An ADTRAN White Paper

ADTRAN

Defining Broadband Speeds: Deriving Required Capacity in Access Networks

Defining Broadband Speeds: Deriving Required Capacity in Access Networks

Executive Summary

As part of the Federal Communications Commission's work in developing the National Broadband Plan, the FCC has sought comments on a number of subjects related to broadband access. In NBP Public Notice # 1 [1], the FCC sought comments regarding the definition of broadband. ADTRAN, in responding to that request, referenced a previous white paper [2] in which the definition of broadband was linked to capacity in the access network.

More recently, the FCC has sought comments (in NBP Public Notice # 11 [3]) regarding middle and second mile access, including the amount of second-mile and middle-mile capacity required to provide adequate broadband Internet access to the end users of the network. This white paper directly addresses that question and expands on the relationship between capacity and the speed experienced by users in access networks. The paper discusses the importance of sustainable speed as experienced by the user – as opposed to other definitions such as peak or advertised speed – to a number of widely used application classes. It then addresses the relationship between capacity and demand that enables sustainable speeds on the access network.

The paper also provides updated projections for traffic demand and required capacity, based on new estimates and projections in Cisco's Visual Networking Index [4, 5]. The traffic volumes provided in the index are converted to estimates of per-household traffic, with scaling added to account for diurnal patterns. The mean traffic is then scaled to account for self-similar traffic distributions, resulting in per-user capacity requirements. The data generated from the 2009 VNI shows growth rates for North American consumer Internet traffic that are significantly higher than the corresponding rates from last year's source data.

This paper has been updated to reflect feedback from service providers. The first version scaled mean traffic volume by 4:1 (2:1 for burstiness and a second 2:1 scaling for node-to-node variation) to generate required capacity. The current paper eliminates the second 2:1 scaling in favor of monitoring and upgrading capacity on a node-to-node basis in deployed networks.

1 Introduction

As part of the Federal Communications Commission's work in developing the National Broadband Plan, the FCC has sought comments on a number of subjects related to broadband access. In NBP Public Notice # 1 (released August 20, 2009), the FCC sought comments regarding the definition of broadband. While many of the comments received in response to that notice discussed speed as one of the defining characteristics of broadband, there was not a consensus regarding how speed should be defined. In addition to several comments which made no attempt to define speed,[1] a number of comments proposed that speed be defined in terms of the following modifiers: "peak" rate,[2] "advertised" rate,[3] "configured" rate,[4] "average" rate,[5] and "delivered" rate.[6]

The above terms have different value to the users of broadband services. "Peak," "advertised," and "configured" rates are note very useful to subscribers as they may not ever be experienced by an individual user accessing a broadband service, depending on the design of the access network. "Average" or "delivered" rates are more meaningful to the user, but they still may not ensure broadband performance appropriate for widely used applications. For the definition of a speed-related metric to be meaningful, it needs to be derived based on the requirements of the application classes that are broadly used or expected to grow significantly. We review these application classes and their requirements, and derive such a metric in Section 2 of this paper.

The definition of a meaningful speed metric generates another question. How is the performance of an access network to be evaluated against the metric both before and after the network is deployed? While measuring speed in a deployed network may be relatively simple, predicting it accurately prior to deployment can be difficult or impossible, depending on the network architecture. In many networks, the individual user speed varies significantly with the overall demand placed on the network by the pool of users served by it. This variation can be sensitive to minor changes in the distribution of user traffic, which can change unpredictably.

Network capacity, as opposed to speed, is independent of traffic loading. The capacity of a wireline access network[7] can generally be determined by inspection of the appropriate

[1] Comments in response to NBP Public Notice # 1 from: Allied Fiber; ARRL; Covad; Internet2; OPASTCO; Qwest; Rural Cellular Association; TDI; and Time Warner Cable.

[2] Comments in response to NBP Public Notice # 1 from Qualcomm.

[3] Comments in response to NBP Public Notice # 1 from: FTTH Council; Hughes; NCTA; and Verizon.

[4] Comments in response to NBP Public Notice # 1 from Comcast.

[5] Comments in response to NBP Public Notice # 1 from Clearwire.

[6] Comments in response to NBP Public Notice # 1 from: CenturyLink; Frederick Maia; Free Press; Google; Native Public Media / National Congress of American Indians; Utopian Wireless; and Windstream.

[7] The capacity of a wireless access network can be considerably more difficult to determine than that of a wireline network. It is possible to estimate capacity for wireless networks, however, given some simplifying assumptions. That task is outside the scope of this paper.

network parameters. While speed cannot be determined directly from capacity, the relationship between capacity and the demand placed on a network is one of the factors determining whether the desired speeds will be supported. In that sense, capacity is an "enabler" of speed.

This white paper generates projections for the capacity that will be required on a per-subscriber basis for consumer broadband access networks over the next few years. The requirements are based on the traffic projections in Cisco's Visual Network Index 2008-2013 with updates from Cisco [4, 5]. Cisco's projections are converted from continental totals to per-subscriber values than can then be applied to first, second, or middle-mile capacity requirements for access networks. As such, the values are directly applicable to the questions asked in NBP Public Notice # 11 [3] regarding the network components of broadband capacity.

The required capacity projections in this paper are updated relative to similar projections provided in a previous white paper [2], which covered some of the same material. The differences between the earlier figures and the current ones are discussed, as well as the implications of those differences regarding the need to update projections on a regular basis as additional data becomes available.

2 Speed

There are many ways to define speed in an access network. Most service offerings advertise the maximum speed available on the connection, which may or may not be available when there are multiple users trying to access the network at the same time. In general, the speed experienced by an individual user on a broadband access network is likely to vary depending on how many other users are trying to use the same network, what they are trying to do, the architecture of the network, and other factors such as rate caps that may be placed on tiered service offerings by the access network service provider.

Even if we avoid marketing terms such as "advertised" or "maximum" speeds, defining speed in the most meaningful way is not trivial. As noted above, speed is not a fixed parameter – it varies based on the momentary demand being placed on the network. So while we can generate a single measurement of speed, for instance by using one of the many publicly available speed testing services,[8] that measurement tells us little about the overall performance of the connection over time.

Even if we take multiple measurements, we are faced with how to combine them into a meaningful metric. Take the example in Figure 1, which shows ten speed test results from a hypothetical access network.

[8] One example of a testing service is www.speedtest.net.

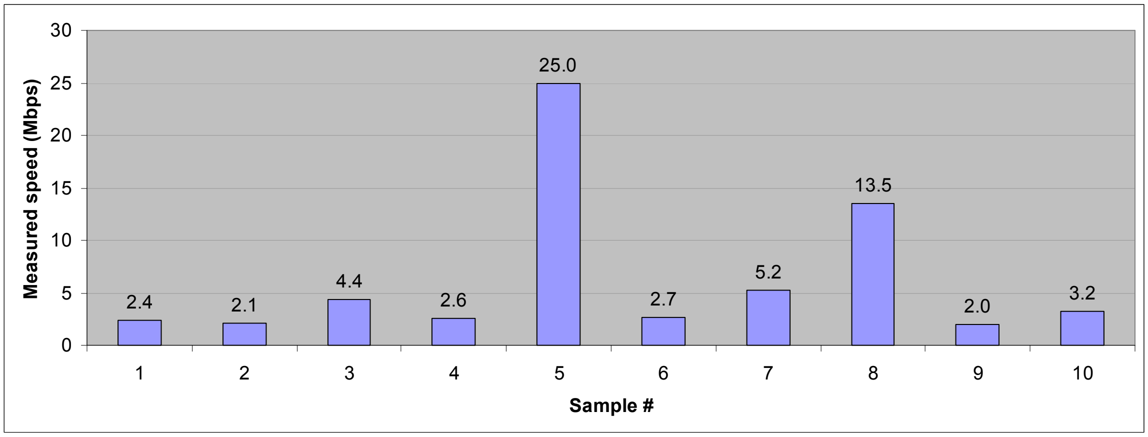


Figure 1 – Example speed test results

- The maximum sample is 25 Mbps. No other test result approaches the maximum sample value – in fact, only one other sample exceeds half the maximum value.
- The mean (or average) of the ten samples is 6.3 Mbps. While taking the mean is a straightforward approach, the resulting value in this case is higher than 80% of the speed test results.
- The median value is approximately 3.0 Mbps. Since the median by definition is exceeded by half the samples, we can expect to achieve this speed with about 50% probability.
- The minimum sample is 2.0 Mbps.

