Growth of the Internet

K. G. Coffman and A. M. Odlyzko
AT&T Labs - Research
kgc@research.att.com, amo@research.att.com

Preliminary version, July 6, 2001

Abstract

The Internet is the main cause of the recent explosion of activity in optical fiber telecommunications. The high growth rates observed on the Internet, and the popular perception that growth rates were even higher, led to an upsurge in research, development, and investment in telecommunications. The telecom crash of 2000 occurred when investors realized that transmission capacity in place and under construction greatly exceeded actual traffic demand. This chapter discusses the growth of the Internet and compares it with that of other communication services. Internet traffic is growing, approximately doubling each year. There are reasonable arguments that it will continue to grow at this rate for the rest of this decade. If this happens, then in a few years, we may have a rough balance between supply and demand.

Growth of the Internet

K. G. Coffman and A. M. Odlyzko
AT&T Labs - Research
kgc@research.att.com, amo@research.att.com

1. Introduction

Optical fiber communications was initially developed for the voice phone system. The feverish level of activity that we have experienced since the late 1990s, though, was caused primarily by the rapidly rising demand for Internet connectivity. The Internet has been growing at unprecedented rates. Moreover, because it is versatile and penetrates deeply into the economy, it is affecting all of society, and therefore has attracted inordinate amounts of public attention.

The aim of this chapter is to summarize the current state of knowledge about the growth rates of the Internet, with special attention paid to the implications for fiber optic transmission. We also attempt to put the growth rates of the Internet into the proper context by providing comparisons with other communications services.

The overwhelmingly predominant view has been that Internet traffic (as measured in bytes received by customers) doubles every three or four months. Such unprecedented rates (corresponding to traffic increasing by factors of between 8 and 16 each year) did prevail (within the US) during the crucial two-year period of 1995 and 1996, when the Internet first burst onto the scene as a major new factor with the potential to transform the economy. However, as we pointed out in [CoffmanO1] (written in early 1998, based on data through the end of 1997), by 1997 those growth rates subsided to approximate the doubling of traffic each year that had been experienced in the early 1990s. A more recent study [CoffmanO2] provided much more evidence, and in particular more recent evidence, that traffic has been about doubling each year since 1997. (We use a doubling of traffic each year to refer to growth rates between 70% and 150% per year, with the wide range reflecting the uncertainties in the estimates.)

Other recent observers also found that Internet traffic is about doubling each year. The evidence was always plentiful, and the only thing lacking was the interest in investigating the question. By the year 2000, though, the myth of Internet traffic doubling every three or four months was getting hard to accept. Very simple arithmetic shows that such growth rates, had they been sustained throughout the period from 1995 (when they did hold) to the end of 2000, would have produced absurdly high traffic volumes. For example, at the end of 1994, traffic on the NSFNet backbone, which was well

instrumented, came to about 15 TB/month. Had just that traffic grown at 1,500% per year (which is what a doubling every three months corresponds to), by the end of 2000, there would have been about 250,000,000 TB/month of backbone traffic in the U.S. If we assume 150 million Internet users in the U.S., that would produce a data flow of about 5 Mb/s for each user around the clock. The assumption of a doubling of traffic every four months produces traffic volumes which are only slighly less absurd.

The table below shows our estimates for traffic on the Internet. The data for 1990 through 1994 is that for the NSFNet backbone, and so is very precise. It is incomplete only to the extent of neglecting what is thought to have been small fractions of traffic that went completely through other backbones. The data for 1996 through 2000 are our estimates, and the wide ranges reflect the uncertainties caused by the lack of comprehensive data.

Table 1.1. Traffic on Internet backbones in U.S.. For each year, shows estimated traffic in terabytes during December of that year.

year	TB/month
1990	1.0
1991	2.0
1992	4.4
1993	8.3
1994	16.3
1995	?
1996	1,500
1997	2,500 - 4,000
1998	5,000 - 8,000
1999	10,000 - 16,000
2000	20,000 - 35,000

Table 1.2 presents our estimates of the traffic on various long distance networks at the end of 2000. The voice network still dominated, but was likely to be surpassed by the public Internet within a year or two. (For details of the measurements used to convert voice traffic to terabytes, and related issues, see [CoffmanO1].) In terms of bandwidth, the Internet is already dominant. However, it is hard to obtain good figures, since, as we discuss later, the bandwidth of Internet backbones jumps erratically. In terms of dollars, though, voice still provides the lion's share (well over 80%) of total revenues. We concentrate in this chapter (as in our previous papers, [CoffmanO1, CoffmanO2]) on the growth rates in Internet traffic, as measured in bytes. For many purposes, it is the other measures, namely bandwidth and revenues, that are more important. The reason we look at traffic is that we find more regularity there, and in the long run, we expect that there will be direct (although not linear) relations between

traffic and the other measures. In particular, based on what we have observed so far, we expect capacity to grow somewhat faster than traffic.

Table 1.2. Traffic on U.S. long distance networks, year-end 2000.

network	traffic (TB/month)
US voice	53,000
Internet	20,000 - 35,000
other public data networks	3,000
private line	6,000 - 11,000

The studies of [CoffmanO1, CoffmanO2] led to the proposal of a new form of Moore's Law, namely that a doubling of Internet traffic each year is a natural growth rate. This hypothesis is supported by the estimates of Table 1.1, as well as by evidence presented in [CoffmanO1, CoffmanO2] of many institutions whose data traffic has been growing at about that rate for many years. This "law" is discussed further in Section 8. It is not a law of nature, but rather, like the Moore's Law for semiconductors, a reflection of the complicated interactions of technology, economics, and sociology. Whether this "law" continues to hold or not will have important implications for the fiber optic transmission industry.

Much of this chapter, especially sections 6-8, is based on our earlier studies [CoffmanO1, CoffmanO2]. In Section 2, we present yet more evidence of how often popular perception and subsequent technology and investment decisions are colored by myths that are easy to disprove, but which nobody had bothered to disprove for an astonishingly long time. In Section 3, we look at historical growth rates of various communication services, and how they compare to the much higher growth rate of the Internet. Section 4 is a brief review of the history of the Internet. Section 5 discusses some of the various types of growth rates that are relevant in different contexts. Section 6 presents the evidence about Internet traffic growth rates we have been able to assemble. Section 7 is devoted to new sources of traffic that might create sudden surges of demand, such as Napster. Section 8 discusses the conventional "Moore's Law" and the analog we are proposing for data traffic. Section 9 suggests a way of thinking about data traffic growth, based on an analogy with the computer industry. Finally, Section 10 presents our conclusions.

2. Growth myths and reality

Internet growth is an unusual subject, in that it has been attracting enormous attention but very little serious study. In particular, the general consensus has been that Internet traffic is doubling every three

or four months. Yet no real evidence of that astronomical rate of growth was ever presented. As we discuss later, Internet traffic did grow at such rates in 1995 and 1996, but before and since it has been about doubling each year.

At this point, we would like to point out the need for careful quantitative data in evaluating any claims about growth rates. Some examples of public claims that do not match reality are presented in [CoffmanO2]. Here we discuss another case, this one concerning the widely held belief that any capacity that is installed will be quickly saturated. The British JANET network, which provides connectivity to British academic and research institutions, will be discussed in more detail later. What is important is that it is large (with three OC3 links across the Atlantic at the end of 2000), and has traffic statistics going back several years available at ⟨http://bill.ja.net/⟩. A press release, available at ⟨http://www.ja.net/press_release/archive_announce/index.html⟩ as "Increase in Transatlantic Bandwidth - 28 May 1998" (but actually dated 3 June 1998), described what happened when JANET's transatlantic link was increased from a single T3 to two T3s:

> With effect from Thursday 28 May 1998, JANET has been running a second T3 (45 Mbit/s) link to the North American Internet, bringing the total transatlantic bandwidth available to JANET to 90 Mbit/s. ... Usage of the new capacity has been brisk, with the afternoon usage levels reaching in excess of 80 Mbit/s. This is of course evidence of the suppressed demand imposed by the single T3 link operating previously. The fact that usage has risen so quickly on this occasion is also indicative of the improved domestic infrastructures ... that now exist.

This quote certainly appears to support the claim that demand for bandwidth is inexhaustible. One could easily conclude that traffic essentially doubled as soon as capacity doubled. The quote is imprecise, though, since it does not say how often those "afternoon usage levels" are "in excess of 80 Mbit/s," nor does it say how those usage levels are measured. The usage statistics for JANET, available at ⟨http://bill.ja.net/⟩, enable us to obtain precise information. Table 2.1 shows the data the transfer volumes on the more heavily utilized U.S. to U.K. part of the link, for several days before and after the doubling of capacity of the link. (No data for May 27 is available, and the figures for May 28, the day the second T3 was put into operation, are suspiciously low, probably reflecting incomplete measurements, so those are not included.)

What we observe is that although there was substantial growth in traffic after the capacity increase,

Table 2.1. Traffic from U.S. to the JANET network during late spring 1998, when the capacity was doubled.

day	GB	utilization
Wed 5/20	272.7	58.8%
Thu 5/21	275.5	59.4
Fri 5/22	265.1	57.1
Sat 5/23	202.7	43.7
Sun 5/24	189.8	40.9
Mon 5/25	211.2	45.5
Tue 5/26	267.2	57.6
Wed 5/27		
Thu 5/28		
Fri 5/29	286.6	30.9
Sat 5/30	209.7	22.6
Sun 5/31	199.9	21.5
Mon 6/01	318.1	34.3
Tue 6/02	319.2	34.4
Wed 6/03	295.9	31.9
Thu 6/04	343.2	37.0
Fri 6/05	322.4	34.7
Sat 6/06	208.3	22.4
Sun 6/07	202.7	21.8
Mon 6/08	338.0	36.4
Tue 6/09	307.2	33.1

suggesting that the transatlantic link had been a bottleneck, this increase was far more moderate than the popular Internet growth mythology or the JANET press release would make one think. While capacity doubled, traffic increased by less than a third.

3. Growth rates of other communication services

Telecommunications has been a growth industry for centuries, but growth rates have generally been modest, except for a few episodes, such as the beginnings of the electric telegraph (cf. [Odlyzko3]). For example, the number of pieces of mail delivered in the U.S. grew by a factor of over 50,000 between 1800 and 2000, but that was a growth rate of about 5.6% per year. (If we adjust for population increase, we find a growth rate of about 3.5% in the mail volume per capita.) The number of phone calls in the U.S. grew by a factor of over 230 between 1900 and 2000, for a compound annual growth rate of 5.6%. (The per capita growth rate was 4.2% during this period.) Long distance calls grew faster, about 12% per year between 1930 and 2000, and transatlantic calls faster yet. (There was just one voice circuit between the U.S. and Europe in 1927, when service was inaugurated. It used radio to span the ocean.

This single low quality link grew to 23,000 voice circuits to Western Europe by 1995, for a compound annual growth rate of capacity of 16%.)

One communications industry that has been growing very rapidly recently is wireless communication. Table 2.2 shows the growth of the U.S. cell phone industry, with the number of subscribers as of June of each year, and the revenue figures obtained by doubling those of the first six months of each year (and thus seriously understating the full-year figure). In many other countries, wireless communication has developed faster and plays a bigger role than it does in the U.S.. Still, even in the U.S., at the end of 2000, there were close to 100 million cell phones in use, and the rate of growth was far higher than for traditional wired voice services.

Table 2.2. Growth of U.S. cell phone industry

year	number of subscribers (millions)	revenues (millions)
1985	0.20	$ 352
1986	0.50	721
1987	0.89	959
1988	1.61	1,772
1989	2.69	2,813
1990	4.37	4,253
1991	6.38	5,307
1992	8.89	7,267
1993	13.07	9,639
1994	19.28	13,038
1995	28.15	17,499
1996	38.20	22,388
1997	48.71	26,270
1998	60.83	30,573
1999	76.28	38,737
2000	97.04	49,291

The cell phone example is worth keeping in mind, since it shows that volume of traffic or even the number of users has only a slight correlation to value. In the U.S. (unlike several other countries), there were more Internet users than cell phone subscribers at the end of 2000 (around 150 million vs. about 100 million). However, the revenues of the cell phone industry were far higher than those of the Internet. If we take a rough estimate of 60 million residential Internet users, and assume they pay an average of $20 per month (both slight overestimates), we find that the total revenues from this segment come to about $15 billion. Business customers, with dedicated connections to the Internet,

pay considerably less than that. For example, the 2000 revenues from business Internet connections of WorldCom (whose UUNet unit has the largest backbone in the world, often thought to carry over 30% of the total backbone traffic) were just $2.5 billion (up from $1.6 billion in 1999).

