CHAPTER 6

Can the Internet Hit a Wall?

I predict the Internet will soon go spectacularly supernova and in 1996 catastrophically collapse.

—Bob Metcalfe, Ethernet inventor and 3Com founder, in his *InfoWorld* column in 1995

The current prediction of bandwidth overload has a name—*exaflood*—and the potential inability of the network to handle such a load is the stuff of both cocktail-party conversations and serious boardroom discussions. While some may say it is doomsday, the-sky-is-falling talk, others warn that to not pay attention to the numbers would be foolish. As with most things, the appropriate level of concern is probably somewhere smack-dab between the extremes. And, of course, the entrepreneur sees it as a magnificent opportunity.

Let's start where everyone agrees: the global communications infrastructure, underpinned by the network, will continue to grow in importance and the volume of traffic will continue to grow at a significant rate. Whether the growth is linear or exponential will most likely be determined by the country and economic conditions, but it is reasonable, if not extremely conservative, to conclude that individuals, businesses, and governments alike will increasingly rely on networks worldwide.

There are predictions that have Internet traffic growing up to 11 times between 2008 and 2015. For example, take voice connections. The *Wall Street Journal* reported that as the 20 exabytes of data generated every year via telephone are transferred to video, those 20 exabytes could multiply "by a factor of 100 or more."[1] There are estimates that the annual run-rate of traffic in late 2012 will be 522 exabytes per year.[2] Exabytes may seem a bit extreme (an exabyte, if you're dying to know, is 1 quintillion bytes, or as Grant Gross in *PC World* estimated, "50,000 years of DVD quality video"), but when you consider that more and more content is being digitized and transferred online, it doesn't look like an unreasonable prediction. In fact, a zettabyte, which is 1,000 exabytes, could very well be the new term of the digiterati beyond 2012.[3]

Take the Laboratory of Neuro Imaging (LONI), which is a leader in the development of advanced computational algorithms and scientific structures for the comprehensive and quantitative mapping of brain structure and function. Located at the University of California, Los Angeles, LONI has the largest neuroimaging database in the world, with more than 25,000 unique scans, at nearly a petabyte in size (1,000 petabytes equals 1 exabyte). Hundreds of researchers work with datasets ranging from 20 megabytes to several hundred gigabytes.[4] And that is just one organization in one field of network use. The uses and bandwidth requirements are similarly grand and growing in almost every industry and aspect of our daily lives. A survey conducted by the Aberdeen Group found that companies, on average, expect to increase their bandwidth by 108% over a 12-month period.[5]

Some would argue that today's networks are simply not designed to support traffic increases of this magnitude. They hypothesize that the networks have not been built out to adequately handle the range of new, media-rich applications that are exploding in popularity today. Consider for a moment the scale and requirements of all the current *and* future users, devices, and applications—and they could be right.

In the recent past, network service providers created service-specific networks, meaning they created a network designed for a single application, such as voice or video, then offered you, the customer, a single service. All of these providers (such as your cable, local phone, or mobile provider) built their networks to deliver one specific service and prided themselves on their ability to do it well, making sure it was always available with no interruption.

Several years ago, that all started to change. As service providers added Internet access to their offerings, it opened up the network to all the different voice, video, and data applications consumers could access (this is often referred to as *triple-play*: voice, video, and data from one service provider). Boundaries between service providers began to blur. Email became accessible on a mobile phone, TV shows were viewable online, phone calls could be made via the Internet, a movie stored on a home device could be watched from a computer in an airport terminal. The lines between services, which had once been easy to discern, began to blur.

In response, there has been a great deal of convergence in the industry, both among the service providers themselves and within their siloed, service-specific networks. Many traditional phone, cable, and mobile phone providers acquired, partnered, or merged with one another in efforts to combine their service-specific networks and provide multiple quality services to customers. At the same time, these providers began to evolve their siloed networks into multiservice networks.

The problem is that because of the predictable performance of the service-specific networks, consumers have come to expect a certain quality. We rely on always hearing a dial tone whenever we pick up the phone; we expect emergency 911 service to always be available; we want a clear picture when watching TV; we require reliable access to email. While we are willing to make allowances for some of the obvious differences among the devices we use to access the content—for example, a picture on a 40-inch flatscreen HDTV is going to be different than the picture on a 12-inch laptop screen—we have little tolerance for any degradation of the overall experience. Let's face it: regardless of how we access a movie, we expect the picture to not be jumpy and the sound to not be choppy. And if it is, we jump ship from one provider to another.