Which of the above choices, if any, provides the most appropriate description of network performance? Before we can generate a meaningful answer, we must understand the underlying requirements of the applications being used on the network.

2.1 Applications

Most of the data on current and forecasted application traffic volumes is taken from Cisco's Visual Networking Index 2008-2013 [4], updated by correspondence with Cisco [5].[9] Additional historical data comes from Cisco's VNI 2007-2012 [6], which contains estimates for traffic going back to 2006, and from other sources cited inline.

Figure 2 shows the estimated and projected consumer Internet traffic for North America for the years 2008-2013. The applications classes shown in the figure are discussed in the sub-sections below.

[9] The Cisco VNI 2008-2013, as published in June of 2009, contained inconsistent values for Consumer Internet traffic for North America in Table 3 and Tables 4-10. Cisco has since provided updated traffic values by email correspondence, and has indicated that they intend to update the published version of the VNI with the new values.

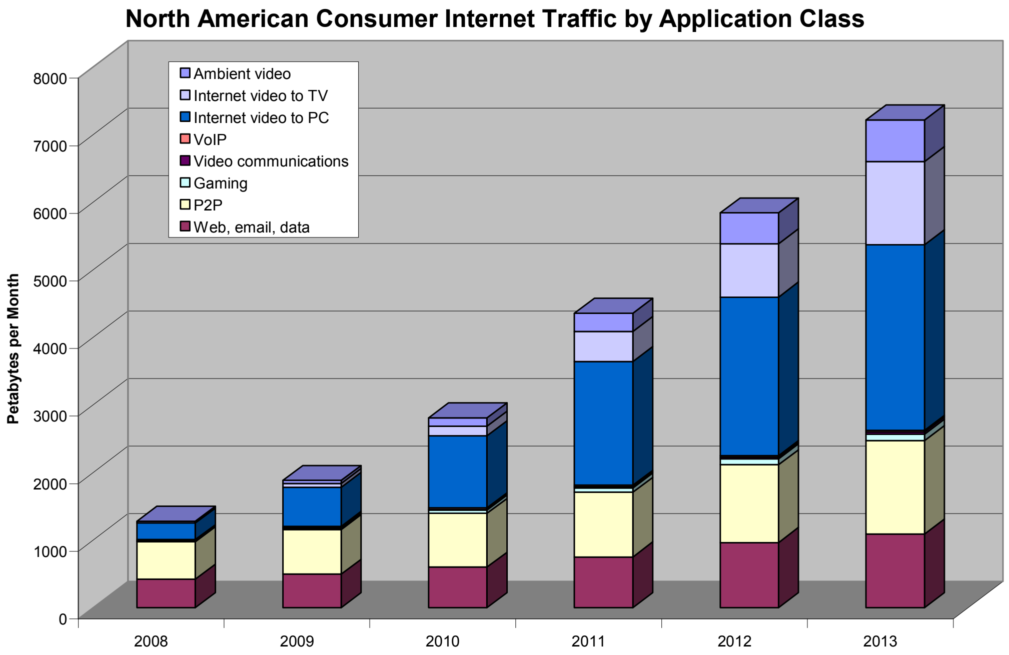


Figure 2 – North American consumer Internet traffic

2.1.1 Internet Video

Internet video is defined as video content which is accessed over the Internet via a subscriber's High Speed Internet Access (HSIA) service (as opposed to IPTV, which is sourced by a subscriber's service provider as a service separate from HSIA). While the best known examples of Internet video come from sites such as YouTube and Hulu, there are many different video sources, including broadcast and cable TV network web sites, social networking sites, movie delivery services, and educational web sites.

Internet video is widely considered to be the single largest factor in the growing requirement for bandwidth in broadband data services. Usage of the application has increased tenfold since 2006 and is expected to experience another six fold increase by 2013. The growth of this application is changing traffic characteristics for HSIA in the following ways:

- It triggers a corresponding increase in raw volume. Current playout rates for Internet video range from approximately 300 kbps (YouTube standard definition) to approximately 2 Mbps ("HD" content from network web sites), for videos that may range from tens of seconds to hours in length. As true HD content at 5 to 6 Mbps becomes more widely available it will continue to drive volumes higher yet [7].
- When video is streamed for immediate playout, the playout rate is tied to a near-real time requirement for content delivery. Subject to the size of the receive buffer, viewing a video file in near-real time requires a data transfer rate at or above the playout rate, which must be sustained with little interruption for the duration of the video. If the transfer rate drops below the playout rate long

enough to starve the playout buffer, the video will "freeze" until the buffer refills. Repeated "freezes" can make streaming videos unwatchable.

- These video applications are driving higher usage statistics. As people increasingly turn to Internet video instead of traditional sources for video entertainment, the percentage of subscribers who are actively using the service at a given time grows.

Most sources classify Internet video differently from video content downloaded using peer-to-peer (P2P) applications for purposes of traffic analysis. The former is generally accessed via a client-server model, from sites such as YouTube, and watched in near-real time as it is being streamed (although some video content can be downloaded and stored for later viewing). Peer-to-peer traffic, discussed in Section 2.1.3, has different characteristics. Some applications such as Joost (discussed in Section 2.1.7), combine aspects of both of the above categories.

2.1.1.1 Video to PC and Video to TV

Most streaming video to date is watched directly on the user's PC. There is a growing market, however, for devices that allow Internet video to be played out on a television rather than a computer monitor [8, 9, 10]. As this market matures, it will reinforce the growth of Internet video by making the viewing experience more familiar, regardless of the source (*e.g.*, families watching movies or TV episodes via Internet video rather than broadcast or cable).

Cisco breaks Internet video into the separate categories of Video-to-PC and Video-to-TV. Video-to-TV content is identified as "film and television content" as opposed to the user-generated content common on sites like YouTube.

2.1.2 Ambient Video

Ambient Video is an emerging application class, first identified in this year's VNI. It consists of consumer generated content such as video from home security cameras, "nannycams," pet cameras, etc., which is forwarded from homes to other locations such as workplaces.

The rates required for ambient video vary, since video monitor sources of this type can be configured for low frame rates. Total volume is relatively low to date, but is expected to grow at a compound annual growth rate (CAGR) of 95% through 2013. As with Internet video, the application requires near-real time performance. Unlike Internet video, the traffic model is subscriber-to-subscriber (rather than client-to-server).