The conclusion of the previous paragraph is that even in the U.S., basic Internet transport revenues are less than half those of cell phones. Yet volumes of traffic are far higher on the Internet. The average daily time spent by a subscriber on a cell phone in the U.S. is about 8 minutes. If we count wireless communication as taking 8 Kb/s (since compression is used), we find that the total volume of traffic generated by cell phone users in the U.S. at the end of 2000 was only about 1,500 TB/month, a tiny fraction of the 20,000 to 35,000 TB/month traffic on U.S. Internet backbones. (Moreover, this comparison overestimates wireless traffic, since most of the mobile calls are local, whereas backbone traffic is by definition long distance.)

The comparison of revenues from Internet connectivity to those of the cell phone industry leads naturally to the next topic, namely a comparison with the entire phone industry. As we saw above, Internet revenues were under $25 billion in the U.S. in 2000. On the other hand, the revenues of the entire telephone industry (including wireless communication and data services such as private lines leased by corporations) were around $300 billion that year. Thus in terms of revenues, the Internet is still small. Furthermore, it is so intimately tied to the phone industry that it is difficult to see what its role is. The basic technologies (fiber transmission, SONET, and so on) that are used for Internet transport were developed initially for voice telephony, but were easily adopted for data. (Some, such as SONET, will likely turn out to be redundant, but are still widely used.) At the transport level, voice has been carried as bits for a long time. What happened is that during the late 1990s, the long distance telecommunications infrastructure has changed. It used to be dominated by the demands of voice transport, and data was a small part of what it carried. Now, however, its development is driven by data, especially Internet. For quite a long time, the volume of data was extremely small, so that even though the growth rate was higher than for voice, this did not affect the overall growth rate of the infrastructure. That was one reason the telecommunications industry has repeatedly been surprised by the demand for bandwidth in the 1990s. Moreover, the transition from voice to data domination was complicated by the presence of several types of data, with substantially different growth rates. We discuss this in more detail below.

Another reason that the recent upsurge in demand for bandwidth was a surprise is that there had been several previous false predictions that data traffic was about to explode. The excitement of the early 1990s about the “telecommunications superhighway” and “500 channels to the home,” to be

accomplished through technologies such as hybrid fiber-coax, certainly led to large financial losses and serious disappointments (cf. [Noll2]). However, there were even earlier periods of extremely rapid growth followed by sudden deceleration. For example, the number of modems in the U.S. grew between 1965 and 1970 at about 60% per year, to over 150,000 at the end of that period [WalkerM]. Had that growth rate been maintained, we would have had about 200 billion modems in the U.S. by the end of 2000, clearly an absurd number. Instead, it appears that growth in the 1970s followed the projections made around 1970 (p. 297 of [WalkerM]), which predicted annual increases of 25 to 30%. It is interesting to read the speculations in [DunnL] about the supposedly rosy prospects for electronic cash, distance education, and other data services (as well as for Picturephone) that were supposed to power the growth of networks. In general, predicting what communications services society will accept, and how it will use them, has been very hard, cf. [Lucky1, Odlyzko2]. In particular, even recent history is littered with technologies that seemed extremely promising at one point, such as ISDN (cf. [Kleinrock, WuL]) or SMDS (Switched Multimegabit Data Services - a high speed packet switched WAN tecnology), but never attained more than a marginal role.

There are two aspects of the inability to forecast the prospects of communications technologies that are worth discussing at greater length. One goes back to the earlier discussion of wireless telephony, and how the mobility offered by cell phones appears to be more important for many people than broadband Internet access. Sometimes, though, higher bandwidth did prevail. In the early days of telephony, there was widespread lack of appreciation of how attractive it would eventually prove to be. The telephone was used primarily for business purposes, and the telegraph appeared to be adequate for that to many. Yet it was the phone that won, even though it appeared to use bandwith very wastefully when compared to the telegraph, and even though it encouraged what was often dismissed as “idle chatter.” The attractions of instantaneous personal interactions turned out to be crucial in leading to an almost universal penetration of the telephone in industrialized countries. In the last four decades of the 20th century, though, the telecommunications industry several times attempted to extend its success with the voice telephone by introducing videotelephony. This service appeared to offer the attraction of an even deeper level of communication than voice. Yet prospective users have not only not embraced it, but have in many cases treated it with hostility. There is a growth of videoconferencing, but even that is far slower than its proponents had forecasted. For a variety of reasons that have not been completely explained, videotelephony does not appeal to people for person-to-person communication. On the other hand, mobile narrowband voice flourishes.

The other aspect of the dismal record in forecasting the prospects of communications technologies

that we now consider is that of the nature of traffic carried. Data networks, which in commercial settings go back about four decades, have spent essentially all this time in the shadow of the much larger voice telephone network. (They also benefited from being able to use the infrastructure of the phone network, and were also constrained by its limitations, but that is less relevant for us here.) It was therefore natural for networking experts to continuously think of voice traffic, and in particular of the possibility of eventually carrying it as data. Looking further out, to a stage where the progress of technology appeared to offer the possibility of data networks becoming much larger than the phone networks, it was also natural to think of enriching the communications medium through the addition of video. (See the projections of Estill Green [Green, Lucky2] and Hough [Hough], for example.) Later the huge volume of broadcast data (radio and especially television) offered further possibilities for traffic that could be carried on data networks. The key point is what was seen as eventually filling data network was streaming multimedia traffic. The Internet's rise to dominance was a surprise for many reasons, but one of the main ones was that it did not fit this model. Although much current work on Internet technologies is devoted to streaming multimedia, there are good reasons, to be discussed later, why such traffic is not likely to dominate.

Although it has proven hard to forecast which technologies will be widely adopted, once a service had been successfully introduced, it often showed regular growth rates for extended periods of time [Odlyzko2]. The approximately 30% annual growth rate that had been projected in 1970 for data transmission (or, to be more precise, for the proxy for actual transmission that is offered by the number of modems) appears to have held not just in the 1970s, but in the 1980s and most of the 1990s as well. There are no comprehensive statistics (and there are measurement problems, in that private lines, whose bandwidth is often taken as a measure of the data traffic, can also be used for voice transmission). However, there are a few pieces of evidence supporting those growth rates around 1980 in [deSolaPITH]. Those same growth rates appeared to also hold for long distance private line transmission in the mid 1990s [CoffmanO1], and for local data bandwidth in the late 1980s and most of the 1990s [Galbi]. The comprehensive data summarized in [Galbi] is especially interesting. During that period, installed computer power came close to doubling each year, and the new "Information Economy" was taking root, but this was not reflected in the volume of data traffic. This low rate of growth in data transmission may have come from the high cost and poor quality of data transmission, of from other causes, such as lack of uniform standards that would enable easy data communication between companies. It may also have been caused to a large extent by the slow rate at which computation and communication technologies were adopted. Whatever the reasons, this low growth rate of approximately 30% a year

(low by comparison to growth of computing power) in data transmission was higher than that of voice networks. Hence by the mid-1990s, the bandwidth of long distance data networks (primarily private lines, used for intra-company communication) was already comparable to that of the voice network [CoffmanO1].

The Internet has historically had a growth rate of close to 100% per year in the traffic it carried. As Table 1.1 shows, it was growing with striking regularity in the early 1990s at this rate. Then it experienced a period of astronomical growth in 1995 and 1996, and then reverted to an approximate doubling each year in 1997, and has continued growing at about that rate through the end of 2000. The big question is how fast it will grow in the future. While the overwhelming preponderance of opinion all through the end of 2000 was that Internet traffic was doubling every three or four months, by early 2001 the consensus started changing. Some analysts even began projecting declines in the growth rates to the 50% per year range by around 2005. And indeed, some sources of growth did dry up. With the crash of telecom stocks (caused largely by the realization that expected demand and revenues were not materializing), investments slowed, and many dot-coms that had been busily filling transmission pipes with their content disappeared. In a related development, corporate managements started asking for detailed justifications for new data networking expenditures, instead of rushing to endorse any proposals that came along. At various enterprises, the growth rates of data traffic, which had been close to doubling every year in the late 1990s, began to slow down towards doubling every 18 or 24 months. It is not inconceivable that overall data traffic growth may be moving back to its historical rate of around 30% per year. We do not think this will occur, but before considering the reasons why (presented in detail in sections 6 to 9). we look at the general history of the Internet and its growth rates.

At this point we just remark that the dominant role of the Internet in communications, whether in terms of bandwidth of networks, or popular consciousness, is a fairly recent phenomenon. There had been extensive discussions of the “Information Superhighway” and the “National Information Infrastructure” for a long time. Leading thinkers foresaw the possibilities for much improved communication offered by new technologies, and there was tremendous effort devoted to various systems. However, the general expectation was that the “Information Superhighway” would be composed of a very heterogeneous collection of (interconnected) networks. This was true even as late as the beginnings of the Clinton presidency, in 1993 and 1994 (cf. [NII]). It was only in the mid to late 1990s that the Internet was perceived as evolving towards the all-encompassing network, carrying all types of traffic.

4. Internet history

The past 5-10 years have witnessed not only an explosion of activity, but the creation of entirely new sectors within the optical industry. As the concept of WDM began to emerge, many new companies developing WDM transport equipment came into existence. The newer enterprises pushed the older established equipment vendors to more aggressive deployment schedules and a constant downward trend for the corresponding prices of WDM transport equipment followed. In what appeared to be an almost insatiable demand for more bandwidth, a situation arose that allowed the creation of the new companies and the accompanying innovation. Not only did new equipment vendors emerge, but also new national scale carriers were created. This trend is continuing as the concept of optical layering/networking is gaining acceptance and new optical equipment companies are being formed on a regular basis. They deal not only with "traditional" WDM transport equipment, but also with terrestrial ultra long haul systems, regional and metro optimized systems, and various incarnations of optical cross connects.

There were hundreds of developments and contributions enabling this burst of activity. Many of the technical innovations are described in this book and its predecessors. However, perhaps the greatest single factor that fueled this phenomena was the belief and perception that traffic and hence needed capacity were growing at explosive rates. This is a remarkable fact, especially when one recalls that around 1990, both the traditional carriers and most of their equipment vendors still expected the traffic demands to not vary much from the voice demand growths (which historically was around 10% per year). In fact both carriers and equipment vendors were arguing that WDM would not be needed and that going to individual channel rates of at most 10 Gb/s would be adequate. Also, around 1995, the conventional wisdom was that 8 channel WDM systems would suffice well into the foreseeable future. Now it almost appears as if the pendulum has swung the other way. Is too much capacity being deployed and are many of the reported traffic growth rates correct, and if so will they continue?

As we explained in the previous section, the early skepticism about the need for high capacity optical transport was rooted in the reality of the telecommunications networks. Up until 1990, they were dominated by voice, which was growing slowly. Then, by the mid-1990s, they became to be dominated (in terms of capacity) by private lines, which were growing three or four times as fast. And then, in the late 1990s, they came to be dominated by the Internet, which was growing faster still.

Before we go through the analyses for the traffic growth on the Internet we must first at least define the Internet and describe the history and structure of it. This is paramount in helping put much of later

described growth analyses into perspective.

When one now speaks of the Internet, it is usually described as an evolution from ARPANET to NSFNET, and finally to the commercial Internet that now exists. Arguably, the phenomenal growth of the Internet started in 1986 (more than 17 years after its "birth") with NSFNet. However, the path was very complicated and full of many twists and turns in its roughly 40 year history [Cerf, Hobbes, Leiner].

