ms, so it isn’t considered in the values below.

If we assume a 100 ms de-jitter buffer (which for the variation in Figure 2 would capture about 993 out of 1000 frames) and network IPTD values from section 6.2.2, the total delay for continental US calls can be estimated at

$$L_{total} = 25 + 35 + 50 + 15 + 50 = 175 \text{ms.} \qquad (12)$$

IPDV across the access network is an unknown in this application, because it is highly dependent on the other best effort traffic being generated by the user. If a voice packet must queue up behind a dozen 1500-byte data packets to get across the access link, that packet is likely to be delayed beyond the reach of any de-jitter buffer. Depending on the traffic loading, data congestion in the access networks may also significantly affect the mean values of L_{A1} and/or L_{A2}, causing excessive IPTD across the network.

6.3.2 Conversational video

IP video calls are very similar to the VoIP-to-VoIP calls shown in Figure 6 in terms of the elements in the end to end path. Video calls are virtually always IP-to-IP, so traffic transits two access networks. Consumer video calls also suffer from the same disadvantage as “Internet VoIP,” in that they are treated by the network as best effort traffic.

Video frame periods vary based on individual call setup parameters, but typically range from 33 ms (for 30 frame per second [fps] video) to 100 ms (for 10 fps video). If we hypothesize that buffering and lookahead requirements apply to video frames in the same way as to voice packets, we can estimate the video encoding, decoding, and transmit buffering requirements as roughly 2.5 times the frame size, or 83 to 250 ms.[6] Assuming that users who care about video call quality will opt for 30 fps video, we’ll concentrate on that end of the range. However, after adding 50 to 100 ms for de-jitter buffer delay to the 83 ms for encoding/decoding/frame buffering, there is little or no time for network delay at the preferred delay value of 150 ms.

If we test against the limit delay value of 400 ms and assume a de-jitter buffer delay of 100 ms, the overall network delay requirement becomes

$$L_{A1} + L_N + L_{A2} \leq 400 - 83 - 100 = 217\,\text{ms}. \qquad (13)$$

Using the values for core and access network delay from section 6.2.2, it looks like the 400 ms limit for video delay should be supportable for continental US and for most international video calls. However, we still need to consider serialization delay, which is dependent both on the packet size and on the connection speed. Video bandwidth can vary widely, depending on the speed of the end-to-end connection and the quality desired by the users. At a midrange rate of 500 kbps and a frame rate of 30 fps, video frames exceed 1500 bytes in length. Assuming a 1500-byte MTU as the maximum packet size, the serialization time for video can be significantly longer than for voice. At 1 Mbps, serialization delay would add approximately 11 ms to overall delay relative to VoIP.

Assuming the above parameters and a 1 Mbps connection on access networks designed per section 6.2, L_{A1} and L_{A2} could have worst case values of 46 and 26 ms, respectively. These values would support video traffic over a core network delay of up to 145 ms. So, we can still say that the 400 ms limit for video delay should be supportable for continental US and most international video calls.

[6] This hypothesis was tested by completing an IP video call at 30 fps between adjacent laptops on an Ethernet LAN, verifying that there was synchronization between video and audio, and recording and measuring the difference in one-way audio delay between through-the-air propagation and the IP video call. The measured difference was 167 ms, which makes the hypothesis reasonable assuming de-jitter buffer delay of 50 to 100 ms.

As with "Internet VoIP," congestion in the access network due to best effort traffic is unavoidable and may cause unacceptably large values for IPDV. It may also affect the mean value of IPTD.

6.3.3 Interactive games

Multiplayer online interactive games represent a significant challenge with regard to network delay, since the path that traffic must take before an "instantaneous" response is perceived by the user is the longest of the cases examined. Figure 8 shows an example. In order for the response to an input from User A to be displayed back on User A's game client console, the following actions must occur:

1. The game console generates the traffic necessary to communicate the input.
2. The traffic transits User A's access network and the core network to the game server (blue path in Figure 8).
3. The game server processes the input and generates the appropriate output.
4. The output traffic transits the core network and User A's access network (red path).
5. The response is processed for display by game client A.
6. Similar traffic must be transmitted from the game server to other clients (green path) before User A's action is observed on their consoles.