2.1.3 Video communications

While video communications (two-party video calls or video conferencing) is not yet widely adopted, it may be close to emerging as a significant application due to three enabling factors:

- Widespread broadband access,

- Widely available, inexpensive and easily installed webcams (frequently integrated in new laptops), and
- Free, widely available video communications features added on to VoIP and instant messaging applications.

Cisco projects that traffic for video communications will ramp up significantly in the 2013-2018 time frame. Once that growth does occur it will drive significant requirements for both symmetric and real time traffic volume.

The data transfer rate for a video conference connection is near constant when averaged over seconds, but may vary with the encoded video content when measured over a shorter period. The required rate, which is usually symmetric, varies from hundreds of kilobits to several megabits per second. Network performance issues, including drops below the momentary data transfer rate for periods as short as tens of milliseconds, can cause noticeable loss of audio and video quality.

2.1.4 Voice over IP (VoIP)

VoIP applications require constant, symmetric data transfer rates on the order of 100 kbps or less. VoIP has widespread usage but, due to its low bit rate, it drives a small percentage of overall demand. Its main impact on access networks is that its performance is very dependent on congestion. Even momentary congestion will cause noticeable loss in voice quality.

2.1.5 Gaming

Gaming applications have bursty data transfer requirements with rates on the order of 100 kbps. Like VoIP, gaming has stringent real time requirements and can suffer noticeable loss of quality due to momentary congestion.

2.1.6 Peer-to-Peer (P2P)

P2P applications and protocols, and the resulting traffic loads, have been extensively analyzed [11, 12, 13]. Estimates of the amount of traffic generated by P2P applications have ranged as high as almost 80% of all consumer traffic [14]. One of the defining characteristics of P2P traffic is its symmetry – for every peer receiving content over a P2P network, another peer must be sending. Even though newer P2P protocols such as BitTorrent get content from multiple peers to reduce the peak upload burden on any one host, the nature of the application dictates that upload and download traffic is balanced over the entirety of the network. This can strain networks that were designed on the premise of client-server applications and asymmetric traffic loads.

A second characteristic of P2P traffic is that the difference between peak and average daily load levels tends to be less than for other applications [14]. Since most P2P applications deal with non-real time traffic, some users presumably schedule P2P transfers for non-peak traffic periods so as to not interfere with their interactive applications.

While the raw volume of P2P traffic continues to grow, its percentage share is steadily shrinking. Estimated P2P traffic for 2008 was 43% of North American consumer traffic, down from 61% in 2006 (and down further still from the 80% estimated in 2003 [14]). At least part of this trend results from the growing dominance of non-P2P video as a percentage of Internet traffic.

2.1.7 P2P video services (Joost)

Relatively new services like Joost [15, 16] combine some of the challenging characteristics of both streaming video and P2P applications. Joost enables subscribers to download video, including feature-length TV programs and movies, for near-real time viewing (as well as storage and later viewing) using a P2P protocol. As with BitTorrent and other P2P protocols, each file is transferred in "chunks" from multiple peers.

This class of application combines the near-real time requirement of streaming video with the symmetric traffic loading of P2P applications. To the degree that it is adopted, it will affect access network requirements for both capacity and symmetry.

2.1.8 Web browsing, Email, Data

Traditional web browsing, email and other data applications will continue to represent a significant percentage of traffic volume. While speed is certainly a factor in the performance of these applications (especially file transfer), it is frequently less important than other factors such as latency [17].

2.1.9 Applications Summary

The applications described above can be grouped into three categories with regard to their sensitivity to varying speed:

- Near-real time, streaming applications (Video-to-PC, Video-to-TV, Ambient Video). These applications buffer seconds to minutes worth of received data, making them tolerant to short periods in which the data transfer rate drops below the playout rate. If the speed drops for long enough to starve the receive buffer, however, the video will "freeze."
- Real time, interactive applications (Video communications, VoIP, Gaming). The tolerance of these applications to loss of data rate is on the order of tens of milliseconds. The required downstream data rates for VoIP and gaming are below those for most streaming video applications, but the required rates for video communications are as high as or higher than streaming requirements. Required upstream rates for any of these applications can be as high as or higher than for the streaming applications.
- Non-real time applications (P2P, Web Browsing, Email, Data). While these applications all benefit from higher speeds, they are not affected (other than variation in wait time) by variable performance.

The first two categories (near-real time streaming and real time interactive) suffer degraded performance if the network does not deliver the required speed with a high probability. In the streaming category, the requirement may correspond to measured speed that meets or exceeds the data transfer speed on the order of 90% of the time. Real time applications may require measured results that meet data transfer requirements on the order of 99% of the time or better.[10]

The above discussion leads us to a general definition of "speed" that meets the most stringent requirements for consumer applications that enjoy widespread usage. We will refer to this definition as "sustainable speed," which is defined as: ***the speed which a user can achieve with very high (99%) probability.*** Applying this definition to the speed test results of Figure 1, the sustainable speed in that example is no more than 2.0 Mbps and may be significantly less.

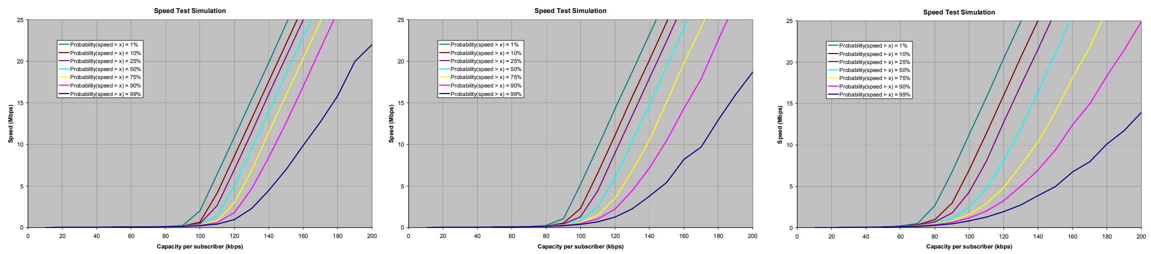
2.2 Predicting Speed

Now that we have reached a useful definition of speed, how do we determine its value in a given network? If the network is already deployed, that task is relatively straightforward. We can perform speed testing which, subject to the constraints of the tests, can provide an estimate of the sustainable speed available to a given user at a given time.

If the network is in the planning stages, predicting sustainable speed is much more difficult. Sustainable speed is dependent on more than just network design parameters, including architecture, link and node capacities, and traffic management design. It is also sensitive to a wide range of usage-related variables that may be largely unknown prior to deployment, including: total average traffic demand placed on the network; the distribution of that average demand between different users; and the mix of applications in use and the protocols over which they run. Even if these parameters can be approximated based on a history of deployments, they change over time, sometimes rapidly. For example, as ambient video emerges as a significant application class, it may change the demand distribution in the upstream direction (which to date has been dominated by a relatively small number of P2P users).

An example of the sensitivity of sustainable speed to small changes in usage parameters is shown in Figure 3. The plots in that figure show the results of three different series of simulations, all with identical parameters except for the shape of the demand distribution. In each case, the simulated network and the average traffic demand was identical, but varying the demand distribution generated significantly different results. (The reader is asked to suspend curiosity regarding details of the simulations, which will be revisited later in the paper. For the moment, the important point is the variation exhibited in the results.)