From the very early research in packet switching, academia, industry, and the US government have been intertwined as partners. Ironically, the beginnings of the Internet can trace itself back to the Cold War and specifically to the launch of Sputnik in 1957. The US government formed the Advanced Research Project Agency (ARPA - the name was later changed to DARPA, Defense Advanced Research Project Agency, and later back to ARPA) the year after the launch with the stated goal of establishing a US lead in technology and science (with emphasis on applications for the military). As ARPA was establishing itself, there were several pivotal works [Klein1, Baran] in the early 1960s on packet switching and computer communications. These works and the efforts they spawned laid many of the foundations that enabled the deployment of distributed packet networks. J.C.R. Licklider (of MIT) [LickC] wrote a series of papers in 1962 in which he "envisioned a globally interconnected array of computers which would enable 'everything' to easily access data and programs from any of the sites". Generically speaking, this idea is not much different from what today's Internet has become. Of importance is the fact the Licklider was the first head of the computer research program at DARPA (beginning in 1962), and in this role he was instrumental in pushing his concept of networks. Kleinrock published both the first paper on packet switching and the first book on the subject. In addition, Kleinrock convinced several key players of the theoretical feasibility of using packets instead of circuits from communications. One such person was Larry Roberts, one of the initial architects for the ARPANET. In the 1965-66 time frame ARPA sponsored studies on "cooperative network of [users] sharing computers"[Leiner], and the first ARPANET plans were begun, with the first design papers on ARPANET being published in 1967. Concurrently the National Physical Laboratory (NPL) in England deployed an experimental network called the NPL Network making use of packet switching. It utilized 768 kb/s lines.

A year before the Moon landing, in 1968, the first ARPANET requests for proposals were sent out, and the first ARPANET contracts were awarded. Two of the earliest contracts went to UCLA to develop the Network Measurement Center, and to Bolt, Beranek and Newman (BBN) for the Packet Switch contract (to construct the Interface Message Processors or IMPs - effectively the routers).

Kleinrock headed the Network Measurement Center at UCLA and it was selected as the first node

on the ARPANET. The first IMP was installed at UCLA and the first host computer was connected in September of 1969. The second node was at Stanford Research Institution (SRI). Two other nodes were added at UCSB and in Utah, so that by the second half of 1969, just months past the first moon landing, the initial four node ARPANET became functional. This was truly the initial ARPANET, and thus a case can be made that this was when the Internet was born. The first message carried over the network went from Kleinrock's lab to SRI. Supposedly the first packet sent over ARPANET was sent by Charley Kline and as he was trying to log in the system crashed as the letter "G" of "LOGIN" was entered.

One of the next major innovations for the fledgling Internet (i.e., ARPANET) was the introduction of the first host-to-host protocol called Network Control Protocal or NCP, which was first used in ARPANET in 1970. By 1972 all of the ARPANET sites had finished implementing NCP. Hence the users of ARPANET could finally begin to focus on the development of applications - another paramount driver for the phenomenal growth and sustained growth of the internet. It was also in 1970 that the first cross-country link was established for ARPANET by AT&T between UCLA and BBN (at the blinding rate of 56 kb/s). By 1971, the ARPANET had grown to 15 nodes and had 23 hosts. However, perhaps the most influential work that year was the creation of an email program that could send messages across a distributed network. (Email was not among the original design criteria for the ARPANET, and its success caught the creators of this network by surprise.) Ray Tomlinson of BBN developed this, and his original program was based on 2 previous ones [Hobbes]. Tomlinson modified his program for ARPANET in 1972, and at that point its popularity quickly soared. In fact it was at this time that the symbol "@" was chosen. Arguably Internet email as we know it today can trace its origins directly to this work. Internet email was clearly one of key drivers for the popularity (and hence the phenomenal traffic growth demands) of the Internet and was the first "killer app" for the Net. It was every bit as critical to the Internet's "success" as the spreadsheet applications were to the popularization of the PC. Internet email provided a new model of how people could communication with each other and alter the very nature of collaborations.

Although there was already considerable work being done on packet networks outside the US, the first international connections to the ARPANET (to England via Norway) took place in 1973. To put the time frame in perspective this was the same year that Robert Metcalfe did his PhD which described his idea for Ethernet. Also during this year the number of ARPANET "users" was estimated to be 2000 and that 75% of all the ARPANET traffic (in terms of bytes) was email. One needs to note that in only 1-2 years from its introduction onto the Internet email became the predominant type traffic. The same

behavior took place several years later for html (i.e., Web traffic), and to a somewhat lesser degree, this was seen for Napster-like traffic within many networks a few years later.

Several other key developments began to take place in the mid 1970s. The initial design specification for TCP published by Vint Cerf and Bob Kahn in 1974 [CerfK]. The NCP protocol which was being utilized at the time, tended to act like a device driver, whereas the future TCP (later TCP/IP) would be much more like a communications protocol. As is discussed later, the evolution from ARPANET's NCP protocol to TCP (which in 1978 was split into TCP and IP) was critical in allowing the future growth and scalability of today's Internet. DARPA had three contracts to implement TCP/IP (at the time still called TCP), at Stanford (led by Cerf), BBN (led by Ray Tomlinson) and UCLA (led by Kirsten). Stanford produced the detailed specification and within a year there were 3 independent implementations of TCP that could interoperate.

It is noted that the basic reasons that led to the separation of TCP (which guaranteed reliable delivery) from IP actually came out of work that was done trying to encode and transport voice through a packet switch. It was found that a tremendous amount of buffering was needed, in order to allow for the appropriate reassembly after transmission was completed. This in turn led to trying to find a way to deliver the packets without requiring a guaranteed level of reliability. In essence, the UDP (User Datagram Protocol) was created to allow users to make use of IP. In addition, it was also in 1978 that the first commercial version of ARPANET came into existence as BBN opened Telenet.

In 1981-82 the first plans were being made to "migrate" from NCP to TCP. It is claimed by some that it was this event (TCP was established as THE protocol suite for ARPANET) was truly the birth of the Internet - defined as a connected set of networks, specifically those with TCP/IP. A few years later (in 1983) another major development occurred, which later enabled the Internet to scale with the "explosive" growth and popularity of the future Internet. This was the development of the name server (which evolved into the DNS) [Cerf, Leiner]. The name server was developed at the University of Wisconsin [Hobbes] This made it easy for people to use the network since hosts were assigned names and it was not necessary to remember numeric addresses. Much of the credit for the invention of the DNS (domain name server) is credited to Paul Mockapetris of USC/ISI [Cerf].

The year 1983 was also the date for two other key developments on ARPANET. The first one was the cutover from NCP to TCP on the ARPANET. Secondly, ARPANET was split into ARPANET and MILNET. Although the road was convoluted, this split was one of the key bifurcations points that later allowed NSFNET to come into existence. Soon thereafter (in 1984) the number of hosts on the ARPANET had grown to 1000, and the next year in 1985 the first registered domain was assigned in

March.

In 1985 NSFNET was created with a backbone speed of 56 kb/s. Initially there were 5 supercomputing centers that were interconnected. One of the paramount benefits of this was that it allowed an explosion of connections (most importantly from universities) to take place. Two years later in 1987, NSF agreed to work with MERIT Network to manage the NSFNet backbone. The next year (1988) the process of upgrading the NSFNet backbone to one based on T1 (i.e., 1.5 Mb/s links) was begun. In 1987 the number of hosts on the Internet broke the 10,000 number. Two year later in 1989 this had grown to around 100,000, and 3 years after that in 1992 it reached the 1,000,000 value. It is noted that if you look at how the number of hosts had been growing from 1984 to 1992 that it was still pretty much tracking a growth curve that was LESS than tripling each year (i.e., doubling every 9 months). In the 1985-86 time frame key decision was made that had very long term impact: that TCP/IP would be mandatory for the NSFNet program.

In the 1988-1990 time frame a conscious decision was made to connect the Internet to electronic mail carriers, and by 1992 most of the commercial email carriers in the US were “like the Internet”. This was still another development that cemented email as the single most important application to take advantage of the Internet.

In 1990 the ARPANET ceased to exist, and arguably NSFNet was the essence of the Internet. The following year Commercial Internet Service Providers began to emerge (PSI, ANS, Sprint Link, to name a few) and the Commercial Internet Xchange (CIX) was organized in 1991 by commercial ISPs to provided transfer points for traffic. NSF’s lifting the restriction on the commercial use of the Net was again one of the pivotal decisions. This was again a key bifurcation point, in that this helped set the stage for the complete commercialization of the Net that would follow only a few years later. In 1991 the upgrading of the NSFNet backbone continued as the work to upgrade to a T3 (i.e., 45 Mb/s links) began. It also interesting to note that it was the next year (1992) than the term “surfing the Internet” was first coined by Jean Armour Polly [Polly], only two years before the ARPAnet/Internet celebrated its 25th anniversary.

It was in the 1993-1995 time period that several major events seemed to emerge which fueled an almost explosive growth in the popularity of the Internet. One of the key ones was the introduction of “browsers” most notable Mosaic. This led to the creation of Netscape that went public in 1995. Even as early as 1994 WWW (i.e., predominantly html) traffic was increasing in volume on the Net. By then it was the second most popular type of traffic, surpassed only by ftp traffic. However, in 1995 WWW traffic surpassed ftp as the greatest amount of traffic. In addition the traditional online dial up systems

such as AOL, Prodigy and Compuserve began to provide Internet access.

In 1996 the net truly became public with the NSFNet being phased out. Soon thereafter major infrastructure improvements were made within the transport part of the Internet. The Internet began to upgrade much of its backbone to OC3-OC12 (up to 622 Mb/s) links, and in 1999 upgrades began for much of the Net to OC-48 (2.5 Gb/s) links.

5. The many Internet growth rates

The Internet is very hard to describe. By comparison, even the voice phone system, which is a huge enterprise, far larger in terms of revenues than the Internet, is much simpler. In the phone system, the basic service is well defined and simple to describe. The users have only limited ability to interact with the system. The Internet is completely different. Users interact with the system in a multiplicity of ways, on wildy different time scales, and there are many complicated feedback loops. The paper [FloydP] is an excellent overview of the problems that arise in attempting to simulate the Internet.

The problems of measuring the Internet are also formidable. There are many different measures that are relevant. In this chapter, just as in the papers [CoffmanO1, CoffmanO2], we will concentrate on traffic as measured in bytes. For the optical fiber telecommunications industry, it is capacity that is most relevant. Unfortunately there are numerous problems in measuring capacity. Much of the fiber is not lit, and even when it is lit, often only a few wavelengths are lit. Finally, much of potential capacity is used for restoration, through SONET or other methods. In addition, even at the levels of links used for providing IP traffic, it is hard to obtain accurate capacity measurements, since few carriers provide detailed data. Further, this type of capacity has a tendency to jump suddenly, as bandwidth is usually increased in large steps (such as going from OC3 to OC12, and then OC48, a phenomenon that contributes to the low utilization of data links [Odlyzko1]). Thus there is little regularity in capacity growth figures. On the other hand, we do find astonoshing regularity in traffic growth, which leads us to propose that a form of “Moore’s Law” applies. In the long run, we expect that capacity will grow slightly faster than traffic, as we explain later.

For many purposes other measures are important, such as the number of users, how they spend their time, how many and what types of commercial transactions they engage in, and so on. There are many sources of such data, and useful references can be found at [Cyberspace, MeekerMJ, Nua].

6. Internet traffic and bandwidth growth

Whether Internet traffic doubles every three months or just once a year has huge consequences for network design as well as the telecommunications equipment industry. Much of the excitement about and funding for novel technologies appear to be based on expectations of unrealistically high growth rates ([Bruno]). In this section we briefly examine a variety of examples in an attempt to understand the traffic growth rates that the Internet has experienced over its lifetime. There are places where the traffic is growing at rates that exceed 100% per year. One such example is LINX (London Internet Exchange). Its online data, available at ⟨http://ochre.linx.net/⟩, clearly shows a growth rate of about 300% from early 1999 to early 2001. There are also examples with growth rates even higher, although those tend to be for much smaller links or exchange points. However, there are also numerous examples of much more slowly growing links. In this section we briefly present growth rates from a variety of sources, and attempt to put them into context. In an earlier study [CoffmanO1] in 1997 we found that the evidence supported a traffic growth rate of about 100% per year (doubling annually). Four years later, the general conclusion is that Internet traffic still appears to be growing at about 100% per year. In other words, we have not found any substantial slowdown in the growth rate.

