The Internet and other networks: Utilization rates and their implications

Andrew Odlyzko

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

Revised version, February 26, 2000.

Abstract. Costs of communications networks are determined by the maximal capacities of those networks. On the other hand, the traffic those networks carry depends on how heavily those networks are used. Hence utilization rates and utilization patterns determine the costs of providing services, and therefore are crucial in understanding the economics of communications networks.

A comparison of utilization rates and costs of various networks helps disprove many popular myths about the Internet. Although packet networks are often extolled for the efficiency of their transport, it often costs more to send data over internal corporate networks than using modems on the switched voice network. Packet networks are growing explosively not because they utilize underlying transport capacity more efficiently, but because they provide much greater flexibility in offering new services.

Study of utilization patterns shows there are large opportunities for increasing the efficiency of data transport and making the Internet less expensive and more useful. On the other hand, many popular techniques, such as some Quality of Service measures and ATM, are likely to be of limited usefulness.

1. Introduction

An extensive study of data networks is documented in [CoffmanO, FishburnO, Odlyzko2, Odlyzko3, Odlyzko4]. This paper presents only a brief summary of the results of that study, and concentrates on their implications for present and future data networks.

Utilization rates of networks have been strangely absent from most papers on the economics of the Internet, such as those in [MacKieM, McKnightB, Varian]. However, these rates determine costs of services, since transmission links are priced by their maximal capacity. Furthermore, utilization rates are the primary means by which network managers determine quality of transmission. Therefore it seemed important to consider current utilization rates on the Internet and the resulting costs. Such a study was carried out in [Odlyzko2, Odlyzko4]. It uncovered a number of surprising results. For example, corporations in the U.S. spend more to transmit large files over their packet networks than they would if they used modems over the switched voice network. The primary culprit behind this phenomenon is the low utilization of most data networks.

A key point of the investigation of [CoffmanO, FishburnO, Odlyzko2, Odlyzko3, Odlyzko4] was the need to consider not just the public Internet, but the full universe of data networks and their role in the economy. For simplicity, only networks inside the U.S. were considered. Although they do attract intense attention, they are a small part of the entire IT (information technologies) industry. While the IT sector of the U.S. economy accounts for about \$600 billion per year [DOC], data communications costs about \$80 billion, and of that, transmission comes to about \$16 billion [Odlyzko4]. (By comparison, the telephone system revenues are close to \$250 billion per year. Even technically sophisticated corporations are still spending more on voice communications than on data.) Thus any modification to data networks has to be considered in light of the total costs it imposes on the economy, not just in terms of what it does to networks.

Data networks are not only still a small part of the economy, they do not operate in isolation. For voice calls, the basic service is easy to describe (Fig. 1), and there is general agreement on desirable quality. On the other hand, data networks increasingly are becoming just enablers (although crucial ones) of other services, and few users care about the network by itself. What is important is how the entire application is perceived by the user, and network transmission is only a part of the system that makes that application possible.

Even in the restricted realm of data networks, the public Internet (those parts of the Internet accessible to general users) is a small fraction of the total system. Measuring networks by their maximal transmission capacity, it was estimated in [CoffmanO] that at the end of 1997 in the U.S., the switched voice network was probably still the largest, but the private line networks were about as large, and the public Internet was considerably smaller. Updating the estimates of [CoffmanO], using the same methodology, yields the following figures for the effective bandwidths (see [CoffmanO] for definition), measured in Gbps (gigabits per second) at the end of 1998:

network	bandwidth (Gbps)
US voice	375
public Internet	150
other public data networks	80
private line	400

Thus looking just at the public Internet does not give a proper perspective on data networks, especially since utilization patterns of private networks are considerably different.

Although data networks are about as large as the voice network in bandwidth, the voice network still dominates in carried load, and is likely to do so for a few more years. The traffic, measured in

TB/month (terabytes per month), through various networks at the end of 1998 was (again updating the estimates of [CoffmanO]):

network	traffic (TB/month)
US voice	43,000
public Internet	5,000 - 8,000
other public data networks	1000
private line	4,000 - 7,000

A comparison of the two tables above shows that there are substantial differences in utilization rates between the voice network and data networks. The question is what this means. It is my contention is that by considering these statistics as well as more detailed ones, we can deduce much about user preferences in data services, and about their willingness to pay for various quality levels.

George Gilder's thesis is that bandwidth supply will soon be increasing so rapidly that we will not have to worry about network congestion. However, that agrument is not entirely convincing. While technology will indeed increase supply, costs and prices are not the same in industries with high fixed costs, regulatory concerns, and substantial barriers to entry. More important, while supply will be increasing, so will demand. Hence it is the balance between the two that will help determine the future of networks.

