Data networks are lightly utilized, and will stay that way

Andrew Odlyzko

AT&T Labs - Research
amo@research.att.com

October 7, 1998.

Abstract. The popular press often extolls packet networks as much more efficient than switched voice networks in utilizing transmission lines. This impression is reinforced by the delays experienced on the Internet and the famous graphs for traffic patterns through the major exchange points on the Internet, which suggest that networks are running at full capacity. This paper shows the popular impression is incorrect; data networks are very lightly utilized compared to the telephone network. Even the backbones of the Internet are run at lower fractions (10% to 15%) of their capacity than the switched voice network (which operates at over 30% of capacity on average). Private line networks are utilized far less intensively (at 3% to 5%). Further, this situation is likely to persist. The low utilization of data networks compared to voice phone networks is not a symptom of waste. It comes from different patterns of use, lumpy capacity of transmission facilities, and the high growth rate of the industry.

1. Introduction

Announcements of new packet networks often lead to news stories claiming IP (Internet Protocol) networks are faster and less expensive than traditional circuit-switched networks (cf. [Keller]). Usually no explanation is offered for this claimed advantage of packet transmission. More technical presentations explain that old-style phone networks reserve two circuits (one in each direction) for a phone call, even though almost all the time only one person is speaking, and that there are frequent pauses during conversations when nothing is being transmitted. In contrast, packet networks transmit data only when there is something to send, and thus it is plausible that they would use transmission capacity more efficiently. Vint Cerf, one of the "fathers of the Internet," made the following comparison of packet versus circuit switching (in the "Telecom Italia" presentation at [Cerf]):

> Circuit (telephony) like reserving bicycle lanes from LA to NY!
>
> Packet (Internet) like sharing of the highway among high speed cars.

That is an appealing analogy. However, it conceals a much more complicated picture. It appears that today most companies are paying more for large file transfers over their private IP-based networks than they would if they used modems over the public switched voice network. This is not an argument for circuit-switched networks over packet ones, since there are other compelling arguments in favor of IP networks (see the companion papers [Odlyzko1, Odlyzko2]). However, it does suggest the need for a more careful investigation of just how data networks are used.

This paper studies average utilization levels of transmission lines in data networks, where the averages are over a full week. Surprisingly, although there is a huge literature on networks, such averages appear to have been little studied, although they are critical to understanding the economics of data networks. One minor reason for concentrating on transmission is that it is the easiest to measure, since switching or routing capacity is notoriously hard to quantify. A much more important reason is that transmission is the most expensive part in a data network. (We concentrate on long distance transport only, and so do not take into account local networking costs, such as those of modems for residential customers of ISPs, which are the bulk of the total cost of Internet services such as America Online.) Typical corporate inter-LAN networks appear to spend around 45% of their operating expenses on transmission, 20% on equipment (depreciation and maintenance) and 35% on people. One regional ISP reports spending 55% of operating funds on transmission and 15% on equipment. Similar estimates that show the dominant role of transmission costs can be found in the cost model for ISPs developed by Leida [Leida]. If data networks are intensively utilized, then we should find transmission lines run at high fractions of their capacity. That is certainly a widespread view.

The impression that packet networks have high utilization levels of transmission and switching facilities is reinforced by the delays observed on the Internet (the "World Wide Wait") and the widely publicized data on usage patterns. Figure 1 (based on Fig. 1.13 of [Ash]) shows the traffic on the U.S. switched voice networks over a two-day period. It is peaked, as folks in Peoria do not like to call their friends or business partners in Poughkeepsie at 3 am. Thus there are long periods when that network is largely idle. During the two days shown in Fig. 1, the average traffic was about 40% of the peak. On the other hand, Fig. 2 shows traffic through the PacBell NAP (Network Access Point), a major exchange point on the Internet, during October 26 and 27, 1997. (Additional data for this NAP, as well as other exchanges, is available through links provided at [CAIDA, NLANR].) This NAP was running full blast almost around the clock. As a fraction of the peak rate observed during those two days, the average throughput was 84% on Monday and 80% on Sunday.

The perception of intensive use even of corporate networks is reflected in frequently heard comments about 70% utilization levels of private lines. These comments are often made without qualification, as if they reflected long-term averages. More experienced people make more precise statements.

For example, Fred Baker of Cisco reports (private communication) that “corporate customers commonly claim their inter-site WAN links are used at 70% of capacity during peak periods.” Brett Leida [Leida] has a model for the load on a typical T1 line from a corporate customer to the Internet which has the peak period load at 70% for several hours each business day, and average load of 34%. Leida obtained his information from members of the MIT Internet Telephony Consortium, which includes many established communications industry players.

This paper presents extensive evidence that average utilization levels are far lower than generally supposed. While the long distance circuit switched voice network has average utilization of about 33%, the Internet backbone links appear to have average utilizations closer to 10% to 15%, and corporate long-haul links (which is where the bulk of data transport capacity is) have utilizations in the 3% to 5% range. A better analogy than Vint Cerf's might be:

> Circuit (telephony) like a lane from LA to NY that is full of well-behaved bicyclists.
>
> Packet (Internet) like sharing of the highway among high speed cars, but with frequent construction detours.
>
> Packet (corporate Intranet) like sharing of a 100-lane highway among a few high speed cars.

At first sight it seems that it should be simple to determine average utilization levels. That is not so, though, since, for privacy reasons, carriers such as AT&T do not monitor how the private lines they lease to customers are used. Individual customers in many cases do not measure their own usage. When they do measure it, they often do not obtain average utilization levels. Even those statistics that are collected are usually regarded as confidential. Thus it is hard to obtain solid estimates, and it is necessary to resort to limited sampling and circumstantial evidence.

The corporate managers who report 70% utilization levels are correct. Their networks do generate such figures, but they are usually misinterpreted. Given the way statistics are collected in many systems, the 70% figure may not even refer to the busy hour over a week, but the busiest 5 minutes over a period of months. Further, it typically applies to only a few links in a system.

Of the various people that I have talked to, the ones who accepted my claims of low utilization levels most readily were designers of private line networks. They are not used to considering utilization rates averaged over a full week. However, once I explained to them that this is what I was after, they typically did a quick mental calculation and said "Of course, this is obvious because of [factors that will be discussed in Section 8 of this paper]. However, such long-term averages are irrelevant."

Low average utilization levels are indeed irrelevant to designers of private line networks. These designers have to provide levels of service specified by their customers at minimal cost, and long-run averages do not matter to them. However, as is shown in Section 9 below and in the companion papers [Odlyzko1, Odlyzko2], average utilization rates are important for understanding such important questions as the profitability of the ISP business, the prospects for packet telephony, and general evolution of data networks, in particular prospects for Quality of Service.

As a sample of the kinds of arguments that can be based on the data in this paper, consider Fig. 4, which shows the usage profile for a corporate 128 Kbps dedicated connection to the Internet. This business uses the Internet both for general connectivity, and also to transmit data between different locations. The average utilization is about 3%, fairly typical for such links. This business could clearly receive all its data on a 56 Kbps link at a cost of suffering delays of at most minutes, and possibly only seconds, in its communications. If email were all that was being transmitted, that would surely be acceptable, and a 56 Kbps link is all that would be in place. That this business pays for a 128 Kbps connection shows that it values the ability to occasionally send or receive data at high rates. The high speed bursts are extremely infrequent, though, and seldom do several collide to saturate the link. Therefore Quality of Service measures would not be of much help. Further, even if 90% of the traffic on this link were frivolous personal usage (stock quotes, cartoons, and so on), banning it would not provide significantly better performance for the high priority applications that justify the cost of the link. When the high priority traffic starts up, it almost always gets the full bandwidth of the link in any case. Note that these arguments would not apply if the link were routinely used at 70% of capacity during business hours, as is commonly believed. In heavy utilization conditions, either Quality of Service measures or banning non-essential traffic would provide better service for the mission-critical applications. That average utilizations on data networks are low shows what kind of connections are desired by customers, and how highly they are valued. In particular, the low utilization rates do throw serious doubt on the advisability of many Quality of Service approaches [Odlyzko1, Odlyzko2], which appear to be motivated by the assumption that networks are heavily congested.

Low utilization rates lead to great opportunities for higher quality or less expensive service from aggregation of traffic. If two business customers have 128 Kbps lines that are used at 70% of capacity during the peak business hours, relatively little can be gained by combining their traffic streams. One would still need 256 Kbps of capacity. On the other hand, if they both behave like the business of Fig. 4, aggregating their traffic on a 192 Kbps circuit would give each one the perception of having a dedicated 192 Kbps link. On larger scales, with more customers involved, the benefits are much greater, and they underlie the economics of public networks.

Section 8 presents quantitative analyses of the reasons for low average utilization rates of data networks, and argues that such rates will persist. The companion paper [Odlyzko1] suggests some ways to increase those utilization rates to some extent. However, it is unlikely that data network utilization rates will ever approach those of the switched voice network. The key point is that low utilization may be technologically inefficient, but it may often be economically efficient when the total system cost is considered. If a newspaper doubles the capacity of the private line between its editorial offices and the printing plant, the utilization rate will drop in half. However, the staff may gain an extra half hour to work on the edition before it goes to press, the half hour that is cut from the transmission time of the electronic layout. Whether that is worthwhile or not has to be decided by the managers of the business, and the utilization rate is irrelevant. When we see companies routinely paying for lightly utilized networks, we can conclude that they do value the ability to send data in high speed bursts, and that should guide us in the design and operation of networks.

