



nemertes
RESEARCH
INDEPENDENCE INTEGRITY INSIGHT

```
<link rel="alternate" type="application/rss+xml" title="Nemertes Research" />  
<style type="text/css" media="all">@import "/modules/system/defaults.css";</style>  
<style type="text/css" media="all">@import "/modules/system/system.css";</style>  
<style type="text/css" media="all">@import "/modules/user/user.css";</style>  
<style type="text/css" media="all">@import "/sites/all/modules/event/event.css";</style>  
<style type="text/css" media="all">@import "/sites/all/modules/img_assist/img_assist.css";</style>  
<style type="text/css" media="all">@import "/sites/all/modules/logintoboo.../node.css";</style>  
<style type="text/css" media="all">@import "/sites/all/modules/new.../node.css";</style>  
<style type="text/css" media="all">@import "/sites/all/modules/new.../node.css";</style>
```

```
<script type="text/javascript">  
</script>  
<script type="text/javascript">  
</script>
```

```
<link rel="stylesheet" type="text/css" href="/sites/all/themes/nemertes/images/common/logo.css" />
```

```
<div id="content">  
<div id="header">  
<div id="logo">  
  
<a href="/" title="Nemertes Research">Nemertes Research</a>  
</div>  
</div>
```



The Internet Singularity, Delayed:

Why Limits in Internet Capacity Will Stifle Innovation on the Web



http://www



Independence. Integrity. Insight.

The Internet Singularity, Delayed:
Why Limits in Internet Capacity Will Stifle
Innovation on the Web

Produced by Nemertes Research
Fall 2007

Table of Contents

TABLE OF CONTENTS	2
TABLE OF FIGURES	4
1 ACKNOWLEDGEMENTS	5
2 EXECUTIVE SUMMARY.....	6
3 OVERALL FRAMEWORK: DEMAND, INFRASTRUCTURE, AND INVESTMENT ...	8
4 MODELING USER DEMAND.....	12
4.1 The Application-Centric Demand Model	13
4.2 The Innovation-Centric Demand Model	14
4.3 Moore’s Law for Internet Applications?.....	15
4.4 Behavior and Biological Curves.....	17
4.5 The Emerging Virtual Generation	18
5 MODELING SUPPLY	20
5.1 Optical.....	20
5.1.1 Optical Methodology.....	20
5.1.2 North America.....	24
5.2 Switching and Routing.....	25
5.2.1 Protocols and Layers.....	25
5.2.2 Methodology	26
5.2.3 Core.....	26
5.2.4 Connectivity.....	27
5.2.5 One Box Two Trunks	27
5.2.6 Access.....	31
5.2.7 Wireless: Building Footpaths Across the Digital Divide.....	36
6 INVESTMENT.....	37
6.1 Methodology for Determining Investment	37
7 KEY FINDINGS: THE COMING BANDWIDTH CRUNCH	41
7.1 Global and NA Supply and Demand Curves	41
7.1.1 Demand vs. Supply Overall.....	41
7.1.2 When Is Supply Not Really Supply?	43
7.2 Investment Gap: What It Takes To Prevent The Crunch.....	45
7.3 Sensitivity Analysis	45
7.4 Comparison With Other Studies	47
8 DOES THE INTERNET EVER BREAK?.....	50
8.1 Access Circuit Saturation.....	50
8.2 Router Issues	51

Table of Contents

8.2.1	Router Congestion	51
8.2.2	Addressing and Route Table Expansion.....	52
9	CONCLUSIONS AND RECOMMENDATIONS	54
10	APPENDIX A: DETAILED METHODOLOGY	55
10.1	Detailed Demand Methodology.....	55
11	BIBLIOGRAPHY AND SOURCES.....	59
11.1	Sources	59
11.2	Bibliography.....	59

Table of Figures

Figure 1: Nemertes Internet Model Influence Diagram	9
Figure 2: Global Internet Capacity Curves.....	10
Figure 3: North American Demand.....	16
Figure 4: Global Optical Revenues	21
Figure 5: Incremental Global Optical Capacity.....	22
Figure 6: Projected Global Incremental Optical Investment.....	23
Figure 7: Global Incremental Optical Investment	23
Figure 8: Global Optical Capacity	24
Figure 9: North American Optical Capacity	25
Figure 10: North American Switching Capacity.....	30
Figure 11: Growth in Global Switching Capacity.....	31
Figure 12: North American Broadband Access Lines	32
Figure 13: Projected North American Access Lines	33
Figure 14: Total North American Access Lines	34
Figure 15: Global Access Lines	34
Figure 16: North American Access Capacity	35
Figure 17: World Access Capacity.....	35
Figure 18: Global Internet Infrastructure Investment.....	38
Figure 19: Global Investment in Infrastructure.....	38
Figure 20: North American Investment in Infrastructure	39
Figure 21: North American Investment in Infrastructure	40
Figure 22: North American Capacity versus Demand	42
Figure 23: Global Capacity versus Demand	43
Figure 24: North American Demand Compared to Access Limits	44
Figure 25: Utilization Sensitivity.....	46
Figure 26: Measured vs. Modeled Data.....	48
Figure 27: Internet User Population	55
Figure 28: Total Internet Capable Devices	56
Figure 29: Internet Capable Devices by Region.....	56
Figure 30: Total North American Demand.....	57
Figure 31: North American Nominal Demand.....	58

1 Acknowledgements

Nemertes would like to thank the many outside researchers who have generously discussed their insight and findings with us or reviewed our approach, models, or data, including:

-- Noel Chiappa, Internet researcher and an inventor of the multiprotocol router (1981)

-- kc claffy, PhD, principal investigator for the distributed Cooperative Association for Internet Data Analysis (CAIDA), and resident research scientist based at the University of California's San Diego Supercomputer Center.

-- Jeffrey Cole, PhD, director and Michael Suman at The Center for the Digital Future at The USC Annenberg School

-- Andrew Odlyzko, PhD, Director, The Digital Technology Center at The University of Minnesota.

-- Arielle Summits, Wilson Craig and the rest of the developers of the Cisco Global IP Traffic Forecast and Methodology, 2006-2011, Cisco Systems Inc.

-- The 70 subject-matter experts at major enterprise organizations, service providers, equipment vendors, venture capital and financial firms, and content providers who shared details of their companies' traffic patterns, revenue breakdowns, and future investment strategies in confidence with us.

Your assistance has been invaluable. Thank you.

In addition to the above, we crosschecked our research against statistics provided by the Pew Internet Research and Internet Gatekeepers Inc. as well as against statistics generated by market research firms and provided as part of other research projects (see the bibliography for details). We also drew upon five years of results from Nemertes' proprietary benchmark studies that involved the personal participation of IT executives at roughly 500 enterprise organizations. We are grateful for your collective insight and wisdom.

Please note that providing us with research or reviewing our findings does not in any way imply agreement with our findings or confer responsibility for any errors. Our findings, and any errors, are solely our own.

Finally, the authors of this report would like to thank our fellow Nemerteans and our families for their insight, support, and patience during this project.

2 Executive Summary

In this research study, Nemertes performed an in-depth analysis of Internet and IP infrastructure (which we call capacity) and current and projected traffic (which we call demand) with the goal of understanding how each has changed over time, and determining if there will ever be a point at which demand exceeds capacity.

To assess infrastructure capacity, we reviewed details of carrier expenditures and vendor revenues, and compared these against market research studies. To compute demand, we took a unique approach: Instead of modeling user behavior based on measuring the application portfolios that users had currently deployed, and projecting deployment of those applications in future, we looked directly at how user consumption of available bandwidth has changed over time.

In other words, we assumed that users had consumed, or would consume, a certain amount of bandwidth, and that the rate of change of that bandwidth consumption was the metric that mattered, rather than the specific portfolio of applications. This is similar to the way that Moore's Law focused on the rate of improvement of processing power, rather than the specific portfolio of technology innovations that enabled that rate to occur. We call this the innovation-centric demand model, and we believe it's the only reliable way of "predicting the unpredictable"—obtaining reasonably accurate projections about future scenarios without needing to know the specific innovations that made it possible.

We then validated our findings with the best available data. To gather that data we consulted academic and industry research organizations and conducted primary interviews with more than 70 enterprises, vendors, service providers and investment companies, as well as drawing on our established base of five years of several hundred benchmarked enterprise organizations.

This resulted in the first-ever study that assessed both infrastructure investment and current/projected traffic patterns independently, and compared the two. It is also the first study to apply Moore's Law (or something very like it) to the pace of application innovation on the 'Net—and validate that it appears to conform to the available data so far.

Our findings indicate that although core fiber and switching/routing resources will scale nicely to support virtually any conceivable user demand, Internet access infrastructure, specifically in North America, will likely cease to be adequate for supporting demand within the next three to five years. We estimate the financial investment required by access providers to "bridge the gap" between demand and capacity ranges from \$42 billion to \$55 billion, or roughly 60%-70% more than service providers currently plan to invest.

It's important to stress that failing to make that investment will *not* cause the Internet to collapse. Instead, the primary impact of the lack of investment will be to throttle innovation—both the technical innovation that leads to increasingly newer and better applications, and the business innovation that relies on those

Executive Summary

technical innovations and applications to generate value. The next Google, YouTube, or Amazon might not arise, not because of a lack of demand, but due to an inability to fulfill that demand. Rather like osteoporosis, the underinvestment in infrastructure will painlessly and invisibly leach competitiveness out of the economy.

One could even whimsically speculate—as we did in the title—that the lack of investment could be holding back the time at which the Internet reaches a “singularity” (a point at which accelerating change creates an unpredictable outcome, such as the Internet becoming independently sentient).

More seriously, we did not set out with an agenda, or to prove or disprove a particular point (singularity or otherwise). We modeled capacity and demand using the best tools at our disposal and validated the findings as fully as possible against the best available data. The result is, as results of all models necessarily are, a projection, the accuracy of which we can only improve with better data and more refined modeling techniques, and we welcome suggestions on the latter and access to the former.

We did come away with the overwhelming conviction that there is a deep industry need for better (more comprehensive and accurate) data in this area. The Internet is almost opaque to serious researchers, even those with the necessary technical skills, integrity and desire, for the simple reason that carriers and content providers refuse to reveal their inner workings. The reasons for this are good—service and content providers are reluctant to reveal their proprietary competitive advantages, or accidentally breach their customers’ privacy—but the need for open, honest, and comprehensive information exchange is acute. So we conclude by urging content and service providers to cooperate with researchers in sharing data.

3 Overall Framework: Demand, Infrastructure, and Investment

Internet modeling exercises have typically concentrated on demand and neglected capacity (or vice versa). Moreover, traffic demand and growth measurements almost inevitably assess traffic that's already on the network, for the very logical reason that it's much easier to capture aggregated traffic statistics from devices in the core, rather than monitoring devices at the edges attempting to inject traffic into the network to see how well they succeed. (One is reminded of the old joke about the drunk looking for his keys under the streetlight, not because that's where he left them but because it's where the light is.)

While this approach can provide useful predictions of the ways in which loads evolve over time, it does not really give any insight into the ways in which lack of capacity degrades service, or actively limits demand. Moreover, it provides virtually no insight into how demand is generated—thus missing one of the most critical pieces of the demand/capacity equation. Since one of the hypotheses Nemertes wanted to explore was the possibility that capacity might be limiting demand either at present or in the future, we needed an approach that allowed us to do so.

We therefore modeled demand and capacity independently. This independence is important because it allows us to decouple the impact of capacity on demand. If they are not decoupled, the model will fail to capture a scenario in which capacity is limiting demand. Researchers may be able to note that demand has slowed, but won't be able to tell if this is a true measure of actual demand (users conveniently want no more capacity than happens to be available to them) or a case in which capacity is in fact the limiting factor. Our model was designed to capture reality as exactly as possible, given the rather considerable constraints inherent in a situation in which there is limited reliable data, and future projections are heavily dependent on the emergence of unknown applications and capabilities.

Nemertes approached its analysis by considering the Internet as an exercise in supply (expressed by infrastructure) and demand (expressed in terms of the Internet usage patterns). The first step in doing this was to build an influence model that would serve to direct our data gathering (Please see Figure 1: Nemertes Internet Model Influence Diagram).

The demand side of the model is driven by users, which we segregated by geography. For the purposes of this analysis, we grouped them by North America, Europe, Latin America, Asia Pacific and Africa Middle East. Each of these groups has differing levels of access to a series of Internet-attached devices, each of which run a range of applications. The degree to which a device can generate load is proportional to the amount of time the user desires to do so as well as the degree to which the device in question is physically capable of pushing data. Each of these applications, then, drives data through the five geographic areas and ultimately generates a load that is felt by the Internet as a whole.

Overall Framework: Demand, Infrastructure, and Investment

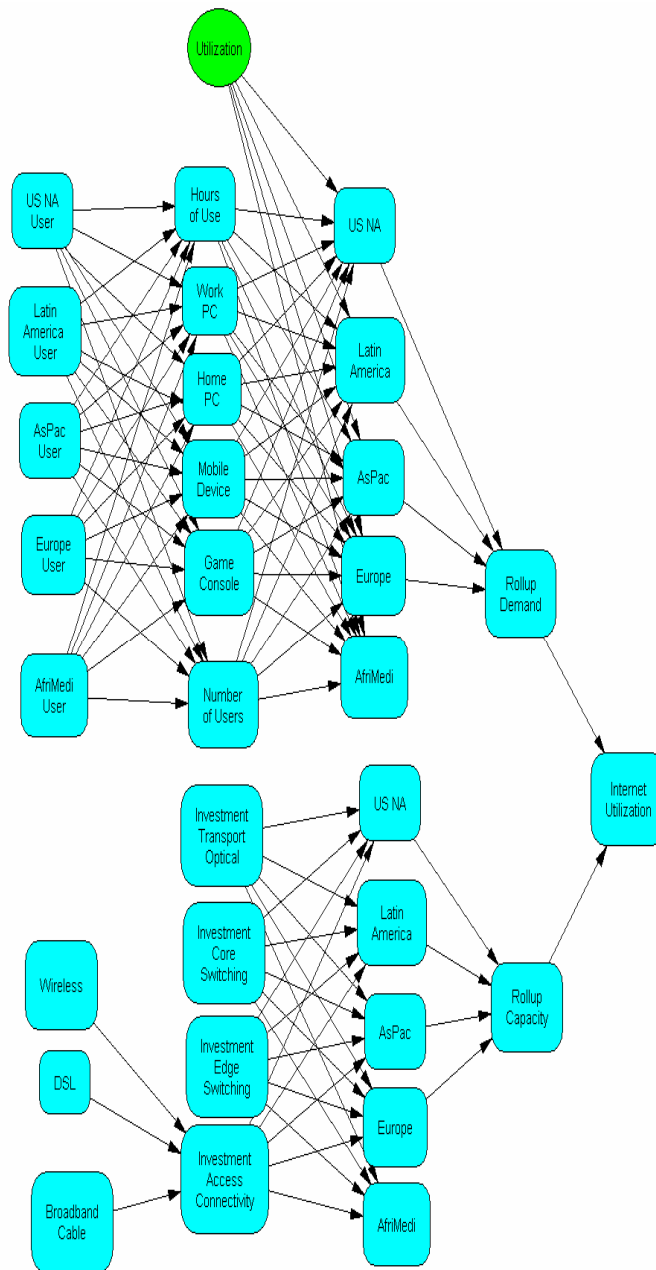


FIGURE 1: NEMERTES INTERNET MODEL INFLUENCE DIAGRAM

It bears noting that Nemertes did not try to model a typical user for each geography. However, the approach selected allowed us to model a maximal demand profile for each region, which sets the upper bound for demand, without having to survey thousands of users. In telephony terms, Nemertes' approach focused on the absolute busy hour characteristics of a user group rather than trying for a comprehensive, user demand profile over time.

Overall Framework: Demand, Infrastructure, and Investment

The capacity side posed its own problems. Most Internet modeling ignores the supply side entirely, or simplifies it considerably, not without reason. True capacity is defined as the maximum throughput measured over some time period. It turns out that this is a complex undertaking when the thing whose throughput you are measuring is characterized by billions of nodes with billions of potential paths from one node to another node. The degree of complexity is compounded when you realize that many different routing protocols are being used such that determining the path that may be used is somewhat indeterminate.

Virtually all the available research literature that attempted to model such a problem was concerned with deriving an algorithm that could be used, rather than one that worked in a practical sense for sizing the Internet. As a consequence, in every case that we examined, the algorithm was far too complex to actually solve. Instead, we opted for a more simplistic approach that fundamentally treats the Internet as a series of containers for holding bits. This approach allows us to simply count up the devices that generate bits, multiply them by their maximal bit rate and then add up the capacities to obtain the total capacity.

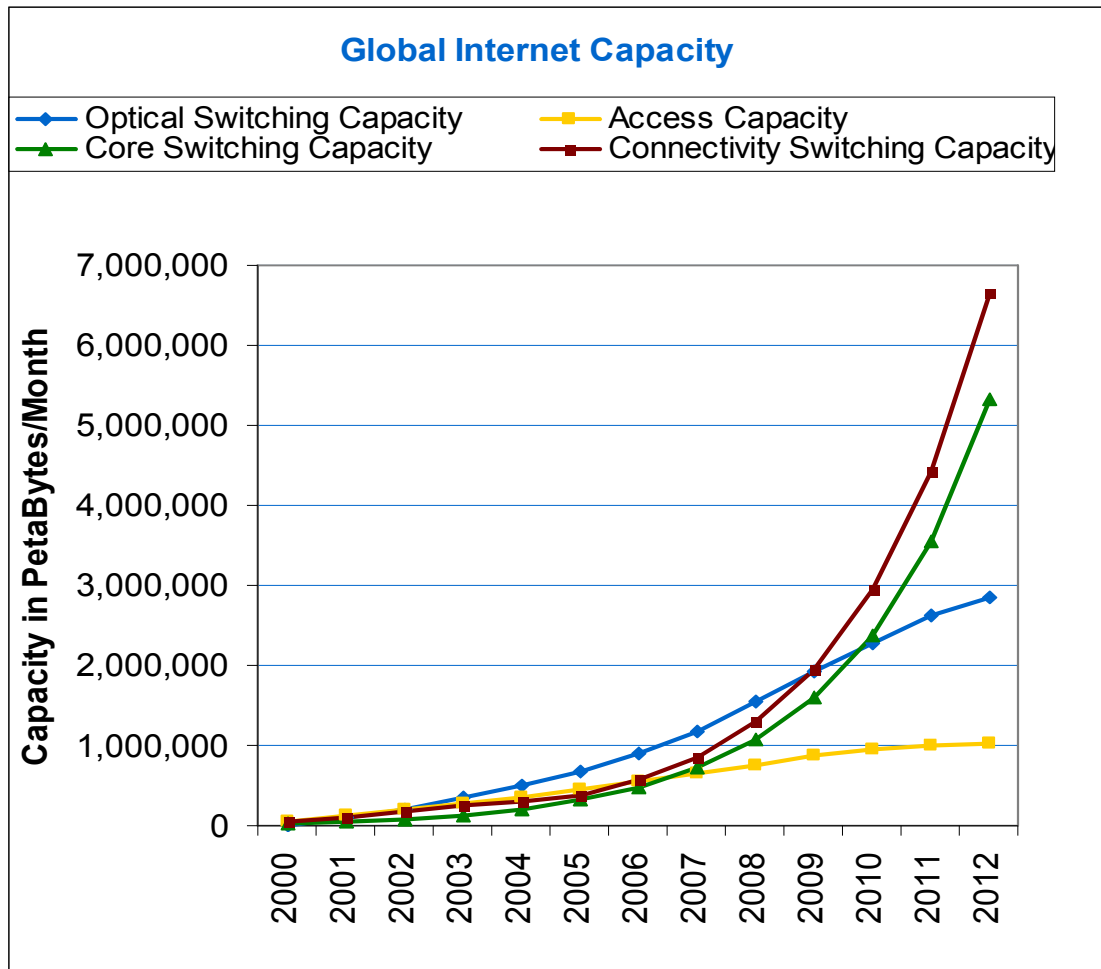


FIGURE 2: GLOBAL INTERNET CAPACITY CURVES

Overall Framework: Demand, Infrastructure, and Investment

The model approaches demand from the perspective of several domains: core switching, optical backbone, access lines and connectivity switching (Please see Figure 2: Global Internet Capacity Curves). With the exception of access lines, which are physical and wireless transmission media, the balance are represented by investment in electronics. In each case, we researched the annual capital equipment expenditure, translated that investment into capacity, and computed a total incremental investment/capacity. These figures, then, became the discrete capacities against which we compared demand.