Section 2 is devoted to disproving a variety of common myths about the Internet. It is based on evidence that is quantitative, although not as precise as one might hope for. Section 3 presents an evaluation of the reasons for the different utilization levels of voice and data networks. Later sections are increasingly speculative, dealing with the likely evolution of the Internet.

Section 4 uses the observations about the utilization patterns of current data networks to explain the failure of ATM and the poor prospects for many Quality of Service (QoS) schemes. Section 5 discusses the opportunities for increasing the quality and lowering the costs of the Internet by learning from the experience of the switched voice network. Section 6 deals with the role of differential service levels and usage sensitive pricing. Section 7 speculates on the most important factors that are likely to influence the evolution of the Internet. Finally, Section 8 presents some final conclusions and predictions.

2. Common wisdom or common misconceptions?

Much of the "folk knowledge" about the Internet is simply false. This section discusses the most important examples.

Packet networks are not necessarily more efficient than the switched voice network. A key point, to be addressed in Section 3, is what is meant by efficiency. In general publications, though, it is often asserted without qualification that packet networks are less expensive then the switched voice network. Some of the new packet-only carriers have been showing comparisons in which IP transport saves more than 90% over the cost of traditional switched networks. In particular, savings on transport costs are widely perceived as the main advantages of carrying voice over packet networks. On the other hand, when one considers existing networks, it turns out that most corporations spend more on transferring large files over their internal IP networks than they would if they used modems over the public switched voice network. This is an astounding result, since modems use only a small fraction of the bandwidth of the digital channel that is provided for voice calls. Further, network costs of voice calls are small compared to the prices charged. The estimates for the cost of transmitting a megabyte of data over various networks are estimated in [Odlyzko4] as follows:

network	dollars/MB
modem	0.25 - 0.50
private line	0.50 - 1.00
Frame Relay	0.30
Internet	0.04 - 0.15

This table suggests an obvious question: Why don't corporations junk their private networks and send data via modems over the public switched voice network? The answer is that the cost estimates of the table apply only to large file transfers, and do not take into account other factors, such as latency. As an example, a credit card authorization involves transfer of only a few hundred bytes, and so would cost far more over a modem than the table might suggest. It would also take far longer, tens of seconds instead of seconds, and thus lead to lower productivity of the sales force and customer dissatisfaction. There are thus unbeatable advantages to packet networks, but they are not in network costs, but in flexibility.

The public Internet is small relative to other data networks. Although it is the public Internet that has caught all the attention, it is still dwarfed in transmission capacity and costs by the private line networks, as was shown in the tables in the Introduction. It is also far smaller than the switched voice network. However, it is growing much faster, about 100% a year, than either the voice network, which is growing at around 10% per year, or the private line networks, which are growing at around 20-30% per year. (See [CoffmanO] for details.) Therefore in a few years the public Internet will be the dominant communication network, but it is not that yet.

Few data networks are congested. A surprising fact is that even though it provides high quality service, the switched voice network has considerably higher average utilization than any large collection of data networks. There is a general perception that the public Internet is hopelessly crowded, and even most network experts believe that private line networks are congested as well. Reality is different, as is shown in [Odlyzko2] and summarized in the table below. (The utilization rates in the table above, and elsewhere in this paper, refer to averages over a full week.)

network	utilization
local phone line	4%
U.S. long distance switched voice	33%
Internet backbones	10-15%
private line networks	3-5%
LANs	1%

Some parts of the Internet are highly congested, especially the public peering points, the NAPs and MAEs. Many university links to the public Internet are also heavily loaded, which may have persuaded generations of students that all networks are heavily utilized. However, the backbones of the Internet are relatively lightly loaded. The estimates of their utilization rates in [Odlyzko2] (based partially on estimates of sizes of various networks in [CoffmanO]) are consistent with recent measurements which show that as long as transmission stays on a single backbone, latency and jitter are not a problem. What is congested are many of the feeder links to the backbones from smaller ISPs, especially those that aggregate modem traffic. Fig. 2 shows the traffic pattern on a T1 line (1.5 Mbps) belonging to an ISP. It runs at a high fraction of its capacity for large parts of the day, but still manages to provide relatively high quality service, with minor delays and packet losses. (Average utilization is in the 40-45% range.) Other ISP links show even higher utilization, frequent saturation, and high packet loss rates. (Other examples of traffic patterns of ISPs, as well as those of other users, are in [Odlyzko2, Odlyzko4].)