As a result, the providers started announcing plans to build out their IP *next-generation networks* (NGNs) to evolve their current infrastructure to better handle the exploding growth of all the emerging IP applications and the network traffic being generated. This gets back to the question of whether the service provider networks are capable of keeping up with the exploding demands of all these new applications as they converge. And what about the ones we haven't even imagined yet? The Sustainable Network Law (Chapter 5) states that the more broadband made available, the faster network innovation occurs in a somewhat snowballing effect. So, provide a next-generation network, and in a few months you get next-generation applications.

It's highly probable that networks not originally designed to support multiple services will have problems dealing with the new demands of these applications, such as web conferencing and HD video, which require considerably more bandwidth than the pure data applications of even just 10 years ago. About 40 hours of HD video represent as much traffic as 1 million email messages.[6] A single iPhone can eat up as much bandwidth as 5,000 simultaneous voice calls, and HD video takes 35,000 times the bandwidth of an average web page.[7]

In fact, video is often cited as one of the applications driving network utilization and rising bandwidth demands. Take the popular video-sharing website YouTube. In 2007, users uploaded 65,000 new videos and viewed more than 100 million YouTube videos *daily*, representing more than a 1,000% increase from just one year earlier. Only a year later, in August 2008, YouTube was the world's number three site in terms of global minutes, and the number two global search engine, with close to 10 billion search queries a month.[8] In one month in the U.S. alone, users watched 12.6 billion videos, translating to 591 million hours online.[9] (The most popular video was downloaded 88 million times by people around the world by June 2008.)[10]

Given these types of numbers, it probably comes as no surprise that in 2008, Internet video made up approximately one-quarter of all consumer Internet traffic.[11] As video makes the move to HD (and perhaps even 3D at some point in the future), the requirements on the network go up significantly. There are estimates that sites such as YouTube could, by themselves, produce enough data to more than double all Internet traffic.[12]

The 2007 Bernstein report predicts that approximately 157 million Americans will watch video on the Internet at least once a month by 2010.[13] Others predict that in the foreseeable future, 100% of U.S. Internet users will view online videos regularly.[14] IP video is predicted to represent 80% of all video minutes in 15–20 years.[15] This represents a substantial amount of video content, particularly if you consider that the average time a U.S. home used a TV set during the 2007–2008 television season was up to 8 hours and 18 minutes per day, a record high since Nielsen started measuring television in the 1950s (and a statistic worth exploring another day!).[16] Then there are the mobile phone users, who spent three hours a month watching mobile video, a trend that many anticipate will grow significantly, especially in remote countries.[17]

As more people turn to the network, and more content is digitized, and the capabilities of the devices themselves continue to increase to support more of this content, it is perfectly feasible that the network infrastructure could start to strain under the demands. Keeping up with the bandwidth demands—in other words, being able to appropriately process all of the traffic at adequate speeds—definitely poses a challenge.

Keep in mind, however, it is a challenge that those who provision the network—telephone, cable, and mobile carriers—have been dealing with for some time now. Network traffic has been growing by about 50% year over year for the last decade,[18] and at this rate, the amount of traffic in 2020 will be 100 times what it is today.[19] Yet carriers have kept up over the past 10 explosive years.

The amount of Internet traffic in 2020 will be 100 times what it is today.

To see what carriers and providers in the U.S. have been dealing with since 2000, take a look at the predicted traffic demand shown in Figure 6-1. Can you imagine if your 401k graphed that well? The growth isn't that surprising when you consider people in the U.S. who used the Internet were online 27 hours a month (*www.tvturnoff.org*),[20] with predictions that 61% of the connected U.S. population will use the Internet daily.[21] Throw in all the things users do while online, from surfing the Web to watching video or playing online games (there are predictions that 33% of that U.S. Internet population will play games online),[22] and is it any wonder that this has been occupying the world's service providers for the last decade? It hasn't exactly snuck up on them. Their data centers and network operations hubs have been bursting year after year after year.