[10] How speed should be measured and how those measurements should be interpreted deserves its own white paper. For instance, most speed measurements are actually measuring throughput, or the time it takes to transfer a defined volume of test data. Variations in speed during the measurement period are usually averaged out. However, that subject is outside the scope of this paper, which (despite its emphasis to this point on speed) will focus on capacity.



a) α = 1.8 **b) α = 1.5** **c) α = 1.2**

Figure 3 – Simulated speed test results

In short, sustainable speed cannot be predicted based solely on network design parameters, and its sensitivity to small changes in usage parameters makes it difficult to predict at all.

3 Capacity

Compared to the above discussion of speed, definition of the capacity of a network is straightforward.[11] Unlike speed, capacity is dependent only on the network design, and not on the variable demands placed on the network by its users. As a result, capacity is a constant value for any given network.

While speed is the rate at which a user's traffic is sent across the network, ***capacity is defined as the ability of the network to carry that traffic***. The relationship between traffic, capacity and speed is illustrated in Figure 4. Since individual subscribers send or receive traffic only intermittently, the traffic (per unit time) averaged over all subscribers, including both active and inactive subscribers, is much lower than the actual speed at which traffic is transferred when an individual subscriber is active. The actual speed for an individual subscriber changes from one moment to the next, due both to changes in the individual's momentary demand and to changes in the overall demand placed on the network.

The network's capacity is shown prorated per subscriber in Figure 4. The prorated capacity can be thought of as each subscriber's "fair share" of the bandwidth available on the network. If every subscriber served by the network was active and trying to get as much bandwidth as possible at the same time, the actual speed for each subscriber would equal the prorated capacity.

In real networks, all of the subscribers virtually never compete for bandwidth at the same time, so the actual speed for the subset of subscribers who are active is generally higher than the prorated capacity. If the capacity is sufficient to handle the average traffic (again, averaged over all subscribers, both active and inactive) with margin for the burstiness inherent in data traffic, then the sustainable speed experienced by active subscribers will be much higher than the prorated capacity.

[11] As noted earlier, determination of capacity for wireless networks is considerably more difficult than for wireline networks.

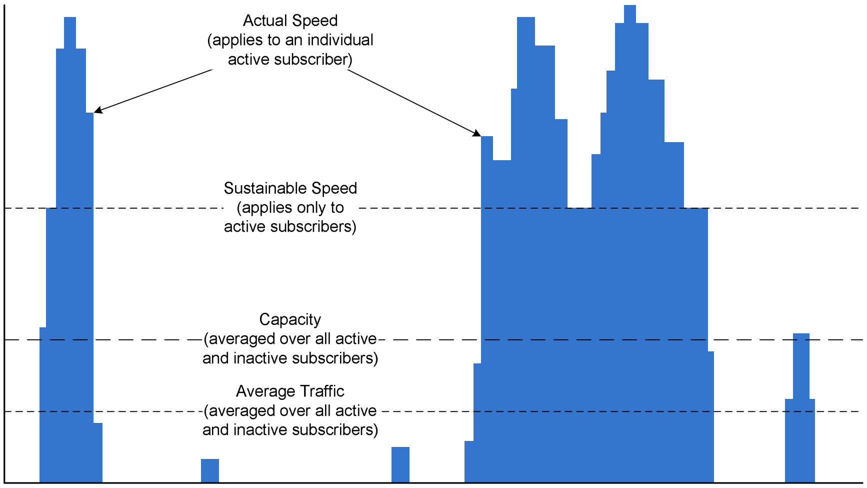


Figure 4 – Capacity and speed

3.1.1 Quantitative Definition of Capacity

The capacity of an access network is expressed quantitatively as the bandwidth available between the core network (the Internet) and the subscribers served by the access network. The available bandwidth is defined by the “weakest link,” or the segment of the access network between the core network and the subscribers with the least amount of bandwidth. As an example, Figure 5 shows a network where the available bandwidth between points A and B is 2 Gbps, defined by the bandwidth between Nodes 3 and 4.

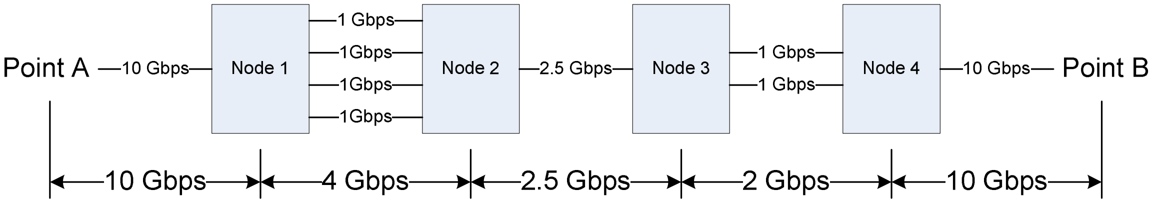


Figure 5 – Available bandwidth

Since an access network (or a portion thereof) can serve anywhere from tens to thousands of subscribers, we will express capacity on a per-subscriber basis. The capacity for a given subscriber is the prorated share of the bandwidth available between the core network and the subscriber. The capacity per subscriber can be different in different portions of a network. For example, the network in Figure 6 has three access nodes, each of which directly serves 100 subscribers over a dedicated 20 Mbps link. The capacities for the subscribers served by each access node are summarized in Table 1.

It is important to note that when capacity is prorated per subscriber, the prorating occurs across all subscribers served by the network, not just the active ones. The demands used in Section 4 to generate projected capacity requirements are averaged over all subscribers, regardless of whether they are active or not at any given instant.

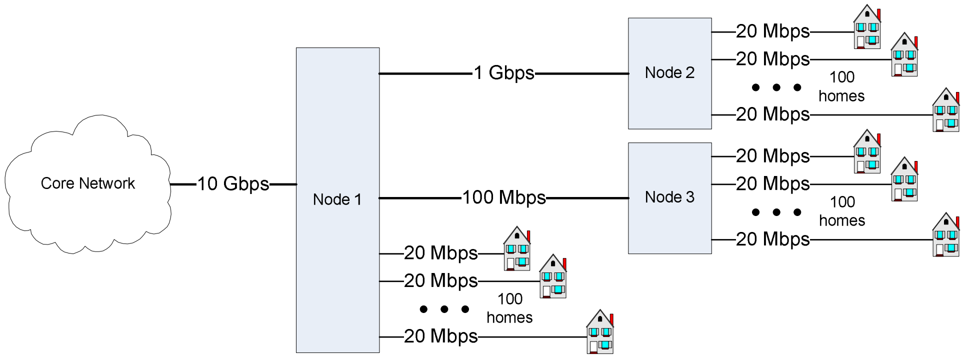


Figure 6 – Network capacity example

Table 1 – Capacities in Figure 6 example

Network Section	Capacity per Subscriber
Node 1	20 Mbps
Node 2	10 Mbps
Node 3	1 Mbps

Consider the specific case of the subscribers served from Node 3. From left to right in the figure, there are three links between the core network and each of those subscribers:

1. The 10 Gbps link from the core network to Node 1. This link is shared by 300 subscribers, making the prorated capacity (10 Gbps / 300) = 33 Mbps per subscriber.
2. The 100 Mbps link between Nodes 1 and 3. This link is shared by 100 subscribers, so its prorated capacity is 1 Mbps per subscriber.
3. The dedicated 20 Mbps link from Node 3 to the subscriber. Since the link serves a single subscriber, its prorated capacity equals its rate.

The prorated capacity for the subscribers served from Node 3 is the smallest of the individual link capacities, or 1 Mbps.