We then plotted each connection modality against the respective geographical regions, and summed the resulting capacities to arrive at the total supply. We plotted both supply and demand as petabytes per month. Although a petabyte per month is a misleading measurement, since it tends to mask things like peak usage and imply a continuous use dynamic, we nevertheless adopted it since most of the literature has settled on this convention.

It turns out that taking this approach is not a bad match to the approach we used for assessing demand. Since we approached both from a maximal perspective, any difficulties in servicing demand would be expressed at the margin and allow easy testing of assumptions and sensitivity. We reasoned that if the model indicated a problem developing in a situation of maximal demand serviced by maximal capacity, it would be a simple thing to back off demand to see at what point the demand could be serviced, then compare the findings to reality.

4 Modeling User Demand

To model user demand, we first looked at the maximum possible demand, which essentially measures how much data users could hypothetically generate on their Internet-attached devices, given the following:

- a) the types of Internet-connected devices (home PCs, work PCs, mobile Internet devices, interactive gaming consoles, and IP-enabled TV)
- b) the port speeds of those devices (e.g. the maximum possible data rate of the network interfaces)
- c) the number of devices.

To create a maximum possible demand curve, we assessed the number of each type of device available to each user for the years 2000 through 2012 based on publicly available data and projections. (Please see Appendix A: Methodology, for further details). We found that, counting all devices and all port interfaces, an Internet-attached North American user in 2007 is theoretically capable of generating approximately 61 Mbit/s of traffic, equating to 20 Petabytes of traffic per month.

It's important to note that by this measure, not only does maximum possible demand exceed Internet capacity today, but it has always exceeded capacity and very likely always will, because of the fact that port speeds on Internet-attached devices tend to be within an order of magnitude or so of Internet circuits themselves (and there are far more devices than circuits).

For example, back in the 1990s when the Internet backbone consisted of T1 (1.5 Mbit/s) and T3 (45 Mbit/s) circuits, the LAN speed of the typical Internet-connected host was 10 Mbit/s (an order of magnitude larger than the backbone). Today, a typical Internet-connected PC has a 100 Mbit/s or Gbit/s Ethernet link, and the Internet backbone circuits are generally OC-768 (38 Gbit/s).

As Andrew Odlyzko and K.G. Coffman noted back in 2001, "For the foreseeable future a handful of workstations will in principle be able to saturate any given Internet link. A few thousand machines will continue to be capable of saturating the entire Internet." (Coffman and Odlyzko, 2001).

Moreover, access circuits (the "last mile" typically connecting Internet-attached devices to the Internet) tend to run at speeds several orders of magnitude lower than the devices they serve. A typical DSL circuit, for example, delivers a maximum of 1.5 Mbit/s to one or more Internet-attached devices in each household (and each device, keep in mind, has a port speed of 100 Mbit/s or up). This means that the Internet-attached devices in a typical household are fully capable of saturating the household's Internet connection; the same holds true for business sites.

However, simply because the devices *can* saturate the 'Net doesn't mean they *will*. In practice, network-attached devices very rarely (in fact, almost never) generate traffic at 100% of port capacity for a sustained period of time. The actual port utilization depends heavily on the machine's CPU, the types of applications running on the machine and the habits of the user.

Modeling User Demand

So, the second step in constructing a demand model lies in determining how much of that hypothetical capacity is being used. To gain this perspective, we consider the fact that applications are running on devices and, more importantly, people are operating the devices. This will not always be the case; as many researchers have observed, machine-to-machine traffic will begin to become significant over time. However, in the window of this study (2000 – 2012) we still foresee the majority of Internet traffic being associated with people operating devices and driving applications.

4.1 The Application-Centric Demand Model

There are fundamentally two approaches to constructing an Internet traffic demand model. The first approach is to try to create one or more “typical user profiles” that describe which applications each user is running, for how long, and how much bandwidth a typical application consumes as it’s being run by a typical user. (This is more difficult than it sounds, because application usage can often be highly idiosyncratic, particularly for consumer applications: Some people change channels 50 times in an hour and download many more movies than they can possibly watch; others watch nothing but their daily half-hour news shows.)

Once the researcher creates user profiles, he or she then attempts to measure historical usage patterns for the application to arrive at total traffic volume. Finally, the researcher projects current trends into the future on an application-by-application basis to estimate how traffic volumes might evolve.

This approach has a couple of significant advantages, including granularity (researchers can distinguish among different application types), defensibility, and comprehensiveness (if the list of applications is comprehensive, researchers can feel relatively certain they’ve accurately captured the majority of the traffic that’s out there, assuming the profile modeling is correct).

Most likely for these reasons, this approach is favored by several consequential researchers in this space, most recently an internal research team at Cisco Systems Inc., which deployed this approach to model consumer application deployment in (Cisco, “Global IP Traffic Forecast and Methodology, 2006-2011”, 2007). The Cisco model assessed all known consumer applications generating significant amounts of Internet and IP traffic in the summer of 2007, and extrapolated these out through 2011. (Cisco also looked at enterprise applications, but with less granularity, which is reasonable as these tend to be more difficult to classify meaningfully into categories).

There is one significant problem with this approach, which is that by definition, it can’t easily account for innovation. In particular, it does a poor job of meaningfully predicting sudden shifts in user behavior, or the emergence and rapid adoption of new applications. As J. Licklider has stated: “People tend to overestimate what can be done in one year and to underestimate what can be done in five or 10 years.” ([Licklider] J. C. R. Licklider, *Libraries of the Future*, MIT Press, 1965.)

Modeling User Demand

This is because in order to factor these in to such a model, researchers have to have some idea in advance of what these shifts and applications might be. This essentially forces researchers to become futurists. Yet predicting the future is notoriously difficult, even for those whose livelihoods depend on doing so, and the difficulty goes up dramatically with the specificity of the required prediction (it's easier to correctly predict that the economy will go up sometime within the next five years than that a specific technology will emerge in, say, autumn 2010).

Moreover, getting into the game of predicting specific applications weakens the predictive effectiveness of the model, because if an application doesn't emerge precisely as planned (something that's highly likely) the model will be less accurate. An application-centric approach wouldn't have predicted the rise of YouTube, for example, which emerged in 2005, and which Cisco says was already responsible for roughly 27 Petabytes/month in 2006—about as much traffic as traveled on the Internet in total in the year 2000.

Skeptics may point out that as impressive as that sounds, 27 Petabytes/month represents a relatively small percentage of current Internet traffic, and that's certainly true. As we'll discuss further in Section 7, however, the stage is already set for technologies to emerge as rapidly as YouTube (well within the window of our study horizon, in fact) which may have a significant effect on overall Internet traffic.

4.2 The Innovation-Centric Demand Model

The second approach is to assume, a priori, that innovation occurs. That is, user pattern shifts and new applications will emerge, and the rate and impact of these new behaviors and applications can be both modeled historically and projected into the future without knowing the specific details of these changes in advance. This approach is particularly useful when modeling technical innovation—it allows researchers to predict *that* a certain event will occur without requiring them to predict precisely *how* this will happen.

The well-known Moore's Law is an example of this type of predictive model: Back in 1965, Intel co-founder Gordon Moore observed that the number of transistors that can be inexpensively placed on an integrated circuit is increasing exponentially, doubling approximately every two years. The trend has continued for more than half a century and is not expected to stop for at least the next decade.

The significance of Moore's Law is that Moore did not have to describe precisely *which* technical innovation would lead to this doubling between, say 1988 and 1990—all he had to do was validate that such innovation was in fact occurring at an exponential rate. So long as the law continued to apply, he could be confident the innovation would occur.

The strength of an innovation-centric demand model is that it's the best available mechanism for modeling the unknown. The main weakness (other than the loss of a certain amount of granularity) is that it applies only under conditions

Modeling User Demand

of continuous innovation, and it's difficult to know in advance what those conditions are, or at what point they will cease to occur.

As has been noted extensively elsewhere (Kurzweil, 2006) innovation-centric models tend to result in exponential increases, and in the real world exponential increases tend to be halted by external environmental forces (otherwise the world would be knee-deep in rabbits, which like many other animals, reproduce exponentially).

The weakness of an innovation-centric approach is that it might not actually apply, which makes validation against measured data extremely important for models of this type. (Unfortunately, as we note later on, through no fault of the researchers, credible measured data is impossible to come by in the area of Internet traffic measurement. This is one of the key issues that researchers must address to enable the industry to make informed decisions).

The challenge in selecting the right model for the demand curve, then, comes down to choosing between the application-centric and innovation-centric approaches. Again, application-centric models, if well-executed and validated, do a fine job capturing the current state and enabling near-term projections of that state. But they do a poor job at projecting future states that could be highly dependent on innovation. We believe that this fundamental weakness of the application-centric approach precluded its use here.

4.3 Moore's Law for Internet Applications?

For these reasons, the Nemertes model deployed instead takes an innovation-centric approach to projecting user demand. Essentially, we applied a Moore's Law-like approach not to Internet traffic volumes, but to the applications and devices that generate such traffic.

As noted, we created a user profile that included an aggregate set of Internet-attached devices (each with an associated port speed). The devices available to each user, as well as the port speeds associated with each device, can be reliably validated from 2000 to 2006 (based on market figures for availability of such devices), and projected for the years 2007-2012. It's worth noting that both the number of Internet-attached devices, and the speeds of those devices, associated with each Internet-connected user increase dramatically over time (see Methodology for details).

Then we assessed the limited data available on current utilization of users accessing the Internet through unlimited-bandwidth pipes. We converted those utilization figures to traffic volumes, and as will be discussed shortly, computed those traffic volumes as a percentage of maximum theoretical throughput.

We then developed a growth rates for these utilization figures, to account for the fact that measured utilization is unlikely to remain constant for the 12-year period under study. In fact, as will be discussed, we have good reason to think that it will rise dramatically. We applied these growth rates retroactively to estimate the volume of Internet traffic for the years 2000-2007. Finally, we

Modeling User Demand

projected forward to estimate the amount of user demand for the years 2007 through 2012.

The utilization rate that we computed is based on measurements and assertions that we believe are close to being accurate. In 2001, an evaluation of the Internet conducted by Coffman and Odyzko concluded that the Internet traffic could not be more than 85 Petabytes per month. We also had assertions by the carriers that since 2000 they have seen growth in demand that approached 100% a year.

An additional datum is that several researchers reached independent estimates of Internet usage for 2006 in the 2000 Petabyte-per-month range. If the demand increases that the carriers saw and the measurement seen in 2000 and 2006 were correct, the implication was that demand was really growing at rates approaching and in some cases exceeding 100% year over year.

Growths of this magnitude translate into utilization rates that, in 2006 stand at .066 % or about 350 Megabytes per day of usage. This seems imminently reasonable. It also agrees with what one would expect of an average user of the Internet. This is equivalent to downloading about an hour of Internet video, or multiple hours of working, emailing, talking, sharing, uploading, downloading and watching video—often at the same time. Based on the most recent figures from the Annenberg Center for The Digital Future, the typical Internet user spends roughly four hours per day “actively using” the Internet (as opposed to being merely online), spread across multiple devices, and the majority of that time is spent heavily multitasking, often running streaming multimedia applications in the background while focusing on less bandwidth-intensive applications in the foreground.

As a final step, we projected demand curves into the future based on estimates of utilization and growth that show the influences of Moore’s Law behavior as described above, resulting in the following demand chart for North American traffic (Please see Figure 3: North American Demand).

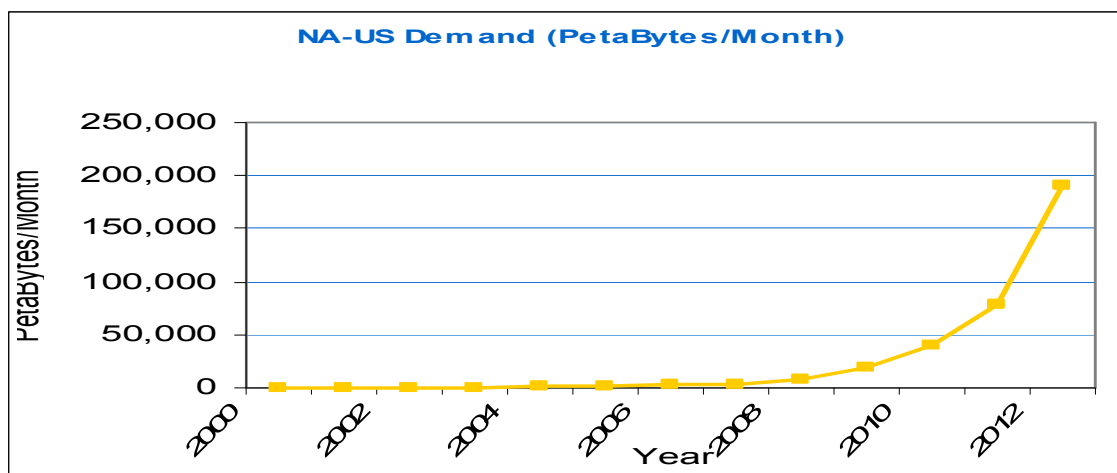


FIGURE 3: NORTH AMERICAN DEMAND

4.4 Behavior and Biological Curves

At first blush, this projection might seem a bit eyebrow-raising. Two hundred Exabytes (200,000 Petabytes) after all, equates to an individual user consuming or generating 26 Gbytes/day by 2012. Even spread across multiple devices, this would seem to be extreme—it equates to each and every user deploying nearly seven hours of high-definition interactive video per day.

And before we explain how we arrived at such a projection, and why we believe it's reasonable, it's worth stating clearly: *We are not predicting that this level of traffic will occur.*

In fact, we are virtually certain that it will *not*, because it can't. As we'll discuss at greater length in Section 7, even if such a demand existed, it could not be satisfied, because the infrastructure to support such a demand is lacking.

However, we believe that the demand *could* occur, and in the absence of constraints imposed by infrastructure, that it would be at least likely, if not guaranteed. Here's why: As noted, an innovation-centric model is the best possible way to model innovations that can't be precisely predicted.

The challenge lies in two areas: First, determining the appropriate measures that are increasing (or possibly increasing) exponentially. Second, validating that the exponential innovation model actually applies at a particular point in time (as opposed to somewhere off in the future).

Regarding the first, as noted the measure we settled on was utilization, specifically utilization of the local bandwidth (e.g port speeds) available to the user. This is distinct from utilization of WAN access circuits, which is something we will discuss shortly. Our key finding is that utilization of available port speeds appears to have been increasing exponentially over time (and is in fact responsible for a considerable part of the high traffic growth measured by Odzko and others in the earlier part of the decade). Given that our analysis matches the best-available data, we can be reasonably sure that it has been accurate at least thus far.

In short, users have been demonstrably using more and more of the capacity that's available to them—and we believe this is a growth rate that will continue. Coupled with the fact that there's dramatically more bandwidth available to them going forward (again, talking in terms of port speeds, not necessarily access), this results in the startling growth projections reflected here.

That brings us to the second point: *Why* do we believe the utilization growth rate will continue, barring external constraints? Why wouldn't it slow down or even decrease?

This is simply another way of asking ourselves, “Does the exponential innovation model apply at present?” We believe it does for the following reasons. First, the assumption matches the best available data so far (with a caveat discussed in Section 7). As noted earlier, agreement with measured data is one of the key validations required for this type of model.

4.5 The Emerging Virtual Generation

Second, we believe behavioral conditions favor it. It's true that there's often a significant time lag between the point in time when an innovation is introduced and when it reaches widescale deployment. But as has been discussed copiously elsewhere, the millennial generation (18 and under) that is coming of age is comprised of people with demonstrated capacity to adopt networked applications in widespread fashion in record time—12-18 months or less. (Examples include YouTube, Facebook, etc.) Based on the best available data, including that from the Center for Digital Life, Internet users today, and younger users in particular, are:

- Multitasking (2/3 say they run more than one application at one time)
- Running IP-enabled devices at once (in other words, it's not "either/or" it's "both/and")
- Switching preferentially from non-IP applications to IP-based versions of the same applications (for example, listening to radio streams on the 'Net rather than via satellite or radio, and watching television shows on the 'Net instead of on TV).

All these trends drive up utilization, and all are accelerating.

Third—and quite importantly—we believe that a key cluster of technologies have matured to a point that enables rapid, low-cost development of a broad range of applications. These technologies include Web 2.0 programming and development tools, HD displays, low-cost cameras and recorders, and data storage. According to a recently published IDC/EMC study, the amount of data globally in 2006 was some 161 Exabytes, proceeding to 988 by 2010, with much of the data-generating applications existing on Internet-attached devices (phones, PCs). (IDC, 2006).

Although we're deliberately trying to avoid predicting specific applications, it's worth considering the following: MIT researchers have already developed devices that allow users to set up a variety of "always on" video connections between friends and family members. These can be displayed on wall-mounted high-definition displays, or carried as low-definition keychain-sized trinkets (with wireless access to the 'Net). Either way, they enable users to set up sustained interactive connections between remote parties.

Why would anyone do this? Consider the following trends:

- Increasing oil prices, raising the cost of travel
- Aging baby boomers caring for elderly parents, with college-age children
- A far-flung, highly distributed population.

A wall-mounted HD display with high-resolution camera and an easy-to-use interface could serve as a "virtual window" between families and friends. (Users could also use the system to download a changing display of photos and artwork from the 'Net, as a high-end restaurant in Dallas already does).

Such a "virtual window" would enable adults to keep better track of their aging parents and far-off children, and maintain long-distance friendships. This

Modeling User Demand

is something that early adopters already do (though typically via low-res Webcam connections from hotel rooms while traveling).

But what really moves this concept from “nice-to-have” to “need-to-have” is the rising cost of energy, which is already beginning to limit extraneous travel. A steep increase in oil prices (something that’s certainly not unthinkable) could drive widespread adoption of such an application within the next five years in a “tipping point” effect. Nemertes is already seeing similar trends in the business world, with over 80% of enterprises defining themselves as “virtual workplaces” and nearly 40% planning to use video over IP within the next 12 months, with travel avoidance cited as a key driver (Source: Nemertes Benchmarks).

And again, deployment of this sort of “virtual window” technology deployment is possible well within the timeframe of this study. As noted earlier, the key technologies—low-cost, high-definition displays, inexpensive cameras, copious storage—already exist. Within this context, the idea that the typical user will be running hours of HD interactive video by 2012 suddenly doesn’t seem so farfetched.

As noted, we aren’t in the business of predicting particular applications, and we don’t mean to imply that just because this particular application *could* happen, it necessarily *will*. The point, again, is that technologies enabling this type of application, and many others, already coexist with a ‘Net-savvy generation predisposed to use them and an economic environment that increasingly favors their use.

In sum, we believe that the environment necessary for a Moore’s-law increase in application utilization exists today. Or as Vint Cerf put it recently to the Washington Post: “Once you have very high speeds, I guarantee that people will figure out things to do with [them] that they haven't done before.” (Cerf 2007). Specific details on how we constructed our demand model are described in the Methodology. But to see how the model plays out in the context of the current capacity environment, we have to first examine infrastructure capacity.

5 Modeling Supply

5.1 Optical

Internet capacity, at the highest level, is defined by how much traffic the optical backbone can carry. A significant part of the Nemertes model is devoted to assessing the capacity of this optical infrastructure.

The optical backbone that makes up the long haul capability and, to an increasing extent the distribution to the user, is composed of optical fibers driven by opto-electronics. These fiber runs, for all practical purposes, have inexhaustible bandwidth. The limit to what can be carried is set by the opto-electronics.

For this reason, determining the carrying capacity of the fiber is less an exercise in counting fibers, than it is in inventorying the optical hardware that drives them then multiplying by the speed at which they can operate.

5.1.1 Optical Methodology

While inventorying hardware in the Internet is not really practical, from a direct examination standpoint, it is nevertheless possible to review the amount of equipment being deployed into the network annually and make an informed assumption as to the degree to which that investment translates into optical capacity.

Various financial filings by optical electronic vendors were examined to determine the annual revenue of the largest optical equipment manufacturers. (Please see Figure 4: Global Optical Revenues), in the year 2006, the total revenue for these manufacturers worldwide amounted to more than \$6.3 billion.

While the revenue of the manufacturers is not precisely the same as the investment in the network, it is still true that the investment would necessarily be tightly correlated with the revenue of the manufacturers. The missing piece would be the amount of vendor revenue that is derived from non-equipment sales. This varies by vendor, but tends to be de minimis, when compared to the revenue generated from the sales themselves. For the purposes of this study, Nemertes assumed that, for all practical purposes, they are the same.