Some recent reports and projections conclude that Internet traffic is only about doubling each year, but claim that it was growing much faster until recently, and that its growth rate will continue to slow down. In that view, the telecom crash of 2000 was associated with a sudden decline in the growth rate of traffic. As far as we can tell, that is not accurate. The general rate of growth of traffic appears to have been remarkably stable throughout the period 1997-2000. As one of the most convincing pieces confirming this claim, we cite the news story [Cochrane], based on official figures from Telstra, the dominant Australian telecommunications carrier. This story reports that Telstra's IP traffic was almost exactly doubling each year between November 1997 and November 2000. (The printed version of this news story, but not the one available online at the URL listed in [Cochrane], shows a very regular growth, about 100% per year, from the beginning of 1997 to November 2000.) Hence our conclusion is that the problem the photonics industry is experiencing are not caused by any sudden slowdown in traffic, but rather by a realization that the astronomical growth rates that people had been assuming were phantasies.

Most of this section is drawn from the more detailed account in [CoffmanO2]. There are only a few new pieces of information. For example, the China Internet Network Information Center has statistics (at ⟨www.cnnic.net.cn/develst/e-index.shtml⟩) of the Internet bandwidth between China and the rest of

the world. It grew from 84.64 Mb/s in June 1998 to 2,799 Mb/s in December 2000, for a compund growth rate of 305% per year. Thus even in a rapidly growing economy like that of China, where the Internet penetration is low, and which is trying to catch up with the industrialized world, traffic is only doubling about every six months.

The comparison of the international bandwidth for Australia and China is instructive. In December 2000, Telstra had about 1,000 Mb/s to the rest of the world, about a third of Chinese bandwidth. Thus, making allowances for other Australian carriers, we can speculate that Australia may be exchanging half as much traffic with international destinations as China does, even though the latter has over 60 times the population. This shows the degree to which countries can differ in their intensity of Internet usage. The data in [Cochrane], showing that Telstra's IP traffic in November 2000 reached about 270 TB/month also shows that our general estimates for U.S. backbone traffic are reasonable, since the U.S. is not only larger than Australia, but also richer on a per capita basis and has a better developed telecommunications infrastructure.

In the remainder of this section we examine some of the data and trends from ISPs (Internet Service Providers), exchange points, and residential traffic patterns, along with traffic from "stable sources" (such as corporate, research, and academic networks).

It is noted that the data for the first two sources (ISPs and exchange points) is not nearly as complete nor reliable as only a few years ago. However, much better data is available for the "stable sources", and several are examined in much more detail later. As a brief note on conversion factors, traffic that averages 100 Mb/s is equivalent to about 30 TB/month. (It is 32.4 TB for a 30-day month, but such precision is excessive given the uncertainties in the data we have.)

Unfortunately the largest ISPs do not release reliable statistics. This situation was better even a couple of years ago. Much of the older data was used in previous studies [CoffmanO1]. For example, MCI used to publish precise data about the traffic volumes on their Internet backbone. Even though they were among the first ISPs to stop providing official network maps, one could obtain good estimates of the MCI Internet backbone capacity from public presentations. These sources dried up when MCI was acquired by WorldCom, and the backbone was sold to Cable & Wireless. As was noted in [CoffmanO1], the traffic growth rate for that backbone had been in the range of 100% a year before the change.

Today, one can obtain some idea of the sizes (but not traffic) of various ISP networks through the backbone maps available at [Boardwatch]. However, even those are not too reliable. The only large ISP in the U.S. to provide detailed network statistics is AboveNet, at ⟨http://www.above.net/traffic/⟩. Therefore, we looked at this ISP in moderate detail. We have recorded the MRTG (Multi-Router

Traffic Grapher) [MRTG] data for AboveNet for March 1999, June 1999, February 2000, June 2000, November 2000, and April 2001. The average utilizations of the links in the AboveNet long-haul backbone during those four months were 18%, 16%, 29%, 12%, 11%, and 10%, respectively. (The large drop between February and June 2000 was caused by deployment of massive new capacity, including four OC48s. One of the reasons we concentrate on traffic and not network sizes in this chapter is that extensive new capacity is being deployed at an irregular schedule, and is often lightly utilized. Thus it is hard to obtain an accurate picture of the evolution of network capacity.) If one just adds up the volumes for each link separately, one finds that between March 1999 and April 2001, the total volumes of traffic increased at an annual growth rate of about 200%. However, this figure has to be treated with caution, as actual traffic almost surely increased less than 200%. During this period, AboveNet expanded geographically, with links to Japan and Europe, so that at the end it probably carried packets over more hops than before. Since we are interested in end-to-end traffic as seen by customers (which can be thought of as the ingress and/or egress traffic into and/or out of "the network"), we have to deflate the sum of traffic volumes seen on separate backbone links by the average number of hops that a packet makes over the backbones (perhaps around 3). Even when there is reliable data for a single carrier, such as AboveNet, some of the growth seen may be coming from gains in market share, both from gains within a geographical region, and from greater geographical reach, and not from general growth in the market.

We next look at Internet exchange points. When the NSF Internet backbone was phased out in early 1995, it was widely claimed that most of the Internet backbone traffic was going through the Network Access Points or NAPs (which are effectively interconnection vehicles), which tended to provide decent statistics on their traffic. Currently it is thought that only a small fraction of backbone traffic goes through the NAPs, while most goes through private peering connections. Furthermore, NAP statistics are either no longer available, or not as reliable. This is in sharp contrast to the situation in 1998 [CoffmanO1]. As documented elsewhere [CoffmanO2], there is very little that can be reliably concluded about current growth rates of Internet traffic by examining the statistics of the public NAPs in the U.S.

However, the situation was slightly better when we examined a large number of international exchange points. These included LINX (the London Internet exchange), AMS-IX (the Amsterdam Internet exchange), the Slovak Internet exchange, HKIX (a commercial exchange created by the Chinese University of Hong Kong, BNIX (located in Belgium), the INEX (an Irish exchange), and FICX (the Finish exchange). Some of these show growth rates of only about doubling per year while others show

much faster growth rates. Traffic interchange statistics are hard to interpret, unless one has data for most exchanges, which is virtually impossible to obtain. Much of the growth one sees can come from ISPs moving from one exchange to another, moving their traffic from one exchange to another, or else coming to an exchange in preference to buying transit from another ISP. Consider the specific case of LINX. A large part of its growth is almost surely caused by more ISPs exchanging their traffic there. Between March 1999, and March 2000, the ranks of ISPs that are members of LINX have grown by about two thirds, based on the data on the LINX home page. Hence the average per-member traffic through LINX may have increased only around 120% during that year.

The traffic from residential U.S. customers will probably begin to increase at a faster rate in the near future. The growth in the number of users is likely to diminish, as we reach saturation. (You cannot double the ranks of subscribers if more than half the people are already signed up!) However, broadband access, in the shape of cable modems and DSL (and to a lesser extent fixed wireless links) will stimulate usage. The evidence so far is that users who switch to cable modem or DSL access increase their time online by 50 to 100%, and the total volume of data they download per month by factors of 5 to 10. A 5 or 10-fold growth in data traffic would correspond to a doubling of traffic every four months if everyone were to switch to such broadband access in a year. However, that is not going to happen. At the end of 1999, there were about 3 million households in the U.S. with broadband access. The most ambitious projections for cable modem and DSL access call for about 13 million households to have such links in 2003, and between 50-60 million in the year 2007. That is approximately a doubling each year. (There was apparently almost a tripling in the ranks of households with broadband access in 2000, but the telecom crash that wiped out many of the ADSL providers has led to a slowdown in the pace of deployment in 2001.) The traffic from a typical residential broadband customer is likely to grow beyond the level we see today, as more content becomes available, and especially as more content that requires high bandwidth is produced. Still, it is hard to see average traffic per customer among those with broadband connections growing at more than 50% a year, say. Together with a doubling in the ranks of such customers, this might produce a tripling of traffic from this source. Since the ranks of customers with regular modems are unlikely to decrease much, if any, and since their traffic dominates, it appears that the most likely scenario will be for the total residential customer traffic growing no faster than 200% per year, and probably closer to 100% per year. (Access from information appliances, which are forecast to proliferate, is unlikely to have a major impact on total traffic, since the mobile radio link will continue to have small bandwidth compared to wired connections.)

We next consider traffic at various stable institutions, corporate, academic, and governmental.

Growth in traffic can be broken down into growth in the number of traffic sources, and growth in traffic per source. For LINX, much of the increase in traffic may be coming from an increase in member ISPs. For individual ISPs, much of the increase in traffic may also be coming from new customers. Yet in the end, that kind of growth is limited, as the market gets saturated. The rest of this section focuses on rates of growth in traffic from stable sources. Now nothing is completely stable, as the number of devices per person is likely to continue growing, especially with the advent of information appliances and wireless data transmission. Hence we will consider growth in traffic from large institutions that are already well wired, such as corporations and universities. Most corporations do not publicize information about their network traffic, and many do not even collect it. However, there are some exceptions. For example, Lew Platt, the former CEO of Hewlett-Packard, used to regularly cite the HP Intranet in his presentations. The last such report, dated September 7, 1998, and available at ⟨http://www.hp.com/financials/textonly/personnel/ceo/rules.html⟩, stated that this network carried 20 TB/month, and a comparison with previous reports shows that this volume of traffic had been doubling each year for at least the previous two years. (As an interesting point of comparison, the entire NSFNet Internet backbone carried 15 TB/month at its peak at the end of 1994.) Several other corporations have provided data showing similar rates of growth for their Intranet traffic, although some indicated their growth has slowed down, and a few have had practically no growth at all recently.

Internal corporate traffic appears to be growing much more slowly than the public Internet traffic. Data for retail private lines as well as for Frame Relay and ATM services show aggregate growth in bandwidth (and therefore most likely also traffic) in a range of 30-40% per year. The growth is slow for retail private lines, and fast for Frame Relay and ATM. These rates are remarkably close to the growth rate observed in the late 1970s in the US, which was around 30% per year [deSolaPITH]. Thus, it is the corporate traffic to the public Internet that is growing at 100% per year. It is also important to note that in the year 2000 over two thirds of the volume on the public Internet appeared to be business to business. Thus, the acceleration on the overall growth rate of data traffic to about 100% per year from the old 30% or so a year appears to be a consequence of the advantages of the Internet, with its open standards and any-to-any connectivity.

For the remainder of this section we concentrate on publicly available information, primarily about academic, research, and government networks. These might be thought of as unrepresentative of the corporate or private residential users. Our view is just the opposite, in that these are the institutions that are worth studying the most, since they normally already have broadband access to the Internet, tend to be populated by technically sophisticated users, and tend to try out new technologies first. The

spread of Napster through universities is a good example of the last point. We believe that Napster and related tools, such as Gnutella and Wrapster, are just the forerunners of other programs for sharing of general information, and not just for disseminating pirated MP3 files. As we explained elsewhere, there is already much more digital data on hard disks alone than shows up on today's Internet. Further, this situation is likely to continue.

The prevalent opinion appears to be that in data networks, "if you build it, they will fill it." Our evidence supports this, but with the important qualification that "they" will not fill it immediately. That certainly has been the experience in local area networks, LANs. The prevalence of lightly utilized long distance corporate links was noted in [Odlyzko1]. That paper also discussed the vBNS (very High Speed Backbone Network) research network, which was extremely lightly loaded. Here we cite another example of a large network with low utilizations and moderate growth rates. Abilene is the network created by the Internet2 consortium of U.S. universities [Dunn]. Its backbone consists of 13 OC48 (2.4 Gb/s) links. Moreover, most of the consortium members had OC3 links to it. The average utilization in June 2000 was about 1.5%, and by April 2001 it had grown to about 4.1%. Thus in spite of the uncongested access and backbone links, traffic did not explode.