As the tables in the Introduction show, most of the data transmission capacity is in private corporate networks. Their traffic patterns tend to be far different from those of the ISP line profiled in Fig. 2. Fig. 3 shows utilization of a corporate T1 line. The average utilization of this line is slightly under 1%. Comparing the graphs of figures 2 and 3, it is easy to grasp that the performance of those two lines will be different, and that traffic control algorithms suitable for one might not fit the other one.

Utilization of the T1 line in Fig. 3 is lower than the average 3-5% cited for corporate networks in the table. Thus a more typical link would show somewhat higher traffic than Fig. 3, but it would still look very low compared to the traffic in Fig. 1. It is worth noting that the line in Fig. 3 does have occasional spikes of higher utilization, but they tend to be shorter than the 5-minutes averaging interval. On some days it also has 5-minute spikes much higher than those of Fig. 3, which represents a pretty typical business day for that link.

While most corporate networks are run at low average utilizations, there are many exceptions. The most prominent are international lines, such as the one profiled in Fig. 4. The average traffic shown there is 59 Kbps, or 46% of capacity during the day that is profiled. On this particular link there is little traffic in the reverse direction, so average utilization of the entire line (which, as is always the case in current data network as a legacy of the switched voice network, consists of two one-directional links), during a business day is around 25%. Over a full week, average utilization might therefore be expected to be around 20%. Some international links have higher utilizations, over 30%. I do not have enough data to be certain, but it appears that the average utilization of trans-oceanic corporate private lines might be in the 10-20% range.

Congestion is not necessarily the biggest problem on the Internet. The "World Wide Wait" is often caused by problems other than lack of bandwidth. A study carried out in 1997 by Christian Huitema [Huitema] about accessing some popular servers showed that 20% were not reachable. Among the 80% that could be reached, 42% of the delays were caused by network transmission, with DNS accounting for 13% and servers for the remaining 45%. Further, there are some indications that in the last couple of years, the performance of the backbones has improved, while servers are falling behind.

If bandwidth were the critical resource in corporate networks, then surely it would be monitored closely. Yet the study of utilization levels in [Odlyzko2] was hampered by lack of data. Existing data was hard to obtain, since it is usually regarded as sensitive. In most cases, though, there was no data to release, since utilizations are not even measured consistently. Hence bandwidth is presumably just one of many problems that network managers have to deal with, and not necessarily the most important.

In general, although the Internet is acclaimed as the simple network, it is extraordinarily complicated and hard to keep operational. A simple example is the description in [WiltziusBD] of the difficulties in getting a network operational, but any network operator can provide a multitude of stories of problems that crop up. One of the major problems with the Internet is that most of the work of keeping it running has to be done at the edges, and so is wastefully duplicated, with network administrators at thousands of institutions doing essentially the same planning, implementation, and operational tasks [Odlyzko3]. Often (and perhaps almost always) networks are not optimized because there are not enough skilled experts to do it.

"The tragedy of the commons" may not be an insurmountable threat for the Internet. It is widely believed that queueing by congestion is the only way to run the Internet, since demand, driven by flat rate pricing, is insatiable. There are many situations where adding capacity does not help. For example, road congestion in metropolitan areas can be relieved only temporarily by building more highways. The problem is that when travel speeds increase, people move further away, to take advantage of less expensive housing, opportunities to be closer to family, and other reasons [Gibbs, SchaferV]. However, in a dynamic environment with growing bandwidth, this argument is questionable. The "tragedy of the

commons” argument is even suspect in slower growing industries. Consider, for example, the following statistics on local calling in the United States, based on data in [FCC]. Such calling is overwhelmingly paid for by a fixed monthly fee, and is thus insensitive to the number and length of calls.

year	lines (millions)	local calls (minutes per day per line)	local calls (minutes per day per person)
1980	102.2	39	17.5
1988	127.1	39	20.2
1996	166.3	40	25.1
1997	173.9	42	27.3

The amount of calling per line has not changed appreciably in 17 years, although the “tragedy of the commons” analogy suggests it should have grown. The standard reply to this is that people have limited time, and so their demand for phone service has already reached saturation. Yet that is not the case, since the amount of calling per person has grown vigorously, over 55% between 1980 and 1997. (This includes fax and modem calls). Individuals and institutions have voted with their pocketbooks for low utilization, presumably because it was perceived to provide higher quality of life and work.