This paper documents the low utilization levels of data networks mentioned above (and summarized in Table 1). It is likely that some people in the communications industry understand this already. For example, given the aggregate size of private line networks (see [CoffmanO]), the only way that the MCI prediction (see Vint Cerf's presentations at [Cerf]) of data traffic overtaking voice traffic around the year 2002 can be correct is if private line networks are extremely lightly utilized.

Sections 2 and 3 discuss what networks are to be measured, and the units of measurement. Section 4 presents data about switched voice networks, to serve as a benchmark in comparing various data networks. Section 5 discusses the backbones of the public Internet (i.e., those backbones that are accessible to general users). Section 6 presents data about some research networks. Section 7 is devoted to evidence about utilization of private line networks. Section 8 discusses the reasons that data networks are likely to stay underutilized. Finally, Section 9 closes with some comments and conclusions.

2. What is to be measured, and why

The focus of this paper is on long-term average utilization of long-haul lines in the data and voice networks, the DS0, T1, T3, OC3, and similar lines that customers such as ISPs lease from telecommunications carriers. (Some carriers, such as AT&T, MCI, and WorldCom, both own such lines and use them to offer Internet services to their own customers, and also lease such lines to other carriers.)

Corporations building private line networks and the majority of ISPs depend on such leased lines, and it is the economics of this business that I wish to explore. I will not deal with the utilization of the fiber network that is used to provide these provisioned T1, T3, and other circuits (a fascinating subject in its own right).

I will consider only U.S. data networks, although there will be some data about international links and institutions. The U.S. not only accounts for more than half of the traffic, but it also has much lower transmission costs [ITU, GMLCOBRS]. Therefore its data network behavior is likely to foreshadow what will be seen in other countries in the near future, as they expand their telecommunications infrastructure and reduce prices.

Only long distance links will be considered. For the voice phone network this will mean not looking at utilization of access links, such as the copper wire from a house to the nearest central office or the links from the central office to long distance switches. For data networks, LANs (Local Area Networks) will also not be considered in detail. They are an important part of the picture, and are discussed at some length in [Odlyzko1], but in this paper they will be mentioned only briefly.

The main reason for not considering local links is that their utilization patterns differ substantially from those of long-haul facilities. It is widely recognized that LAN utilization is extremely low. Few people appreciate just how low it is. There are no comprehensive statistics, but we will cite as one example the University of Toronto network [Toronto]. The main reason for selecting this academic institution is that its network is unusually well instrumented, with statistics collected for all important segments, and displayed with the MRTG program of Oetiker and Rand [MRTG]. Toronto is not profligate with network resources, as its Internet link is unusually congested (as will be discussed later), and so are many of its internal WAN links. Still, the average utilization of its 173 Ethernets, during the week ending at 4 pm on Sunday, March 8, 1998, was 1.1%. Only 24 Ethernets had average utilization levels over 2% during that week. Graphs do show occasional spikes in usage (the reason for having all that bandwidth), but they tend to be short. Even if we take the maximal utilization level for each Ethernet during any 30-minute period over that week, and average it over the 173 Ethernets, we find it is only 8.7%.

The graphs of network usage that are included in this paper are typically for Sunday and Monday. The reason is to show the different time of day and day of the week patterns of traffic loads on various networks. The implications of the similarities and differences in such patterns are be explored at greater length in [Odlyzko1, Odlyzko2].

3. Conversion factors

It will be convenient to state some conversion factors between different units and between the bandwidth of a connection and the traffic carried by that connection.

Voice on the phone network is carried in digitized form at 64,000 bits per second. We will be using the computer industry notation in which Kbps = kilobit per second, 1024 bits per second. To keep the presentation simple, we will say that each channel takes 64 Kbps. The inaccuracy this will introduce is minor.

Each voice call occupies two channels, one in each direction, so takes up 128 Kbps of network bandwidth. Thus one minute of a voice call takes 60*128*1024 bits, or 960 KB (kilobytes). Rounding this off, we get

1 minute of switched voice traffic $\overset{=}{\sim}$ 1 MB.

(Compression can reduce that to a much smaller figure, and is used to some extent on high-cost international circuits, as well as on some corporate private line networks. As far as the network is concerned, though, it is carrying 1 MB of digital data for each minute of a voice call.)

A T3 (or DS3) line operates at 45 Mbps in each direction, so that if it were fully loaded, it would carry 90 Mbps. Over a full month of 30 days, that comes to 29 TB (terabytes, 10^{12} bytes). We will say that

full capacity of a T3 link $\overset{=}{\sim}$ 30 TB/month.

A T1 line (1.5 Mbps) is 1/28-th of a T3, and we will say that

full capacity of a T1 link $\overset{=}{\sim}$ 1 TB/month.

4. Switched voice networks

It is interesting to not only estimate utilization levels of various data networks, but also to compare them with the circuit switched network. The book [Keshav] is an excellent source that contrasts the technologies involved in these types of networks. However, no comprehensive description of how they are used appears to exist.

Figure 1 shows the typical traffic pattern on U.S. switched voice networks. It is derived from Fig. 1.13 of [Ash]. This graph aggregates all the phone calls over the four time zones of the continental U.S., as well as the comparatively small number of calls to Hawaii, Alaska, and other places. (For more data, including calling patterns in smaller regions, see [Ash].) Voice networks, such as that of AT&T, are engineered to provide a low-cost solution to all normal demands. This means that many calls may get blocked in cases of an earthquake, say, but even peak hour demands during the busiest days, such as Mother's Day or the Monday after Thanksgiving, are accommodated. For example, to cite a small sample of the data in [Ash], on Monday, Dec. 2, 1991, which was the busiest day for the AT&T network until then, of 157.5 million calls, only 228 were blocked on intercity connections. In spite of this, the average utilization of long distance links in the switched voice network is close to 33%, as is explained in [CoffmanO], based on data from [Ash]. This efficiency comes from careful engineering (using techniques such as RTNR, Real Time Network Routing [Ash], that route calls between New York City and Philadelphia through Chicago when spare capacity is available on those routes), from the smoother and more predictable nature of voice traffic in general, and the predictable growth in demand for voice services. An important contributor to the high average utilization of voice networks is the sharing of this network among several classes of users with different calling patters, a point explored at greater length in [Odlyzko1, Odlyzko2].

Average utilizations are far lower if one considers the entire telecommunicationis network. There are extensive circuits that exist to provide service in case of fiber cuts and similar outages. These circuits have large capacity, but they are used to protect data circuits as well as voice lines, and are outside the scope of this paper.

5. The public Internet

The Internet is slow, as anyone who surfs the Web can attest. However, it has proved impossible so far to produce a simple description of where the problems lie. (For the most thorough statistical study of Internet performance currently available, see [Paxson].) Many of the problems are with the servers. However, the general impression is that the backbones are seriously congested. This view is supported by studies of comparative backbone performance, which do show substantial differences in performance among different ISPs. This view is also strengthened by data such as that of Fig. 2, showing traffic through a major public exchange point on the Internet. The flat service profile seen there is characteristic of demand exceeding supply. Similar flat service profiles are seen in the data for other public exchange points (available through [CAIDA, NLANR]), as well as for some other congested networks (see Fig. 8 later, which shows saturation on the link from the U.S. to Switzerland between 9 in the morning and 7 in the evening, Swiss time). There are reports of packet loss rates of over 30% during peak periods when transiting the NAPs and MAEs, although there is disagreements as to whether these losses are caused by packets being dropped at these transit points, or delays at those points causing timeouts in various TCP implementations. Traffic patterns on large backbone links appear to follow the same flat pattern suggestive of saturation, as is shown in Fig. 3 (based on data from [ThompsonMW]).

While the data cited above does suggest extreme congestion, some of it raises questions. For example, the traffic profile on the MCI OC3 link of Fig. 3 is flat, but the average traffic (averaged over the full week of August 24 to 30, 1997, including data not shown in Fig. 3, but presented in [ThompsonMW]) is 30.0 Mbps in one direction and 32.7 in the other. Since an OC3 has capacity of 155 Mbps in each direction, the average utilization of this link is only 20%! Even if one looks at the 5-minute averages, the highest seen on this OC3 link during the week covered by [ThompsonMW] is 60.3 Mbps, less than 40% of capacity. (For the trans-Atlantic T3 link in [ThompsonMW], average utilization is about 42% for the U.K to U.S. direction, and 56% the other way, with many 5-minute averages showing saturation of the eastward link.)

Kerry Coffman and I have studied the publicly available information about Internet backbones [CoffmanO]. Our estimate was that at the end of 1997, the traffic through these backbones totaled between 2,500 and 4,000 TB/month, and that the effective bandwidth was around 75 Gbps, which gives average utilization of between 10% and 16%. (Effective bandwidth was computed by adding up the capacity of the backbone links, which came to 2,100 T3 equivalents, and dividing by 2.5, to account for a typical packet traversing 2.5 backbone links between source and destination.)

There are many uncertainties about the estimates in [CoffmanO]. However, they appear to be in the right range, based on feedback from various sources in the industry. They also appear to fit estimates made for some networks separately. For MCI, their publicly declared traffic of 170 TB/week at the end of 1997, together with the estimate of a backbone of about 400 T3 equivalents, produces an average utilization estimate of 15% (again assuming 2.5 backbone hops per packet). (The MCI Internet transport is provided by their ATM network, so I am taking some liberties in interpreting it as if it were a routed network of point-to-point circuits.)