Even on more congested links, it often happens that an increase in capacity does not lead to a dramatic increase in traffic. This is supported by several examples. Such examples include the University of Waterloo, the SWITCH network, the NORDUNet network, the European TEN-155 network, the Merit network, the University of Toronto, Princeton University, and the University of California at Santa Cruz [Coffman02] . Below we go into moderate detail for these networks. Figure 6.1 shows statistics for the traffic from the public Internet to the University of Waterloo over the last 7 years. Detailed statistics for the Waterloo network are available at ⟨http://www.ist.uwaterloo.ca/cn/#Stats⟩, but Fig. 6.1 is based on additional historical data provided to us by this institution. Just as for the JANET network discussed above, and the SWITCH network to be discussed later, as well as most access links, there is much more traffic from the public Internet to the institution than in the other direction. Hence we concentrate on this more congested link, since it offers more of a barrier. We see that even substantial jumps in link capacity did not affect the growth rate much. Traffic has been about doubling each year for the entire 7-year period. (Overall, the growth rate at the University of Waterloo has slowed down, to about 55% from early 1999 to early 2000. This was at least partially the result of official limits on individual users that were imposed, limits we will discuss later.)

The same phenomenon of traffic doubling each year, no matter what happens to capacity, can be observed in the statistics for the SWITCH network, which provides connectivity for Swiss academic

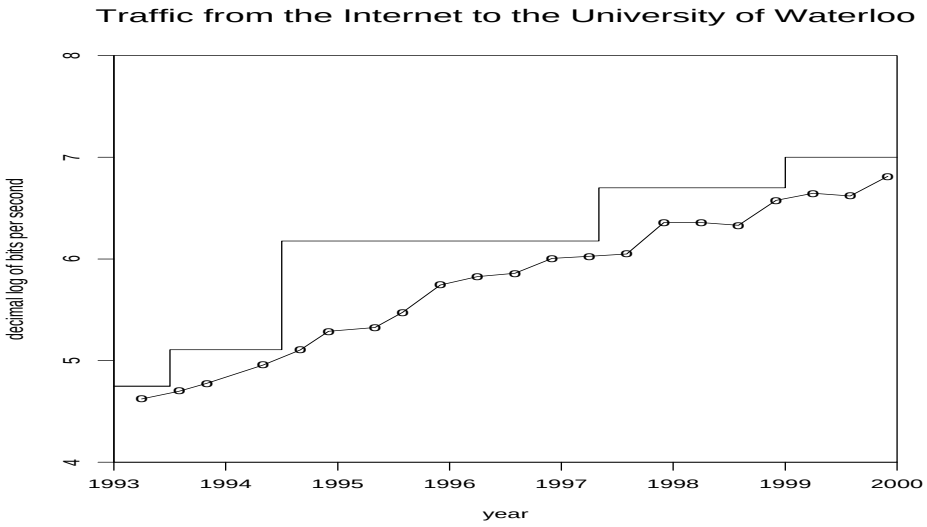

Figure 6.1. Traffic on the link from the public Internet to the University of Waterloo. The line with circles shows average traffic during the month of heaviest traffic in each school term. The step function is the full capacity of the link.

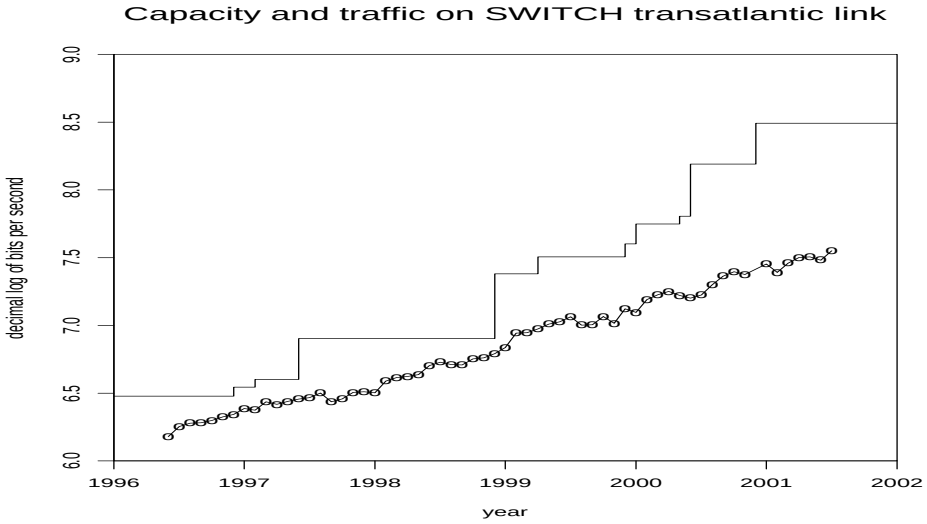

Figure 6.2. Capacity of link between the Swiss SWITCH network and the U.S., and traffic on it towards Switzerland.

and research institutions. The history and operations of this network are described in [Harms, ReichlLS], and extensive current and historical data is available at ⟨http://www.switch.ch/lan/stat/⟩. The data used to prepare Fig. 6.2 was provided to us by SWITCH. As is noted in [ReichlLS], the transatlantic link has historically been the most expensive part of the SWITCH infrastructure, and at times was more expensive than the entire network within Switzerland. It is therefore not surprising that this link tends to be the most congested in the SWITCH network. Even so, increasing its capacity did not lead to a dramatic change in the growth rate of traffic. If we compare increases in volume of data received between November of one year and January of the following year, there was an unusually high jump from Nov. 1998 to Jan. 1999, by 42%. This was in response to extreme congestion experienced at the end of 1998, congestion that produced extremely poor service, with packet loss rates during peak periods exceeding 20%. However, over longer periods of time, the growth rate has been rather steady at close to 100% per year and independent of the capacity of the link. More detailed data about other types of SWITCH traffic can be found at ⟨http://www.switch.ch/lan/stat/⟩, through the "Public access" link. The listings available there as of mid-2000, as well as those from previous years, show that various transmissions tended to grow at 100 to 150% per year. It is worth noting that capacity grew faster than traffic, but not too much faster.

Merit Network is a non-profit ISP that serves primarily Michigan educational institutions. It has data available online at ⟨http://www.merit.net/michnet/statistics/direct.html⟩ that goes back to January 1993. This data was used to construct the graph in Fig. 6.3. The data for January 1993 through June 1998 shows only the number of inbound IP packets. The data for months since July 1998 is more complete, but it is so complete, with details of so many interfaces, that we have not yet determined the best way to use it. Hence we have used only the earlier information for January 1993 through June 1998. The resulting time series is a reasonable although imperfect representation of a straight line, modulated by the periodic variations introduced by the academic calendar. The growth rate is almost exactly 100% per year.

The research networks that were examined have low utilizations. It should be emphasized that this is not a sign of inefficiency. Many novel applications required high bandwidth to be effective. That (along with some additional factors, such as the high growth rate, lumpy capacity, and pricing structure) contributes to the general much lower utilization of data networks than of the long distance voice network [Odlyzko1]

The general conclusion that can be drawn from the examples listed in this section (along with numerous other examples) is that data traffic has a remarkable tendency to double each year. There are

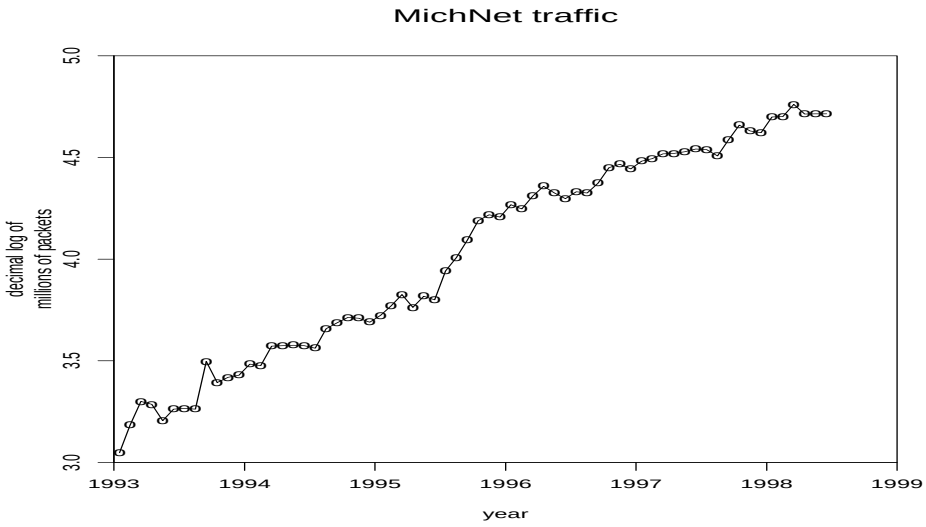

Figure 6.3. Traffic from Merit Network to customers.

of course slower and faster growth rates. Overall, though they tend to cluster in the vicinity of 100% per year. To date the authors have not seen any large institutions with traffic doubling anywhere close to three or even four months.

The growth rates that are cited here, are often affected strongly be restrictions imposed at various levels. As described elsewhere [CoffmanO1, CoffmanO2] , some of the explicit limits are imposed by network administrators.

The arrival of Napster (discussed in section 7) led many institutions to either ban its use or else limit traffic rates to some parts of the campus (typically student dormitories). Push technologies were stifled at least partially because enterprise network administrators blocked them at their firewalls. Email often has size restrictions that block large attachments (and in some cases all attachments are still banned). Teleconferencing is only slowly being experimented with on corporate intranets, and even packetized voice sees very limited (although growing) use.

Similar constraints apply to most of the content seen on the Web. As long as a large fraction of potential users have limited bandwidth, such as through dial modems, managers of Web servers will have an incentive to keep individual pages moderate in size. Thus, one can see that Internet traffic is subject to a variety of constraints at different levels. Some are applied by network managers, others by individual users, and the interaction of these constraints with the rising demands is fundamental what

produces the growth rates observed.

The ability to sustain the high growth rate of Internet traffic will require the creation of new applications that will generate huge traffic of volumes. At current growth rates, by 2005 there will be 8 times as much Internet as voice traffic (on the US long haul networks). If voice were packetized, in all likelihood the voice traffic would only account for about 3% of the Internet traffic. Thus, voice traffic will not fill the pipes that are likely to exit, and neither will traditional Web surfing. This will create a dilemma for service providers, network administrators, and equipment suppliers: to sustain the growth rates that the industry has come to depend on, and to accommodate the progress in technology, new technologies are needed. Such applications will appear disruptive to network operations today, and as such they often have to be controlled. However, tin the long run, they must be encouraged.

7. Disruptive innovation

It is often said that everything changes so rapidly on the Internet that it is impossible to forecast far into the future. The next "killer app" could disrupt any plans that one makes. Yet there have been just two "killer apps" in the history of the Internet: email and the Web (or, more precisely, Web browsers, which made the Web usable by the masses). Many other technologies that had been widely touted as the next "killer app," such as Push technology, (Push technology allows the sending of information directly to one's computer instead of the computer needing to actively go out and obtain it.) have fizzled. Furthermore, only the Web can be said to have been truly disruptive. From the first release of the Mosaic visual browser around the middle of 1993, it apparently took under 18 months before Web traffic became dominant on Internet backbones. It appears overwhelmingly likely that it was the appearance of browsers that then led, in combination with other developments, to that abnormal spurt of a doubling of Internet traffic every three or four months in 1995 and 1996.

What were the causes of the 100-fold explosion in Internet backbone traffic over the two-year period of 1995 and 1996? We do not have precise data, but it appears that there were four main factors, all interrelated. Browsers passed some magic threshold of usability, so many more people were willing to use computers and online information services. Users of the established online services, primarily AOL, CompuServe, and Prodigy, started using the Internet. The text-based transmissions of those services, which probably averaged only a few hundred bits per second per connected user, were replaced by the graphics-rich content of the Web, so transmission rates increased to a few thousand bits per second. Finally, flat rate access plans led to a tripling of the time that individual users spent online [Odlyzko3], as well as faster growth in number of users.