While capacity can be straightforward to determine, its usefulness depends on its ability to be related to speed or other network performance parameters. Unfortunately, we cannot relate capacity directly to sustainable speed with a simple equation – if we could, we could apply that equation directly to speed and bypass the entire discussion of capacity. We can, however, establish an approximate bounding relationship between capacity, demand, and speed which we can use to determine how much capacity is required to enable sustainable speeds.

3.2 Capacity, Demand, and Speed

To illustrate bounding relationship between capacity, demand, and speed, we return to the simulations first mentioned in Section 2.2. Figure 7 shows the results of a series of Monte Carlo simulations of individual speed test results across a resource shared by 500 users. The average demand is kept constant at 100 kbps per user, and the overall capacity of the shared resource is varied from one set of simulations to the next. Each set consists of 10,000 simulations in which the demand placed by each user on the network is defined by a Pareto distribution with the shape parameter α = 1.2. (This value comes close to the classic 80/20 distribution of demand mentioned in Section 3.3.3.) In each simulation, a random user runs a speed test.

The results in Figure 7 show the probability with which the speed test exceeded a given speed under a given set of conditions. The horizontal axis shows the capacity of the network, prorated per subscriber. The center value (100 kbps) is the average demand placed on the network, also prorated per subscriber. Each line plotted on the graph shows the rate achieved in the speed test at a given probability. For instance, with the network capacity at 120 kbps per subscriber, the sustainable speed (the speed that can be achieved with 99% probability) is 1.9 Mbps – meaning that a user can achieve 1.9 Mbps or better 99% of the time. At the same capacity, a user can achieve at least 8.1 Mbps 50% of the time, and at least 20.5 Mbps 1% of the time.

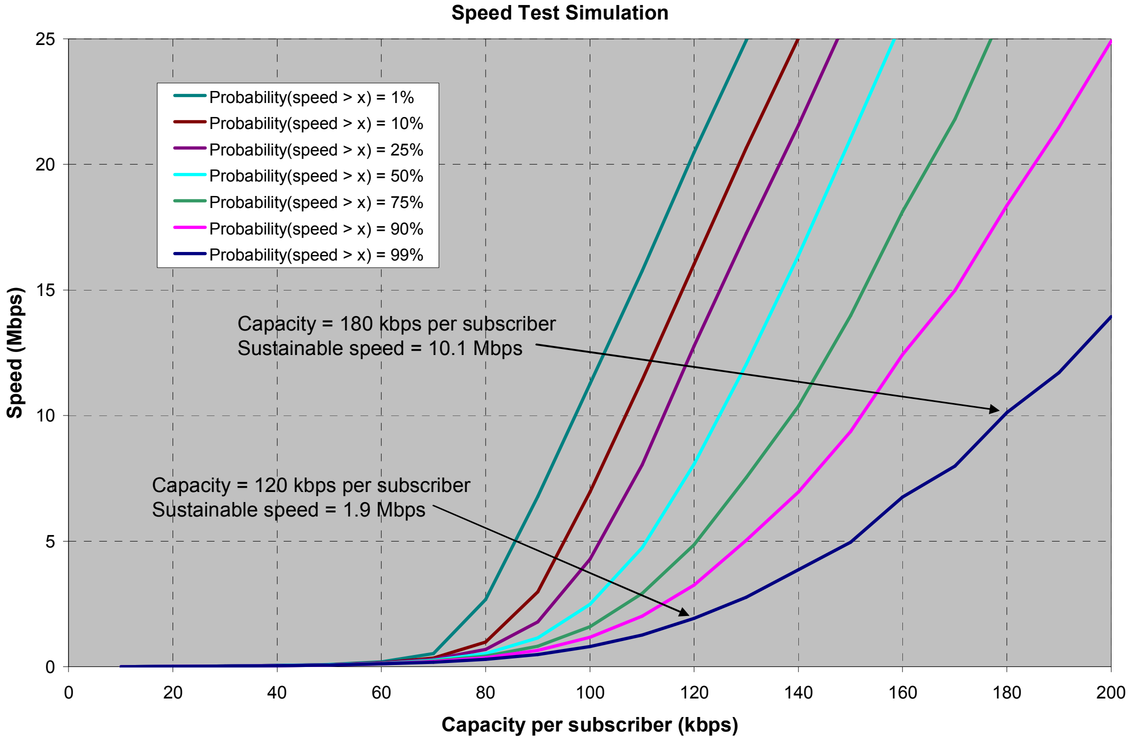


Figure 7 – Speed test simulation, α = 1.2

A quick inspection of Figure 7 reveals two obvious results. First, when network capacity is lower than demand (the left side of the chart), the network is congested and performance is minimal. Second, when network capacity exceeds demand by a sufficient margin (the right side of the chart), performance is excellent.

A third observation is less obvious: as capacity increases in the uncongested region of the chart, ***the increase in an individual user's sustainable speed is proportional to the increase in the overall capacity of the resource.*** This result has a multiplicative effect – if there are many users sharing a resource, the benefit of increasing the overall capacity of that resource is not divided among them – it increases the sustainable speed for each user by a significant fraction of the overall increase.

For the specific simulations documented in Figure 7, an increase in capacity from 120 to 180 kbps per subscriber represents an increase of 30 Mbps in the total shared resource. The corresponding increase in sustainable speed (using the 99% percentile point) is from 1.9 to 10.1 Mbps, for an improvement of over 8 Mbps. If the effect was shared proportionally between all 500 users of the resource, the increase would be only 60 kbps!

It is important to note that the simulation described is not intended to be representative of any specific network, nor is it detailed enough to solve for a specific relationship between capacity, demand and speed in a realistic scenario. It makes no attempt, for instance, to simulate protocol behavior or traffic management techniques other than max-min fairness [21]. As noted earlier, actual demand distributions in deployed networks (while they may resemble Pareto in some respects) are largely unknown and are subject to change over time.

While the above caveat reminds us that we cannot directly predict speed using capacity, the multiplicative effect of increases in shared capacity allows us to establish a bounding relationship which, while it may be conservative in that sustainable speeds may be higher than necessary, will not force the shared resource to be exorbitantly expensive. If we provide enough capacity margin relative to demand – that is, if we stay far enough towards the right side of charts like Figure 7 – we can expect that the sustainable speed experienced by active subscribers will be much higher than the prorated capacity on the network.

3.3 Margin

The preceding paragraph begs the question – how much capacity margin should we provide relative to expected demand? To answer this question, we need to examine the sources of variation in demand and to determine how best to account for each source.

3.3.1 Diurnal variation

Internet traffic volume experiences a daily pattern reflecting user activity cycles [14, 25, 26]. While business activity peaks during normal weekday office hours, consumer activity peaks during evening hours, with much less variation between weekdays and weekends.

Since the diurnal pattern is well documented and predictable, we can account for it by estimating the mean traffic volume during the busiest period of the day. We do so in the quantitative estimation of household traffic developed in Section 4.

3.3.2 Variation due to Self-Similar Traffic

It is generally understood that Internet traffic is bursty – that is, individuals actively send or receive traffic intermittently rather than at a continuous rate. As traffic from different subscribers is combined (or aggregated), this burstiness appears as variation over time in the traffic volume.

One well documented characteristic of Internet traffic related to burstiness is known as self-similarity [18, 19]. A process which is self-similar will exhibit similar patterns of variation across different scales. One of the effects of self-similarity is fluctuations in momentary traffic that exceed those predicted by models such as Poisson arrivals. Mori *et al.* [20] measures the skewness and marginal distributions of Internet traffic on a number of network links. The results show positively skewed distributions with momentary loads that can exceed twice the mean volume. The paper notes that the variation in momentary loads holds for measurement intervals ranging from 0.01 seconds to 10 seconds.