In data traffic, growth has been much faster than in telephone lines, and links have tended to get saturated. This is often cited as a prototypical “tragedy of the commons” problem. Yet the situation is less clear than it might seem. Growth has been orderly. Fig. 6 shows the average traffic from the public Internet to the University of Waterloo. (See [CoffmanO, Odlyzko4] for more details.) Although the capacity of the link has had several sudden jumps, usage has grown at a pretty steady 100% a year. Similar steady growth rates have been seen in other networks, see [CoffmanO]. Thus these networks have not in general had to cope with sudden surges in demand that saturated new capacity as soon as it became available. Even when such surges materialized (as they did at the University of Waterloo when student dorms were hooked up to the campus Ethernet), they were contained by simple local measures.

The orderly, although very rapid, growth in data traffic is probably caused by a combination of several factors. Available content, user skills and awareness, and the network infrastructure all have to develop in parallel. The last factor is especially important. Users don’t just go and exploit the Internet. Except for cases of malice, they usually need to have a substantial local infrastructure in order to make use of the wider networks, and this limits the growth in demand for connectivity.

Not only is growth of data traffic steady, actual traffic is generally predictable once it is sufficiently aggregated. Several of the graphs in this paper, as well as many of those in [Odlyzko2, Odlyzko4] combine displays of traffic for several days. It is noteworthy that the traffic patterns are generally consistent from week to week, with Monday through Thursday usually behaving the same, and Friday, Saturday, and Sunday each having its own particular load graph. This is the same behavior that has been observed on the switched voice network, and goes counter to the claim that data traffic is chaotic and is only constrained by congestion.

The "bursty nature of data traffic" is not the culprit behind low utilization rates of data networks. Data traffic does not smooth out as well as switched voice traffic, and it shows long range dependence [FeldmannGWK, LelandTWW]. However, that does not mean that high utilization cannot be achieved. Figures 4 and 5 show that it can. In Fig. 5, we see essentially full utilization over 9 hours during the business day, and in Fig. 4, for a much smaller link, more than 80% utilization over a comparable period. Most of data transport uses TCP, which fills available bandwidth and can produce high load factors. Thus low utilization has to come from a different source, and the next section will be devoted to this topic. As a preliminary step, let us note that high utilization carries a penalty. For example, during the hours of peak usage in Fig. 4, the average packet drop rate was around 5%, so service quality was substandard (and the throughput figure was deceptively high, since it included substantial retransmissions). High utilization thus had to come from explicit or implicit choices about the quality of data transport to be provided.

3. Economic efficiency versus engineering efficiency

Economic efficiency means satisfying customers' demands. Engineering efficiency means providing a service with the minimal amount of resources. The two are often in conflict. The switched voice system provides a compromise that is more efficient than the Internet in both economic and engineering aspects. It gives a high quality service to its customers any time they want to use it, and yet manages to have a higher utilization rate of the transmission facilities. It does so at a rather high monetary cost, though, and offers little flexibility.

There are basic reasons for low utilization rates of data networks that are caused by the rapid growth rate of data traffic and the economies of scale in purchases of transmission links. These reasons are discussed in [Odlyzko2]. However, the main cause of this low utilization rate of private line networks (and thus of most of the Internet) is the attempt to satisfy user needs. Business customers with traffic patterns such as the one in Fig. 3 could carry all their traffic on 56 Kbps lines instead of T1s. Their decision not to do so shows what they find desirable and affordable. The manager of a branch lab of a major software producer described the private line from that lab to company headquarters as follows:

> I see peak bandwidth as the basic commodity I buy. ... When we had a 256Kb data line

> it was too slow (it interfered with productivity). With a T1 line, no one has complained. I guess our T1 line is less than 1% utilized. ... I would not go for a T3 line (it would not improve our productivity) but I would not cut back on the T1 line.

Note that this manager does not care about average utilization of the line, in common with most people in similar positions. (The previous section noted already how seldom utilization data is collected.) What he cares about is that he and his coworkers get what they need quickly. Thus latency is the key issue. However, it is not packet latency, the concern of most of networking literature, that matters. It is transaction latency, the time it takes to complete whatever one cares about (transmitting an email message or downloading a Web page), that is important. Transaction latency can be lowered to a large extent by going for higher bandwidth.

Since most of the traffic on the Internet is HTTP, it is often dismissed as just Web-surfing. However, Web-surfing includes customers downloading product information or placing orders, as well as employees accessing corporate databases. Therefore it often deserves high priority. (In general, trying to assign priorities to packets based just on the application is unlikely to be productive.) However, Web-surfing does produce traffic patterns such as that of Fig. 3, with low average utilizations, when it meets user demands.