The Internet was able to support this explosion in use because it was utilizing the existing infrastructure of the telephone network. At that time, the Internet was tiny compared to the voice network. It is likely that the data network that handles control and billing for the AT&T long distance voice services by itself was carrying more traffic than the NSF Internet backbone did at its peak at the end of 1994. Today, by contrast, the public Internet is rapidly moving towards being the main network, so quantum jumps in traffic cannot be tolerated so easily.

In late 1999, a new application appeared that attracted extensive attention and led to many predictions that network traffic would see a major impact. It was Napster. At the time numerous articles in the press cited Napster's ability to "overwhelm Internet lines", and have claimed that it has forced numerous universities to ban or limit its use. The impression one got from those press reports was that Napster was causing a quantum jump in Internet traffic, and was driving the traffic growth rates well beyond the normal range. However, upon close examination this does not appear to be completely accurate, and the use of Napster has not increased growth rates much beyond the annual doubling or tripling rates, even within university environments, where Napster is most popular. That is not to say that is has not resulted in huge amounts of traffic, nor that it has not had serious impact on several major networks.

Napster provides software that enables users connected to the Internet to exchange and/or download MP3 music files. The Napster (web) site matches users seeking certain music files with other users who have those files on their computer. The Napster system preferentially uses as sources of files machines that have high bandwidth connections. This means that universities are the primary sources, since other organizations with fast dedicated links, mainly corporations, do not allow such traffic. The result is that although college students are often cited as the greatest users of MP3 files, it is the traffic from universities that gets boosted the most. (Since that direction of traffic is typically much less heavily used than the reverse one, the impact of Napster is much less severe than if the dominant direction of traffic were reversed.) Regular modem users are usually not affected, since their connections are too slow. However, the proliferation of cable modems and DSL connections that have "always-on" high bandwidth connectivity is leading to problems for some residential users, especially since the uplink is the one that invariably has the more limited bandwidth.

A key reason that Napster is of great interest to us is that similar types of sharing applications effectively turn consumers of information into providers of information. (The World Wide Web was designed for such information sharing, but for some types of files Napster and its kin are preferable.) These applications will effectively turn the traditional consumer PCs into Internet servers which will

output large amounts of traffic to other users. In Napster's case this has been predominantly MP3 music files, but other programs, such as Gnutella, work with more general data. It is highly probable that such applications could be one of the key applications that fuel the continued annual doubling or tripling of data traffic.

Napster first became noticeable in the summer of 1999. Its share of the total Internet traffic on many of the university networks has grown from essentially nothing to around 25% of the total traffic by mid to late 2000. In [CoffmanO2] the traffic generated by Napster and its impact on various networks was examined. The amount of Napster traffic that is reported by several university networks (such as UC Santa Cruz, University of Michigan, University of Michigan, Indiana, UC Berkeley, Northwestern University, and Oregon State University to name a few) range from around 20% at some to as high as 50%. However, the reported numbers are often very preliminary, and in some cases they compare Napster traffic to total traffic, while in others it appears that the high values may represent a comparison only to the out traffic. In any event this is a phenomenal growth rate for any single application. Since it started from zero and our data only goes out to about a year from that time, it is risky to extrapolate this initial explosion out indefinitely. In most cases [CoffmanO2] Napster has had a noticeable effect on the growth rate of traffic on this campus, but not an outlandish one.

Several networks, such as that of the University of Wisconsin-Madison that report Napster traffic making up as much as 30% of the total are not doing anything to limit Napster since they claim that they still have plenty of bandwidth. Others have imposed limits on the total bandwidth available to the dormitories.

Aside from Napster, occasionally even a large institution will experience a local perturbation in its data traffic patterns caused by one particular application. For example, the SETI@home distributed computing project, ⟨http://setiathome.ssl.berkeley.edu⟩, uses idle time on about three million PCs (as of mid-2001) to search for signs of extraterrestrial intelligence in signals collected by radio telescopes. This project is run out of the Space Sciences Institute at the University of California at Berkeley, and within a year of inception accounted for about a third of the outgoing campus traffic [McCredi]. (Moreover, this was extremely asymmetrical traffic, with large sets of data to be analyzed going out to the participating PCs, and small final results coming back. That most of the data went away from campus made this application less disruptive than it would have been otherwise.) Its disruptive effect is moderated by limiting its transmission rate to about 20 Mb/s. At the University of California at Santa Cruz, a complete copy of the available genome sequence was made available for public download in early July 2000. This, combined with coverage in the popular press and on Slashdot, led to an

immediate surge in traffic, far exceeding the effects of Napster. If the interest in this database continues, it will require reengineering of the campus network.

The SETI@home project is interesting for several reasons. It is cited in [McCredie] as a major new disruptive influence. Yet it contributes only about 20 Mb/s to the outgoing traffic. An increasing number of PCs and workstations are connected at 100 Mb/s, and even Gigabit Ethernet (1,000 Mb/s) is coming to the desktop. This means that for the foreseeable future, a handful of workstations will in principle be capable of saturating any Internet link. Given the projections for bandwidth, a few thousand machines will continue to be capable of saturating all the links in the entire Internet. Thus control on user traffic will have to be exercised to prevent accidental as well as malicious disruptions of service. However, it seems likely that such control could be limited to the edges of the network. In fact, such control will pretty much have to be exercised at the edges of the network. QoS will not help by itself, since a malicious attacker who takes over control of a machine will be able to subvert any automatic controls.

Finally, after considering current disruptions from Napster and SETI@home, we go back and consider browsers and the Web again. They were cited as disruptive back in 1994 and 1995. (Mosaic was first released unofficially around the middle of 1993, officially in the fall of 1993, and took off in 1994.) However, when we consider the growth rates for the University of Waterloo, for MichNet [CoffmanO1], or for SWITCH (which apparently had regular growth throughout the 1990s, according to [Harms]), we do not see anything anomalous, just the steady doubling of traffic each year or so. If we consider the composition of the traffic, there were major changes. For example, Fig. 7.1 shows the evolution of traffic between the University of Waterloo and the Internet. (It is based on analysis of traffic during the third week in each March, and more complete results are available at ⟨http://www.ist.uwaterloo.ca/cn/Stats/ext-prot.html⟩.) The Web did take over, but much more slowly than on Internet backbones. There are no good data sets, but it has been claimed that by the end of 1994, Web traffic was more than half of the volume of the commercial backbones. On the other hand, the data for the NSFNet backbone, available at ⟨ http://www.merit.edu/merit/archive/nsfnet/statistics/.index.html⟩, show that Web traffic was only approaching 20% there by the end of 1994, a level similar to that for the University of Waterloo. Thus at well-wired academic institutions such as the University of Waterloo and others that dominated NSFNet traffic, the impact of the Web was muted.

Perhaps the main lesson to be drawn from the discussion in this section is that the most disruptive factor is simply rapid growth by itself. A doubling of traffic each year is very rapid, much more rapid than in other communication services. Fig. 7.1 shows email and netnews shrinking as fractions of the

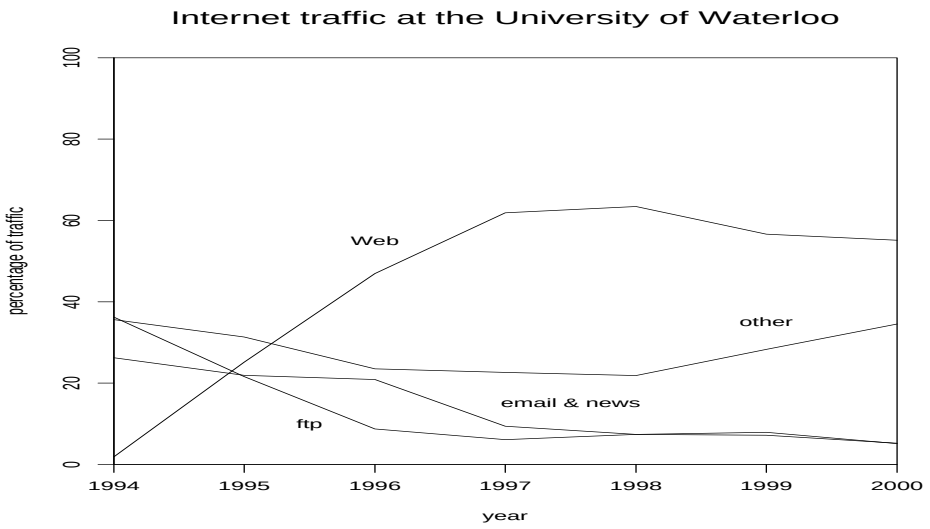

Figure 7.1. Composition of traffic between the University of Waterloo and the Internet. Based on data collected in March of each year.

traffic at the University of Waterloo, from a quarter to about 5%. Yet the byte volume of these two applications grew by a factor of 12 during the 6 years covered by the graph, for a growth rate of over 50% per year, which is very rapid by most standards. If we are to continue the doubling of traffic each year, new applications will have to keep appearing and assuming dominant roles. An interesting data point is that even at the University of Wisconsin in Madison, which analyzes its data traffic very carefully, about 40% of the transmissions escape classification. That is consistent with information from a few corporate networks, where the managers report that upwards of half of their traffic is of unknown types. (A vast majority of network managers do not even attempt to perform such analyses.) This shows how difficult coping with rapid growth is.

8. Moore's Law for data traffic

The approximate doubling of transmission capacity of each fiber that as described in [CoffmanO2] is analogous to the famous "Moore's Law" in the semiconductor industry. In 1965, Gordon E. Moore, then in charge of R&D at Fairchild Semiconductor, made a simple extrapolation from three data points in his company's product history. He predicted that the number of transistors per chip would about double each year for the next 10 years. This prediction was fulfilled, but when Moore revisited the

subject in 1975, he modified his projection for further progress by predicting that the doubling period would be closer to 18 months. (For the history and fuller discussion of "Moore's Law", [Schaller].) Remarkably enough, this growth rate has been sustained over the following 25 years. There have been many predictions that progress was about to come to a screeching halt (including some recent ones), but the most that can be said is that there may have been some slight slowdown recently. (For example, according to the calculations shown in [ElderingS], the number of transistors in leading-edge microprocessors doubles every 2.2 years. On the other hand, the doubling period is lower for commodity memories.) Experts in the semiconductor area are confident that Moore's 1975 prediction for rate of improvement can be fulfilled for at least most of the next decade.

Predictions similar to Moore's had been made before in other areas, and in [Licklider] they were made for the entire spectrum of computing and communications. However, it is Moore's Law that has entered the vernacular as a description of the steady and predictable progress of technology that improves at an exponential rate (in the precise mathematical sense).

Moore's Law results from a complex interaction of technology, sociology, and economics. No new laws of nature had to be discovered, and there have been no dramatic breakthroughs. On the other hand, an enormous amount of research had to be carried out to overcome the numerous obstacles that were encountered. It may have been incremental research, but it required increasing ranks of very clever people to undertake it. Further, huge investments in manufacturing capacity had to be made to produce the hardware. Perhaps even more important, the resulting products had to be integrated into work and life styles of the institutions and individuals using them. For further discussions of the genesis, operations, and prospects of Moore's Law, [ElderingSE, Schaller]. The key point is that Moore's Law is not a natural law, but depends on a variety of factors. Still, it has held with remarkable regularity over many decades.

While Moore's Law does apply to a wide variety of technologies, the actual rates of progress vary tremendously among different areas. For example, battery storage is progressing at a snail's pace, compared to microprocessor improvements. This has significant implications for mobile Internet access, limiting processor power and display quality. Display advances are more rapid than those in power storage, but nowhere near fast enough to replace paper as the preferred technology for general reading, at least not at any time in the next decade. (This implies, in particular, that the bandwidth required for a single video transmission will be growing slowly.) DRAMs are growing in size in accordance with Moore's Law, but their speeds are improving slowly. Microprocessors are rapidly increasing their speed and size (which allows for faster execution through parallelism and other clever

techniques), but memory buses are improving slowly. For some quantitative figures on recent progress [GrayS]. From the standpoint of a decade ago, we have had tidal waves of just about everything: processing power, main memory, disk storage, and so on. For a typical user, the details of the PC on the desktop (MHz rating of the processor, disk capacity) do not matter too much. It is generally assumed that in a couple of years a new and much more powerful machine will be required to run the new applications, and that it will be bought for about the same price as the current one. In the meantime, the average utilization of the processor is low (since it is provided for peak performance only), compression is not used, and wasteful encodings of information (such as 200 KB Word documents conveying a simple message of a few lines) are used. The stress is not on optimizing the utilization of the PC's resources, but on making life easy for the user.