This data indicates that a 2:1 margin relative to the mean traffic volume can minimize, if not eliminate, congestion due to the self-similarity of Internet traffic. We will apply this margin when estimating required capacities in Section 4.

3.3.3 Node-to-Node Variation

The above two sections each reference "mean traffic volume." Section 3.3.1 discusses how to account for diurnal variation in the mean, and Section 3.3.2 discusses how to account for variation over shorter time periods. We must still consider variation in the mean from one access node to another.

In Section 4 we calculate mean traffic volume on a per-subscriber basis using source data representing the entirety of North America, estimated to contain 70 million households with broadband access in 2009 (we refer to this as the "national mean value"). In contrast, a typical access node (such as a DSL Access Multiplexer [DSLAM] or a wireless base station) may serve anywhere from a few dozen to a few thousand broadband subscribers. The "mean traffic per subscriber" as applied to a single access node is not necessarily the same as the "national mean value." There can be substantial differences in the mean traffic volume for one access node as compared to another, because the population served by each access node is made up of individual subscribers with differing levels of activity.

The traffic demands that different subscribers place on broadband access vary over a tremendous range. As is the case in many populations, a minority of subscribers consumes a majority of the resources. One study [14] indicated that 2.9% of subscribers (in a pool of over 100,000) accounted for over 40% of the traffic on the network, and that the top 20% of subscribers accounted for slightly over 80% of the traffic.

The concentration of demand in a small percentage of the subscriber population significantly increases the expected variance in demand in access networks, where the subscriber pool is smaller than in aggregation or core networks. In a network providing access to 100 subscribers, the usage characteristics of less than 20 households can be expected to dominate the traffic. The addition or loss of only two or three heavy users could result in a change of 50% or more in the mean traffic volume.

If this node-to-node variation is not accommodated, performance on the nodes with the heaviest mean traffic volume can experience mean utilization near 100% and suffer a corresponding loss in performance. One way to accommodate the variation is to apply an additional fixed margin when estimating required capacity – however, that solution forces the extra capacity on all nodes, not just the ones that need it. Instead, many service providers monitor utilization and adjust capacity as necessary on a node-by-node basis. We assume that service providers are adjusting capacity in this way, and do not apply any additional fixed margin when estimating required capacities in Section 4.

4 Traffic and Capacity Projections

The traffic volume values provided in [4] are monthly totals for consumer Internet traffic in North America. In this section we relate those figures to usage on a per-household basis in the belief that the resulting figures may provide some guidance for scaling shared capacity in access networks.

As a quick check on the source data, the total monthly volume reported by the Minnesota Internet Traffic Studies [22] for the US was from 1,200 to 1,800 Petabytes. This is in line with Cisco's estimate of 1279 Petabytes per month for consumer traffic in North America in 2008.

The monthly estimates and forecasts for North America Internet traffic by sub-segment for the years 2006 through 2012 are taken from different tables in [4, 5] and compiled in Table 2. This is the same data shown graphically in Figure 2.

Table 2 – North American consumer Internet traffic by application class

By Sub-Segment (PB per month)	2008	2009	2010	2011	2012	2013	CAGR 2008–2013
Web, email, data	421	494	599	750	964	1,089	21%
P2P	555	662	795	956	1,150	1,384	20%
Gaming	9	19	50	64	88	92	59%
Video communications	3	6	11	18	24	34	63%
VoIP	18	21	22	23	23	23	5%
Internet video to PC	246	579	1,063	1,830	2,345	2,744	62%
Internet video to TV	3	56	146	444	789	1,233	233%
Ambient video	22	45	120	271	456	614	95%
Totals	1279	1881	2807	4357	5839	7213	41%

The following discussion is based on 2008 figures. Based on US population estimates and the most recent census figures for persons per household [23], there were approximately 117 million households in the US. Approximately 55% of US adults had broadband Internet access and another 10% had dial-up access [24]. Assuming that the Pew survey did not include more than one person per household, we can infer a high correlation between personal and household Internet access since access is normally provided on a per-household basis. So, approximately 65 million households had broadband connections, which should account for the vast majority of consumer traffic (with broadband having 5.5 times the number of dialup connections at over 10 times the speed and longer average session times, we can safely assume that dialup volume was relatively minor).

From Table 2, total volume in 2008 was about 1279 Petabytes per month. Spreading 1279 Petabytes per month across 65 million households gives us a long term average (measured over days) volume of approximately 60 kbps per household.

Table 3 shows the above analysis applied to the data in Table 2. The broadband adoption rate for year 2008 is from [24]. The rate for 2009 is 4% higher than 2008 based on 3% gains reported over 9 months, and the rate for 2010-2013 is extrapolated at a linear increase of 3% per year based on survey results in [24] indicating that the rate of broadband growth may be slowing.

Table 3 – Internet traffic long term average traffic per household

Estimated broadband adoption	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
No. of households (million)	117.5	118.6	119.8	120.9	122.1	123.3
Broadband adoption rate	55%	59%	62%	65%	68%	71%
No. of broadband households (million)	64.6	70.0	74.3	78.6	83.0	87.5
Traffic by Sub-Segment (kbps per household)						
Web, email, data	20.1	21.8	24.9	29.4	35.8	38.4
P2P	26.5	29.2	33.0	37.5	42.7	48.8
Gaming	0.43	0.84	2.08	2.51	3.27	3.24
Video communications	0.14	0.27	0.46	0.71	0.89	1.20
VoIP	0.86	0.93	0.91	0.90	0.86	0.81
Internet video to PC	11.8	25.5	44.2	71.8	87.2	96.7
Internet video to TV	0.14	2.47	6.07	17.43	29.33	43.47
Ambient video	1.05	1.99	4.99	10.64	16.95	21.65
Totals	61.0	83.0	116.6	171.0	217.0	254.3

The above data includes both upstream and downstream traffic averaged over a long period (greater than 24 hours). As is well documented [14, 25, 26], traffic volume exhibits a diurnal pattern reflecting user activity cycles. While business activity peaks during normal weekday office hours, consumer activity peaks during evening hours, with much less variation between weekdays and weekends. A diurnal pattern with similar peak times of day applies to different categories of traffic, although different application classes exhibit different excursions from the mean [14].

We use data from [14] to estimate upstream and downstream volumes during peak daily periods. Modeling the diurnal excursions from the mean as approximately symmetric,[12] the corresponding average loads during peak traffic hours can be estimated as

$$P = M\left(1+\frac{r-1}{r+1}\right), \qquad (1)$$

[12] This assumption of symmetry probably results in underestimation of the peak period averages. The diurnal patterns for consumer traffic in [14] look approximately symmetric, but those in [26] look like they exhibit positive skewness, which would make the peak period volumes somewhat higher than those calculated here.

where: *P* = the average load during peak periods,
M = the long term average load, and
r = the diurnal max/min traffic ratio.

The same study provides upstream vs. downstream traffic ratios for traffic from different application classes. These values are incorporated for the year 2008 data in Table 4.