From the data in [CoffmanO], it appears that average utilization of Internet backbones has decreased between the middle of 1996 and the end of 1997. This is consistent with reports that national backbones have become less of a problem and are providing high quality service on their networks with low latencies, low jitter, and low packet loss rates. This may have been a result of ISPs deliberately trying to provide better service. They may also have overestimated traffic demand and overbuilt their networks, since Internet traffic grew much faster in 1996 than in 1997 [CoffmanO]. They may also be preparing for much greater traffic in the near future, with the development of new applications such as packet telephony, and a large scale shift of private line traffic to the Internet. Some of this buildup may also be caused by many more ISPs building national backbones, and moving to high speed links in order to meet competitive pressures.

6. Research networks

The previous section discussed the public Internet, namely those parts of the Internet accessible to general users. We next look at a mixed case, namely the Internet as it was transitioning from a research network to a commercial enterprise, and then at some past and current research networks.

NSFNet provided the Internet backbone until the phasing out of that program in April 1995. Hearsay suggests, but there do not seem to be any firm statistics to substantiate this, that through the end of 1994 NSFNet was carrying almost all of the non-military backbone traffic. (Carriers such as UUNet, PSINet, and BBN started to build new private backbones and expand existing ones at that time. There were also restricted research networks in existence then, but they appear to have been smaller, and access to them was much more restricted than to NSFNet, which was already carrying much commercial traffic.) Statistics on NSFNet's performance are available at [NSFNet]. They show that at the end of 1994, the 19 T3s in the NSFNet backbone were operating at about 5% average utilization. The T3s replaced T1s completely by the end of 1992, and given the 100% annual growth rates of NSFNet traffic, they must have been utilized at about 1% of capacity initially.

A more representative view of NSFNet's operation is probably that presented in the study [ClaffyPB], based on the NSFNet's T1 backbone in May 1992. This appears to be the only careful study of utilization patterns on NSFNet (and the only study of this kind since the work of Kleinrock and Naylor [KleinrockN] on ARPANet, the precursor of NSFNet, two decades earlier). The average utilization rate of all the T1s was 15.5% during the week of May 10-17, 1992. The maximum 15-minute average load on the entire T1 network was 27.1%. Considering single links separately, the highest weekly average utilization rate was 35%, and the highest 15-minute average load was 89%.

The [ClaffyPB] study was carried out on the T1 network while NSFNet was transitioning from T1s to T3s. The statistics in [ClaffyPB], when compared to those for the entire NSFNet at [NSFNet] show that the T1s carried about a third of the NSFNet backbone traffic in May 1992. Given the growth in traffic on NSFNet, it appears that the load on just the T1s in May 1992 was comparable to that on the entire NSFNet towards the end of 1990, which is when the whole network consisted just of T1s. Thus it seems that an average utilization rate of around 15% was regarded as tolerable, but that higher rates would have produced inadequate performance in that environment.

Finally, we consider a modern experimental network. When NSFNet was privatized in 1995, NSF established the vBNS network for research projects in high performance communications. It appears to have the largest capacity among research networks, with OC12 bandwidth on most connections, and total bandwidth of all links around 250 T3 equivalents. In comparison, there were about 2,100 T3 equivalents in all the commercial Internet backbones at the end of 1997, while the NSFNet backbone had only 19 T3s in 1994, and several corporate private line networks have over 20 T3s today. vBNS does provide excellent performance, with round trip times between East and West coasts of 70 milliseconds. That latency is sufficient for all voice and video applications that are being developed, provided it can be obtained on a sustained basis. (The speed of light through fiber puts a lower bound of 40 milliseconds on such round trip times. Thus there is little point in dreaming up applications that require smaller latencies. Laws of nature have to be obeyed!) vBNS appears to provide such latency consistently. On many days, the maximal round trip time recorded is under 100 milliseconds. (For details on testing and performance of vBNS, see the paper [MillerTW] and the statistics on the Web page [VBNS].) What vBNS has not established yet is whether the excellent performance it provides can be scaled to larger and more heavily utilized networks. (The traffic on vBNS is not typical of the public Internet, and in particular has many fewer distinct flows, which helps the underlying ATM network provide good service.)

vBNS is lighty utilized, although traffic is growing, with more institutions joining. All traffic goes through the ATM interfaces to Cisco routers, which in early 1998 were all of OC3 speeds, 155 Mbps. During the week ending on May 10, 1998, the highest weekly average utilization was in Chicago (12.8% incoming and 24.0% outgoing). The average over the 16 interfaces was 4.5% for incoming and 5.6% for outgoing traffic. (On vBNS, as well as on other networks, incoming and outgoing traffic volumes do not have to be equal, since multicasting is a large factor.) Since these are OC3 interfaces to an OC12 network, it appears that if the average packet took the equivalent of two hops on the backbone (this is a bit of stretch, first because of multicasting, and second because vBNS traffic is carried by the MCI ATM network, but we can imagine how the network would run if it went through routers), then the average utilization rate of the links was under 3%.

7. Private line networks

Little has been published about utilization of private lines, even though they form the bulk of the long distance data networking "cloud," as is shown in [CoffmanO]. Existing sources that do mention utilization rates explicitly tend to claim that these rates are high. For example, as was mentioned in the Introduction, [Leida] estimates that dedicated business connections to the Internet are run at 34% of capacity. Some passages in [TeleGeography] imply that at least for international private lines, utilization is very high. On the other hand, there are also some indications that corporate data networks are lightly utilized. For example, the article [Roberts] reports that the network of GMAC Mortgages had less than 5% utilization even during peak periods (although this was supposed to be a temporary condition). Several other articles in magazines such as *Data Communications* or *Network Computing* mention successful implementations of IP telephony over private line or Frame Relay networks that were lightly utilized. Thus it appears from these publications that uncongested networks might not be uncommon. This section shows that uncongested networks are not only uncommon, but are the rule.

Most of the evidence for low utilization of data networks that I have collected has come from network managers that wish to identify neither themselves nor their employers. The main exception is Bill Woodcock of Zocalo, a regional ISP based on the West Coast, who provided extensive statistics on dedicated business lines to the Zocalo network for several months in the fall of 1997. Table 2 shows the utilization rates for all such lines coming in to one particular Zocalo Point of Presence (PoP) in Northern California during the week ending November 29, 1997. (To protect the privacy of Zocalo customers and also Zocalo's competitive position, the exact location is not disclosed.) The bandwidth-weighted average utilizations for the lines in Table 2 are 1.6% for receive and 1.2% for the transmit sides.

It is very hard for a single set of statistics, such as that of Table 2, to represent fairly the complicated picture of private line utilization. Zocalo data, as well as data from other service providers, shows that there is one class of customers who consistenly use their dedicated Internet access lines at high rates, namely ISPs. By aggregating traffic from many sources, they can obtain much higher average utilizations. Dial ISPs (those which service residential dial-up customers) sometimes also overload their lines, when they do not worry about providing high quality of service. Table 2 contains data for just one dial ISP line (the last entry, with the heaviest T1 usage in this collection), and this particular customer has an unusual configuration that leads to erratic usage patterns, typically heavier than that for the week covered by that table. Fig. 6 shows the traffic pattern from another dial ISP with a 768 Kbps line. That ISP has average utilization rate of about 40%.

Additional data from Zocalo and other ISPs suggests that average utilizations for dedicated business connections to the Internet are higher than those of Table 2 (even if one excludes ISP customers), closer to 3% over a full week. However, there is tremendous variation.

In addition to ISPs, there is at least one other class of customers who often have reasonably high utilization rates, namely academic institutions. As an extreme example, consider Fig. 8. It shows the traffic pattern on the trans-Atlantic 8 Mbps link from the SWITCH network that serves Swiss academic and research institutions [Harms, SWITCH]. The direction from the U.S. to Switzerland is saturated for many hours each day, and weekly utilization in March 1998 was about 50% for that side of the link, and 20% for the reverse direction. (Other SWITCH links, to European networks, are much less congested, presumably reflecting lower costs. See [SWITCH].) Fig. 9 shows the traffic pattern for the University of Toronto connection to the Internet, which in February 1998 had weekly average utilizations of 57% for the receive and 45% for the transmit side. Such high utilizations in academic settings, which are experienced by large populations of students and faculty, and which are also much more readily available than corporate traffic statistics, may be contributing to the widespread impression of general heavy utilization of private line networks. However, even in academia there are many examples of low utilizations (even aside from experimental networks like vBNS, discussed in Section 6). For example, Fig. 10 shows the traffic pattern on the T3 link to the Internet from Columbia University. There the average weekly utilization is about 11% for the receive and about 9% for the transmit side. A similar picture can be seen in the Princeton University statistics at [Princeton], whose two Internet links with aggregate capacity of 31 Mbps had average utilizations in May 1998 of 13.4% on the incoming sides and 6.2% on the outgoing sides. (It is worth emphasizing once more that low utilization rates are not necessarily a symptom of waste. Given the pricing schedules for Internet access, it may very well be less expensive for Columbia to have a lightly utilized T3 than several heavily loaded T1s. This point is covered more extensively in Section 8.)