To make life easy for the end user, though, clever engineering is employed. Because the tidal waves of different technologies are advancing at different rates, optimizing user experience requires careful architectural decisions [GrayS, HennessyP]. In particular, since processing power and storage capacity are growing the fastest, while communication within a PC is improving much more slowly, elaborate memory hierarchies are built. They start with magnetic hard disks, and proceed through several levels of caches, invisibly to the user. The resulting architecture has several interesting implications, explored in [GrayS]. For example, mirroring disks is becoming preferable to RAID (Redundant Arrays of Inexpensive Disks) fault tolerant schemes that are far more efficient but slower.

Table 8.1. Worldwide hard disk drive market. (Based on Sept. 1998 and Aug. 2000 IDC reports.)

year	revenues (billions)	storage capacity (terabytes)
1995	$21.593	76,243
1996	24.655	147,200
1997	27.339	334,791
1998	26.969	695,140
1999	29.143	1,463,109
2000	32.519	3,222,153
2001	36.219	7,239,972
2002	40.683	15,424,824
2003		30,239,756
2004		56,558,700

The density of magnetic disk storage increased at about 30% per year from 1956 to 1991, doubling every two and a half years [Economist]. (Total deployed storage capacity increased faster, as

the number of disks shipped grew.) In the 1990s, the growth rate accelerated, and in the late 1990s increased yet again. By some accounts, the densities in disk drives are about doubling each year. For our purposes, the most relevant figure will be total storage of disk drives. Table 8.1 shows data from an IDC study, which shows storage capacity shipped each year just about doubling through the year 2000, and then slowing down. However, that study was prepared in 1998, and since then IDC has revised upwards its estimates for disk storage systems towards a continuation of the doubling trend. Similar projections from Disk/Trend (⟨http://www.disktrend.com/⟩) also suggest that the total capacity of disk drives shipped will continue doubling through at least the year 2002. Given the advances in research on magnetic storage, it seems that a doubling each year until the year 2010 might be achievable (with some contribution from higher revenues, as shown in Table 8.1, but most coming from better technology). After about 2010, it appears that magnetic storage progress will be facing serious limits, but by then more exotic storage technologies may become competitive.

It seems safest to assume that total magnetic disk storage capacity will be doubling each year for the next decade. However, even if there is a slowdown, say to a 70% annual growth rate, this will not affect our arguments too much. The key point is that storage capacity is likely to grow at rates not much slower than those of network capacity. Furthermore, total installed storage is already immense. Table 8.1 shows that at the beginning of the year 2000, there were about 3,000,000 TB of magnetic disk storage. If we compare that with the estimates of Table 1.1 for network traffic, we see that it would take between 250 and 400 months to transmit all the bits on existing disks over the Internet backbones. This comparison is meant as just a thought exercise. The backbones considered in Table 1.1 are just those in the U.S., whereas disks counted in Table 8.1 are spread around the world. A large fraction of the disk space is spare, and much of the content is duplicated (such as those hundreds of millions of copies of Windows 98), so nobody would want to send them over the Internet. Still, this thought exercise is useful in showing that there is a huge amount of digital data that could potentially be sent over the Internet. Further, this pool of digital data is about doubling each year.

An interesting estimate of the volume of information in the world is presented in [Lesk]. It shows that already in the year 1997 we were on the threshold of being able to store all data that has ever been generated (meaning books, movies, music, and so on) in digital format on hard disks. By now we are well past that threshold, so future growth in disk capacities will have be devoted to other types of data that we have not dealt with before. Some of that capacity will surely be devoted to duplicate storage (such as a separate copy of an increasingly bloated operating system on each machine). Most of the storage, though, will have to be filled by new types of data. The same process that is yielding faster

processors and larger memories is also leading to improved cameras and sensors. These will yield huge amounts of new data, that had not been available before. It appears impossible to predict precisely what type of data this will be. Much is likely to be video storage, from cameras set up as security measures, or else ones that record our every movement. There could also be huge amounts of data from medical sensors on our bodies. What is clear, though, is that "[t]he typical piece of information will *never* be looked at by a human being" [Lesk]. There will simply not be enough of the traditional "content" (books, movies, music), nor even of the less formal type of "content" that individuals will be generating on their own.

Huge amounts of data that is machine generated for machine use suggests that data networks will also be dominated by transfers of such data. This was already predicted in [deSolaPITH], and more recently in [Odlyzko2, StArnaud, StArnaudCFM]. Given an exponential growth rate in volume of data transfers, it was clear that at some point in the future most of the data flying through the networks would be neither seen nor heard by any human being. Thus we can expect that streaming media with real-time quality requirements will be a decreasing fraction of total traffic at some point within the next decade.

There will surely be an increase in the raw volume of streaming real-time traffic, as applications such as videoconferencing move onto the Internet. However, as a fraction of total traffic, such transmissions will not only decrease eventually, but may not grow much at all even in the intermediate future. (Recall that at the University of Waterloo over the last 6 years, the volume of email grew about 50% a year, but as a fraction of total traffic it is almost negligible now.) The huge imbalance in volume of storage and capacities of long distance data networks means that even the majority of traditional "content" will be transmitted as files, and not in streaming form. For more detailed arguments supporting this prediction [Odlyzko2]. This development (in which "content" is sent around as files for local storage and playback) is already making its appearance with MP3, Napster, and related programs.

The huge hard disk storage volumes also mean that most data will have to be generated locally. There will surely also be much duplication (such as operating systems, movies, and so on that would be stored on millions of computers). Aside from that, there will surely be huge volumes of locally generated data (such as from security cameras and medical sensors) that will be used (if at all) only in highly digested form.

The examples in [CoffmanaO2] support the notion that there is a "Moore's Law" for data traffic, with transmission volumes doubling each year. Even at large institutions that already have access to state-of-the art technology, data traffic to the public Internet tends to follow this rule of doubling each

year. This is not a natural law, but, like all other versions of "Moore's Law," reflects a complicated process, the interaction of technology and the speed with which new technologies are absorbed. A "Moore's Law" for data traffic is different from those in other areas, since it depends in a much more direct way on user behavior. In semiconductors, consumer willingness to pay drives the research, development, and investment decisions of the industry, but the effects are indirect. In data traffic, though, changes can potentially be much faster. A residential customer with dial modem access to the Internet could increase the volume of data transfer by a factor of about five very quickly. All it would take would be installation of one of the software packages that prefetch Web sites that are of potential interest, and which fill in the slack between transmissions initiated by the user. Similarly, a university's T3 connection to the Internet could potentially be filled by a single workstation sending data to another institution. Thus any "Moore's Law" for data traffic is by nature much more fragile than the standard "Moore's Law" for semiconductors, for example. Thus it is remarkable that we see so much regularity in growth rates of data transfers.

Links to the public Internet are usually the most expensive parts of a network, and are regarded as key choke points. They are where congestion is seen most frequently at institutional networks. Yet the "mere" annual doubling of data traffic even at institutions that have plenty of spare capacity on their Internet links means that there are other barriers that matter. The obvious one is the public Internet itself. It is often (some would say usually) congested. A terabit pipe does not help if it is hooked up to a megabit link, and so providing a lightly utilized link to the Internet does not guarantee good end-to-end performance. Yet that is not the entire explanation either, since corporate Intranets, which tend to have adequate bandwidth, and seldom run into congestion, tend to grow no faster than a doubling of traffic each year. There are other obstructions, such as servers, middleware, and, perhaps most important, services and user interfaces. People do not care about getting many bits. What they care about is the applications. However, applications take time to be developed, deployed, and adopted. To quote J. Licklider (who probably deserves to be called "the grandfather of the Internet" for his role in setting up the research program that led to the Internet's creation),

> A modern maxim says: "People tend to overestimate what can be done in one year and to underestimate what can be done in five or ten years."
>
> (footnote on p. 17 of [Licklider])

"Internet time," where everything changes in 18 months, has a grain of truth, but is largely a myth.

Except for the ascendancy of browsers, most substantial changes take 5 to 10 years. As an example, it is at least four years since voice over IP was first acclaimed as the "next big thing." Yet its impact so far has been surprisingly modest. It is coming, but it is not here today, and it won't be here tomorrow. People take time to absorb new technologies.

What is perhaps most remarkable is that even at institutions with congested links to the Internet, traffic doubles or almost doubles each year. Users appear to find the Internet attractive enough that they exert pressure on their administration to increase the capacity of the connection. Existing constraints, such as those on email attachments, or on packetized voice, or video, as well as the basic constraint of limited bandwidth, are gradually loosened. Note that this is similar to the process that produces the standard Moore's Law for PCs. Intel, Micron, Toshiba, and the rest of the computer industry would surely produce faster advances if users bought new PCs every year. Instead, a typical PC is used for three to four years. On one hand there is pressure to keep expenditures on new equipment and software under control, and also to minimize the complexity of the computing and communications support job. On the other hand, there is pressure to upgrade, either to better support existing applications, or to introduce new ones. Over the last three decades, the conflict between these two pressures has produced a steady progress in computers. Similar pressures appear to be in operation in data networking.

In conclusion, we cannot be certain that Internet traffic will continue doubling each year. All we can say is that historically it has tended to double each year. Still, trends in both transmission and in other information technologies appear to provide both the demand and the supply that will allow a continuing doubling each year. Since betting against such "Moore's laws" in other areas has been a loser's game for the last few decades, it appears safest to assume that data traffic will indeed follow the same pattern, and grow at close to 100% per year.

9. Further economic and technical considerations

A frequently asked question concerns the elasticity of demand for data transmission capacity. However, for long-range projections it might be more useful to think of analogies with the computer industry. In that industry product managers clearly do think about elasticities in the short or intermediate terms. From a long-range perspective, though, what dominates are the effects of Moore's Law. Table 9.1 (drawn from [FishburnO]) shows a dozen years from the history of Intel. The leading microprocessor sold for roughly a constant price all during this period. However, its power was increasing at the exponential rate given by Moore's Law. Intel's total revenues (and profits) grew, as more processors were being sold, but this growth rate was considerably more modest than that of the computing power.

Users found the increasing computational power of new PCs sufficiently attractive that they not only bought new PCs, but increased their total spending. They did this even though most of that power was sitting idle, and it was only the occasional bursts of recomputing a spreadsheet or bringing up a presentation package that mattered. A similar evolution might take place in networking. Total spending may (subject to business cycles) increase at a moderate pace, while the bandwidth and traffic grow at rates determined by technological progress. If that happens, we are likely to see traffic and capacity about doubling each year, with capacity growth faster than that of traffic.

Table 9.1. Intel and its microprocessors. For each year lists the most powerful general purpose microprocessors sold by Intel, its computing power, price at the end of the year (in dollars), and Intel's revenues and profits for that year (in millions of dollars).

year	processor	mips	price	revenue	net profit
86	386 DX (16 MHz)	5	300	1265	-173
87	386 DX (20 MHz)	6		1907	248
88	386 DX (25 MHz)	8		2875	453
89	486 DX (25 MHz)	20	950	3127	391
90	486 DX (33 MHz)	27	950	3922	650
91	486 DX (50 MHz)	41	644	4779	819
92	DX2 (66 MHz)	54	600	5844	1067
93	Pentium (66 MHz)	112	898	8782	2295
94	Pentium (100 MHz)	166	935	11521	2266
95	Pentium Pro (200 MHz)	400	1325	16202	3566
96				20847	5157
97	Pentium II (300 MHz)	600	735	25070	8945

10. Conclusions

Much of the almost hyperactivity within the optical fiber telecommunications industry over the past few years can be traced to the perceived and real growth of the traffic on the Internet. We maintain that the overall growth rate of the Internet for most of its existence (despite some excursions) was remarkably close to "doubling every year", and we anticipate that this rate will continue into the foreseeable future. In effect we see a type of Moore's law associated with the growth of data traffic. This type of growth rate is in sharp contrast to the historical growth rates of various methods of communications (including conventional mail, telegraph service, and traditional voice phone service) that tended to be no greater (and typically much less) than about 10% per year. Still, even though a doubling each year represents very fast growth, it is only comparable to the rate of progress in transmission capacity. Hence

we are unlikely to see the huge increases in spending on optical communication that many business plans had been based on.