Table 4 – Traffic during peak hours, 2008

Application class	Web	P2P	Video to PC	Video to TV	Ambient video	Other (1)	Total
Long term mean *M* (kbps)	20.1	26.5	11.8	0.14	1.05	1.43	61.0
Diurnal max/min *r* (2)	5	2	5	2	2	4	
Mean volume *P* (peak time)	33.5	35.4	19.6	0.2	1.4	2.29	92.3
Down/up ratio (3)	8	1	8	8	1	1	
% downstream	89%	50%	89%	89%	50%	50%	
Downstream (peak time)	29.8	17.7	17.4	0.2	0.7	1.1	66.9
Upstream (peak time)	3.7	17.7	2.2	0.0	0.7	1.1	25.4

Notes on Table 4:

1. The Other class includes the gaming, video communications, and VoIP sub-segments.
2. [14] states that the maximum to minimum diurnal load ratio is about 2 for P2P traffic and about 5 for Web browsing traffic. For this analysis, the Web browsing ratio is applied to interactive categories and the P2P ratio is applied to categories in which files can be scheduled for off-peak download. Video-to-PC, which consists primarily of shorter clips at lower bit rates, is places in the interactive category. Video-to-TV, which includes feature length films at high playout rates, is placed in the off-peak category. Ambient video is assumed to be less subject to variation due to the background nature of video monitoring. While all the sub-segments in the Other class are interactive, VoIP and video calling may be somewhat more distributed in time so the ratio applied is reduced slightly.
3. Downstream/upstream ratios in [14] are approximately 8 for client/server applications that primarily download data, and approximately 1 for symmetric applications. For this analysis, all video traffic except ambient video is assumed to follow the client/server model. Increased adoption of P2P video (*e.g.*, Joost) could push upstream rates higher.

Table 4 shows how values for traffic loading during peak usage times, as they would be measured over reasonably short time frames of several minutes, are derived. Totals for the same parameters are provided in Table 5 for the years covered by the current Cisco forecast. The same data is shown graphically, and broken out by application class, in Figure 8 and Figure 9.

Table 5 – Average traffic scaled per household during peak usage hours

Direction	2008	2009	2010	2011	2012	2013	CAGR 2008-2013
Down (kbps per household)	67	95	138	206	261	303	35%
Up (kbps per household)	25	32	42	57	71	83	27%

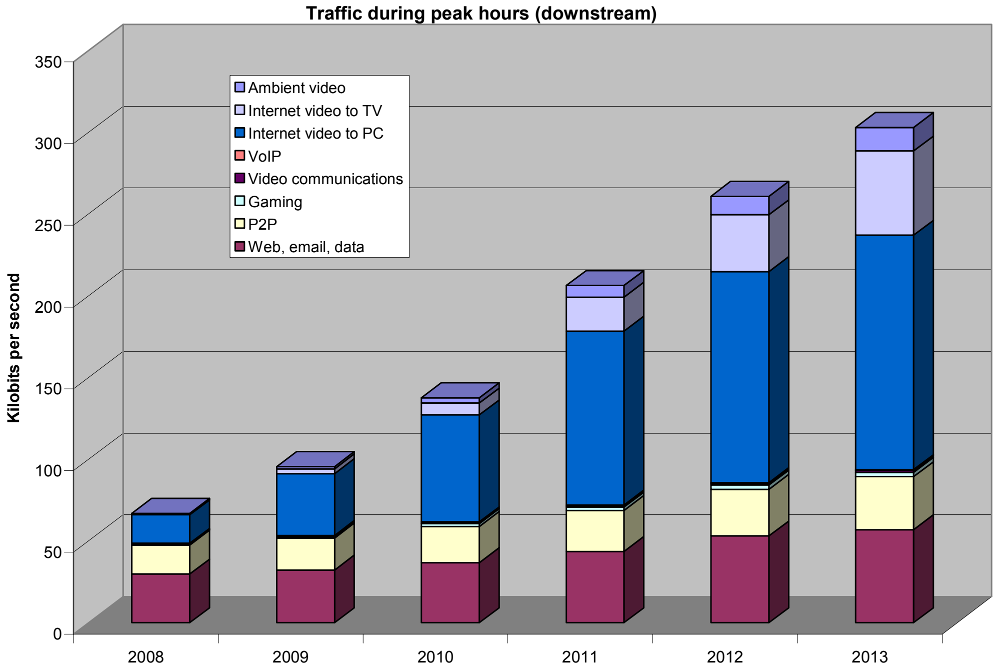


Figure 8 – Average traffic per household (peak hours, downstream)

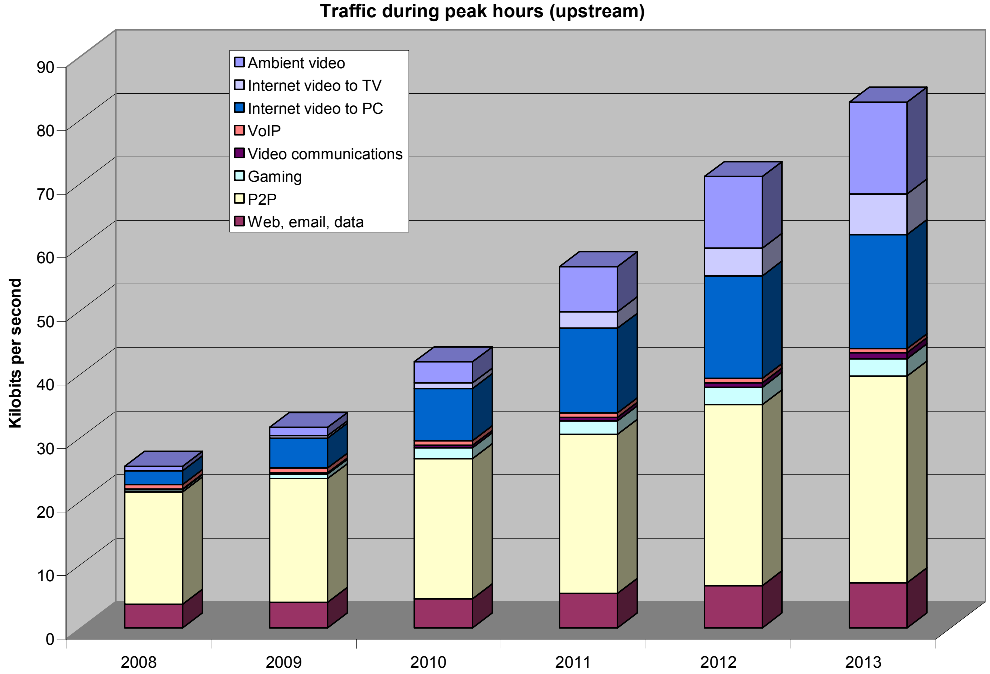


Figure 9 – Average traffic per household (peak hours, upstream)

At this point, we need to step back and list the accumulated caveats regarding the above numbers.

- Despite the scaling, the average values shown in Table 5, Figure 8 and Figure 9 are obviously not intended to be applied to individual households. Subject to the caveats below, they are scaling factors applicable (with the addition of appropriate margins) to large numbers of households in a given population.

- The extrapolation of future broadband adoption in this analysis may deviate significantly from the assumptions made by Cisco when generating their total volume forecasts. Differences in those extrapolations could have a significant impact on the per-household CAGR values.
- Finally, we need to remember that the forecast data from Cisco may not reflect actual future trends, considering the volatile history of the Internet [27]. While some characteristics of the Internet are relatively invariant, such as diurnal patterns and traffic self-similarity, other characteristics can change almost literally overnight. Rapid adoption of new applications, protocols or technologies could render the forecast obsolete, even in the relatively limited span that this forecast covers.
- Compared to the warnings listed above the following items seem almost trivial, but they are included for completeness. The analysis used forecast numbers for North America (including Canada) against population figures for the United States only, which inflated the per-household figures by an estimated 10%. Partly offsetting that, the forecast values do not include signaling traffic, acknowledged in [4] to add about 3% to overall volumes.