The relatively flat usage patterns of academic institutions such as those in figures 9 and 10 may also be contributing to the impression that such patterns predominate among all data networks. However, most corporate networks show patterns such as those of figures 4 and 7, with most of the traffic concentrated during the business day. Even the SWITCH network of Fig. 8 shows this pattern, either because Swiss students and faculty have different habits than North American ones, or else because its traffic is dominated by commercial research establishments. The implications of such patterns of use are explored further in [Odlyzko1, Odlyzko2].

The main reason for discussing Internet links so extensively is that I was able to obtain extensive collections of statistics on them. I have much less data about traditional private line networks. In particular, some people claim that SNA networks (the traditional method for carrying mainframe traffic) might have higher utilizations than IP networks, but so far I have no solid evidence of that. For IP networks, the evidence points to utilization rates in the 3-5% range. As an example, the large corporate IP network profiled in Fig. 7 has average utilization of about 4% over a full week.

Most of the private line networks that I was able to obtain statistics for were actually composed of Frame Relay links, probably because Frame Relay networks tend to be better instrumented. The Frame Relay networks are semi-public, meaning that the traffic from many customers is carried on the same network from a service provider like AT&T or MCI, but almost always connects sites within the same organization. (Although some carriers are introducing SVCs, switched virtual circuits, almost all traffic is currently carried on PVCs, permanent virtual circuits, which provide point-to-point connections only.) The Frame Relay business is growing at about the same rate as the Internet, namely 100% per year, and is doing that partially by cannibalizing traditional private line business. Customers pay for a port to the network, which imposes an absolute limit on the rate at which they can send data into the Frame Relay network, and for CIR (Committed Information Rate), which is the rate that the service provider promises to carry successfully to the destination. (Bursts above the CIR may be discarded if the network is congested.) Typical arrangements are that the CIR is half or a quarter of the port speed. (For more details, and the advantages and disadvantages of Frame Relay services, see [Cavanagh].) The average utilization of ports appears to be around 3%. The highest utilization I have seen was 12%. It occurred in the very expensive international (multi-continental, even) network of Frame Relay links for a large corporation, where there are strong incentives to utilize transmission capacity heavily, even at the cost of quality of service.

Most of the hard evidence I have collected supports estimates of average utilization rates for private line networks of around 3% or at most 4%. I am more comfortable making an estimate of 3-5% to compensate for several factors. One is the the lack of knowledge of some networks, such as SNA ones, which may be more heavily utilized. Another one is that although Frame Relay ports are utilized only around 3% of their capacity, their much lower cost compared to traditional private lines, and higher latency aparently often lead customers to use a port larger than the private line it replaces (see [Cavanagh]). This suggests that leased lines might be utilized more heavily than Frame Relay ports.

So far I have presented arguments for low utilization rates for private lines based on measurements for some networks and extrapolations from that to the entire data networking universe. Another strong argument in favor of the estimate of low utilization rates for private lines comes from looking at the total amount of data traffic. The bandwidth of all the private lines is large, comparable to that of the voice network [CoffmanO]. If those lines were utilized much more heavily than the 3-5% rate estimated above, there would be a huge amount of data traffic. However, most of the private lines are used by the large companies, those in the Fortune 500. Although their data traffic is growing explosively, it is not all that large yet. Lew Platt, the CEO of Hewlett-Packard, stated in a Sept. 1997 press release that the HP Intranet carried about 10 TB/month. (A similar statement by Platt a year earlier claimed 5 TB/month, showing that HP experienced the common 100% annual growth rate in their traffic.) Nortel was carrying about 15 TB/month at the end of 1997, with growth rates of 80% for the previous three years (private communication from Terry Curtis, who is in charge of Nortel networks). There are several other corporations with networks about as large as HP's or Nortel's. The collective revenues of the Fortune 500 are around $5,000 B, while those of HP are about $40 B, with Nortel (which is not included in the Fortune 500 as it is a Canadian company) at $15 B. Extralopating from these and other examples where I have estimates for total corporate traffic, it appears unlikely that there could be more than 3,000 to 5,000 TB/month of traffic inside all corporations in the U.S.. However, that 3,000 to 5,000 TB/month estimate is exactly what one obtains by combining the 3-5% estimated utilization rate of this paper with the bandwidth estimate for all private line networks of [CoffmanO].

Another argument for low utilization rates for private lines is based on pricing. This is discussed at greater length in [Odlyzko1]. If private line utilization rates were high, costs of transporting data over them would be very low, much lower than over the Internet or even over Frame Relay. However, all communications industry sources agree that Frame Relay is usually less expensive than private line, and that VPNs (Virtual Private Networks) over the public Internet are even less expensive.

8. Data networks will stay lightly utilized

Although higher utilizations than are prevalent today should be achievable (as is discussed in [Odlyzko1, Odlyzko2]), it seems likely that data networks will continue to be utilized much less intensively than the switched voice network. Some of the inherent inefficiency of data networks comes from their voice heritage. A phone call is given two symmetric channels, each of 64 Kbps. Although normally only one person speaks at a time, it is the few moments when both do that are often most important in conveying information. Thus having a full channel for each person was a reasonable choice when calls invariably meant voice calls and technology was not up to doing compression effectively. As a result of that early decision, data lines are also symmetric. This leads to substantial inefficiencies in a world where a data line connects two computers. This can be seen in Fig. 8. Clearly SWITCH customers would be much better off if instead of having 8 Mbps of capacity in each direction across the Atlantic, they had 12 Mbps going East and only 4 Mbps going West. Another example is that of off-site emergency backup lines. Typically these are run at night, and carry data from a university or corporate site to some distant storage facility. The return path is almost never used, but crucial when disaster strikes, and data has to be restored. In such a setting, a half-duplex link would be much more efficient.

The inefficiency created by forced symmetry of data lines is less of a problem in large backbone data networks like those of the Internet, where a mix of traffic sources produces a rough balance, but it is still a problem. Noticeable imbalances can be seen even on large trunks, such as on the MCI OC3 Internet backbone link profiled in [ThompsonMW], where the patterns of traffic to the south and to the north do differ. (The imbalance in the two directions is huge on the US-UK T3 link described in [ThompsonMW]. This imbalance is attributed to most Web servers being located in the US.) However, the inefficiencies resulting from such traffic imbalances are hard to eliminate.

Symmetry of data lines is probably a minor contributor to the overall inefficiency of data networks. Much more important are the nature of data traffic and the extraordinarily high rates of change and growth in the industry.

Data traffic is much burstier than voice traffic. During a peak hour, the U.S. voice networks carry around two million simultaneous calls, with tens of thousands of calls being processed by a single switch. Under those conditions, addition of one more call has a minor effect on the behavior of the network. On the other hand, a single workstation can generate data traffic in the tens of megabits per second, which is noticeable when most of the Internet backbone trunks are 45 Mbps or 155 Mbps. The bursty nature of traffic on corporate data networks can be seen in Fig. 4, which was already discussed in the Introduction. The traffic profile on the line pictured in Fig. 4 looks smoother when one considers hourly averages, as is done in Fig. 6 of [Odlyzko1]. It is still very bursty, and one might think this burstiness is caused by the low capacity of the line (128 Kbps). However, even high capacity lines do not have smooth traffic profiles when one considers short time scales For example, Fig. 5 shows traffic on an OC3 link in the MCI Internet service when averaged over 5-minute and one hour intervals. (This is the same link for which hourly averages in the reverse direction are shown in Fig. 3, and the data shown here are those in [ThompsonMW].)

Even when individual computers limit their data transfer speeds, the resulting traffic is not as nicely behaved as voice traffic. It is now widely accepted that data traffic is self-similar [LelandTWW] (see [FeldmannGWK] for latest results and more complete references). This means that as transfers from many sources are aggregated, there is some smoothing, but much less than on the voice network. It seems that there are fundamental limitations on the efficiency that can be achieved on data networks.

The work on self-similarity of data traffic shows that the usual procedure of looking at just 5-minute or 1-hour averages of traffic is not adequate to understand what goes on. One should study traffic on millisecond time scales, but that is currently done only in a few experimental setups. Networks are engineered based on cruder averages, and the usual rules one hears about in high quality networks are of the form "a T1 link has to be upgraded if hourly averages exceed 50% of the capacity over more than 5% of the business hours." For Internet backbones, a common rule [Gareiss] is that "during peak periods, an ISP should have at least 30 percent to 40 percent of spare bandwidth. The good news is that most providers have 50 percent or more." (Unfortunately there are many subtleties in defining spare bandwidth, so it is hard to interpret these claims precisely.)

In corporate networks, data traffic is concentrated during regular business hours, as can be seen in Fig. 7 (and figures 6 and 7 of [Odlyzko1]). The usual rule of thumb is that the busy hour carries about one sixth of the day's traffic. Since there is very little weekend traffic, this means that the traffic carried in a 168-hour week is equivalent to that carried over about 30 hours of running at peak hour utilization. If average peak hour utilization were 50%, that would produce average utilization over the full week of 9%. This figure would go up to 12% if peak hour utilization of 70% could be tolerated.

A common rule among network managers appears to be to upgrade a T1 link when its peak hour utilization exceeds 50% or 55%, and a T3 when its utilization exceeds 70%. Any large network typically has some links running close to these thresholds. As a result, managers usually overestimate how heavily their networks are used and that may be one source for the common perception of 70% utilization. (Network managers also appear to overestimate the utilization of their LANs, again because they react to the "hot spots" that require action, and do pay less attention to the bulk of their facilities.)