The missing portion of the equation, then, is the amount of capacity that a given investment dollar buys. We computed this capacity amount on a historical basis by dividing the investment by the transmission capacity of the devices being purchased. This was distributed among optical equipment operating at the OC 12, OC 48, and OC 192 rates. For the purposes of this analysis, the bandwidth figures that were used to represent the OC transmission rates were 622 megabits per second for an OC 12, 2,488.2 megabits per second for an OC 48, and 9,953.28 megabits per second for an OC 192.

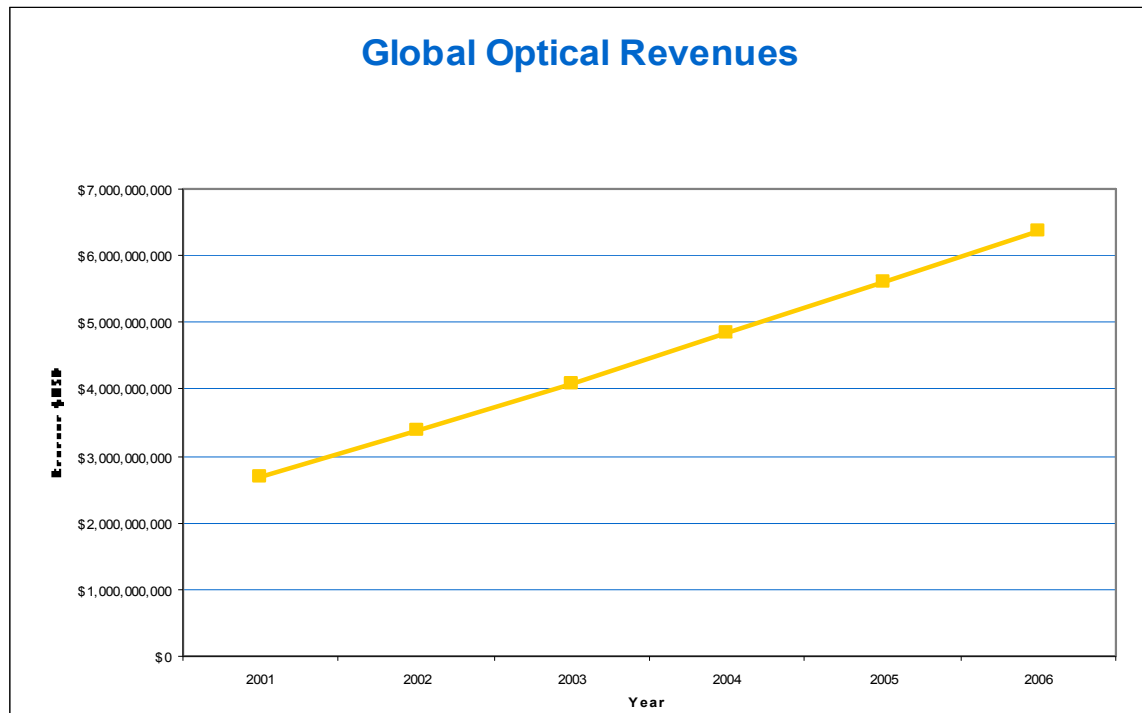


FIGURE 4: GLOBAL OPTICAL REVENUES

These data rates are deployed over various optical transmission/detection technologies, primarily dense wave-division multiplexing (DWDM), multiservice provisioning platforms (MSPP), and reconfigurable add-drop multiplexers (ROADM). DWDM uses different wavelengths of light to carry different transmission channels. In DWDM, this can be as high as 80 channels with 50 GHz spacing. DWDM is most usually deployed to drive single mode fiber, whose core diameter is 9 micrometers.

MSPP, on the other hand is most usually found in metro rings. MSPP is used to drive SONET (Synchronous Optical Network) protocols to multiple add/drop multiplexers. Such setups frequently use multi-mode fiber with core diameters of up to 62.5 micrometers.

Finally, ROADMs are rapidly being superseded by MSPP and is expected to generally disappear from the metro environment by 2012.

The following (Please see Table 1: Optical Investment by Technology Type) shows the investments in the various technologies. Knowing the relative amount of optical data rates deployed in each technology, allows us to compute the optical capacity for each year. This gives us an incremental worldwide optical carrying capacity of 2.78 petabits per second in the year 2006 (Please see Figure 5: Incremental Global Optical Capacity).

Modeling Supply

Technology	2001	2002	2003	2004	2005	2006
Revenue						
DWDM (Core)	\$8,700,000,000	\$7,250,000,000	\$5,800,000,000	\$4,350,000,000	\$2,900,000,000	\$1,450,000,000
DWDM (Metro)	\$0	\$0	\$0	\$0	\$0	\$0
ROADM (Metro)	\$0	\$0	\$561,000,000	\$374,000,000	\$187,000,000	\$0
MSPP (Metro+Core)	\$0	\$0	\$0	\$0	\$0	\$0
Total Investment	\$8,700,000,000	\$7,250,000,000	\$6,361,000,000	\$4,724,000,000	\$3,087,000,000	\$1,450,000,000

TABLE 1: OPTICAL INVESTMENT BY TECHNOLOGY TYPE

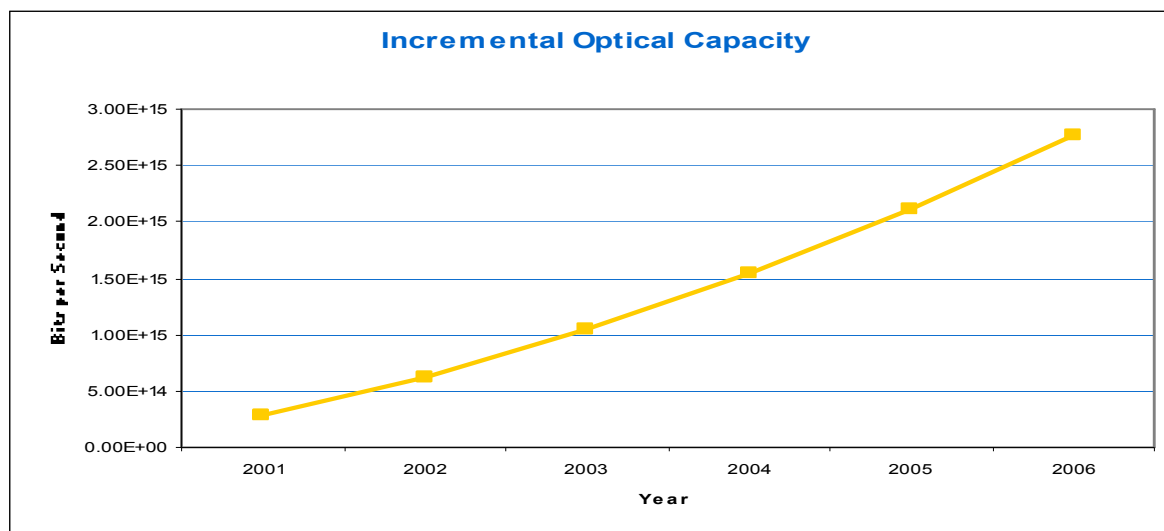


FIGURE 5: INCREMENTAL GLOBAL OPTICAL CAPACITY

The essential problem, though, is not how much capacity was available in the past or is available now, but how much will be available in the future. While this is not precisely predictable, we can look at the revenue projections for the various manufacturers, look at the projections made by the various industry trade groups and look at recent historical trends.

Additionally, we can look at the predictions being made for relative revenues for each of the optical technologies. Predictions made by Lightwave in 2006 showed Metro MSPP technology investment flattening out in 2009, while ROADM investment would essentially disappear by 2012.

When these predictions are factored into the linear curves that simple trending would produce, the result is the following set of investment lines (Please see Figure 6: Projected Global Incremental Optical Investment). When we add these, the following chart results. (Please see Figure 7: Global Incremental Optical Investment).

Modeling Supply

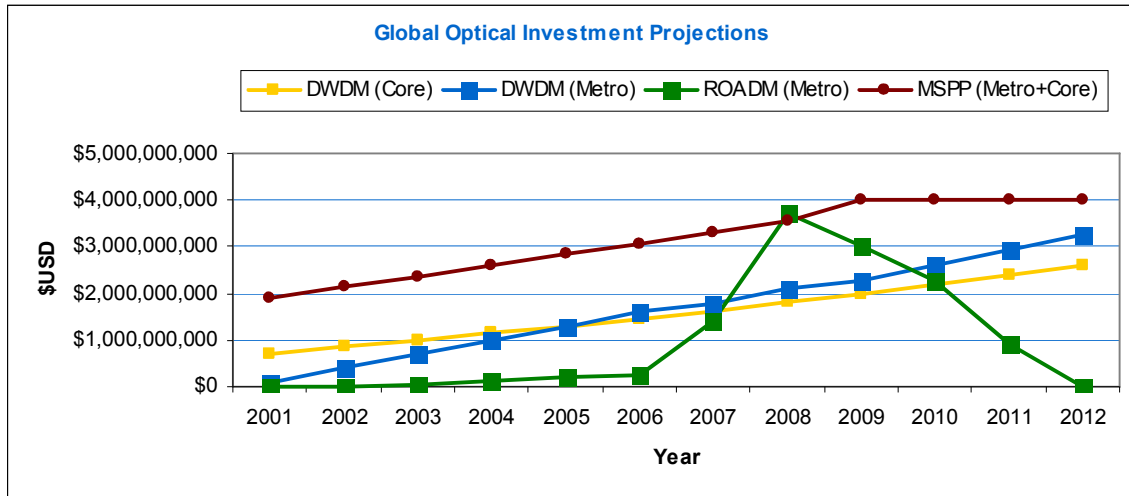


FIGURE 6: PROJECTED GLOBAL INCREMENTAL OPTICAL INVESTMENT

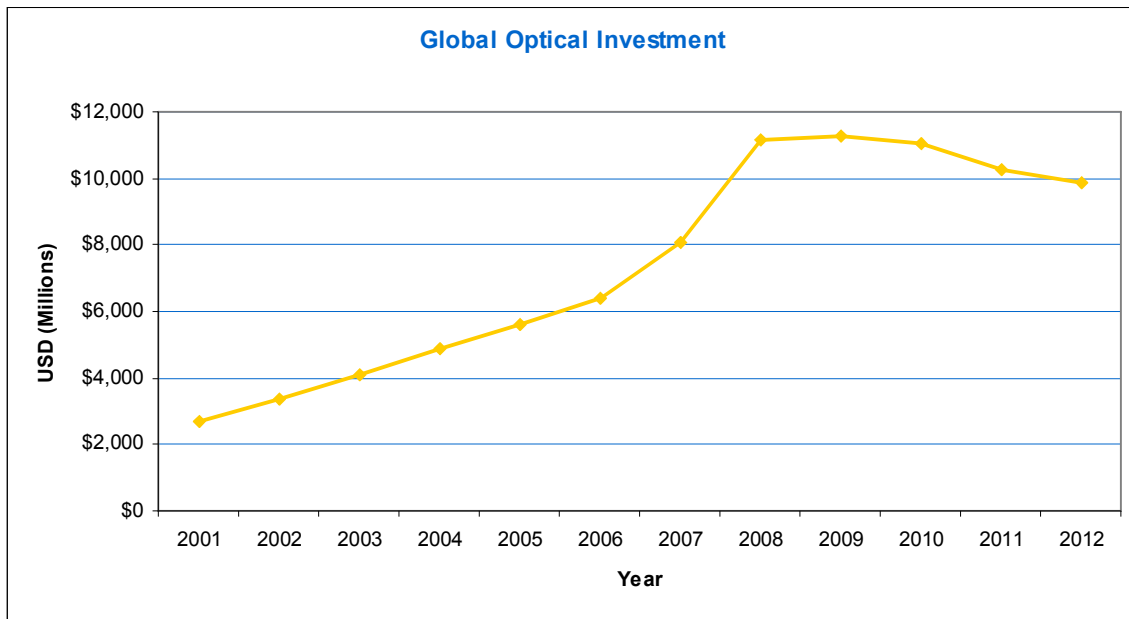


FIGURE 7: GLOBAL INCREMENTAL OPTICAL INVESTMENT

As can be seen, the trend line predicts that by the year 2012 approximately \$9.8 billion of new optical investment per year will be made worldwide. This investment, when converted to capacity, shows the following curve (Please see Figure 8: Global Optical Capacity). As shown, the total capacity of the backbone is approximately 5 million Petabytes per month or 5000 Exabytes.

Modeling Supply

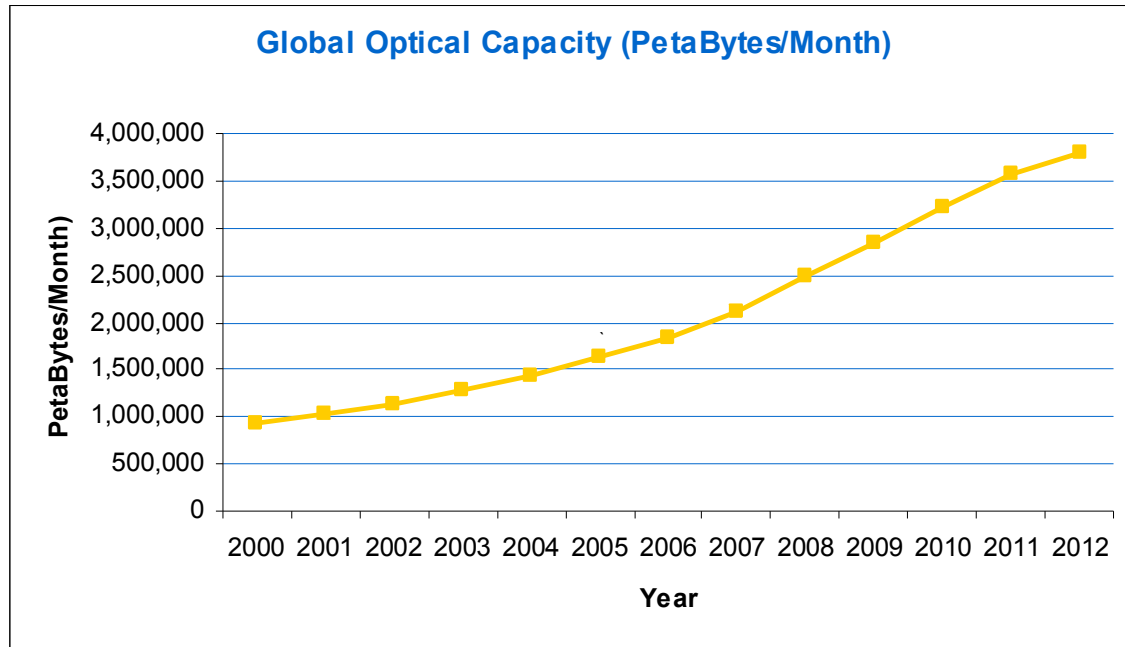


FIGURE 8: GLOBAL OPTICAL CAPACITY

5.1.2 North America

As the discussion above indicates, we can approximate the capacity for the optical backbone of the worldwide network by looking at worldwide investment. This yields an answer that seems reasonable on its surface, but which gives no insight into the situation in North America. To arrive at an approximation for North American capacity, it is necessary to utilize an abstraction. This is just the simple assumption that North American investment is proportional to the percentage of Internet usage generated by North America, and this in turn, is proportional to the relative number of North American users. The notion here is that investment tends to match usage, which users drive.

Fortunately, we know the relative number of North American users. In 2006, the United States Census estimated that out of approximately one 1.2 billion Internet users, 234 million of them were American. This yields a factor that allows us to index the total optical investment and compute the relative investment in North America (Please see Figure 9: North American Optical Capacity).

This approach may overstate the investment in optical capacity in North America, especially in future years, since it does not distinguish between optical distribution such as fiber to the home from optical backbones such as submarine cable. Still, it probably does a pretty good job of estimating the optical capacity available to North American users since an increasing amount of Internet traffic generated by North America is directed to sites overseas, which make use of the optical capacity implicit in the international backbones. Consequently, the optical capacity available to North America is actually higher than a pure investment figure would indicate.

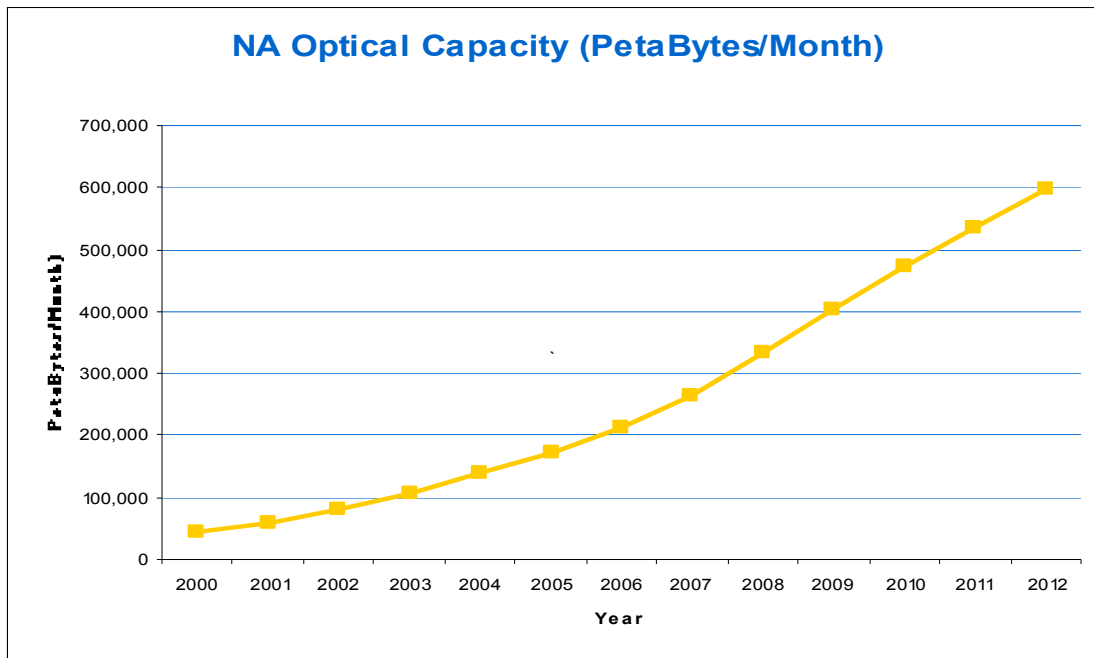


FIGURE 9: NORTH AMERICAN OPTICAL CAPACITY

5.2 Switching and Routing

5.2.1 Protocols and Layers

The Internet is based on the first four layers of the ISO seven-layer model: physical, data link, network and transport. The previous section on optical infrastructure covered the physical layer and some of the data-link layer. This section focuses on the data link, network and transport layers. Each layer communicates with layer above and below and each layer has its own set of protocols that define the communications interface and behavior of data transport. Most of the action in the Internet is focused on the TCP/IP protocol. TCP, Transmission Control Protocol operates at the transport layer (layer 4) and IP operates at the network layer (layer 3). Despite people talking about them as a single unit, TCP and IP are two distinct communications protocols. Another significant protocol is UDP (User Datagram Protocol) that also operates at level 4. UDP is a significant protocol for VOIP communications, video, chat and some peer-2-peer applications.

Typically, when people refer to switching and routing, they are referring to switching activities taking place at layer 2 - i.e. Ethernet – and, routing activities taking place at layer 3 – IP routing. Unfortunately, there is a lot of overlap in switching/routing equipment since most equipment does some level of both switching and routing.

Modeling Supply

The goal is to calculate an estimate of the total switching and routing capacity of the Internet infrastructure. Our research indicates that the last time anyone had firm data on Internet capacity was back in the early 1990s when the bulk of the Internet was still shared by research institutions and the core capacity was provided by the National Science Foundation (jointly with IBM, MCI and managed by Merit Networks). To start with the measurements done in the 1990s offers no value to today's calculations; all of the equipment has been replaced and the NSF component of the network has been dwarfed by newer private capacity connections.

Another way to measure capacity is to interview all of the Internet service providers and add up all of their inter-switch trunks and lines to get an aggregate Internet capacity. This approach, too, is not possible. For this research, we interviewed a group of Internet service providers and asked them to share with us (for this research) their network architectures. None of the ISPs would disclose complete details on their infrastructures, however. We were able to glean tidbits of information from annual reports and briefings but it became a process of trying to put together a jigsaw puzzle with only a few pieces and a second-hand sketch of what the end-puzzle might look like.

We took a novel approach for this research: We postulated that all Internet switching and routing capacity must involve switching/routing equipment. We looked at the switching and routing equipment market and 85-90% of the market share is based on the shipments of Cisco, Juniper, Alcatel and Ericsson (Redback). Therefore, if we could track the shipments of equipment then we could start to build a network capacity model.