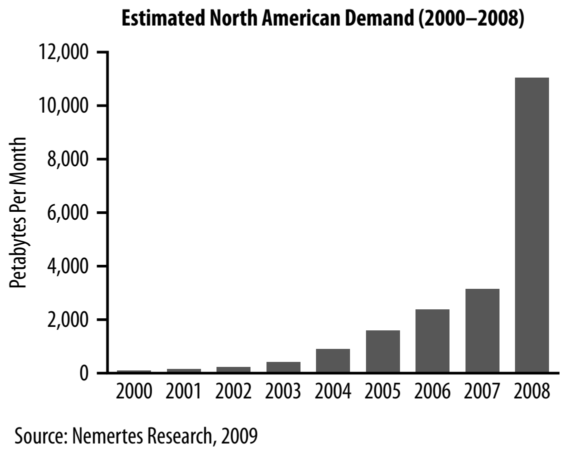

Figure 6-1. *North American Internet traffic from 2000 to 2008 increased slightly; the growth in 2008 makes one wonder if the presidential election had a little to do with it. (In case you forgot, 1 petabyte equals 1 quadrillion bytes.)*

In fact, an overbuild of the network took place in the late 1990s and early 2000s that some providers never recovered from when the dot-com bubble burst. At that time, however, there was really only the promise of the adoption of users, devices, and a plethora of innovative IP applications and services. As this book postulates, the users, devices, applications, and services seem to have caught up. It isn't the "if we build it, they will come" field of dreams of the past. Today, there are already lines, thousands of rows deep, at the ticket booth trying to get into the game.

And while the network providers have been preparing for this moment, at least in their backbone and edge networks, the problem is that the rate of adoption and the ensuing network implications, particularly within the last mile (of the access network), may still have taken some by surprise. With IP traffic growth expected to continue, fueled by an explosion in video and other rich-media content as well as by mobile broadband growth, analysts are predicting bottlenecks and traffic jams on the network by 2012, especially in the U.S. Nemertes, an analyst research group, reported that demand on the network, which it estimated as growing roughly 140–150% year over year through 2013, could be gated by network capacity starting in the 2011–2012 timeframe. *Demand* represents what people would use *if* they could get it, while the *capacity* is the amount of traffic the network can actually handle (the supply). As with anything, when demand overtakes supply, problems arise. In this case, it can manifest itself in slow response times, disruptions to audio and video delivery, and even total unresponsiveness.[23]

We have already seen leading indicators. When pop legend Michael Jackson died in 2009, there was a huge surge in online traffic, with some sites seeing a fivefold increase in traffic that resulted in periods of slowness and even downtime.[24] Now, these were site-specific slowdowns, not network-wide, presumably due to the overload of those sites' web servers and their inability to adequately handle all the requests for information, but they illustrate how spikes and rapid traffic growth can impact the overall quality of network access.

It could be indicative of things to come, particularly for mobile carriers, which are potentially more susceptible to these bottlenecks because of the verifiable explosion of traffic resulting from the rapid adoption of smartphones. For example, it is not unusual for an iPhone user to consume up to 1 GB of data per month, which equates to the consumption of approximately 5,000 simultaneous traditional voice users. For network providers who have been accustomed to supporting those 5,000 users, to suddenly have to support a large number of iPhone users places incredible strains on the network (multiply 2 million new iPhones by 5,000, and those 2 million new smartphone users have just replaced 10 billion cell phone users, or about double the world's actual population). When that happens, real-time users can suffer from latency and lag time, slower downloads, and an overall less-than-stellar user experience.

Providers are working on reducing the amount of traffic flowing through the network. For instance, they are starting to offload smartphone traffic from the mobile network to a fixed one (phone/Internet service) as soon as the user enters her office or home. This automatic switch is only now possible due to the previously discussed convergence of the fixed and mobile networks that is taking place within and among network providers. The benefit is that users can receive a more resilient, consistent connection. We are also seeing developers work on the strain that devices themselves place on the network. For example, RIM has a network operations center that formats web content before it is forwarded to its BlackBerry PDA device. It seems to work, given that the typical BlackBerry user only consumes about 20 MB per month. It remains to be seen, however, whether this approach will maintain its benefits as more and more 3G BlackBerry devices reach the market.

There were some who predicted that if sales of 3G smartphones kept going strong, mobile carriers would run out of radio wave spectrum, which is required to add capacity to the mobile network, by the end of 2008. Now that timeline has come and gone, and there was no major meltdown.