Designers of private line networks usually estimate average utilization better than network managers do. The reason is that they tend to rely on design rules that specify peak hour utilization of 15%, 20%, or 30% (to quote some common figures that I have heard, which vary depending on expected applications and link capacities). If the peak hour utilization is 20%, then in a corporate setting the average weekly utilization will be under 4%.

Why would one plan for peak hour load of 20%, when even T1s commonly behave well with 50% loads? Data traffic is not only bursty, but it grows much faster and in less predictable ways than voice traffic. While the load on the switched voice network has been growing about 8% a year recently, capacity of private line networks (and therefore presumably traffic on them) has been growing around 15% to 20% a year [CoffmanO]. The Internet appears to be growing about 100% a year now, and has grown at that rate for at least a decade, with the exception of 1995 and 1996 when it appears to have grown about 1,000% a year. Several corporations, such as HP and Nortel, both mentioned earlier, report that their internal IP traffic has been growing about 100% a year. Not only is that growth far more rapid than in the switched network, but it typically is uneven inside a corporation, as new services are deployed. Furthermore, installing new capacity is a slow process, with waits of up to a year reported for private line T3s, and some orders lost or simply not filled. In this environment, where internal customers are constantly screaming about their "mission-critical" applications requiring better service, it is prudent to overprovision. If capacity is too high, that is just some extra money. If capacity turns out to be too low, one can lose important business and get fired.

The natural tendency to build in adequate safety margins is aggrevated by the lumpy nature of network capacity. What happens when a T1 becomes overloaded (which probably means its average utilization over a week approaches 10%)? Typically a second T1 is put in. This reduces traffic load to half of what is considered tolerable. Let us assume that traffic increases smoothly at 100% a year. Then, after a year both T1s will be full. At that stage a third T1 will be put in, and after a further 7 months, a fourth one. At the end of the second year all four T1s will be full. At that stage, however, usually a T3 will be put in and the T1s removed (unless there is need for redundant links for higher reliability). The reason is that most routers currently cannot balance the load on more than four T1s. This requires a replacement of four T1s operating at full load (i.e., 10% of capacity) by a single T3 operating at 1.4% of capacity. After one year, the utilization level on that T3 will be up to 2.8%, after another year at 5.6%, and after yet another year, it will be time to put in a second T3. However, if we look at the entire 5-year period, starting with a single overloaded T1 and ending with a single overloaded T3, a simple calculation shows the average utilization level (weighted by capacity) will be much less than the 10% one might have expected, closer to 5%.

The extreme example above is caused partially by the deficiency of current routers. However, even after this defect is eliminated (as it is supposed to be soon) a similar problem will exist in a form that is only slightly milder. A T3 has traditionally cost about 8-10 times as much as a T1. (The recent shortage of capacity appears to have pushed up T3 prices, but let us not take this into account, as this is likely a temporary condition. Similar jumps in price by 8 or 10 are observed in going from 56 Kbps circuits to T1s, presumably indicating the reduced costs of providing high capacity lines. See [FishburnO] for data on private line prices and further discussion.) This means that one would not install (except for redundancy reasons) more than 7 T1s. In practice, given the cancellation fees in terminating a T1, as well as the lead time for installing T3s, one would probably never go to more than 6 T1s before switching to a T3. However, when traffic from 6 T1s, each operating at 10% of capacity, is moved to a T3, capacity utilization drops to 2%. The traffic profile for Columbia University, shown in Fig. 10, suggests that (at least for the two days shown there) it could be accomodated by 7 T1s. However, if fractional T3 access is not available, then it is less expensive to have a mostly empty T3.

The voice phone network does not suffer from the lumpy capacity of data lines. Additional connections between 4E switches are added in T1 increments (24 voice lines), and since a 4E has tens of thousands of lines, capacity is almost a continuous variable. However, there are other aspects in which the voice network also has lumpy capacity inefficiencies (in switching, for example). Furthermore, the small increments of transmission capacity in the voice network carry their own heavy burden, since they make it impossible to lower costs by going for higher bandwidth pipes.

In less than a decade, NSFNet went from 56 Kbps circuits to T1 and then to T3 trunks. Such jumps by factors of almost 30 in each case are large, and mean that the upgraded links will be underutilized for a long time. This underutilization can be overcome to some extent by using fractional T1 and T3 connections (for example, the University of Waterloo went from 56 Kbps to 128 Kbps, to T1, and more recently to 5 Mbps, [Waterloo]), but those are used less often than one might expect, if one judges by the statistics in [VS], which show relatively few fractional T1 links. As we move above T3 speeds to OC3, OC12, OC48, and OC192, gaps become smaller, making the likely underutilization from this source less severe. It will still be present, though. When MCI upgraded their Internet backbone from OC3 to OC12, their average utilizations must have dropped substantially. It is even conceivable that not all the intermediate speeds on the OC hierarchy will be used. As was mentioned above, another important factor appears to be at work that leads to low utilization rates. The marketplace appears to favor constructing systems out of a few basic building blocks, even when those block sizes are not ideal for the task at hand. LANs are increasingly dominated by 10 Mbps Ethernet, 100 Mbps Fast Ethernet, and (just starting) 1,000 Mbps Gigabit Ethernet, with 10,000 Mbps Fast Gigabit Ethernet under discussion already. Doesn't this lead to massive mismatches in capacity? Shouldn't we have 4 Mbps and 16 Mbps devices? Well, we sort of did, with various Token Ring technologies, for example, but they all seem to be fading, and the few flavors of Ethernet are taking over. Standardization on a few speeds of a single protocol leads to increased efficiency in development and manufacture of devices and software. It also simplifies the tasks of network managers. With only a few speeds to worry about, their task of engineering their networks becomes easier, and they can manage the networks more easily. Overengineering the LANs does waste bandwidth, but saves total system costs.

Similar trading of bandwidth for simplicity of operation is seen in long distance data networks. On a recent day, the average utilizations of the 16 OC3 interfaces to vBNS varied by a factor of 20, and peak 3-minute average utilizations varied by a factor of 40. Such behavior is not seen in the switched voice network.

Large jumps in capacity (such as going from 56 Kbps to 256 Kbps and then to 512 Kbps, and finally to T1) also appear to fit well with the way our perceptual system works. Our eyes, ears, and other senses respond on a logarithmic scale, and so a small jump in the speed of a connection is not perceived as offering much of an advantage. Similarly, it usually takes a large jump in the speed of microprocessors to persuade customers to upgrade. In light of this factor, it is understandable that even when service providers offer a range of speeds with fine granularity, only a few choices, corresponding to a few multiples of some basic speed, are actually used in large numbers. To select intermediate ranges and keep upgrading them as traffic increases would require additional effort from network managers and would not be appreciated by end users.

Network managers always have too much to do. Traffic typically doubles each year, and there are new and unpredictable demands showing up constantly. Further, provision of additional capacity has to fit in with the budget cycle. As a simple example, consider the utilization of Internet links at the University of Toronto and Columbia (figures 9 and 10). Columbia provides a much less congested link. One might easily guess this is because Columbia is richer. On the other hand, in dial-up services, it is the Columbia modem pool that is consistenly saturated for most of the day, whereas that at the University of Toronto hits capacity limits for only brief periods on a regular day. (See figures 3 and 4 in [Odlyzko1].) This inconsistency in provision of data services is likely caused not by relative wealth of these schools, but by the budgetary and hardware cycles. The picture of data networking has to be modified to take into account the dynamic element of the situation.

In the environment of rapid and only roughly predictable growth, maximal efficiency cannot be attained, and simple solutions that work are at a premium. We have seen that in the examples above. Bandwidth is substituted for the careful engineering that makes our voice network efficient. Since data networking will continue its rapid growth, we can expect to see comparable evolution in the future. Bottlenecks like the inability of routers to load-balance more than four T1s will be removed. However, other problems will appear in their place. For the foreseeable future, the data networking scene is likely to resemble the current one, with lots of lightly utilized capacity and a variety of bottlenecks.

The preference for using simple solution that can be made to work right away can be seen at other levels of the networking scene as well. SONET rings waste at least 50% of their bandwidth to provide protection against fiber cuts. A mesh-based solution would be more efficient, but presumably would take too long to engineer. We also run IP over ATM, in spite of at least 20% overhead cost. At another extreme, consider the fax. It is ubiquitous, although one might think that email should have made it obsolete. It also uses the network extremely inefficiently, typically transmitting just at 9.6 or 14.4 Kbps, even though it uses 128 Kbps of bandwidth. Yet it thrives, since it provides a reliable service at low cost, a service that can be used to reach more people than email, say. (And talking of inefficiencies, how come we do not have 28.8 faxes? Also, how come files on the Web are seldom compressed, aside from graphics, even though that would speed downloads?) Such inefficiencies would have been unthinkable in the old voice telephony world, but they are common in the era of rapid growth. Even if these inefficiencies are eliminated, others are likely to take their place.

Were a Martian asked to design a data network for us from scratch using our current technology, we would surely not get what we have. However, a Martian would also have given us neither the NTSC color TV system nor the DOS/Windows operating system.

9. Conclusions

This paper shows that data networks are utilized at low fractions of their capacity, considerably lower than the switched voice network. The question is whether this matters.