5.2.2 Methodology

For our analysis, we divided the switching/routing into two layers: core and connectivity. We picked these not because engineers necessarily think of the Internet in such terms, but because, it is possible to parse the investment in switching equipment into either those functions largely provided by carriers' backbone networks from those that are largely provided by service providers. In telecom terms, we separated long haul from local traffic.

5.2.3 Core

Core Layer – The core layer refers to the core routers/switches that comprise the service provider backbone networks. These are typically very high capacity units with high-speed interfaces, fault-tolerant operating systems, redundant switching and control planes and redundant power. These routers/switches typically provide a peer-to-peer functionality within the Internet that allows traffic to be handed off between networks and for traffic to be collected and routed efficiently worldwide.

Examples of core switch/routers include the Juniper Networks T-Series and the Cisco Systems 12000 Series. Typical capacity (2007) is in the multi-Terabit (1x10E12) range for throughput with interface support up to OC-768 and

Modeling Supply

10 Gigabit/s Ethernet. The performance of such switching equipment has benefited greatly from Moore's Law. The price performance ensures that investments that have generally increased in a linear fashion can show exponentially increasing growth rates of switching capacity.

5.2.4 Connectivity

Connectivity Layer – Many also consider the connectivity layer to be the *Edge* layer. The reason we differentiate is because “edge” implies a sharp demarcation point; the edge of the network. In reality, we see this layer as more of a fluid boundary between the core and the access layers. This layer, as distinguished from the long haul or peering traffic, typically routes traffic within metroplexes or between metroplexes within the same carrier network. Typical equipment in the connectivity layer includes:

Routing - Juniper Networks M-Series Multi-service edge switch

Switching - Juniper Networks MX-Series Metro Ethernet Switch

The connectivity layer is where CLECs, ILECs and MSOs drop-off their Internet traffic. For our analysis, we include equipment that interfaces with DSL, Metro Ethernet, Cable and FTTH. While this tends to blur the boundary between access and connectivity somewhat, it lends itself to relatively easy analysis and the capacity involved is easy to discriminate from access line capacity.

Ethernet Aggregation provides interface at the connectivity layer to cable operators, DSL providers, wireless providers and direct connection to enterprise customers. Ethernet interfaces have been around since the 1980s, Ethernet as a service provider interface is a relatively new phenomena. Acceptance is high since the interface is ubiquitous, high bandwidth and the customer premise equipment is low-cost. Our research indicates a shift from more traditional IP services interfaces to Ethernet interface over the coming years. We reflected this shift in our capacity calculations.

As in core switching, Moore's Law impacts the price/performance of connectivity switching. Over the course of the historical data and as projected into the future, there is a clearly exponential growth observable. This will become apparent as the investment in switching equipment is translated into capacity.

5.2.5 One Box Two Trunks

After extensive review of network capacity modeling, it's clear that the level of model complexity increases at a greater rate than the complexity of the network being modeled. Given the fact that there is no network more complex than the Internet, a true capacity model stretches the limits of current capacity modeling systems. In fact, theoreticians can't even agree on how best to model the capacity of the Internet, let alone actually model the capacity of the Internet,

Modeling Supply

itself. Given this reality, we decided to approach the problem from a very simplistic perspective: one box, two trunks.

We know that every switch and router in the Internet has at least two trunks connecting to other switches and routers. Each trunk operates at a given rate but the rate is not equal to capacity. Network engineers load balance traffic between trunks, and based on discussions with Internet service providers, trunks are typically sized to operate with a peak capacity of no more than 30% loading. As an upper limit, a node with two trunks should have no more than 50% peak loading on either trunk to provide for redundancy.

Our approach is to calculate capacity by assuming that each node has two trunks running at 100% loading. In the real world of the Internet this would be equivalent to the capacity of four trunks running at 50% peak loading, or six trunks running at 33% peak loading, non-stop. This approach, though simple, yields a high-level capacity figure that takes into account the fact that line rate is much higher than actual capacity, yet makes the process of equating line rate to capacity possible.

Estimates for equipment shipped were calculated based on interviews with router manufacturers, researching market research data and analyzing annual reports of both service providers and equipment manufacturers. All information taken together gave us a global estimate of shipments of service provider router/switch equipment, over time.

An estimate of boxes does not directly correlate to Internet capacity for a number of reasons. First, only a percentage of units shipped are ever installed. The rest may be put into spare inventory or disaster recovery inventory; neither ever having an incremental impact on Internet capacity. Second, the switch/router unit is only half of the capacity equation; the other half being the network trunks. Core routers typically have a minimum of two trunks with more trunks added as demand and network routing complexity increase. Finally, there is overlap between enterprise switches and routers and service provider switches and routers. Just looking at total units shipped can be misleading.

To address these issues we did the following:

1. Over time the service provider and enterprise routing equipment has diverged. Today, there are clear product lines destined for service providers and enterprise clients. There is still some overlap where the largest enterprises require low-end carrier class routers and the smallest service providers can use high-end enterprise routing equipment. We believe that this overlap is minimal and not significant to our analysis. For our backward looking analysis, we estimated a percentage of total product shipped was destined for enterprise and we therefore removed it from our capacity calculations.
2. Our goal was to calculate maximum available capacity to match our calculation of maximum potential demand. Therefore, we assumed

Modeling Supply

that all equipment shipped contributes to Internet capacity. This leads to estimation beyond the actual capacity but this over-estimation is mitigated by the maximal demand approach.

3. In reality, the capacity of a given switch/router is directly related to the capacity of the trunks that connect the unit to either peer or higher-level networking components. In fact, the rate of switching capacity - packets per second and backplane throughput- has increased at a greater rate than the increase of speed of the trunks. Therefore, for our calculations, we have assumed that the switch/routers processors are not rate limiting and that capacity is directly related to the capacity of the trunks interconnecting the nodes.

The following table (Please see Table 2: Global Shipments of Core and Connectivity Nodes) shows the estimations of global shipment of core and connectivity nodes (2000 – 2006) and the projected trunk rates for each node. The trunk rates are based on typical connectivity and core trunk rates for each year of the analysis. Given the scope of the assessment, we know that units connected at higher speeds and lower speeds than the rates shown. However, on a global basis, we believe that these rates are a fair estimate of the higher-end trunk rates available at the time.

	2000	2001	2002	2003	2004	2005	2006
Core Units	11400	8300	6660	7744	9594	10666	12071
Connectivity Units	62439	36880	10128	11269	12998	23365	30411
Core Trunk	2.4E+09	4.8+09	9.9E+09	9.9E+09	9.9E+09	1.99E+10	1.99E+10
Connectivity Trunk	1.2E+09	2.4E+09	2.4E+09	4.8E+09	4.8E+09	4.8E+09	9.9E+09

TABLE 2: GLOBAL SHIPMENTS OF CORE AND CONNECTIVITY NODES

We based the calculations for Internet capacity on summing the yearly incremental capacity. We recognize that adding 100 core routers to the network doesn't raise the total number of core routers by 100 – some equipment replaces older equipment. We assume, however, that capacity never decreases. Even when a new router replaces an old router, the aggregate capacity of the new router is greater than the capacity of the existing router since the main reason for upgrade has been, and will continue to be, to increase capacity. Our calculation of incremental capacity for an upgraded node will initially be higher than actual. However, over time as additional trunks and higher speed trunks are turned up, our initial over-calculation will be compensated for.

Modeling Supply

Our research only goes back to 2000, so we needed to determine a starting point for Internet capacity (Please see Figure 10: North American Switching Capacity). In reality, the amount of bandwidth added since 2000 is much greater than any starting bandwidth position, so that the starting point has surprisingly little impact on the final capacity curve, as will be discussed shortly. But, to assess a starting point we made the assumption, based on input from experts (including Andrew Odlyzko) that the Internet capacity in 2000 was probably roughly equal to Internet demand. It's important to note that "roughly equal" does not mean exactly equal since supporting X kbit/s of demand requires $n \times X$ kbit/s of switching and routing capacity. This is related to queuing factors and the need to have much more switching capacity than demand to support a given level of demand.

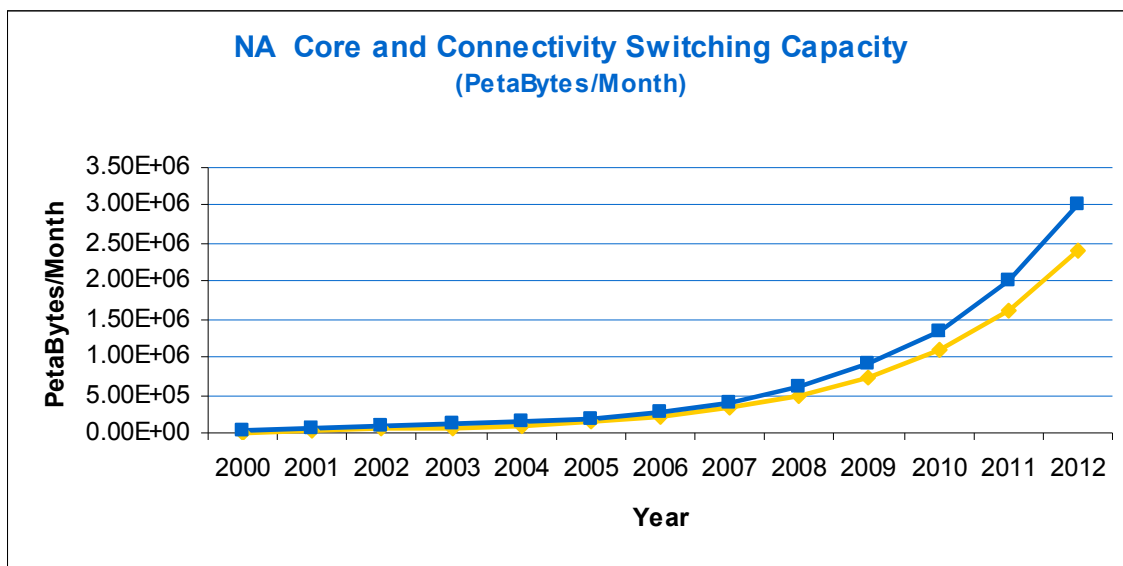


FIGURE 10: NORTH AMERICAN SWITCHING CAPACITY

On a global basis, we estimate that core and connectivity switching capacity in 2000 was 66,000 Petabytes/month. For North America, the estimate is 40,000 Petabytes/month. Over time, core and connectivity capacity have closely tracked each other, and starting in 2007 we project that connectivity capacity will increase at a faster rate than core capacity. From a network design viewpoint, these curves make sense. As higher speed interfaces are extended toward the edge of the network, the connectivity layer will require more total capacity than the core layer. Also, a significant percent of connectivity traffic stays in the region and never touches the core.

We based forward-looking capacity projections on observation of historical increase in capacity as well as primary research on Internet capacity growth (Please see Figure 11: Growth in Global Switching Capacity). Historically, we see capacity growing at a high rate of 61% year-over-year between 2001 and 2006. During this time the core capacity grew faster than connectivity capacity:

Modeling Supply

75% and 53% respectively. However, toward the end of the period we see this gap closing and for forward looking projections, we projected a total capacity growth of 50% per year with the continuation of the trend of connectivity growing slightly faster than core capacity.

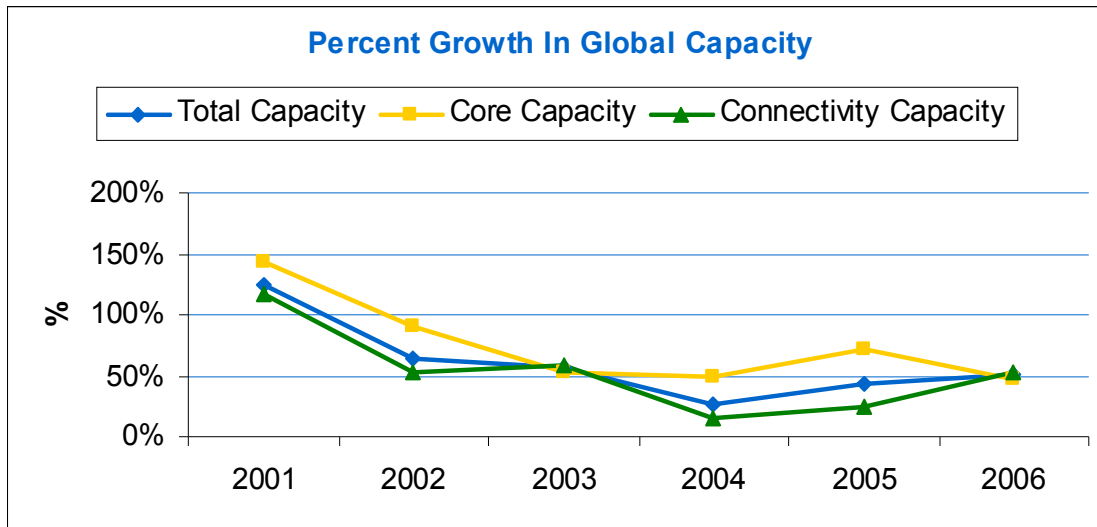


FIGURE 11: GROWTH IN GLOBAL SWITCHING CAPACITY

As a point of reference, our projections fit well with other analyses. In the 1990s there was talk of explosive growth of Internet capacity leading up to the 2000 technology bubble burst. Starting in 2000 there was a significant contraction of investment in Internet infrastructure. Investment didn't stop but as seen in the slope of the curves, the capacity growth rate dropped from 124% year-to-year in 2001 down to a low of 26% year-to-year in 2004. In 2004 we see a reversal of this trend with the year-to-year growth increasing to 43% in 2006 and 50% projected in 2007. As Andrew Odlyzko noted, within any given timeframe the Internet may be growing at an explosive rate, but over time the traffic growth rates are approximately 50% per year. Similarly, we believe that at any time capacity may be growing at an explosive rate, but over time the capacity growth rates are approximately 50% per year. As discussed in the Section 4 (Demand), we are projecting a higher year-over-year growth rate for traffic.

Another way of rationalizing our capacity projections is to compare investment to capacity. As discussed in Section 6 (Investment) we are projecting an annual increase in investment of 10% per year. At the same time, Moore's Law states that the amount of switching capacity per \$ should double every 18 months. Considering the 10% increase, a 50% annual increase in switching capacity is a reasonable assumption based on Moore's Law.

5.2.6 Access

While the Internet is generally thought of as a cloud of connections into which a user plugs, architecturally, the Internet cloud begins where the user

Modeling Supply

accesses that cloud. This access layer is the connection between the user and the ISP (Internet Services Provider) and is usually provided by the local telecommunications carrier.

Access comes in several flavors. Up until the end of the last century, the most common way to access the Internet was through a dial up modem over a telephone line. These dial up connections were limited to 56 kilobits per second and were minimally able to support activities such as Web surfing. Beginning just prior to 2000, though, a significant percentage of users began to acquire broadband access to the Internet, either through DSL (Digital Subscriber Line) technology or cable modems. In the case of the former, the service still traveled over the telephone connection, while the latter was carried over the local cable television cable. Increasingly, newer access technologies based on wireless and optical fiber are being deployed, although their impact on the market is nominal at this time.

It is clear that access plays a significant role in the degree to which users translate their demand to capacity consumption. If the only available access is 56 kilobits per second, then the degree to which packets can be placed on the Internet backbone is far lower than if the connection available is running at a one megabit rate. Because access constraints may throttle demand it is important to determine the actual carrying capacity of this layer.

The analysis of access capacity begins with an inventory of the access lines available. In the case of North America, this is relatively straight forward. The carriers report their broadband access line count routinely to the Federal Communications Commission (FCC). Over time, North American broadband access numbers have followed a relatively linear deployment profile (Please see Figure 12: North American Broadband Access Lines). In 2006, the number reported by carriers was about 99 million.

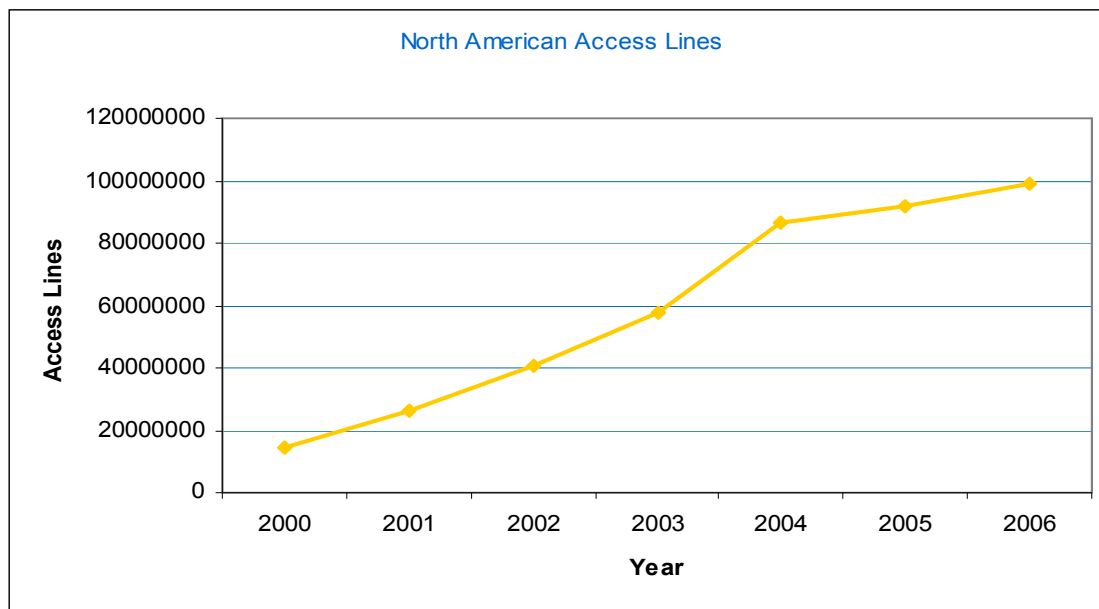


FIGURE 12: NORTH AMERICAN BROADBAND ACCESS LINES

Modeling Supply

Assuming that such deployment continues to follow the same curve, the curve can be projected as shown below (Please see Figure 13: Projected North American Access Lines).