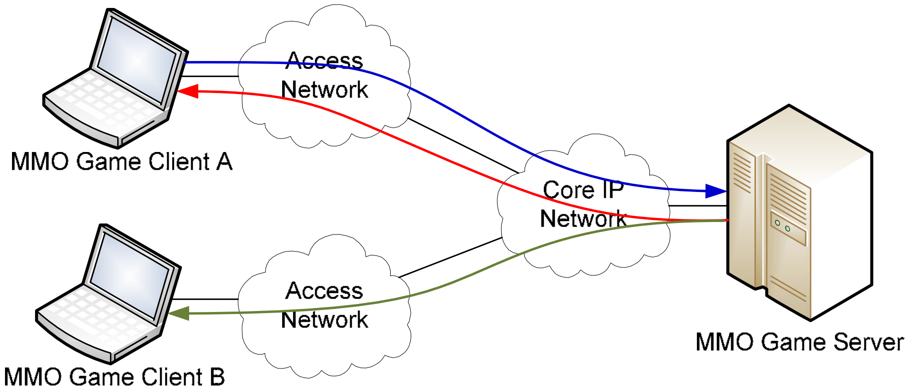


Figure 8 – Multiplayer online gaming client-server data paths

The red and blue data paths shown above are cited in TR-126 as the justification for the 50 ms one way (client to server or server to client) objective for interactive gaming delay specified in that publication. The one way delay, when doubled, corresponds to the 100 ms threshold for the round trip time required for a perceived "instantaneous" response. Similarly, the 75 ms preferred value from TS 22.205 corresponds to a 150 ms round trip time. Although the requirements are specified in terms of one way delays, the users' QoE is dependent in each case on a sum of two paths, one from client to server and a second from server to client.

The requirements include client and server processing time, which varies from platform to platform and for which there are no overall statistics. Assuming near zero time for client and server processing, if 50 ms is allocated for core network delay we can meet the

TS 22.205 preferred value with the 15 ms access network downstream delay but not with the 35 ms upstream delay. However, the target two-way trip value is preserved by the sum of the one-way delays. Assuming 50 ms core network delay, we cannot meet the more stringent TR-126 objective with any non-zero values in the access network.

Fortunately, we can justify examining the TR-126 limit against a core network delay lower than 50 ms. Most First Person Shooters (the subset of games most sensitive to delay) and many other online games provide a choice of servers on which to connect, with geographic and/or latency data available for each server, so players are not forced to connect to a distant server with high delay. Given that information, we can examine the TR-126 value against the average core network delay of 20 ms rather than the worst case figure. Subtracting that from the 50 ms objective gives a target value of 30 ms for the access network. Values lower than 30 ms would allow some non-zero budget for client and server processing. So, delay allocations of 35 ms upstream and 15 ms downstream for the access network should satisfy two-way requirements (although not the upstream one-way requirement) for gaming at the TS 22.205 preferred level on nearly all continental US connections, and also at the tighter TR-126 objective level for connections with core network delay of 20 ms or less.

Since gaming data typically consists of small packets, serialization delay should be similar to that for VoIP.

There is no specific limit provided on jitter for interactive gaming, but user experience is sensitive to excessive jitter [17]. The same caveats apply for this application as for the other delay sensitive applications using best effort CoS: both IPDV and IPTD may suffer momentarily due to excessive congestion.

7 Summary

Delay requirements are reviewed as documented by a number of standards development organizations for different user applications. Most of the applications are highly interactive and time sensitive, but even web browsing performance is shown to be sensitive to small changes in delay. Delay and jitter requirements for different QoS classes, as defined in ITU Recommendation Y.1541, are also reviewed.

The requirements in section 3 are then examined in the context of different residential applications. Web browsing is examined as an application which, while not real-time, frequently fails to meet either the preferred or acceptable requirements specified by the standards bodies. It is shown that web browsing response time is determined primarily by delay, rather than by download or upload speed.