However, there have been hints of the limits of today's mobile networks. On September 17, 2008, the Toronto *Globe and Mail* reported that AT&T, which sells Apple's iPhone in the U.S., is supporting RIM's BlackBerry Bold in 13 countries, but not the U.S. because, "both devices use the same next-generation (3G) network technology, but it now appears AT&T (T) wasn't prepared for the bandwidth-hogging Apple (AAPL) device and doesn't have the resources to launch the Bold until it cures its iPhone issues."[25] In September 2009, the *New York Times* reported that AT&T iPhone users in the U.S. were still experiencing "dropped calls, spotty service, delayed text and voice messages and glacial download speeds," due to the strain the phones were putting on the network.[26]

In early 2009, the *New York Times* wrote a story on the U.S presidential inauguration in the nation's capital that focused on the fact that, "the largest cell phone carriers, fearful that a communicative citizenry will overwhelm their networks, have taken the unusual step of asking people to limit their phone calls and to delay sending photos."[27] They anticipated that the technology-savvy onlookers were going to flood the airways sending and receiving high-resolution pictures and long video cuts, as well as blog posts, tweets on Twitter, and the occasional "it's freaking cold out here!" texts.

Joe Farren, spokesman for the Cellular Telecommunications and Internet Association, was reported by the *New York Times* as saying, "If some of these estimates come true, people should anticipate delays with regards to sending text messages or making phone calls or getting onto the Internet…." According to the article, the group "asked people to send texts rather than make phone calls (text uses less bandwidth than speech) and to send photos only after the event."[28]

The Internet's top 40 sites slowed by as much as 60% by the time the inauguration ceremony started at 11 a.m.,[29] as unprecedented numbers of people logged in to try to watch the historical event online. This points to the potential capacity limits of the current network infrastructure, which strained and sometimes even froze under the spike in general traffic. There were no reports, however, of major issues with mobile connectivity during the January 2009 U.S. presidential inauguration. This could be due in part to the capacity that cell phone carriers feverishly added in anticipation of the onslaught of usage. They provisioned additional access points on cell towers and the necessary landlines to carry the extra traffic from the tower to the provider's backhaul and core networks. This begs the question of how long we can stay ahead of demand with just-in-time buildouts. Events and issues such as these may serve as a wake-up call for the entire industry.

Predictions place the telecommunications service provider infrastructure market at over $100 billion by the end of 2008,[30] but analysts place the specific investment in IP infrastructure at less than 10% of all service providers' capital expenditures. Many are even more conservative, placing the percentage of investment in IP at 4–6%.[31] That is well below some of the estimates of the investment needed to support the proliferation of users of Internet-connected devices and IP services and applications. But then again, Moore's Law[32] may come into play here, which describes the ability to double performance approximately every 18–24 months; it could mean that investment can remain fairly stable, yet still increase capacity at a decent rate.

However, Nemertes has predicted that the amount of investment service providers will need to make to build the IP network capacity required just to meet the projected gap (mainly at the edge) between supply and overall IP demand will be between $42 billion and $55 billion in the U.S. and $137 billion globally.[33] Without appropriate investment, some predict user demand could outstrip broadband Internet bandwidth availability in the next two to four years.[34] Will this happen? Maybe, maybe not.

Traditionally, the bulk of investment in the telecommunications industry has come from the network service providers themselves. However, new models may emerge for sharing the burden of the investment or ensuring that the telecommunications industry is given the proper incentives to adequately build out the network. For example, the emphasis on broadband buildouts, which is on the agenda of many governments, could help.

In U.S., the 2009 stimulus package could support $10 billion of investment in one year in broadband networks.[35] In 2008, the Chinese ministry said Chinese carriers were expected to invest about $41 billion (USD) in 3G mobile broadband networks over the next two years, with at least $29 billion (USD) to be spent in 2009. Australia has announced the buildout of a national broadband network with plans to invest $43 billion over eight years to connect 90% of Australian homes, schools, and workplaces.

Only time will tell if these investments are enough to meet and stay ahead of demand. As discussed in different segments of this book, a lot goes into upgrading the network. Investment includes laying more wires and more fiber-optic cables—particularly, in the last mile, which some say is the biggest lag behind demand[36]—as well as investing in the routing, switching, and security infrastructure, among other technologies, to ensure optimal operation. New developments in architectures (data centers), protocols, and standards (such as Multiprotocol Label Switching—MPLS), and advances in addressing (IPv6) and security, will also be critical in the years ahead to ensure that the network can keep up.

However, it must be said that history is on the side of the network, which has been very adept at finding ways to support the ever-increasing number of users, devices, and applications that want to connect. We have just begun to tap into the promise of the network. What it can do is still to be imagined. It's why the network must continue to scale to demand and not hit the wall. The world's societies, economies, and governments have to be up to the challenge.