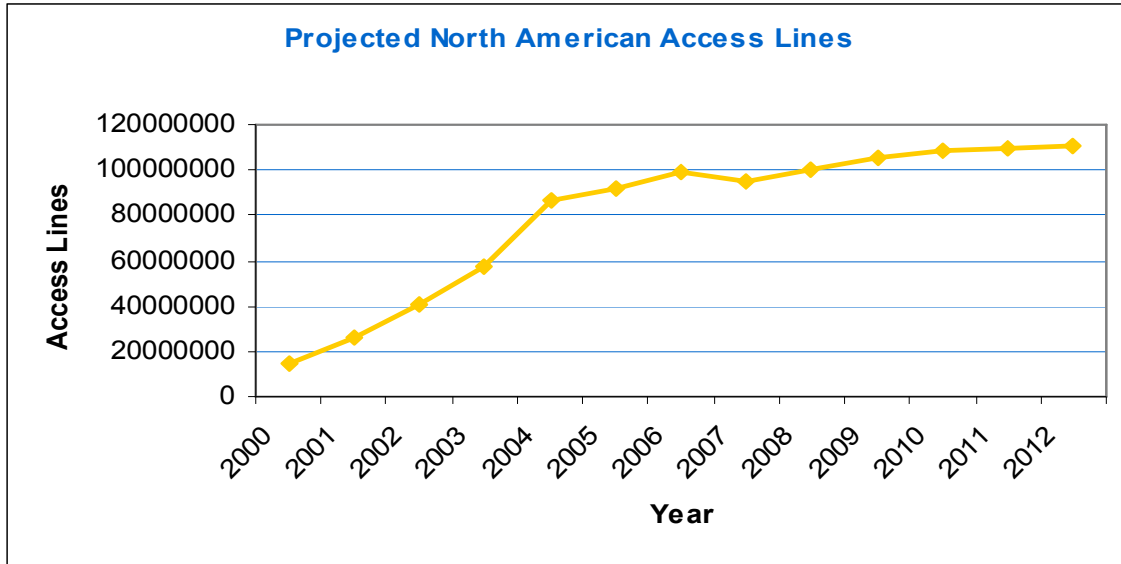


FIGURE 13: PROJECTED NORTH AMERICAN ACCESS LINES

This curve, though, only represents a fraction of the total access capability. In order to be comprehensive, wireless devices must be included as well. Additionally, enterprise access, although small in comparison to the total consumer access needs inclusion as well. When this is done the total broadband access lines very much exceed the total number of potential users, however, this just means that typically users have available to them several different access capabilities simultaneously. When these are included, the curve for North America appears as follows (Please see Figure 14: Total North American Access Lines):

Modeling Supply

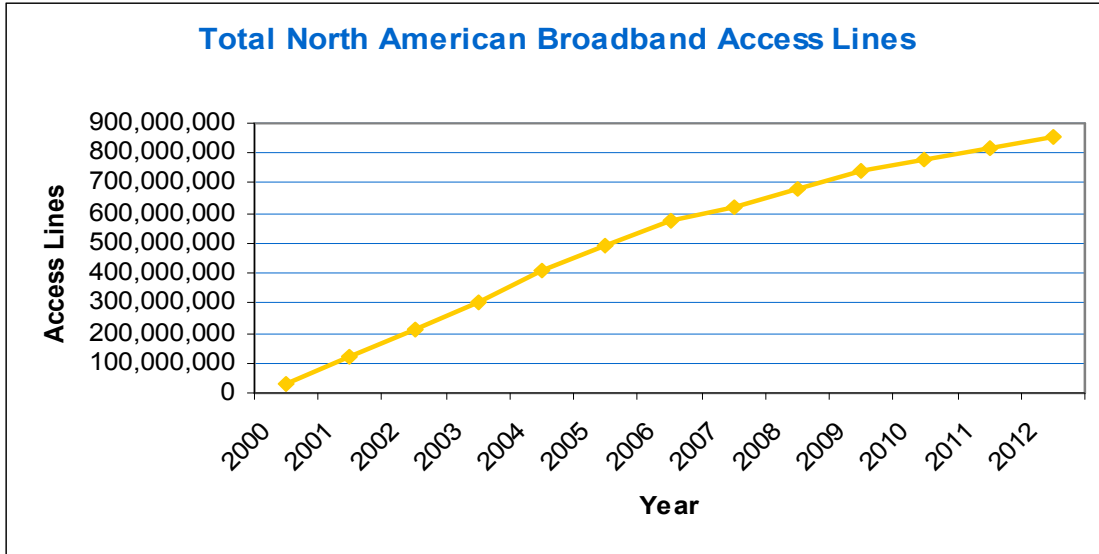


FIGURE 14: TOTAL NORTH AMERICAN ACCESS LINES

This resolves the problem of access lines for North America, but what of the global distribution of access lines? It turns out that North American users as a percentage of total global users has been inventoried by the U.S. Census. When this is combined with data points provided by the EU and countries in Asia and Latin America on the total broadband access lines installed there, it is possible to take the North American access lines and index them to derive the balance of the world's access lines. When this is done, the following curve is derived for the entire Internet (Please see Figure 15, Global Access Lines).

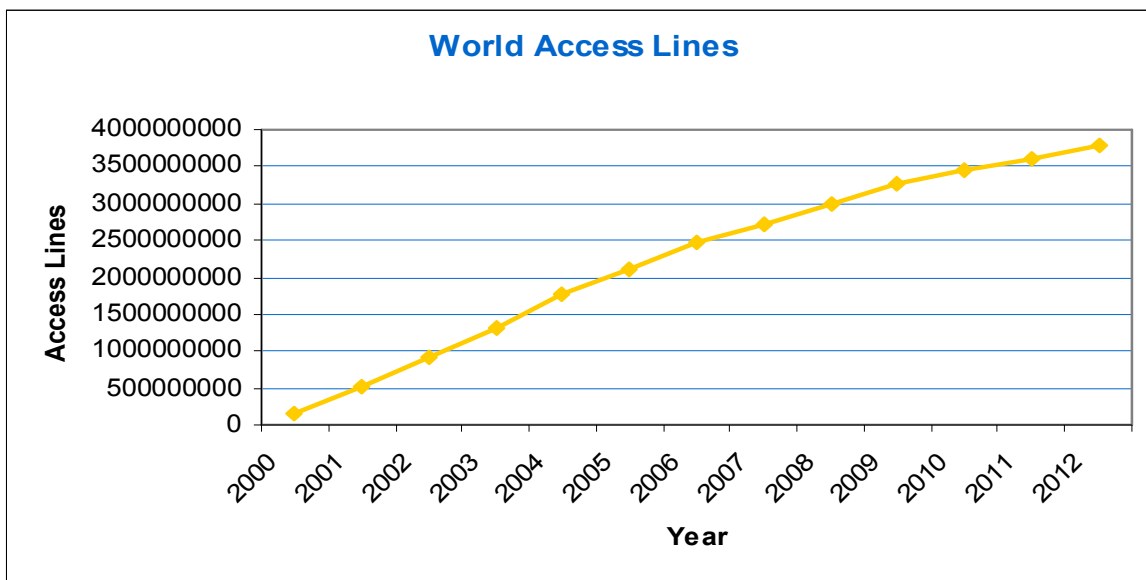


FIGURE 15: GLOBAL ACCESS LINES

Modeling Supply

This information combined with the distribution of access line technology plus the likely data rates provided by that technology, generates the following two capacity curves for North America and the World respectively (Please see Figure 16: North American Access Capacity and Figure 17: World Access Capacity):

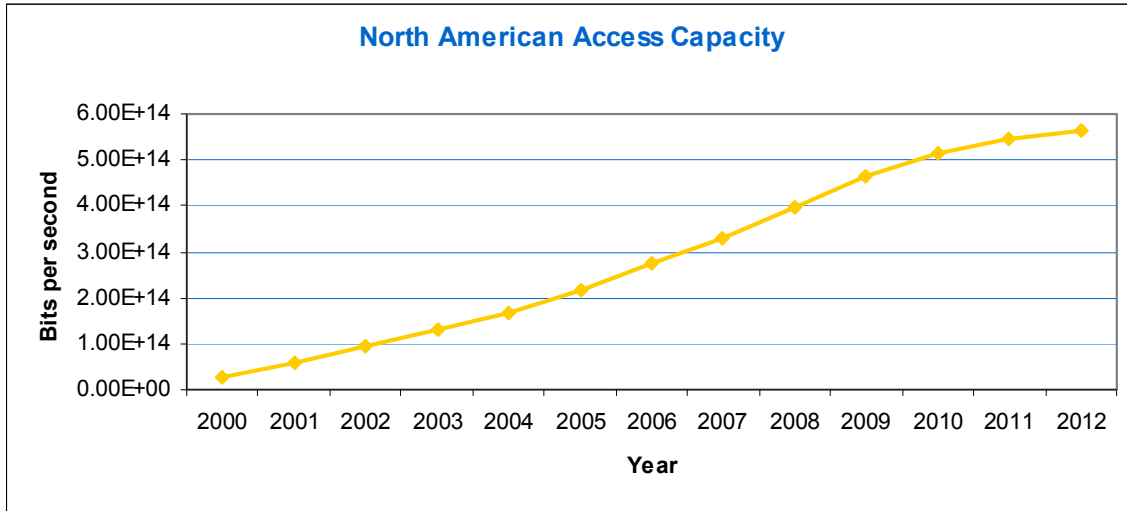


FIGURE 16: NORTH AMERICAN ACCESS CAPACITY

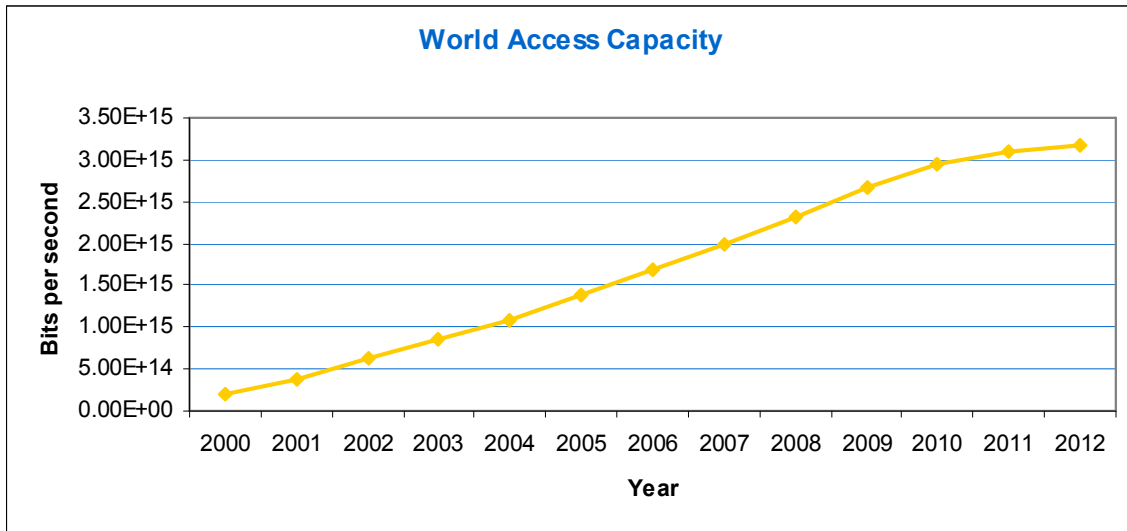


FIGURE 17: WORLD ACCESS CAPACITY

As can be seen, the broadband access capacity grows at an essentially linear rate over time. This curve, however, depends on two assumptions. The first is that world fiber to the home will be increasing over time. The second is

Modeling Supply

that wireless devices will increasingly become a surrogate for fixed access technologies. While the number of mobile devices worldwide is known and the projected uptake can be surmised, the degree to which these devices will be used for data is uncertain. If the technology implicit in wireless devices continues to improve and data rates available to basic devices increases, it is possible that access line capacity could increase by several times on the tail end of the curve, thus inflecting the curve up or, at the very least, straightening it.

5.2.7 Wireless: Building Footpaths Across the Digital Divide

Another consequence of this modeling effort was the realization that access is now evolving past the limitations of fixed infrastructure. While most users of today's Internet access it through landlines, either DSL (Digital Subscriber Line) or broadband cable connections, an increasing number of users, mobility workers in particular, are accessing the Internet through either fixed wireless or mobile wireless technologies.

This has been an evolving process. Once wireless service evolved past simple analog technologies (1G or first generation) to digital (2G or second generation), mobile workers wanted to use their instruments to access data services, especially the Internet. 2G technologies, though, were not sufficiently robust to deliver acceptable data rates. This led to the development of so-called 2.5 technologies which allowed for reliable data transfer, albeit at rather slow data rates; typically no more than 144 kilobits per second.

With the introduction of 3G technology and data rates much higher than 144 kilobits per second, mobile access to the Internet begins to be an attractive possibility. Using protocols such as EV-DO which can achieve burst rates as high as 1.8 megabits per second under the REV A standard or many times that under the REV B standard, land line comparable access is possible. 4G technologies, which are still being developed, are expected to push these rates even higher and will likely provide access rates that exceed many fixed access technologies.

The net impact of this evolution is that the load on the Internet will increasingly be driven by wireless access. While that influence does not significantly influence the results of the model, to the extent that such an influence is more than nominal over the study period, it could have profound implications for network demand. If, for example, new services based on WiMAX, a new 3.5 G wireless data technology, become prevalent, demand could be significantly impacted.

6 Investment

6.1 Methodology for Determining Investment

We derived Internet infrastructure investment from three primary sources:

1. Extensive research of investment information from market researchers, investment firms and publications.
2. Review of manufacturers annual reports to track shipments and revenues.
3. Review of service provider annual reports to discern capital expenditure and investments in various Internet infrastructure technologies.

All investments referred to in this analysis are capital expenditure (CapEx). Operational expenditure (OpEx) is a significant component of the total investment picture. However, given the differences in labor rates on a global basis, we believe that the most reliable and comparable estimates of infrastructure investment are CapEx.

As shown below (Please see Figure 18: Global Internet Infrastructure Investment and Figure 19: Global Investment in Infrastructure) the overall investment curve for Internet infrastructure reflects the overall market for Internet services. In the 1990s there was significant investment, leading up to the technology bubble bursting in the 1999-2000 timeframe. As shown, investment dropped significantly in 2001/2002 and by 2004 the slopes of the investment curves reversed showing continually increased investment over the past three years.

Forward-looking projections of investment are based on the continuation of the increasing investment trend seen over the past three years, as well as input from market analysis reports and primary research. Between 2004 and 2006 we estimate that spending for switching capacity (core and connectivity) increased at an annual rate of 10% per year. Spending for connectivity has been increasing at a higher rate than for core: 20% versus 4%, but we project this to even out, so that connectivity spending will increase at 8-12% and core spending will increase at 8-10%. Overall, spending for core and connectivity is projected to increase 10% year over year.

We project total spending for global Internet infrastructure may grow from \$20.5B in 2007 to \$29.9B in 2012. In other words, capital expenditure for Internet infrastructure may be \$160B over the next five years.

Investment

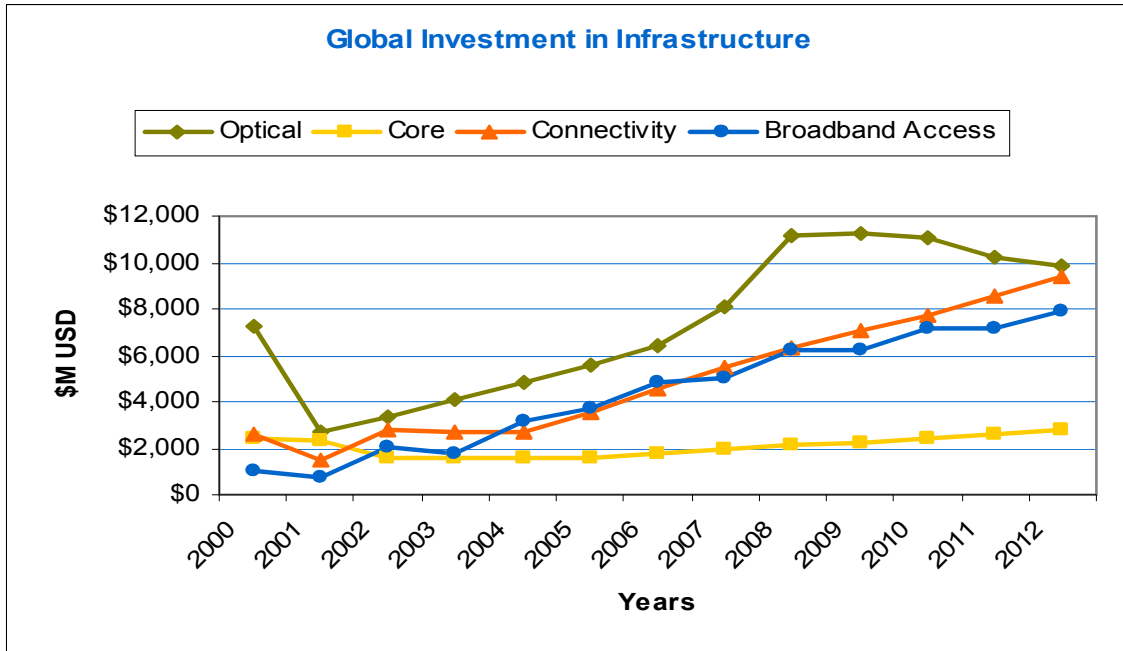


FIGURE 18: GLOBAL INTERNET INFRASTRUCTURE INVESTMENT

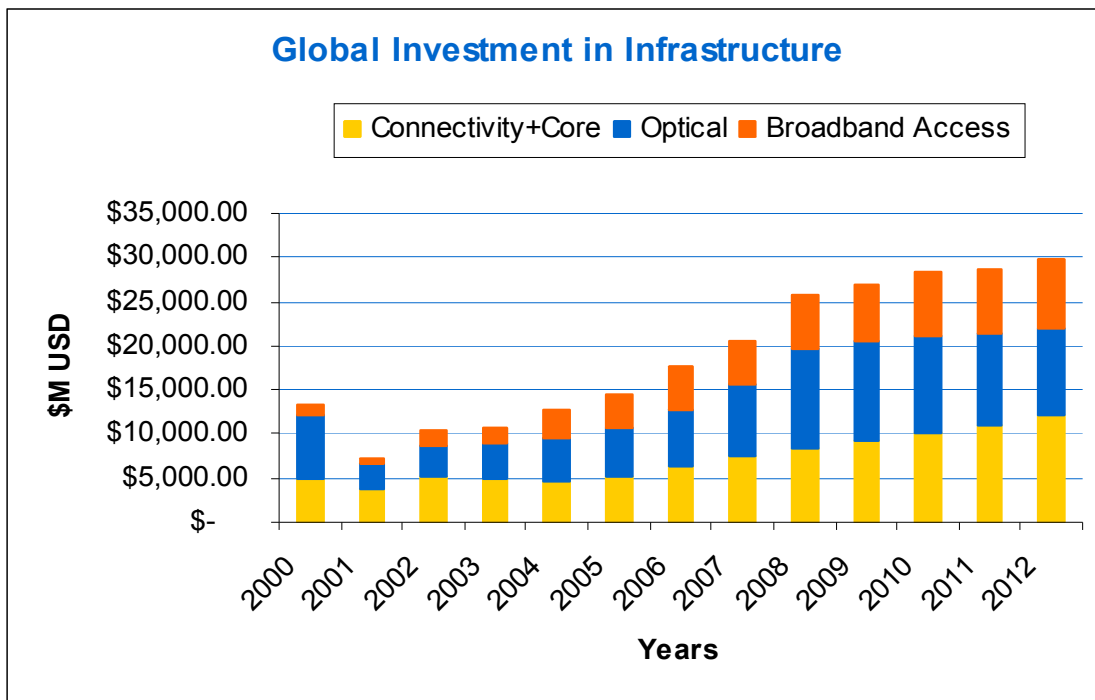


FIGURE 19: GLOBAL INVESTMENT IN INFRASTRUCTURE

Investment

For North America, the investment picture is quite different than the global picture (Please see Figure 20: North American Investment in Infrastructure and Figure 21: North American Investment in Infrastructure). North America has the same growth curve for connectivity as the global investment curve. Optical investment follows the same shape but the relative position of optical to edge are reversed. Core switching investment for North America follows the same pattern as global. The biggest difference is in access line investment. Again, these numbers are CapEx so they reflect the cost of components (modems, DSLAMs, fiber termination equipment, etc.) and not the cost of OpEx (digging trenches, pulling cable and installing devices). However, the curves tell a compelling story: North America is behind the rest of the world in terms of access line investment. For example, according to the ITU, Canada is ranked 7th and the U.S. as 16th, in the world, in terms of broadband penetration per 100 inhabitants. (ITU Report) Granted, the countries that are ranked higher are smaller than both Canada and the U.S. However, North America is losing ground and one would think that the population and size of North America would drive a much higher access line investment profile when compared with the global picture.

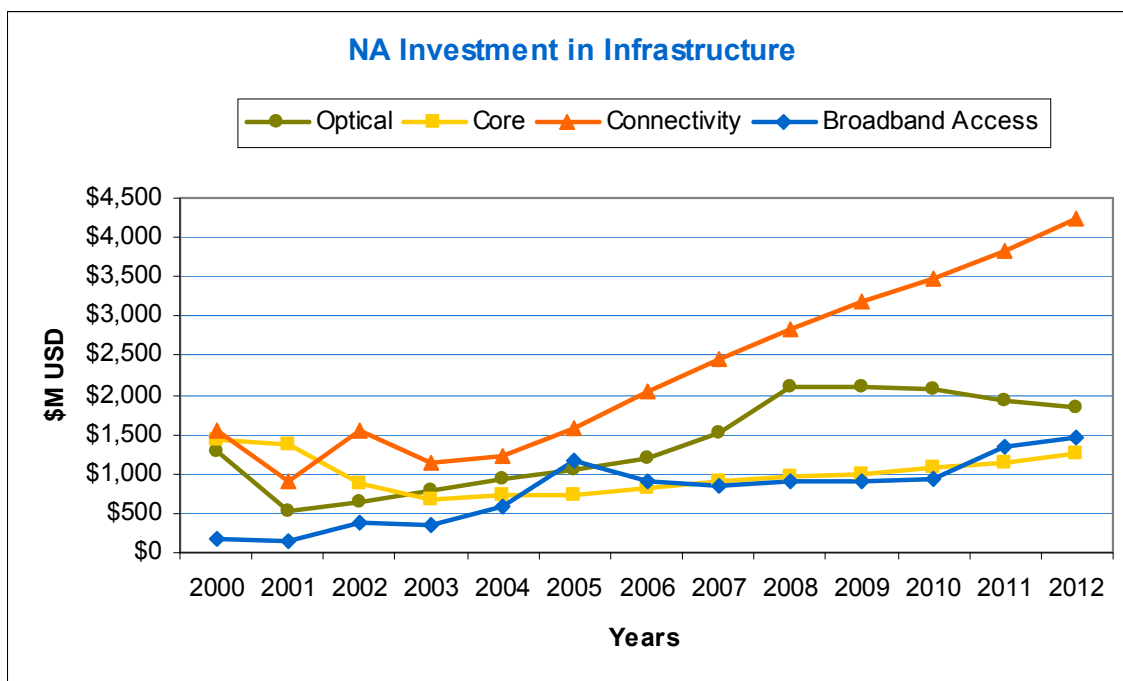


FIGURE 20: NORTH AMERICAN INVESTMENT IN INFRASTRUCTURE

In terms of annual investment, we project the same growth rate for North America as for the global investment analysis: approximately 10% increased spending per year. This equates to \$5.7B estimated in 2007, growing to \$8.8B estimated in 2012. The end-result is we estimate that \$38.6B may be spent in North America for Internet Infrastructure capital expenditure over the next five years.