For designers of private line networks, low average utilization is indeed irrelevant. Their task is to find the most efficient way to provide the connectivity that their clients depend on within the parameters they work in, namely leased lines for exclusive use of those clients. If customers want to accomodate bursty data transmissions, concentrate their traffic during regular business hours, and be free to suddenly generate increased traffic loads with new services, then utilization rates will stay low, and are just part of the price that has to be paid.

On the other hand, from a more global perspective, low average utilizations are important. Here are some examples of what they imply:

(a) With average peak utilizations under 15% on private line networks, there is room for squeezing in packet telephony calls. (If the peak hour utilizations were consistently close to 70%, this would be much more questionable.)

(b) Private line transport is very expensive, and corporations can save by switching over to VPNs over the public Internet. (If average utilizations were high, this economic incentive would be absent, as private lines would be much cheaper.)

(c) The corporate side of the ISP business is much more profitable than estimates such as that of [Leida] show, since they generate much less backbone traffic.

(d) Aggregation of corporate traffic on the public Internet or the semi-public Frame Relay and ATM networks promises much greater savings than would be the case for heavily utilized private line networks.

(e) There is less data traffic than is often estimated on the basis of the aggregate size of data networks (cf. [CoffmanO]).

These and other implications of low utilization rates of data networks (for example, for provision of Quality of Service on the Internet) are considered at greater length in the companion paper [Odlyzko1] and the overview paper [Odlyzko2].

Acknowledgements: I thank Wale Akinpelu, Muayyad Al-Chalabi, Sam Alexander, Jerry Ash, Fred Baker, Vijay Bhagavath, Suparno Biswas, Josh Bottum, Andy Brajkovic, Hans-Werner Braun, Bill Bridges, Ramon Caceres, Douglas Carson, Chris Chase, Bill Chen, Kim Claffy, Kerry Coffman, Terry Curtis, Walter Drechsel, Nikki Dwyer, Bruce Emerson, Norm Farber, Anja Feldmann, Kerry Fendick, John Friedmann, Ira Fuchs, Lou Gabos, Steven Gao, Ehud Gelblum, Bill Gewirtz, Albert Greenberg, Jürgen Harms, Musharoff Hossain, David Houck, Willi Huber, Andrew Hume, Bill Infosino, Pravin Johri, Rao Karanam, Jay Kent, Eleftherios Koutsofios, Russ Kurtz, Tony Lauck, Ken Lofgren, David Lynch, Pat Mariani, Clem McCalla, Chuck McCallum, Greg Miller, Steve Miller, Tom Molfetto, Philip Murton, Peter Olenick, Craig Partridge, Vern Paxson, John Pennewell, Frank Pirz, David Reed, Jennifer Rexford, Ralph Richburg, David Sapir, Henning Schulzrinne, Bill Strain, John Strand, Russell Sutherland, Bruce Tanzi, Kevin Thompson, Bob Tkach, Fred True, Russ Truscott, Les Weppler, Allan Wilks, Chuck Williams, Xavier Williams, Walter Willinger, Pat Wirth, Bill Woodcock, and Eberhard Wunderlich for comments and information that were invaluable in the prepartion of this paper.

References

[Ash] G. R. Ash, *Dynamic Routing in Telecommunications Networks,* McGraw Hill, 1998.

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

[CAIDA] Cooperative Association for Internet Data Analysis (CAIDA), ⟨http://www.caida.org/⟩.

[Cavanagh] J. P. Cavanagh, *Frame Relay Applications: Business and Technical Case Studies,* Morgan Kaufman, 1998.

[Cerf] V. Cerf, PowerPoint slides of presentations, available at ⟨http://www.mci.com/aboutyou/interests/technology/ontech/powerpoint.shtml⟩.

[ClaffyPB] K. C. Claffy, G. C. Polyzos, and H.-W. Braun, Traffic characteristics of the T1 NSFNET backbone, available at ⟨http://www.merit.edu/nsfnet/statistics/⟩.

[CoffmanO] K. G. Coffman and A. M. Odlyzko, The size and growth rate of the Internet, *First Monday,* 3(10) (October 1998), http://www.firstmonday.dk/. Also available at ⟨http://www.research.att.com/∼amo⟩.

[Columbia] Columbia University modem pool statistics, available at ⟨http://www.cs.columbia.edu/acis/networks/modems-usage.html⟩.

[DOC] U.S. Department of Commerce, *The Emerging Digital Economy,* April 1998. Available from ⟨http://www.ecommerce.gov/emerging.htm#⟩.

[FeldmannGWK] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, The changing nature of network traffic: Scaling phenomena, *Computer Communication Review,* 28, no. 2 (April 1998), pp. 5–29. Available at ⟨http://www.research.att.com/∼agilbert⟩.

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

[Gareiss] R. Gareiss, Is the Internet in trouble?, *Data Communications,* Sept. 21, 1997, pp. 36-50. Available at ⟨http://www.data.com/roundups/trouble.html⟩.

[GMLCOBRS] V. Granger, C. McFadden, M. Lambert, S. Carrington, J. Oliver, N. Barton, D. Reingold, and K. Still, Net benefits: The Internet - A real or virtual threat, Merrill Lynch report, March 4, 1998.

[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⟩.

[ITU] International Telecommunication Union, *Challenges to the Network: Telecommunications and the Internet,* Sept. 1997. Can be purchased through ⟨http://www.itu.ch⟩.

[Keller] J. J. Keller, Ex-MFS managers plan to build global network based on Internet, *Wall Street J.,* Jan. 20, 1998.

[Keshav] S. Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network,* Addison-Wesley, 1997.

[KleinrockN] L. Kleinrock and W. E. Naylor, On measured behavior of the ARPA network, in *ATIPS Proceedings, 1974 National Computer Conference*, vol. 43, Wiley 1974, pp. 767-780.

[Leida] B. Leida, A cost model of Internet service providers: Implications for Internet telephony and yield management, M.S. thesis, department of Electr. Eng. and Comp. Sci. and Technology and Policy Program, MIT, 1998. Available at ⟨http://www.nmis.org/AboutNMIS/Team/BrettL/contents.html⟩.

[LelandTWW] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Networking* 2 (1994), 1-15.

[MCI] MCI press release, available at ⟨http://www.mci.com/aboutus/company/news/digipresskits/internetpress.shtml⟩.

[MillerTW] G. J. Miller, K. Thompson, and R. Wilder, Performance measurement on the vBNS, to appear in *Proc. Interop'98 Engineering Conference,* Las Vegas, May 1998. Available at ⟨http://www.vbns.net/presentations/papers/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⟩.

[NLANR] National Laboratory for Applied Network Research, ⟨http://www.nlanr.net/⟩.

[NSFNet] Historical data for NSFNet, available at ⟨http://www.merit.edu/nsfnet/⟩.

[Odlyzko1] A. M. Odlyzko, The economics of the Internet: Utility, utilization, pricing, and Quality of Service. Available at ⟨http://www.research.att.com/∼amo⟩.

[Odlyzko2] A. M. Odlyzko, The Internet and other networks: Utilization rates and their implications, *Selected Papers from the 1998 Telecommunications Policy Research Conference,* S. E. Gillett and I. Vogelsang, eds., Lawrence Erlbaum Associates, 1999. To appear. Also available at ⟨http://www.research.att.com/∼amo⟩.

[Park] R. E. Park, Incremental costs and efficient prices with lumpy capacity, Report R-3723-ICTF, RAND Corp., 1989.

[Paxson] V. Paxson, Measurements and Dynamics of End-to-End Internet Dynamics, Ph.D. thesis, Computer Science Division, Univ. Calif. Berkeley, April 1997. Available at ⟨ftp://ftp.ee.lbl.gov/papers/vp-thesis/⟩.

[Princeton] Princeton University network statistics, available at ⟨http://wwwnet.princeton.edu/monitoring.html⟩.

[Roberts] E. Roberts, Policy-based networking: The new class system, *Data Communications,* Oct. 1997. Available at ⟨http://www.data.com/roundups/class_system.html⟩.

[SWITCH] SWITCH network statistics, available at ⟨http://www.switch.ch/lan/stat/⟩.

[TeleGeography] *TeleGeography 1996/97: Global Telecommunications Traffic Statistics and Commentary,* TeleGeography, Inc., Washington, D.C.

[ThompsonMW] K. Thompson, G. J. Miller, and R. Wilder, Wide-area Internet traffic patterns and characteristics, *IEEE Network,* 11, no. 6 (Nov/Dec 1997), pp. 10-23. Extended version available at ⟨http://www.vbns.net/presentations/papers/MCItraffic.ps⟩.

[Toronto] University of Toronto network statistics, available at ⟨http://www.noc.utoronto.ca/netstats/index.html⟩.

[VBNS] vBNS (very high performance Backbone Network Service) statistics, available at ⟨http://www.vbns.net/⟩.

[VS] Vertical Systems Group, *ATM & Frame Relay Industry Update,* 1997.

[Waterloo] University of Waterloo network statistics, available at ⟨http://www.ist.uwaterloo.ca/cn/#Stats⟩.