The per-household averages in Table 5 represent estimated and projected average traffic volumes. To generate required capacities from these volumes, we scale by the 2:1 margin identified in Section 3.3.2. After performing the required scaling, we arrive at a current (2009) required capacity on the order of 191 kbps downstream and 63 kbps upstream per household for resources in the access network.

We use the CAGRs from Table 5 (and round up in increments of 50 kbps) to generate projected capacity requirements for future years in Table 6. The values for year 2015 in the table should be considered tentative, since they extrapolate the CAGR beyond the forecast numbers provided by Cisco. Given the expense and time associated with deploying broadband infrastructure, however, a projected requirement that looks only three years into the future would result in deployments that are obsolete soon after their introduction.

Table 6 – Approximate required capacity/household in the access network

Direction	2009	2012	2015
Down (kbps per household)	200	500	1200
Up (kbps per household)	100	150	300

As noted in Section 3.3.3, the values in Table 6 are based on mean traffic volumes derived from the total traffic volumes estimated for North America. The actual mean traffic volume measured on a single access node can deviate significantly from this "national mean" value. Service providers must accommodate this variation in node-to-node traffic in their deployments, either by actively measuring traffic and adjusting capacity for each node or by applying margin in addition to the capacity figures provided.

As a final cautionary note, we compare the values in Table 6 with the corresponding values generated from Cisco's VNI for 2007-2012, published in 2008 [6]. The values generated using data from both years are shown in Table 7. As that table shows, updating of the source data by one year has had the result of generating a difference of 2 to 1 downstream, and over 2 to 1 upstream, in the projected required capacities for year 2015. This difference underscores the importance of considering the 2015 values tentative, as well as the importance of revisiting and updating the data on a regular basis.

Table 7 – Comparison of projections from VNIs for 2008-2013 and 2007-2012

Source data	Direction	2009	2012	2015	CAGR
VNI 2008-2013	Down (kbps per household)	200	500	1200	35%
	Up (kbps per household)	100	150	300	27%
VNI 2007-2012 **OUTDATED**	Down (kbps per household)	200	350	600	22%
	Up (kbps per household)	100	100	100	11%

5 Summary

The different ways of defining "speed" in a broadband access network are examined in the context of how speed-related performance affects different classes of widely used consumer applications. Streaming video applications, which drive large and rapidly increasing traffic volumes, are found to be intolerant to excursions in the data transfer rate that go below the playout rate for extended periods. Interactive VoIP and video communications applications are the least tolerant of variable speed performance which may dip below the required rate. Based on the above application classes, we propose that "sustainable speed" be defined as the speed that is achievable with very high probability (on the order of 99%).

While sustainable speed can be measured in existing networks, it is nearly impossible to predict in the planning stages due to its sensitivity to traffic demand parameters. In contrast, network capacity is a network parameter which is independent of demand and which can be determined during network planning. While there is not an explicit relationship between capacity and sustainable speed that would not also be dependent on demand, we show that there is a bounding relationship such that if sufficient capacity plus margin (relative to expected mean demand) is provided in the shared resources, the network should support sustainable speeds that are much higher than the prorated capacity.

Once the necessary margin to derive required capacity from average traffic has been established, the next step is to generate estimates of (and projections for) average traffic. We refine the data provided in Cisco's Visual Networking Index 2008-2013 [4, 5] to generate per-household averages for downstream and upstream traffic during peak daily demand hours, which occur in the evening for consumer traffic. These averages are then scaled to generate projected capacity requirements through the year 2012 (and tentative projections that extend through the year 2015).

6 References

[1] Federal Communications Commission, NBP Public Notice # 1, “Comment Sought on Defining ‘Broadband,’” released 20 August 2009.

[2] Adtran, Defining Broadband Speeds: An Analysis of Required Capacity in Network Access Architectures,” June 2009.

[3] Federal Communications Commission, NBP Public Notice # 11, “Comment Sought on Impact of Middle and Second Mile Access on Broadband Availability and Deployment,” released 8 October 2009.

[4] Cisco, “Cisco Visual Networking Index – Forecast and Methodology, 2008-2013,” 9 June 2009, available at http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html

[5] Email correspondence between Usha Andra (Cisco) and Ken Ko (Adtran) which provided updates to the VNI 2008-2013, 23 October 2009.

[6] Cisco, “Cisco Visual Networking Index – Forecast and Methodology, 2007-2012,” 16 June 2008.

[7] Cisco, “Approaching the Zettabyte Era,” 16 June 2008.

[8] http://www.vudu.com/product_overview.html

[9] http://www.roku.com/default.aspx

[10] http://www.popcornhour.com

[11] Martin, J. and Westall, J., “Assessing the Impact of BitTorrent on DOCSIS Networks,” Fourth International Conference on Broadband Communications, Networks and Systems, 2007

[12] Erman, D and Popescu, A., “BitTorrent Traffic Characteristics,” International Multi-Conference on Computing in the Global Information Technology, 2006

[13] Qi, J., Zhang, H., and Ji, Z., “Analyzing BitTorrent Traffic Across Large Network Cyberworlds,” Proceedings of the 2008 International Conference on Cyberworlds, pp.759-764, 2008

[14] Gerber, A., Houle, J., Nguyen, H., Roughan, M., and Sen, S., “P2P, The Gorilla in the Cable,” 2003, available at http://www.research.att.com/~sen/pub/p2pCable2003.final.pdf

[15] http://www.joost.com/

[16] Alhaisoni, M. and Liotta, A., “Characterization of Signaling and Traffic in Joost,” Peer-to-Peer Networking and Applications, Volume 2, Number 1 / March, 2009, pp. 75-83

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

[18] Erramilli, A., Roughan, M., Veitch, D. and Willinger, W., "Self-Similar Traffic and Network Dynamics," Proceedings of the IEEE, Vol. 90, No. 5, May 2002, pp. 800-819

[19] Yu, B. and Fei, H., "Fractal Analysis of User Sessions Inter-Transaction Time in Social Networks," 4th International Conference on Wireless Communications, Networking and Mobile Computing**,** 12-14 Oct. 2008

[20] Mori, T., Kawahara, R., Naito, S. and Goto, S., "On the Characteristics of Internet Traffic Variability: Spikes and Elephants," Proceedings of the 2004 International Symposium on Applications and the Internet

[21] Jha, S. and Hassan, M., "Engineering Internet QoS," Artech House, 2002

[22] http://www.dtc.umn.edu/mints/home.php

[23] http://www.census.gov/

[24] Horrigan, J., "Home Broadband Adoption 2008," Pew Internet & American Life Project, available at http://www.pewinternet.org/

[25] Marques, H., Rocha, L., Guerra, P., Almeida, J., Meira, W., and Almeida, V., "Characterizing Broadband User Behavior," Proceedings of the 2004 ACM Workshop on Next-Generation Residential Broadband Challenges, October 15-15, 2004

[26] Fukuda, K., Cho, K. and Esaki, H., "The Impact of Residential Broadband Traffic on Japanese ISP Backbones," ACM SIGCOMM Computer Communications Review, Volume 35, Number 1, January 2005

[27] Floyd, S. and Paxson, V., "Difficulties in Simulating the Internet," IEEE/ACM Transactions on Networking, 2001, volume 9, pp 392-403.