Investment

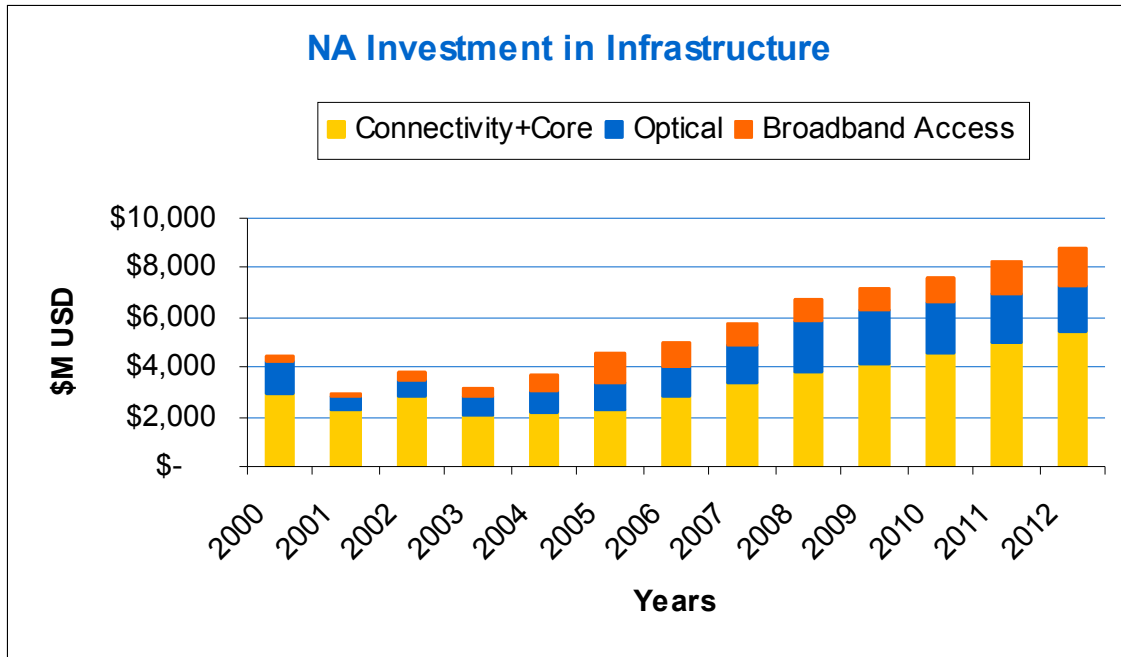


FIGURE 21: NORTH AMERICAN INVESTMENT IN INFRASTRUCTURE

7 Key Findings: The Coming Bandwidth Crunch

As the preceding sections have shown, it is possible to model both the capacity of, and the demand on, the Internet. While extraordinarily interesting as an exercise in analytics, the inevitable question is: so what? In this section, we try to answer that question. However, before we do so, we need to address some of the assertions that others have made about the state of the Internet and, at least say a few words about the use and misuse of capacity demand charts.

Almost since its inception, the Internet has generated hyperbole. In the 1990s, pundits mostly focused on the growth of the Internet in terms of Web pages deployed, as well as the number of logical domains. These metrics, based as they were in the logical rather than physical domain, make the Internet give all appearances of being almost infinite and growing infinitely. Of course, such a dynamic is irrational and ultimately unsustainable.

Beginning in 2000, academicians such as Coffman and Odlyzko began to explore the real meaning of Internet growth. As they began coming up with metrics based on the Internet's capacity to move packets, others began to notice and soon most IT vendors were looking at the Internet, if not in terms of capacity, then at least in terms of the real demand being generated. Cisco's is only the most recent of these studies.

Currently, estimates of Internet demand seem to be centering on the value of 2000 Petabytes per month, and Cisco as well as others have speculated that we may be approaching an era of Exabyte demand. These figures are huge, but are they rational? More importantly, are they supportable? In order to answer these questions, we must compare demand to capacity.

7.1 Global and NA Supply and Demand Curves

7.1.1 Demand vs. Supply Overall

With the preceding as preamble, let us take a look at the supply and demand charts presented in the following figures. As can be seen, the demand lines when plotted against the capacity lines tend to remain well below capacity until the latter part of 2012. At that point, the demand line crosses the access capacity line in North America, and shortly thereafter globally. What does this mean?

The implication is that in 2012, if innovation continues to drive demand, it is likely that the absolute demand will approach the limits of capacity as expressed by absolute access capacity. This does not mean, though, that the Internet will stop working. As will be seen, the Internet is designed to preclude such an occurrence.

In fact, what the charts indicate is that, within the study horizon, the Internet has plenty of spare capacity in every layer except the access layer. This

Key Findings: The Coming Bandwidth Crunch

capacity crunch is rapidly coming to a head and, when demand curves are developed which take into account innovation driven behavior, the Nemertes model indicates that a crisis will evolve where the degradation in Internet performance constrains users' ability to consume more.

This has significant implications for content providers who wish to use the Internet as a delivery vehicle. It also has implications for the evolution of Internet-enabled business, for while it is possible to imagine a business that does not need the Internet to conduct internal business, there are few who can conduct business generally without providing a Web presence for their customers. As the Internet stresses, all of these constituencies will feel the pinch.

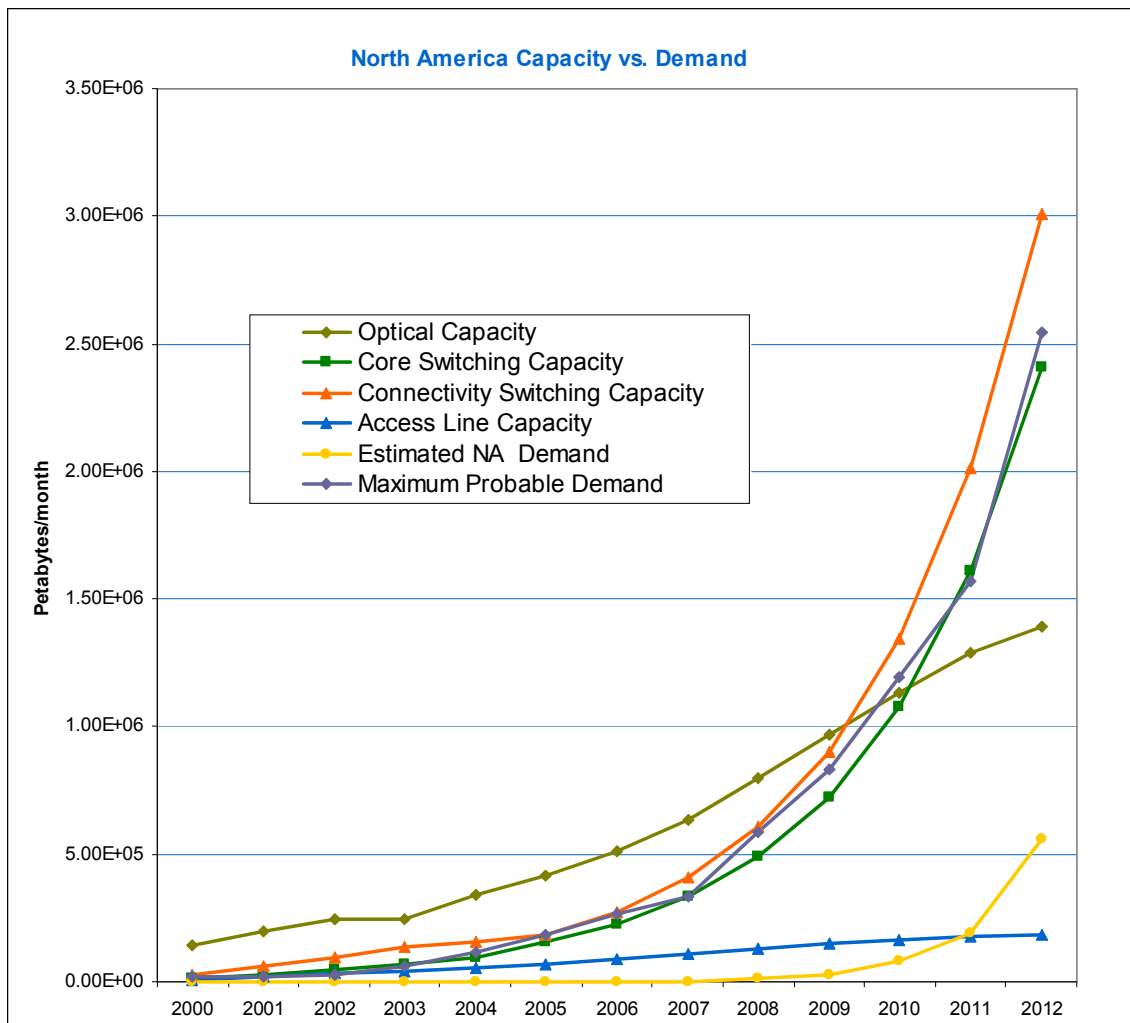


FIGURE 22: NORTH AMERICAN CAPACITY VERSUS DEMAND

Key Findings: The Coming Bandwidth Crunch

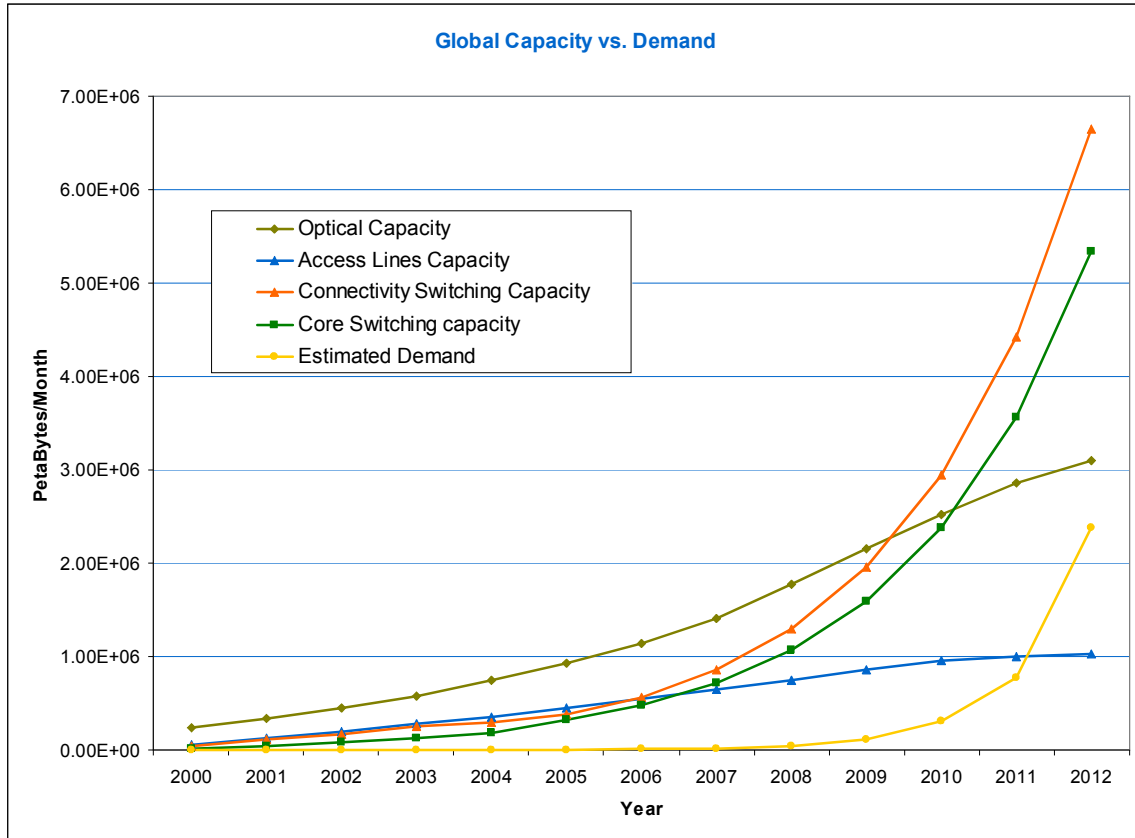


FIGURE 23: GLOBAL CAPACITY VERSUS DEMAND

Of course, it is appropriate to be skeptical of such displays. As we pointed out, demand is highly dependent on the degree to which consumers respond to innovative new applications that allow them to consume bandwidth. If there is no new innovation, then there would not necessarily be new demand drivers. However, as explained previously, Nemertes feels certain that such will not be the case.

7.1.2 When Is Supply Not Really Supply?

It's also important to scrutinize the supply curves, particularly when it comes to predicting what will begin to happen as the demand begins to approach capacity. As noted earlier, we modeled access circuits at their maximum capacity, meaning that we assumed that all capacity was available, bidirectionally, all the time.

However, as anyone with a network engineering background can attest, that's not actually the case. Many of the technologies used to deploy access capacity (such as cable connections) contention-based, and in contention-based environments, the true capacity available is much lower than the overall capacity. Moreover, measured data rates for access types that are theoretically "unlimited" (such as DSL) often exhibit similar characteristics to contention-based technologies.

Key Findings: The Coming Bandwidth Crunch

In contested environments, engineers strive to keep utilization under around 15%, and in any event well below 30%. This is not the case in environments in which we can apply so-called “traffic engineering,” such as carrier cores; in these environments, utilization tends to run much higher. But in contested environments, the real utilization is closer to the bounds discussed,

For these reasons, we assessed what happens at 15% and 30% of available maximum capacity. When we plot our demand line against these two boundaries, the following chart is the result:

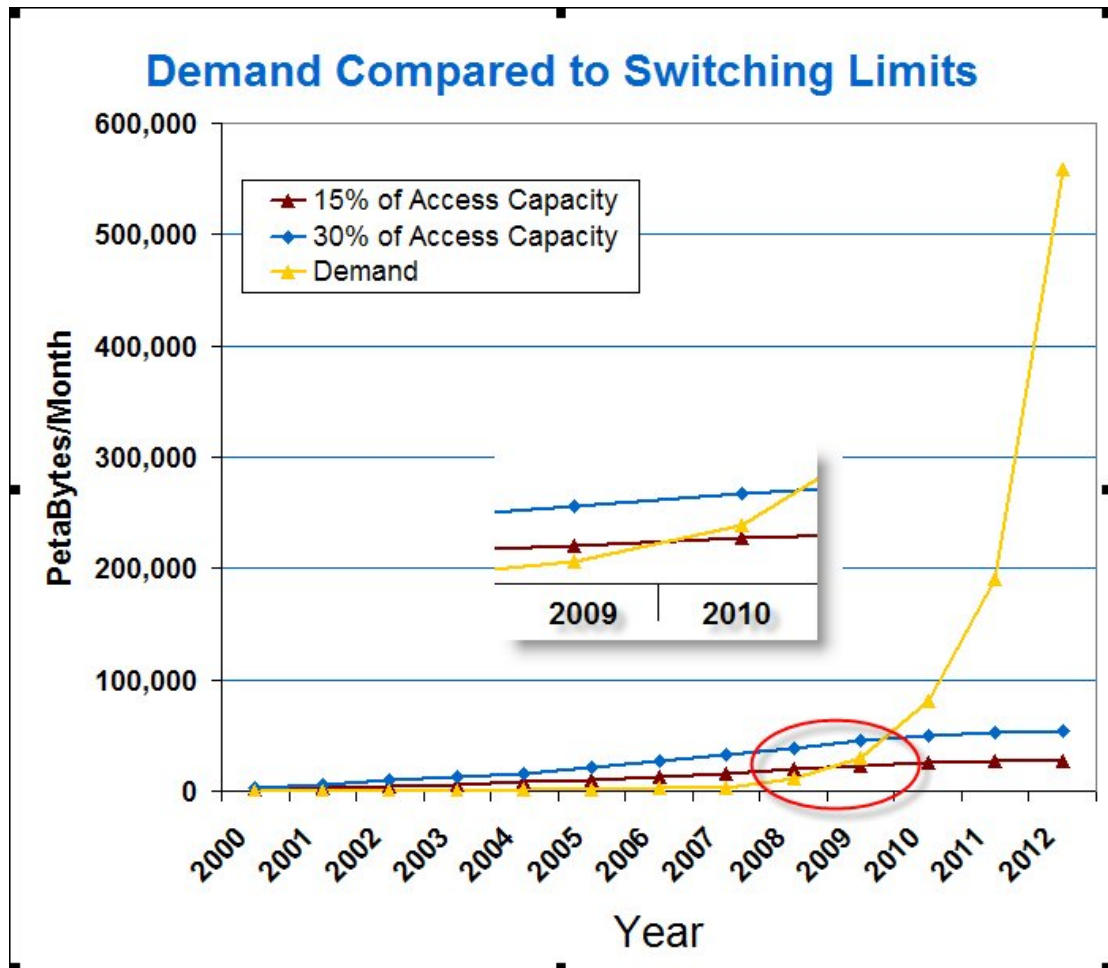


FIGURE 24: NORTH AMERICAN DEMAND COMPARED TO ACCESS LIMITS

As can be seen, when demand is plotted against standard engineering guidelines for shared environments, it is clear that demand crosses supply at 15% in early 2009 and crosses 30% in early 2010. This being the case, the Internet should be showing stress even now (since in any network environment, long before demand reaches a utilization threshold on a sustained basis, there are isolated episodes during which it “spikes” above that threshold).

In fact, we believe this is the case, as we shall discuss shortly. Every user of the Internet is familiar with service degradation that affects file transfers. We collectively expect it and put up with it in an environment where we do not expect

Key Findings: The Coming Bandwidth Crunch

immediate response time for most Internet activities. However, we would surely complain if the application in question depended on the reliable delivery of high bandwidth content in a near real-time environment.

Assuming that 30% is a reasonable proxy for the maximum sustainable utilization of an access circuit (and for the reasons cited above, we believe it is), Internet access essentially runs out of capacity beginning in early 2010.

So, the results of the Nemertes model can be summed up as predicting a looming bandwidth crunch where in which the limitations of access seek to constrain demand the innovation that drives it. Is there anything that can be done?

7.2 Investment Gap: What It Takes To Prevent The Crunch

As noted earlier, much of our capacity predictions are predicated on estimating capacity from investment. Consequently, the gap between desired capacity and available capacity can be expressed as a required investment. When this is done, in the year 2010, just after demand exceeds effective capacity, the investment that would be required then to close the gap in North America amounts to \$43 billion—nearly 60% of the projected carrier investment of \$72 billion.

The only hope to close the gap permanently is that technology will provide access solutions that are less capital-investment intensive. Improvements in DSL technology, the rapid build-out of optical access and the emergence of wireless alternatives the last mile will help, but ultimately, access investment by the carriers will be required in order to address the big crunch from a supply side.

Of course, another way to address unconstrained demand is to constrain it. As noted previously, access almost certainly limits what users expect to do. If access becomes so limiting, then users may decide not to use the Internet and thus reduce the need for additional investment. In such a situation, the market for innovative new Internet-based services would quickly dry up.

7.3 Sensitivity Analysis

As Mark Twain once said, there are three kinds of lies: Lies, damned lies and statistics. Since modeling is in some sense an exercise in statistical analysis, there is always the potential for the model to end up saying what you want it to say rather than what it should say. One way to avoid this situation is through sensitivity analysis.

Sensitivity analysis attempts to assess the model's output in terms of its inputs. If a model reacts in a way that is counter intuitive when the input variables are changed, that is an indication that the model itself is flawed. Nearly 50% of Nemertes' modeling effort was devoted to model testing, and the testing itself was extremely illuminating in terms of what is actually going on in the Internet itself.