Table 1: Average utilization levels

network	utilization
AT&T switched voice	33%
Internet backbones	15%
private line networks	3-5%
LANs	1%

Table 2: Utilization levels (in percent of line capacity) on dedicated business customer lines to a segment of the Zocalo network during the week ending Nov. 29, 1997. Maximal figures refer to highest hourly utilizations.

line rating in Kbps	ave. receive utilization	ave. transmit utilization	max. receive utilization	max. transmit utilization
56	0.10	0.02	3.26	0.48
56	0.85	0.27	11.30	4.74
56	0.93	0.07	13.44	2.61
56	1.20	0.14	11.47	1.14
56	1.26	0.18	6.41	5.96
56	1.34	0.27	6.08	5.39
56	1.37	0.24	12.77	2.50
56	1.43	0.24	17.42	9.38
56	1.52	0.32	8.34	6.91
56	1.57	0.38	68.30	11.28
56	1.60	0.77	33.38	16.75
56	1.61	0.23	16.48	2.60
56	1.90	1.17	23.72	2.90
56	2.03	0.57	19.37	6.90
56	2.03	0.92	62.26	44.90
56	2.24	6.81	21.78	38.61
56	2.57	0.39	51.84	19.72
56	2.67	1.54	67.01	29.22
56	2.89	2.87	15.46	15.73
56	3.15	0.50	54.99	5.11
56	3.47	1.68	33.24	17.66
56	4.38	1.81	51.58	49.62
56	5.21	0.48	68.06	9.71
56	5.41	7.85	47.17	33.42
56	5.54	2.58	38.50	26.21
56	7.75	5.75	41.21	8.19
56	23.56	9.39	67.47	28.32
128	1.28	0.23	14.80	1.57
128	1.62	3.21	12.99	21.13
128	2.03	7.46	14.87	24.91
128	4.56	3.74	69.99	62.35
128	4.57	2.14	55.90	8.65
128	4.69	2.23	42.52	35.65
128	12.31	5.96	83.35	69.38
384	0.58	0.15	4.93	1.19
384	0.90	1.21	12.02	3.95
384	3.95	1.17	59.39	12.64
384	4.75	1.90	28.55	9.98
1536	0.05	0.02	0.49	0.72
1536	0.13	0.06	2.64	3.69
1536	0.23	0.11	2.43	1.44
1536	0.28	0.95	2.00	4.26
1536	0.33	0.09	2.58	2.29
1536	0.50	0.53	4.36	2.82
1536	5.73	4.74	52.70	35.34

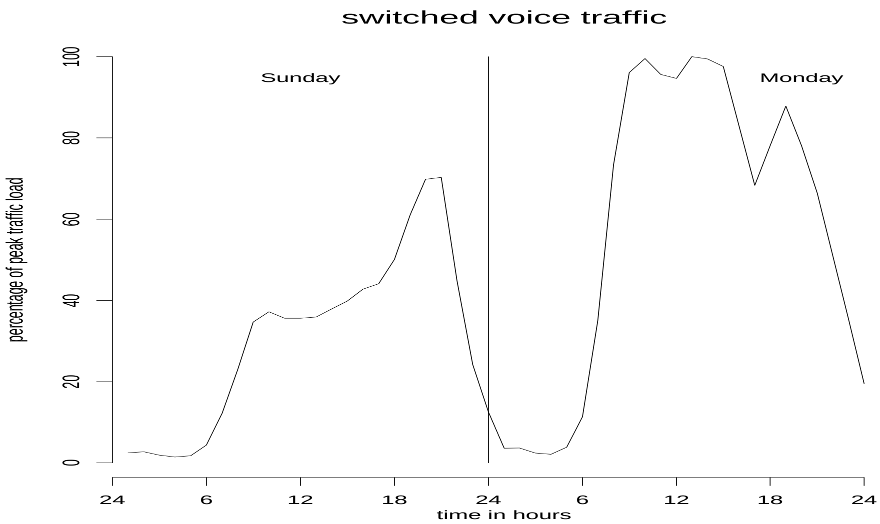

Figure 1: Voice traffic on U.S. long distance networks.

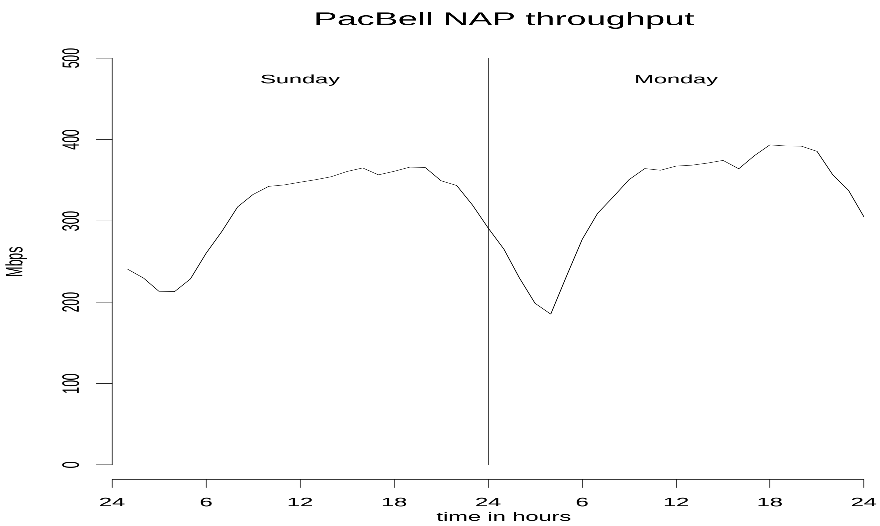

Figure 2: Traffic through the PacBell NAP, in megabits per second, on Oct. 26 and 27, 1997. Pacific Standard Time, 1-hour traffic averages.

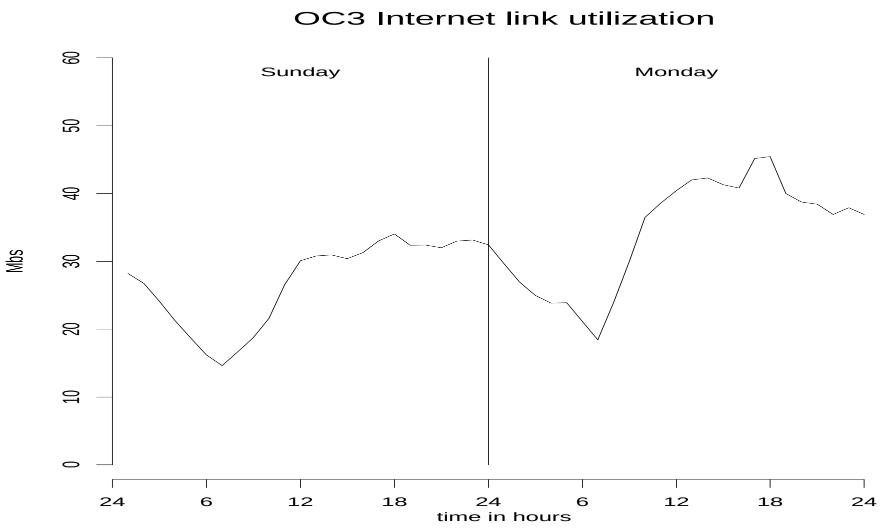

Figure 3: Traffic to the south on an MCI OC3 Internet trunk on August 24 and 25, 1997. Hourly averages, Eastern Standard Time. By permission of MCI.

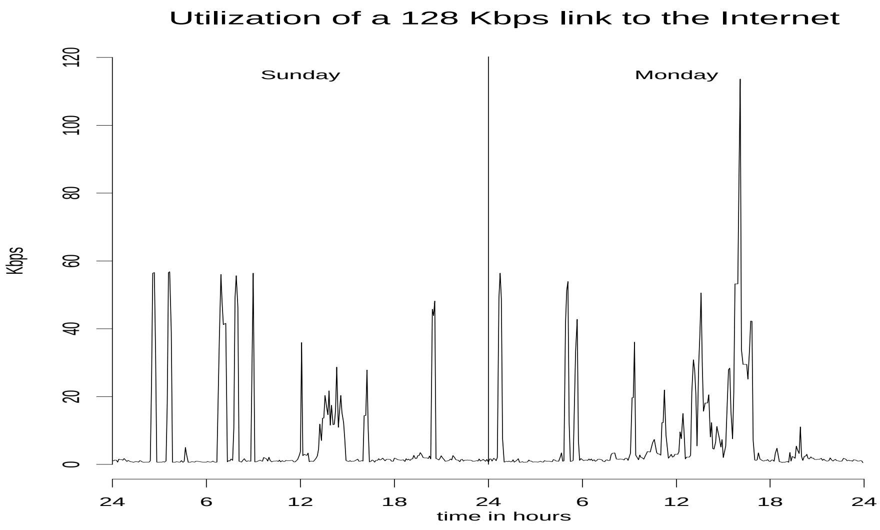

Figure 4: Utilization of a 128 Kbps dedicated business connection to the Internet during February 22 and 23, 1998. Only traffic from the ISP to the customer is shown. 5-minute averages.