Throughout the history of the Internet there have only been two "killer applications": email and the Web (including Web browsers). Several events conspired which allowed an unprecedented explosion (roughly 100 fold increase) in Internet traffic in the 1995-1996 time frame, and the Internet was able to handle this since it made use of the existing telephone industry infrastructure. Since the Internet is quickly approaching the point at which it is the predominant network it is very unlikely that such huge growth rates could be so easily supported in the future.

It also appears that, aside from short-range perturbations, there will be neither a "bandwidth glut" nor a "bandwidth shortage" going into the foreseeable future, in that supply and demand will be growing at comparable rates. As such it is very likely that pricing will begin to play an even more important role in the evolution of traffic. Throughout most of the 1990s data transmission prices were increasing. However, there are recent signs that they are beginning to decrease, and in some cases, especially across the Atlantic and on major trans-continental routes in the U.S., they have decreased dramatically. If they begin to decrease rapidly in general, then many of the constraints on usage that exist today may very likely start to ease. We are likely to see capacity growing somewhat faster than traffic, a continuation of the trend we have already seen in the last few years.

We also believe that "file" transfers and not real time streaming will remain dominant on the network. Streaming real time transmissions will undoubtedly grow in absolute terms, and as a fraction of the total traffic it may increase for a while. However, in all likelihood it will eventually begin to decline as the demand for this type of traffic will not be growing as fast as network capacity. We foresee sharing applications as a likely candidate to fuel traffic growth. One of the first major examples of this was Napster since it effectively turned consumers of information into providers of information. It is extremely likely that such file sharing applications will be some of the key applications that continue to fuel the annual doubling of data traffic.

References

[Abbate] J. Abbate, *Inventing the Internet*, MIT Press, 1999.

[Baran] P. Baran, On distributed Communications Network, IEEE Trans. Comm. Systems, March 1964

[Boardwatch] *Boardwatch* magazine, ⟨http://www.boardwatch.com⟩.

[Bruno] L. Bruno, Fiber optimism: Nortel, Lucent, and Cisco are battling to win the high-stakes fiber-optics game, *Red Herring*, June 2000. Available at ⟨http://www.herring.com/mag/issue79/mag-fiber-79.html⟩.

[Cerf] V. G. Cerf, A brief history of the Internet and related networks, ⟨http://www.isoc.org/internet/history/cerf.html⟩.

[CerfK] V. G. Cerf and R. E. Kahn, A protocol for packet network interconnection, *IEEE Trans. Comm. Tech.*, vol. COM-22, pp. 627-641, May 1974.

[Cochrane] N. Cochrane, We're insatiable: Now it's 20 million million bytes a day, *Melbourne Age*, Jan. 15, 2001. Available at ⟨http://www.it.fairfax.com.au/networking/20010115/A13694-2001Jan15.html⟩.

[CoffmanO1] K. G. Coffman and A. M. Odlyzko, The size and growth rate of the Internet. *First Monday*, Oct. 1998, ⟨http://firstmonday.org/⟩. Also available at ⟨http://www.research.att.com/∼amo⟩.

[CoffmanO2] K. G. Coffman and A. M. Odlyzko, Internet growth: Is there a "Moore's Law" for data traffic?, *Handbook of Massive Data Sets*, J. Abello, P. M. Pardalos, and M. G. C. Resende, eds., Kluwer, 2001, to appear. Available at ⟨http://www.research.att.com/∼amo⟩.

[CTIA] CTIA (Cellular Telecommunications Industry Association), Semi-Annual Wireless Industry Survey, June 1985 to June 2000. Available at ⟨http://www.wow-com.com/wirelesssurvey/⟩.

[Cyberspace] Geography of Cyberspace Directory: Internet Traffic and Demographic Statistics, available at ⟨http://www.cybergeography.org/statistics.html⟩.

[deSolaPITH] I. de Sola Pool, H. Inose, N. Takasaki, and R. Hurwitz, *Communications Flows: A Census in the United States and Japan*, North-Holland, 1984.

[DunnL] D. A. Dunn and A. J. Lipinski, Economic considerations in computer-communication systems, pp. 371-422 in *Computer-Communication Networks*, N. Abramson and F. F. Kuo, eds., Prentice-Hall, 1973.

[Economist] Not Moore's Law, *The Economist*, July 12, 1997.

[ElderingSE] C. A. Eldering, M. L. Sylla, and J. A. Eisenach, Is there a Moore's Law for bandwidth?, *IEEE Communications Magazine*, Oct. 1999, pp. 2–7.

[FishburnO] P. C. Fishburn and A. M. Odlyzko, Dynamic behavior of differential pricing and Quality of Service options for the Internet, pp. 128–139 in *Proc. First Intern. Conf. on Information and Computation Economies (ICE-98)*, ACM Press, 1998. Available at ⟨http://www.research.att.com/∼amo⟩.

[FloydP] S. Floyd and V. Paxson, Difficulties in simulating the Internet, *IEEE/ACM Transactions on Networking*, to appear. Available at ⟨http://www.aciri.org/floyd/papers.html⟩.

[Galbi] D. Galbi, Bandwidth use and pricing trends in the U.S., *Telecommunications Policy*, vol. 24, no. 11 (Dec. 2000). Available at ⟨http://www.galbithink.org⟩.

[GrayS] J. Gray and P. Shenoy, Rules of thumb in data engineering, Proc. 2000 IEEE Intern. Conf. Data Engineering. Also available at ⟨http://research.microsoft.com/∼gray⟩.

[Green] E. E. Green, Communications spectra by the wholesage–2012 A.D., *Proc. IRE*, vol. 50 (1962), pp. 585-587. Reprinted in *Proc. IEEE*, vol. 87 (1999), 1293-1295.

[Harms] J. Harms, From SWITCH to SWITCH* - extrapolating from a case study, *Proc. INET'94*, pp. 341-1 to 341-6, available at ⟨http://info.isoc.org/isoc/whatis/conferences/inet/94/papers/index.html⟩.

[HennessyP] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, Morgan Kaufmann, 1990.

[Hobbes] Hobbes Internet Timeline, ⟨http://www.zakon.org/robert/internet/timeline/⟩.

[Hough] R. W. Hough, Future data traffic volume, *IEEE Computer*, (Sept.-Oct. 1970).

[Kleinrock1] L. Kleinrock, Information flow in large communications networks, *RLE Quartely Progress Report*, July 1961.

[Kleinrock2] L. Kleinrock, *Communication Nets; stochastic message flow and delay*, McGraw-Hill, 1964.

[Kleinrock3] L. Kleinrock, ISDN–The path to broadband networks, *Proc. IEEE*, vol. 79 (1991), 112-117.

[Leiner] B. M. Leiner, V. G. Cerf, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. Wolff, A brief history of the Internet, Version 3.31, Aug. 4, 2000. Available at ⟨http://www.isoc.org/internet/history/brief.html⟩.

[Lesk] M. Lesk, How much information is there in the world?, 1997 unpublished paper, available at ⟨http://www.lesk.com/mlesk/diglib.html⟩.

[Licklider] J. C. R. Licklider, *Libraries of the Future*, MIT Press, 1965.

[LickC] J.C.R. Licklider and W. Clark, On-Line Man Computer Communications, August 1962.

[Lucky1] R. W. Lucky, New communications services–What does society want?, *Proc. IEEE*, vol. 85 (1997), pp. 1536-1543.

[Lucky2] R. W. Lucky, Through a glass darkly–Viewing communications in 2012 from 1961, *Proc. IEEE*, vol. 87 (1999), 1296-1300.

[McCredie] J.McCrddie, UC Berkeley must manage campus network growth, The Daily Californian, March 14, 2000. Available at (http://www.dailycal.org/article.asp?id=1912&ref=news)

[MeekerMJ] M. Meeker, M. Mahaney, and D. Joseph, The Internet user/usage ecosystem framework, Morgan Stanley Dean Witter report, Jan. 24, 2001. Available at ⟨http://www.msdw.com/techresearch/index.html⟩.

[MRTG] The Multi Router Traffic Grapher of Tobias Oetiker and Dave Rand, information and links to sites using it at ⟨http://ee-staff.ethz.ch/∼oetiker/webtools/mrtg/mrtg.html⟩.

[NII] U. S. Information Infrastructure Task Force, reports on *The National Information Infrastructure*, available at ⟨http://www.ibiblio.org/nii/toc.html⟩.

[Noll1] A. M. Noll, *Introduction to Telephones and Telephone Traffic*, 2nd ed., Artech House, 1991.

[Noll2] A. M. Noll, *Highway of Dreams: A Critical Appraisal of the Communications Superhighway*, Lawrence Erlbaum Associates, 1997.

[Noll3] A. M. Noll, Does data traffic exceed voice traffic?, *Comm. ACM*, June 1999, pp. 121-124.

[Nua] Nua Internet Surveys, available at ⟨http://www.nua.com⟩.

[Odlyzko1] A.M. Odlyzko, Data networks are lightly utilized, and will stay that way. Available at ⟨http://www.research.att.com/∼amo⟩.

[Odlyzko2] A.M. Odlyzko, The history of communications and its implications for the Intenet, available at ⟨http://www.research.att.com/∼amo⟩.

[Odlyzko3] A.M. Odlyzko, Internet pricing and the history of communications, *Computer Networks*, vol. 36 (2001), pp. 493-517. Also available at ⟨http://www.research.att.com/∼amo⟩.

[Polly] J. A. Polly, Surfing the Internet: An introduction, *Wilson Library Bulletin*, June 1992, pp. 38-42. Available at ⟨http://www.netmom.com/about/surfing.shtml⟩.

[ReichlLS] P. Reichl, S. Leinen, and B. Stiller, A practical review of pricing and cost recovery for Internet services, to appear in Proc. 2nd Internet Economics Workshop Berlin (IEW'99), Berlin, Germany, May 28-29, 1999. Available at ⟨http://www.tik.ee.ethz.ch/∼cati/⟩.

[Schaller] R. R. Schaller, Moore's law: Past, present, and future, *IEEE Spectrum*, vol. 34, no. 6, June 1997, pp. 52-59. Available through Spectrum online search at ⟨http://www.spectrum.ieee.org⟩.

[StArnaud] B. St. Arnaud, The future of the Internet is NOT multimedia, *Network World*, Nov. 1997. Available at ⟨http://www.canarie.ca/∼bstarn/publications.html⟩.

[StArnaudCFM] B. St. Arnaud, J. Coulter, J. Fitchett, and S. Mokbel, Architectural and engineering issues for building an optical Internet. Short version in *Proc. Soc. Optical Engineering*, (1998). Full version available at ⟨http://www.canet3.net⟩.

[Standage] T. Standage, *The Victorian Internet: The Remarkable Story of the Telegraph and the Nineteenth Century's On-line Pioneers,* Walker, 1998.

[Taggart] S. Taggart, Telstra: The prices fight, *Wired News,* ⟨http://www.wired.com/news/politics/0,1283,32961,00.html⟩.

[WalkerM] P. M. Walker and S. L. Mathison, Regulatory policy and future data transmission services, pp. 295-370 in *Computer-Communication Networks,* N. Abramson and F. F. Kuo, eds., Prentice-Hall, 1973.

[WuL] W. W. Wu and A. Livne, ISDN: A snapshot, *Proc. IEEE,* vol. 79 (1991), pp. 103-111.