Key Findings: The Coming Bandwidth Crunch

As noted earlier, a primary variable that drives demand is utilization. Utilization refers to the degree to which a user actually is able to consume bandwidth. One way in which the model could be tested would be to see what happens in the presence of increased utilization. To test the model, Nemertes applied differing levels of utilization. The following chart shows how the overall demand line increased as a consequence:

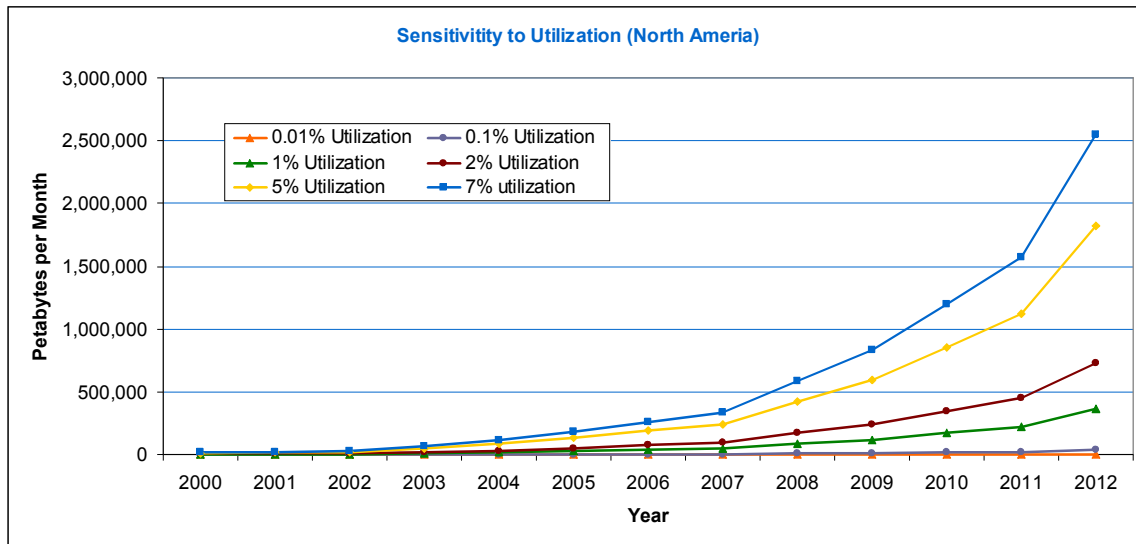


FIGURE 25: UTILIZATION SENSITIVITY

As can be seen, small changes in utilization translate into large increases in demand. The model, consequently, is very sensitive to utilization changes. The question is: Is this rational?

As it turns out, this is precisely what one would expect. Since utilization directly drives the load that a user places on the 'Net, even a slight change in using habits has the potential to drive large changes in the amount of bits being delivered to the 'Net. This is clear when you consider that a small change in bit rate over a sufficiently long enough period of time can amount to large number of bits in total. For example, a 10-bit-per-second change translates into more than two megabytes worth of consumption in a single month. As user habits continue to put value on large media downloads and other high bandwidth applications, we can expect much larger changes in utilization with a commensurate increase in overall network load.

A second point of concern is the degree to which the model is sensitive to initial conditions. In particular, the size of the Internet at the beginning of the study period is highly uncertain. The best idea of what its capacity was in 2000 comes from work done by Coffman and Odlyzko. They estimated that the capacity was just above the demand as expressed at that time; about 85 petabytes per month. However, when we calculated capacity, we came up with a much larger number for capacity even in the access layer, about 61, 000 Petabytes using technology available then.

Key Findings: The Coming Bandwidth Crunch

However, if we were to use the lower figure as a test—e. g. setting capacity at around 85 Petabytes--we find that the change in the point in which demand exceeds supply is roughly the same. This indicates that the model is highly insensitive to the initial capacity. Is this reasonable?

The answer is found in the curve shapes. As can be seen, the capacity lines follow roughly exponential curves as Moore's Law delivers ever greater capacity. Demand, however, is more nearly linear to begin with and then goes exponential as demand increases dramatically starting in 2005. The exception is access capacity, which is roughly linear. As the demand curve goes exponential is it certain to cross it at some point. While raising access capacity enough initially could delay the crossing, it can't prevent it. Lowering the capacity to a level that is inconsistent with levels that must have been available doesn't change the outcome either.

So, in the two areas where sensitivity could be considered to be an issue, the model is returning reasonable results. This would seem to validate the conclusions we derive from it.

7.4 Comparison With Other Studies

That said, that figure places North American Internet traffic at 2469 Petabytes/month in 2006 (the last year for which data is available). This is notably higher than the 800 pbytes/month cited by Andrew Odlyzko, and which represents, as we've stressed repeatedly, the best available data. That's not to say that it's perfectly accurate, a point which he takes pains to highlight on his site, noting that his data sources are "Definitely not a representative sample of all those on the Internet, and great care should be taken in drawing any conclusions based on this data."

Odlyzko's caveat notwithstanding, it's worth comparing data points to determine if there's any degree of possible rationalization. There are two main explanations for the discrepancy between our measured data and that of Odlyzko.

First is the fact that he limits himself to measuring "Internet-only" data. Significantly, this means not including some of the largest, fastest-growing categories of traffic, including commercial IP traffic and potentially IPTV. To understand why this matters, remember that the Internet, by definition, comprises a series of interconnected IP networks, each with greater or lesser degrees of containment. In general, IP traffic that remains on the network of a single entity (carrier or enterprise) is considered "not-Internet." (There are also important distinctions in how the traffic is handled from a routing and addressing perspective).

While cognizant of the distinction, we have considered general IP traffic as "Internet" traffic, in large part because one of the long-term ongoing trends is that the distinction between "Internet" and "non-Internet" is blurring, at least from an infrastructure perspective. A telecommuter, for example, may use the same home-office access circuit to connect both to "big Internet" applications, such as Google and YouTube, and corporate applications such as accounting databases and customer-relationship management systems. (An even newer

Key Findings: The Coming Bandwidth Crunch

trend is for such applications to be provided on sites on “big Internet” sites such as Salesforce.com, blurring the distinctions still further).

Additionally, the carriers that provide Internet services have increasingly consolidated all IP traffic—Internet as well as “non-Internet”—onto common networks. Thus, “big Internet” packets and “non-Internet” packets traverse the same switches, routers, and circuits—and are no longer meaningfully distinguishable from one another.

Other researchers, such as Cisco, have focused primarily on general IP traffic as well, for the same reasons. In some data tables, however, Cisco has segmented out “Internet” from “non-Internet” traffic, with non-Internet traffic equaling roughly 1.47 times that of Internet traffic. Applying that ratio to Odlyzko’s measured traffic yields a figure of roughly 1,976 PBytes/month of general IP traffic in North America—relatively close to 2469. For the period 2000-2004, our figures agree quite closely with Odlyzko’s as the following chart indicates:

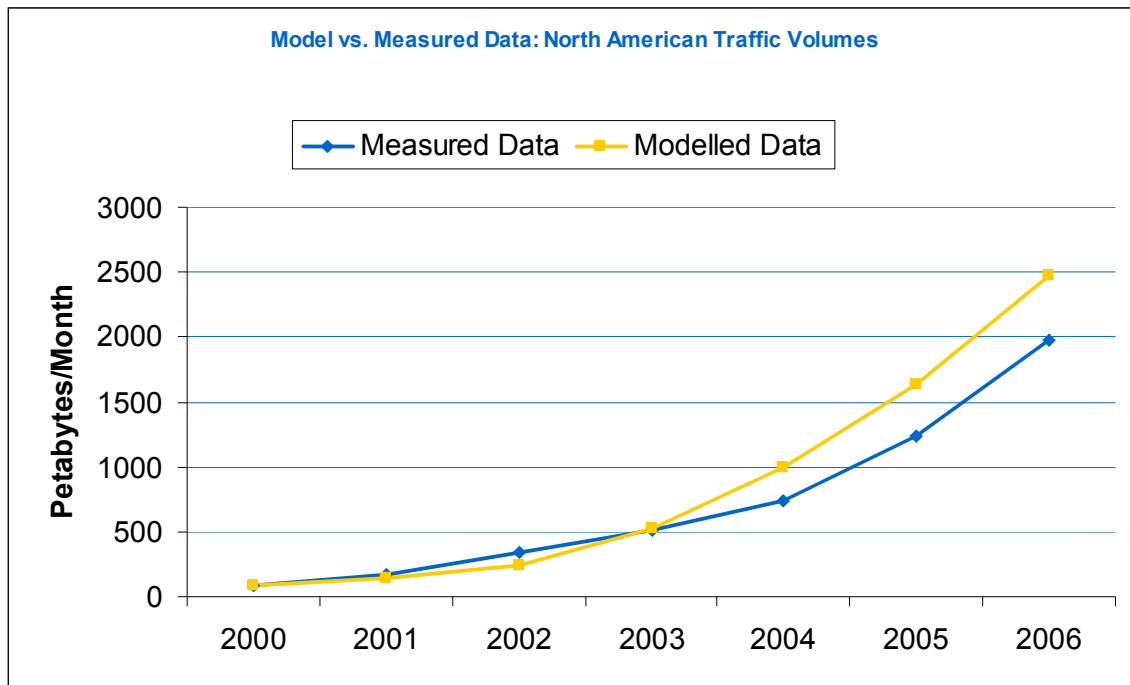


FIGURE 26: MEASURED VS. MODELED DATA

Our numbers run slightly higher than Odlyzkos for the years 2004 through 2006. However, on close inspection it becomes clear that the slopes of the curves are virtually identical for the years 2004-2006---all the difference results from differing projections/measurements for the period 2003-2004.

The difference could reflect true changes in user demand/traffic patterns for those years, or could reflect artifacts in the data, particularly given the fact that the figures for the non-Internet IP portion of Odlyzko’s data is computed, rather than measured.

Again, the key fact is that the *slopes are virtually identical*—which indicates that the underlying trends are the same, which is the more critical

Key Findings: The Coming Bandwidth Crunch

point, given the uncertainties around the measured data and the lack of insight into carrier corers.

The bottom line is that the best available data—including ours, Odlyzko's, and Cisco's—all point to traffic volumes that are, depending on how traffic is measured, in the single-digit exabyte/month range, and growing at roughly the same rate. Our model predicts a much higher rate for the years 2008 and beyond, but that is a result of how we assessed user demand.

And again, the lack of carrier data is forcing us all to extrapolate from a too-limited set of data points and interviews. The best way to arrive at more conclusive measurements than that is for carriers and content providers to cooperate with researchers in providing this data.

8 Does the Internet Ever Break?

As noted previously, our conclusions lead us to believe that the Internet could be facing a crisis in access capacity in the 2010 timeframe. Although this may be an alarming notion, the fundamental question is what happens when that occurs? Will the Internet break?

In a word: NO. It's simply not possible. The Internet's fundamental architecture precludes it. The key to understanding why lies in understanding IP and its related protocols (TCP and UDP) work. Internet Protocol (IP) works by exchanging data in the form of packets, short sequences of bytes consisting of a header and a body. The header describes the packet's destination, which routers on the Internet use to pass the packet along, in generally the right direction, until it arrives at its final destination. The body contains the application data.

IP provides best-effort packet delivery, meaning that it makes no guarantees that packets will arrive at their destinations uncorrupted, in the right order, as duplicates, or in fact, at all. The only thing IP does guarantee (by using a checksum) is that packet's header is error-free, meaning that the receiving router knows from where the packet was sent. Packets whose headers are corrupted are dropped.

Thus, IP requires some flavor of upper-layer protocol to address the issues of reliability, proper ordering, de-duplication, and data integrity. That's where TCP and UDP come in. TCP delivers packets in order, with no duplicates, and guarantees their integrity. It does all this by a simple mechanism: Checking with the sender of each packet to confirm receipt of delivery. This approach is perfect for sending files (which are generally sequences of packets) or other "chunks" of data that can handle small amounts of delay.

It's less effective for real-time traffic, such as voice. Instead, that traffic uses UDP, which doesn't guarantee reliability or ordering in the way that TCP does. Datagrams may arrive out of order, appear duplicated, or go missing without notice—but when they do arrive, they arrive quickly (without the overhead of checking for each packet). Applications that rely on UDP include IPTV and Voice over IP (VoIP).

What all this means, in practice, is that IP assumes the packets will be lost and/or delayed, and has access to several sophisticated mechanisms for coping with that. The real question is what impact these mechanisms have on users, and more broadly on the economy as a whole.

8.1 Access Circuit Saturation

Our model predicts that the far-and-away most likely scenario is that user demand will exceed access capacity (or effective access capacity, which we pegged at 30% utilization). In practice, this simply means that the access line begins to act like a local area network with too much contention—the rate of retransmission goes up, and net throughput goes down.

Does the Internet Ever Break?

At a practical level, the user experiences erratic performance: sometimes files will download quickly, and sometimes they'll drag on, seemingly forever. If large file transfers (such as peer-to-peer) are conflicting with latency-sensitive applications like voice, the voice quality may degrade, although it's worth noting that users can protect against this today by configuring quality-of-service capabilities on their local routers, which generally maintain overall service quality quite well even in congested links. This is not necessarily the case with interactive video, which consumes more bandwidth than voice—if the bandwidth isn't there to begin with, no amount of quality of service can help. But current access circuits can handle current video streams fairly well—the impact on video is likely to be what *won't* happen: no deployment of HD video, no telepresence applications.

Overall, transmitting over a saturated broadband link will feel a lot like the bad old days of dial-up: Long pauses between request and response, with some applications just too painful to bother with.

But the user experience is really just the tip of the iceberg. The real impact is the chilling effect that insufficient capacity exerts on companies that rely upon reliable Internet performance (YouTube, PhotoBucket, Amazon.com, etc) could be faced with a crisis if their customer base simply can't access their “product” in a tolerable manner. New companies that emerge—say to enable high-definition video downloads—may not survive. And finally that plan to rely upon the public Internet (via SSL and IP VPNs) as an increasing component of corporate connectivity may want to reconsider this strategy in light of the potential downstream performance impact.

The Internet, therefore, does not really break in the sense that it will refuse to provide service. Performance degrades to a point at which users are unhappy (and find other things to do that don't involve using the network). Businesses that might generate value from the Internet infrastructure fail to launch, or thrive, and existing businesses that rely on Internet services suffer. In essence, it's as though the Internet protects itself from increasing demand.

8.2 Router Issues

8.2.1 Router Congestion

One thing that's not likely to occur, however, is congestion in core routers (unless it's resulting from other reasons than sheer traffic volume, as noted below). That's because, according to our model, there simply isn't enough traffic being thrown at the routers to cause congestion—the traffic is all getting gated at the edges. And, in fact, the best available evidence indicates that core routers are *not* congested—latency across the core is low, and generally trending downwards.

What this means in practice is that those lucky users who have access to “fat pipes”—high-bandwidth, uncongested links into the ‘Net—will experience none of the delays discussed earlier. In fact, we relied on the measured performance of users with such circuits in calculating how much bandwidth users might use if they had access to a virtually unlimited supply.

Does the Internet Ever Break?

The problem is that such users are disproportionately businesses, who can obtain large amounts of bandwidth at the lowest possible rates, and then further manage it based on their needs. Consumers are often not able to purchase such “fat pipes”—at any price. Thus, an additional effect of the bandwidth crunch at the edges would be to increase the gap between the “haves” (chiefly businesses) and the “have-nots” (chiefly consumers).

8.2.2 Addressing and Route Table Expansion

There are two additional potential “break points” in the Internet that we haven’t discussed in this study so far, but they deserve mention. First is address exhaustion. The majority of today’s Internet relies on IP version 4, which provides for 4,294,967,296 unique addresses—far below projected numbers of Internet devices. Although there are various techniques (such as Network Address Translation, or NAT) that “expand” this number rather considerably, the fact remains that IPv4 is poorly architected to scale to the exabyte volumes of data we’re predicting.

The best solution proposed to date is IP version 6, which expands the number of available to 2^{128} --or more than enough to handle even the most aggressive scenarios. There are three problems with IPv6, however, and two are very closely related. First is that despite having been developed in the early 1990s, it hasn’t yet reached wide-scale deployment, despite a recent (US) government mandate. The reason is that it offers existing users relatively low additional value—essentially, those who are lucky enough to have IPv4 addresses don’t really need IPv6 addresses, and everyone else can’t convince the IPv4 address-holders to go through the pain and frustration of migration for something that doesn’t benefit them (the current IPv4 holders). (And the transition is indeed painful—since IPv6 isn’t backwards compatible with IPv4, the official migration plan requires users to run both sets of protocols—in essence creating an entire duplicate infrastructure, with gateways that translate between the two protocols.)

The statement that IPv6 doesn’t offer enough incremental value to entice IPv4 address-holders to switch has received some pushback in recent years, with claims of superior security, quality of service, etc. compared with IPv4. Discussing the technical ins-and-outs is pointless, however, because there’s clear proof that the benefits don’t exceed the pain of switching: users haven’t done so yet. IPv6 has been available, and heavily promoted, for over a decade—yet users continue to cling to their existing IPv4 addresses. By definition, the market has voted with its feet. Had there been sufficient benefits, the shift would have occurred long ago. The fact that it hasn’t is proof that there aren’t.

However, it’s the third problem with IPv6 that’s most worrisome: Like IPv4, it fails to solve the looming problem of route-table expansion. Although a discussion of routing algorithms is beyond the scope of this study, and in fact is one of the reasons that modeling the Internet is so difficult, it is nevertheless true that when the Internet begins to saturate, routing can become problematic. As

Does the Internet Ever Break?

routers increasingly find it difficult to find viable routes, the routing tables they generate can expand dramatically.

This has three possible consequences. First is that processors inside routing hardware will fail to keep pace. Most experts have concluded that Moore's Law protects us from such an occurrence in the near-to-medium-term future. Router vendors will simply buy faster processing hardware. Second is that routing tables will exceed available memory. Again, Moore's Law saves us: routing vendors can simply buy more (and better) memory.

But the third consequence is more worrisome. As network fragmentation grows, routers need to aggregate more routes, meaning that they can't necessarily store all optimal routes for particular sources and destinations. This means that, over time, routes become less optimal, packets traverse more hops, and core latency grows. Theoretically, the Internet could reach a point where it has so many routes it's no longer capable of supporting real-time traffic.

Now, IPv6 in and of itself doesn't cause this problem—but neither does it cure it. So the broader concern with IPv6 is that while deploying it is likely to be incredibly painful (at least for those in possession of IPv4 addresses) it doesn't ultimately solve the potentially much larger problem of route aggregation.

The bottom line in all this?

Somewhere between 2010 and 2012, if nothing changes, we expect to see significant performance issues for users with low (or lower-speed) access circuits. Additionally, users with IPv6 addresses will find connecting with v4 users complex and challenging (and vice versa). And, there's the outside possibility that route-aggregation issues begin degrading performance at the core. As kc claffy, CAIDA's principal investigator puts it, "The sky's not falling, but parts of it are getting pretty expensive to hold up."

9 Conclusions and Recommendations

We conclude that the evidence is good that demand for Internet and IP services is increasing exponentially, while access investment is proceeding linearly. An exponential curve will always intersect a linear one given enough time, and we believe there's reasonably compelling evidence that the intersection will happen within the next five years, possibly as early as 2010.

The impact of inadequate access infrastructure is likely to be relatively mild when it comes to the experiences of individual users, who will increasingly find themselves encountering Internet “brownouts” or “snow days,” during which performance will (seemingly inexplicably) degrade. But overall, we believe this lack of impact of this inadequate infrastructure will be to slow down the pace of both technical and business innovation. The next Google, YouTube, or Amazon might not arise not because of a lack of demand, but due to an inability to fulfill that demand. Rather like osteoporosis, the underinvestment in infrastructure will painlessly and invisibly leach competitiveness out of the economy.

That said, it's neither our goal nor our role to promote specific policy recommendations with this research. To the degree that our analysis is correct, and an investment gap exists, there are multiple political and economic mechanisms for addressing that gap, and it is beyond the scope of our research to recommend one approach over another (or even any one at all—perhaps the gap need not be filled).

There is, however, another gap that is within scope for us, as researchers, to address, and that is the “data gap.” We have several times noted that the best available data (chiefly that from CAIDA and MINTS) is exceedingly limited, due to the unwillingness of service providers to share details on their infrastructures and capacities.

There are good reasons for this, including concerns about customer privacy (service and content providers open themselves up to lawsuits by revealing data sets that can be deconstructed to reveal identifying details) and competitive advantage (many of the tools and technologies used to monitor network traffic are highly proprietary and represent competitive advantages).

However, we do not believe these problems are insolvable, given the quality of engineering talent employed by these content and service provider firms. And, we further believe that closing the “data gap” would yield immeasurable benefits to the industry as a whole, and provide the necessary platform upon which the policy decisions above—and a host of others—could be based.

So, the single actionable recommendation resulting from this study is a plea for service and content providers to cooperate with researchers in sharing data. We think the future of the Internet depends on it.