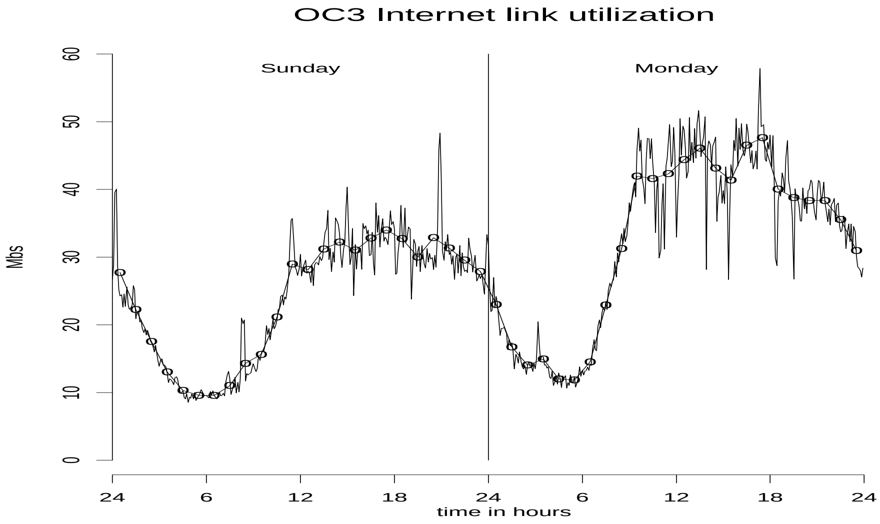

Figure 5: Traffic to the north on an MCI OC3 Internet trunk on August 24 and 25, 1997. Simple line shows 5-minute averages, line with circles hourly averages. Eastern Standard Time. By permission of MCI.

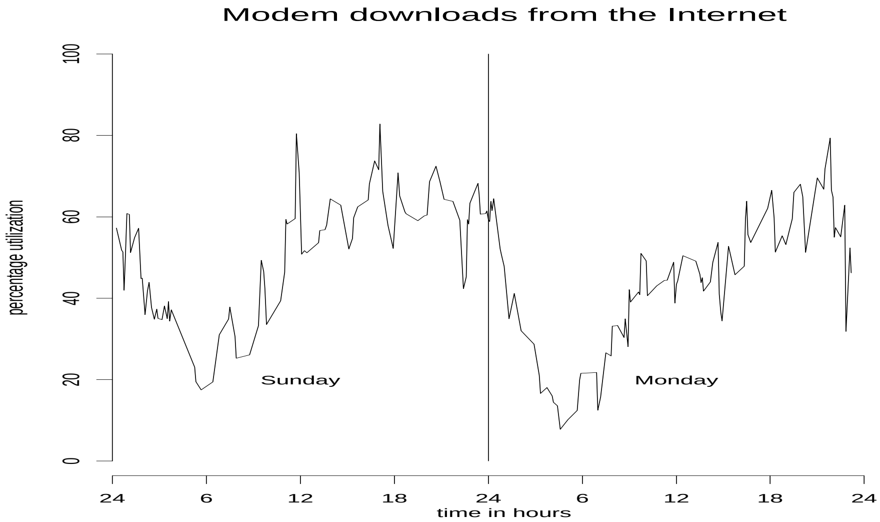

Figure 6: Traffic to a dial ISP in early 1998, 15-minute averages.

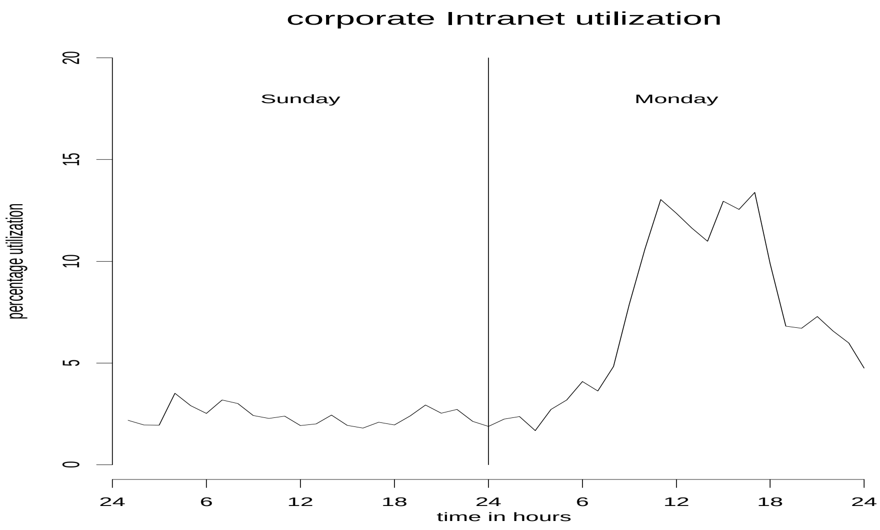

Figure 7: Average utilization of T3 links in a large corporate private line network. Hourly averages.

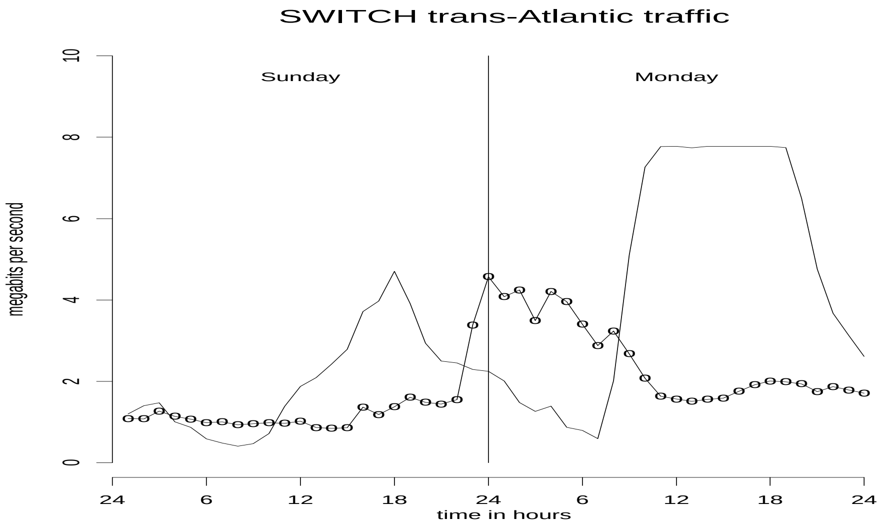

Figure 8: Traffic on the 8 Mbps link between the U.S. and SWITCH, the Swiss academic and research network, during February 1 and 2, 1998. Thin line is the traffic to Switzerland, line with circles traffic to the U.S.. Swiss standard time. By permission of SWITCH.

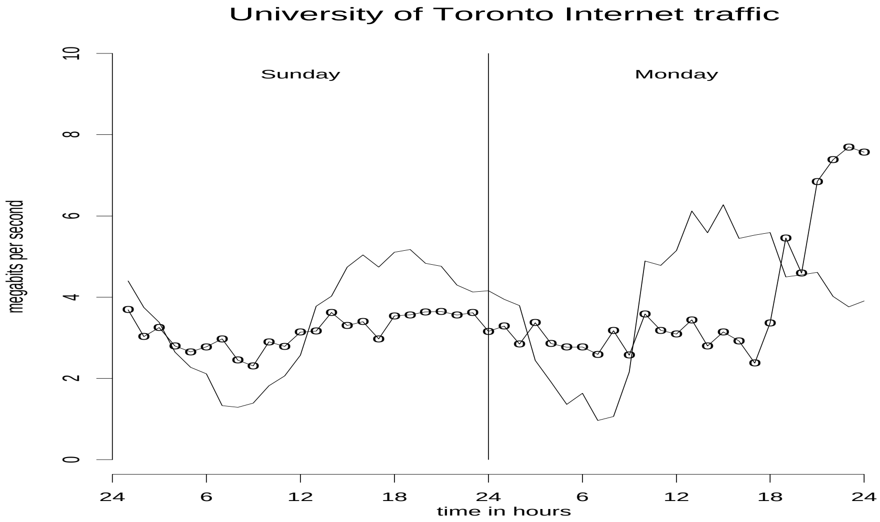

Figure 9: Utilization of University of Toronto's 8 Mbps link to the Internet, January 11 and 12, 1988. Hourly averages, Eastern Standard Time. By permission of University of Toronto.

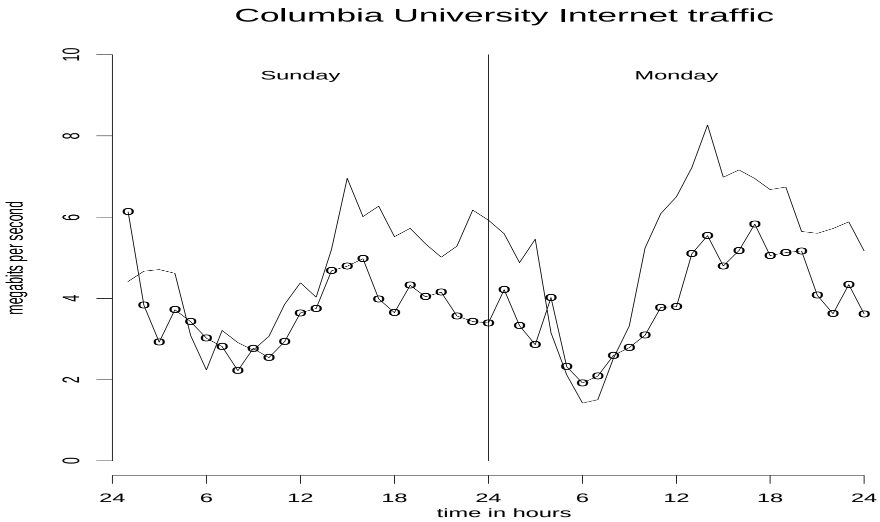

Figure 10: Traffic on Columbia University's T3 link to the Internet, February 1 and 2, 1998. Thin line is the traffic into Columbia, line with circles traffic to the Internet. Hourly averages, Eastern Standard Time. By permission of Columbia University.