We then examine real-time applications along with elements that contribute to delay and jitter, in an attempt to allocate the delay and jitter requirements appropriately to the core and access segments of the network. This is done first with respect to VoIP applications offered by access providers as “home phone” services with CoS and traffic management. It is found that allocating 35 ms to upstream traffic and 15 ms to downstream traffic in the access network supports VoIP applications at the preferred delay value of 150 ms for mouth-to-ear delay on continental US connections, and well within the acceptable value of 400 ms on intercontinental connections. In addition, allocations for jitter in the upstream access, downstream access, and core networks are suggested to support the overall 50 ms requirement for QoS class 0 in Y.1541, and also as values that are attainable for priority traffic.

The values arrived at above are then applied to applications running over best effort HSIA services, with the caveat that traffic from these applications transits the access network as best effort data with no special CoS applied. The resulting delay and jitter performance is highly dependent on both access and core network congestion and cannot be guaranteed.

- “Internet VoIP” – 175 ms delay for worst case continental US routes, which is not too far off the 150 ms preferred value.
- Conversational video – cannot meet 150 ms for all but shortest routes. Meets 400 ms limit value for all but the longest intercontinental routes.
- Interactive games – meets the two-way objective (calculated from the one-way 50 ms objective value in TR-126) for “average” US routes, but not worst case. Meets the two-way objective (calculated from the 75 ms preferred value in TS 22.205) for all continental US routes. While the proposed asymmetric delay targets in the access network do not strictly meet the one way requirements from either standard, the QoE objectives from which those requirements are derived are supported by the sum of the two allocations.

8 References

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

[2] Goodman, S., “Ex Parte Submission -- GN Docket No. 09-40,” submitted to the FCC April 13, 2009.

[3] Adtran, “Defining Broadband Speeds: an Analysis of Peak vs. Sustained Rates in Network Access Architectures,” April 2009.

[4] ITU Recommendation G.114, “One-way transmission time,” May 2003.

[5] ITU Recommendation Y.1540, “Internet protocol data communication service – IP packet transfer and availability performance parameters,” December 2002.

[6] ITU Recommendation Y.1541, “Network performance objectives for IP-based services,” February 2006.

[7] 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.

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

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

[10] ITU Recommendation G.1010, “End-user multimedia QoS categories,” November 2001.

[11] 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.

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

[13] AT&T network performance monitor at http://ipnetwork.bgtmo.ip.att.net/pws/network_delay.html

[14] Global Crossing network performance monitor at http://www.globalcrossing.com/network/network_performance_current.aspx

[15] http://www.eclipse.net.uk/index.cfm?id=home_broadband

[16] http://www.propelpbm.com/support/help/index.html

[17] http://www.q2s.ntnu.no/q2sfc/uploads/online_gaming.pdf

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

[19] 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

[20] www.chineseviking.com

[21] www.ranking.com

9 Appendix: Web page parameters

Wireshark traces were captured for transactions downloading the top 25 most popular web sites for March 2009 as reported by www.ranking.com. Prior to each capture, the web browser cache was cleared. The transactions captured, and relevant parameters from each, are shown in Table 4.