10 Appendix A: Detailed Methodology

10.1 Detailed Demand Methodology

To establish a baseline for Internet users, we focused on U.S. Census data, ITU data and other Web sources (need specific references). Table 27 shows the projected growth of Internet users between 2000 and 2012. There are an estimated 1.24B global users, and North America makes-up 235M, or 19% of the total world Internet population. Internet-connected users in 2012 are expected to exceed 1.32B globally, and 245M in North America. Internet-connected user growth is both a factor of increasing population and increasing availability of Internet access. This is evidenced by the fact that the fastest-growing regions for Internet growth are also the regions with the lowest current penetration of Internet access

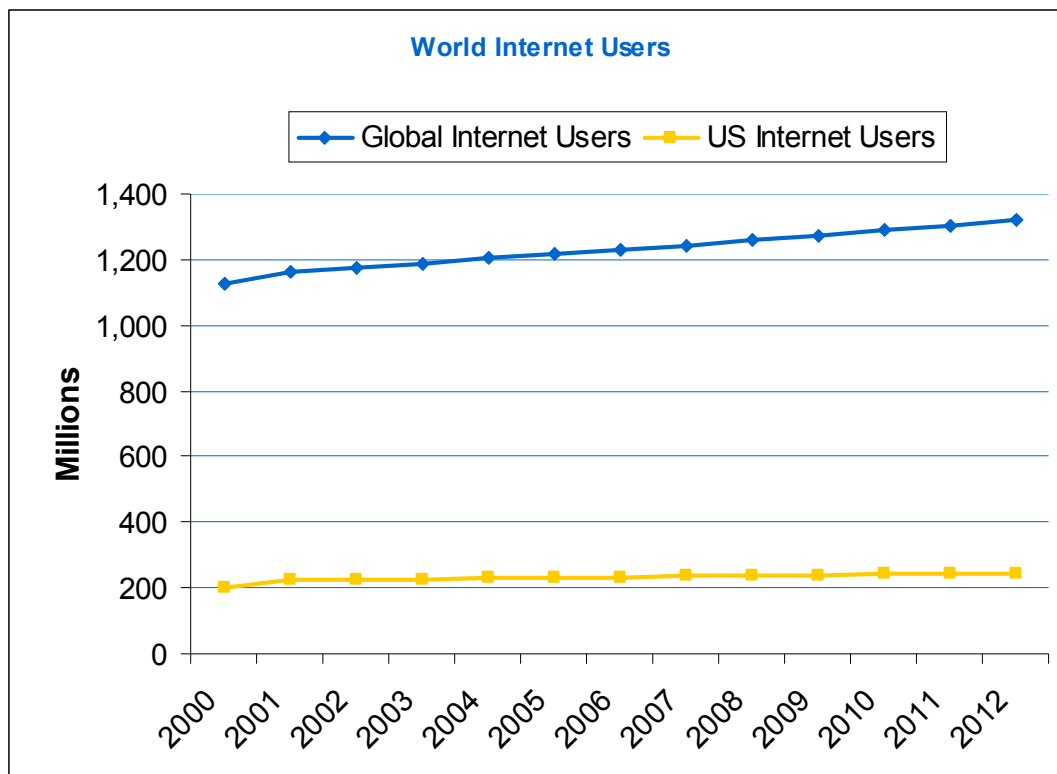


FIGURE 27: INTERNET USER POPULATION

Each of these users connects to the Internet via a range of options: broadband, wireless, cellular, and satellite. Each also is operating (or potentially operating) more than one Internet-enabled device: PC (home, school, and work), PDA/smartphone, game console, or IP-TV set-top box. We obtained the number of devices through a number of sources, including the World Wide Web Consortium, Taylor Nelson Sofres Intersearch (Clickz), and the CIA and other governmental or quasi-governmental entities.

Appendix A: Detailed Methodology

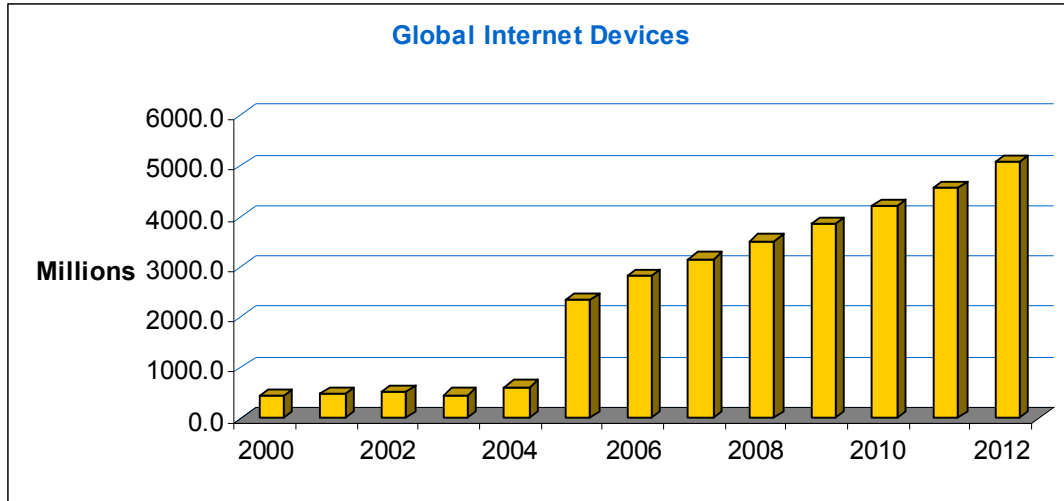


FIGURE 28: TOTAL INTERNET CAPABLE DEVICES

Our estimates show the global number of Internet-connected devices growing from 3 billion today to more than 5 billion by 2012. As astounding as these numbers are, we have seen estimates of up to 15 billion Internet-attached devices today, and 10 trillion in the next 15 years. (Dave Clark, 2007)

Nearly half of all Internet-connected devices in our estimates are Internet-enabled mobile devices. Though each device has relatively low bandwidth consumption, the aggregate demand is significant. Global deployment of 3G/4G wireless will dramatically increase the impact of these devices on Internet demand. This growth is reflected in our model.

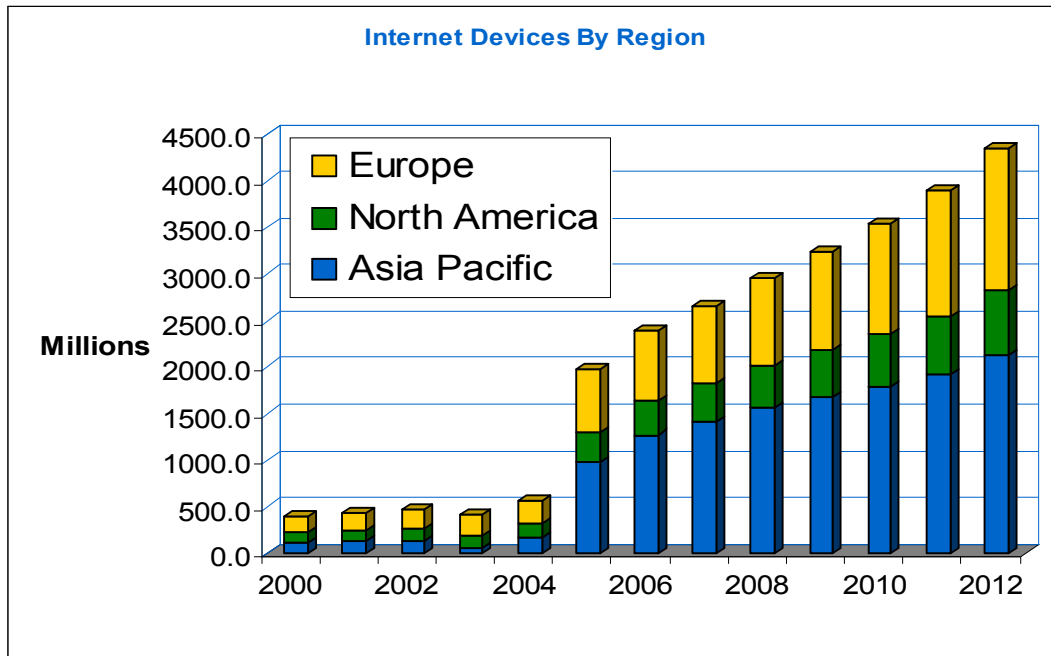


FIGURE 29: INTERNET CAPABLE DEVICES BY REGION

Appendix A: Detailed Methodology

As shown in figure 29, Asia Pacific and Europe dwarf North America in relative number of Internet-attached devices. Population explains part of this difference, but part is due to significant differences in Internet usage behaviors and desires. Our research indirectly reflects these differences by focusing on the relationship between users and devices. As an example, in 2007, we estimate there is a 2:1 ratio of Internet users to devices in Asia Pacific. This compares with an almost inverse relationship, 1:1.74 in North America between Internet users and devices. Europe nearly doubles North America with a ratio of 1:3.55 Internet users to devices. It is this ratio of users to devices, and the type of devices, that allows us to estimate global Internet demand, independent of looking at specific user usage patterns.

In order to do this, it is first necessary to estimate the likely data production capabilities of the devices in question. To do this, we looked at the likely communications port speed for each type of device. In the case of PCs, this ranged from 10 Megabit Ethernet to 1 Gigabit Ethernet. In the case of wireless, this ranged from 2.5G wireless devices running at 56 Kilobits and lower to 3G devices running in excess of 70 Megabits. When the proportion of these devices is factored in, the maximal data rates possible from all of the devices is simply the devices times their port speeds summed over all devices. When this is done, the following chart is generated:

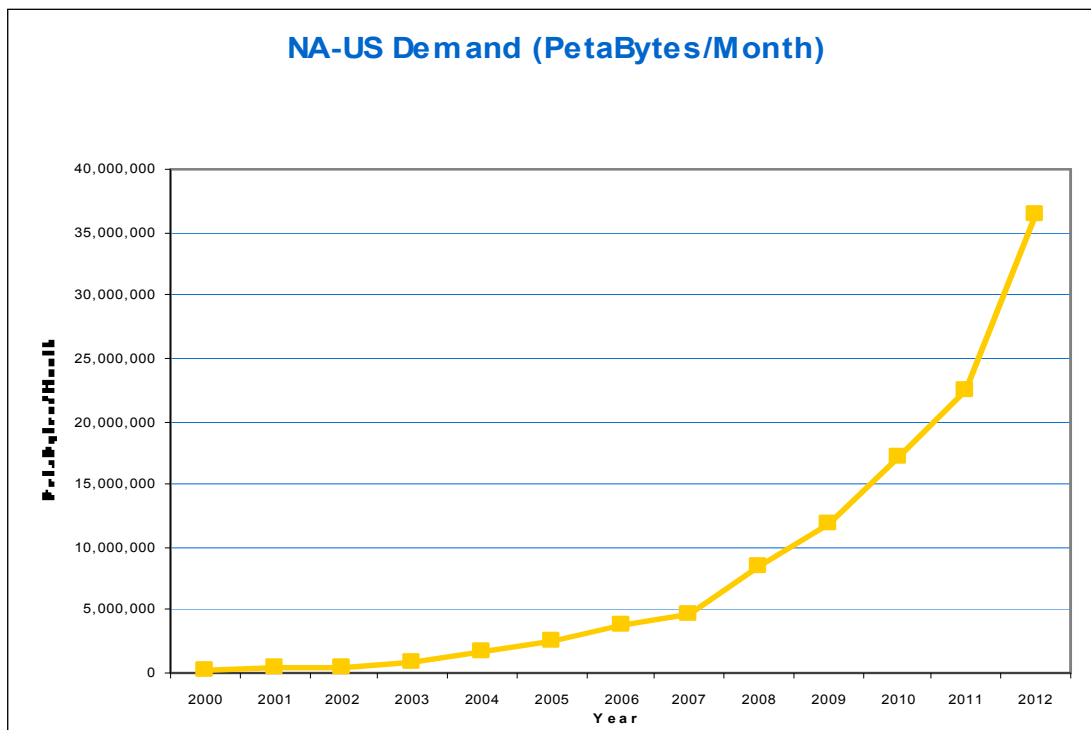


FIGURE 30: TOTAL NORTH AMERICAN DEMAND

This demand represents the total demand if all of the Internet capable devices in every user's possession were to generate data at its maximal capacity.

Appendix A: Detailed Methodology

This, of course, is absurd. No PC can actually pump a Gigabit of data continuously into the network. At the very least, routing protocols both at the LAN and higher in the network ensure that data can't be transmitted continuously. The question is, how do you get from what is theoretically possible at the margin and translate what will really occur normally.

For this, we invented a scalar we call utilization. Utilization translates raw demand into actual demand. The beauty of this approach is that it isn't really necessary that you know what utilization includes as long as you can verify empirically what its value is.

It turns out that data points do exist that allow you to do this in practice. Two known points are the demand value generated in 2001 by Coffman and Odlyzko (approximately 85 Petabytes per month). Another known point is the value generated by Cisco in 2006 (approximately 1900 Petabytes per month).

Additionally, we know that carriers have been claiming that the consumption has been doubling every year for the last two years at least. These three known points allow us to compute a utilization factor of .0657% in 2006. What this equates to is a user in North America using about 353 Megabytes per day. This seems reasonable when factoring all devices and including both work and personal Internet usage.

Utilization changes over time, and generates increases in demand that match the observed demand doubling seen by carriers. Additionally, utilization can be predicted based on projections of the growth of Internet capable devices as well as the growth of the Internet user population. Including these yields the following nominal demand growth curve:

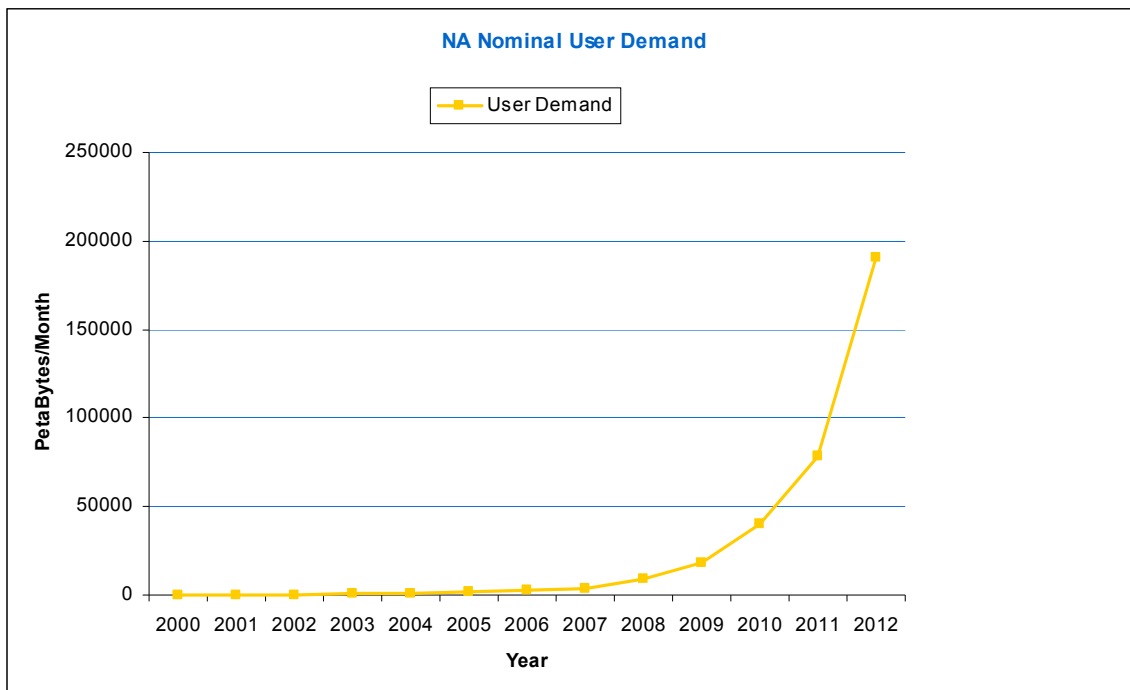


FIGURE 31: NORTH AMERICAN NOMINAL DEMAND

11 Bibliography and Sources

11.1 Sources

The bibliography includes only sources directly cited in the text of this report.

As noted in the acknowledgements, data for the model and charts drew from a wide variety of sources, including:

- Research data and Internet traffic statistics collected by academic organizations such as CAIDA and MINTS
- User demand data from a variety of sources, such as Pew Research and the Center for The Digital Future at the USC Annenberg School
- 70+ confidential interviews with enterprise organizations, equipment vendors, service providers, and investment companies
- Interviews with the several hundred IT executives who regularly participate in Nemertes' enterprise benchmarks.
- Investment figures from service providers and telecom equipment manufacturers.

11.2 Bibliography

Berstein, Marc (2005), The Essential Components of IPTV, Converge!
<http://www.convergedigest.com/bp-ttp/bp1.asp?ID=229&ctgy=>

Broadband Reports.com (2007), Ask DSLReports.com: Why Can't I get Comcast 16Mbps, <http://www.broadbandreports.com/shownews/Ask-DSLReportscom-Why-Cant-I-Get-Comcast-16Mbps-88281>

Cisco (2007), Global IP Traffic Forecast and Methodology, 2006-2011

ClickZ (2006), Internet Usrs Show Their Age, <http://www.clickz.com/3575136>

Coffman, K.G. & Odlyzko (2001), Growth of the Internet, AT&T Labs Research

Congress of the United States (2007), Discussion Draft on a Bill to Provide for a Comprehensive Nationwide Inventory of Existing Broadband Service, and for Other Purposes, http://www.benton.org/benton_files/broadbandcensus.pdf

The Cook Report (2007), Kick Starting the Commons and Preserving the Wealth of Networks, <http://cookreport.com/16.02.shtml>

Bibliography and Sources

Dell'Oro Group (2005), Ethernet Alliance Presentation on 100 Gigabit Ethernet
http://www.ethernetalliance.org/technology/presentations/DesignCon_2006_100G_ND.pdf

Digital TV News (2007), IPTV to reach 103M Homes by 2011,
<http://dtg.org.uk/news/news.php?id=2357>

Europa (2005), Employment in Europe,
<http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/05/383&format=HTML&aged=0&language=EN&guiLanguage=en>

Gantz, J., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A., Manfrediz, A. (2007), The Expanding Digital Universe: A Foretaste of World Information Growth Through 2010, an IDC Paper Sponsored by EMC
http://www.emc.com/about/destination/digital_universe/pdf/Expanding_Digital_Universe_IDC_WhitePaper_022507.pdf

Internet World Stats (2007), Internet Usage in Europe,
<http://www.internetworldstats.com/stats4.htm>

ITFacts (2007), US Console Market to generate \$66 bln by 2012,
<http://www.itfacts.biz/index.php?id=P8689>

Kurzweil, Ray (2006), The Singularity is Near: When Humans Transcend Biology, Penguin Group

NationMaster (2007), Average Size of Housholds by Country,
http://www.nationmaster.com/graph/peo_ave_siz_of_hou-people-average-size-of-households

Netcraft (October, 2007), www.netcraft.com

O'Brien, Kevin (2007), In Europe, a Push by Phone Companies into TV, the New York Times,
<http://www.nytimes.com/2007/08/29/business/worldbusiness/29tele.html>

PVC Forum (2007), Game Sales Charts: Computer and Video Game market Sales,
<http://forum.pcvconsole.com/viewthread.php?tid=15831>

Bibliography and Sources

Reier, Sharon (1998), Computers in the U.S. Do Far More and are Cheaper: Why Europeans Lag in Using PCs, nation Herald Tribune News, <http://www.iht.com/articles/1998/03/19/seurop.t.php>

UNECE (2007), Trends in Europe and North America, <http://www.unece.org/stats/trend/register.htm>

United States General Accounting Office (2006), Telecommunications: Broadband Deployment is Extensive Throughout the United States, but it is Difficult to Assess the Extent of Deployment Gaps in Rural Areas, <http://www.gao.gov/new.items/d06426.pdf>

U S Census (2000), "Home Computers and Internet Use in the United States: August 2000" <http://www.census.gov/prod/2001pubs/p23-207.pdf>

U.S. Department of Commerce (2004), A Nation Online: Entering the Broadband Age, <http://www.ntia.doc.gov/reports/anol/NationOnlineBroadband04.pdf>

WebSiteOptimization.com (2005), China Will Pass US in Broadband Lines by Late 2006, <http://www.websiteoptimization.com/bw/0601/>

Wilson, Carol (2006) Verizon Touts FiOS Market, Cost Cutting, Telephony Online, http://telephonyonline.com/home/news/verizon_touts_fios_092706/