Table 4 – 25 top web sites for March 2009 per ranking.com

URL	Resp. time (s)	Size (bytes)	GETs and DNS requests	Weighted average RTT (s)	Network delay (s)	Eff. turns	Client delay (s)	Notes
www.google.com	1.259	38463	10	0.0734	0.802	10.92	0.435	1
www.yahoo.com	5.162	242922	45	0.0814	3.513	43.14	1.494	2
www.msn.com	5.068	330360	99	0.0737	3.461	46.99	1.207	2
www.live.com	2.378	113794	12	0.0550	0.939	17.07	1.035	1
www.youtube.com	6.787	439647	54	0.0731	3.769	51.59	2.766	2
www.aol.com	8.234	1043341	202	0.0381	4.252	111.70	3.111	2
www.myspace.com	4.62	257722	44	0.0732	2.929	40.03	1.531	2
www.microsoft.com	6.031	374005	62	0.0806	4.044	50.19	1.798	2
www.google.co.uk	1.504	32323	11	0.0672	0.848	12.62	0.626	1
www.facebook.com	5.052	449307	34	0.0711	3.521	49.49	1.287	2
www.ebay.com	5.663	392478	65	0.0728	4.061	55.80	1.385	2
www.ask.com	3.262	155541	34	0.0669	2.024	30.26	1.116	1
www.wikipedia.org	3.933	133056	20	0.0255	0.745	29.29	3.112	1
www.go.com								3
www.casalemedia.com	5.131	150426	31	0.0998	4.318	43.25	0.658	2,4
www.google.ca	2.497	33080	14	0.0566	0.731	12.91	1.725	1
www.findstuff.com	6.063	593571	66	0.0484	3.940	81.48	1.806	2
www.starware.com	1.268	5625	4	0.0920	0.658	7.16	0.606	1
www.att.net	5.878	759963	96	0.0733	4.113	56.12	1.289	2
www.cnn.com	6.763	974945	201	0.0563	4.450	79.07	1.554	2
www.mywebsearch.com	3.525	58940	19	0.0781	2.775	35.53	0.697	1
www.photobucket.com	4.814	534119	74	0.0701	3.594	51.25	0.879	2
www.amazon.com	4.695	481757	104	0.0411	2.706	65.83	1.480	2
www.netzero.net	4.437	195842	49	0.0663	3.282	49.46	1.021	2
www.blogspot.com								5

Notes:

1. Grouped with search engines.
2. Grouped with portals.
3. Go.com could not be evaluated in the same class with the other sites because it started streaming a long video file upon load.

4. Casalemedia.com is not a consumer web site per se, but it appears on the list by virtue of the advertising it serves to many other popular web sites.
5. Blogspot.com could not be evaluated because nearly all of the transaction was encrypted using HTTPS.

For each transaction, the following parameters were extracted from the capture:

- Total response time beginning with the first DNS request or TCP connection, and ending with the last HTTP OK message.
- Total IP bytes received by the client.
- The number of HTTP GET commands and DNS requests, subtotaled by server IP address.
- The RTT to each server IP address.
- Total network delay, including RTTs and server processing time. This was estimated by summing all inter-packet delays greater than 5 ms preceding a downstream packet.
- Total client processing time. This was estimated by summing all inter-packet delays greater than 5 ms preceding an upstream packet. Note that this time is at least as dependent on the client hardware and software as it is on the web site content, and that the values are provided here for comparative purposes only. The client used for the experiment was a mid-range business laptop (Core Duo running at 2.0 GHz, with 2 Gbytes RAM) running Internet Explorer 7.

For each site, the estimated network and client delays were subtracted from the total response time and the remainder was used to estimate download rate as a check on the process. The range of estimated download rates was between 2.2 and 15.8 Mbps. Given the volatility of the check, which was based on a remainder value much smaller than the estimates used to generate it, that range is quite close to the expected actual rates, which would have ranged from approximately 5 to 10 Mbps during the transactions.

The weighted average RTT was calculated for each site by weighting the RTT value for each IP address by the number of commands directed to it, and then averaging the result. The number of effective turns was then calculated as the network delay divided by the weighted RTT.

Both the summed network delays and the RTTs included both network delay and server processing time in the measurements. Considerable sampling variation is expected in the individual RTT values, including variation in server processing times, but averaged over the set of web sites the variation should be reasonably small.

The average parameters for the data set are provided in Table 5. The parameters are grouped by the type of web site as well as for the set as a whole. The web sites tended to fall into one of two categories. Search engines were distinguished by a small number of objects and relatively small size, reflective of their emphasis on fast response time. Portals (the term here is used loosely and includes both portal sites and corporate web sites) were distinguished by much larger download sizes, frequently including more than a hundred objects retrieved from a variety of servers.

Table 5 – Parameters for ranking.com top 25 web sites

Group	Average size (kbytes)	Average requests (GETs and DNS)	Average effective turns	Average response time (s)	Average client processing time* (s)
Search engines	71.4	15.5	19.5	2.453	1.17
Portals	481	81.7	58.4	5.627	1.55
All	339	58.7	44.8	4.523	1.42

*Client processing time is client hardware- and software-dependent.