An important feature of the quote above is that it did not refer to any quantitative measures of network performance. There are certainly studies of what performance is required for tasks such as transaction processing, and how different network technologies compare in satisfying those requirements [Cavanagh]. It appears, though, that increasingly interactions with networks are becoming more complex, and can only be judged by subjective criteria of user satisfaction.

The importance of the subjective factor in judging network performance is also apparent in the choices of circuit speeds. They are usually chosen in simple multiples of some basic rate (such as 56, 128, 256, 512 Kbps) even when intermediate speeds are available. The human perceptual system operates on a logarithmic scale, and therefore it requires large steps to achieve a noticeable improvement in performance.

Low utilizations of data networks should not be surprising. Note that the family car and the phone set are used on average around 4% of the time. The Pentium III PC on an office desktop is idle most of the time, and does little that a 486 machine could not do. However, when it is called upon to typeset a document or recalculate a spreadsheet, it can do so much faster than the 486 machine, and that justifies its purchase. Comparisons of PCs never discuss how much they will be used, and instead concentrate on benchmarks showing how much time those PCs take to perform various tasks. What the resulting low utilizations mean is that the lightly utilized resource is inexpensive enough compared to the value people place on its availability and their time. Note that two or three decades ago, computers were largely mainframes, were kept in computer centers, and were run at high utilization rates. By moving to PCs we have lowered the utilization of the equipment, but have provided much more flexibility.

Low utilizations of corporate lines, such as that of Fig. 3, show the value of high quality transmission. On the other hand, the high utilizations of international links, such as that of Fig. 4, show the limits of what even corporations are willing to pay. When links are as expensive as they are now across the Pacific or the Atlantic, it appears that IT managers either implicitly or explicitly decide to make their users put up with congested networks.

The high utilizations of ISPs links most likely reflect a combination of extreme price sensitivity of the residential modem customers and of the difficulty those customers would have in deriving any benefit from uncongested links. The high latencies and low transmission speeds of modems would guarantee low quality experience in any event.

The comparison of utilization rates of trans-oceanic and domestic corporate links suggests that if prices come down, users will opt for higher quality transmission and thereby lower utilizations. This is what appears to have happened in the LAN environment. I have much less data here than for long haul networks, especially historical data. However, many people say that average utilizations of LANs a decade ago were in the 5% to 10% range. (It is not clear, though, whether this refers to business hours alone or a full day). That is also consistent with a few scraps of hard data, such as the statistics for the Bellcore LANs in the early 1990s in [LelandTWW]. On the other hand, today LAN utilizations at the institutions that I was able to obtain the data for (which is not many, and consist primarily of those places that use the MRTG tool [MRTG]), tend to be around 1%. (This is not to say that this is universal, as there are insitutions with higher rates.) Further, in those institutions that have both 10 Mbps and 100 Mbps Ethernets, the average utilizations of the 100 Mbps links tend to be around half or a third of the rate of 10 Mbps links. LAN equipment has decreased drastically in price, and so it has become easier to satisfy people's desire for bursty transmission. It has also become less expensive to solve problems by tossing bandwidth at them instead of using scarce and expensive network manager time. There is an attempt to economize, and connections get upgraded to 100 Mbps Ethernet only when there is a need for such speeds. In general, though, other pressing network problems are clearly more important than maximizing utilization of network bandwidth.

The general conclusion about low utilization rates of corporate networks is that they are not a sign of waste, but of the value of high quality data communications and of the complexity of running networks. However, these low rates do provide a substantial business opportunity, as will be discussed in Section 5.

4. ATM and QoS in current data networks

Utilization rates and utilization patterns may explain why some technologies have flourished and others have fallen by the wayside. During the 1980s, there was a serious competition between Ethernet and Token Ring technologies for the LAN market. One of the claimed advantages of Token Ring was supposed to be its ability to carry a higher fraction of its peak capacity in routine operations. Another advantage was that Token Ring had QoS support built in. Still, Token Ring lost out. The technical and economic issues were complex, but the primary reason it lost out appears to be that it was more complicated. Its greater engineering efficiency did not save it, and it is easy to see why. In an environment like that of Fig. 3 (note that today LANs operate at about the utilization rate of that figure, although 10 years ago they probably operated at higher rates) average throughput is basically irrelevant. It is only the peak rate that matters.

Similarly, it appears that ATM has failed to take off largely because it is inappropriate for most of today's networks. ATM was conceived with the idea, inspired by voice and multimedia, that traffic would consist of long-lived flows with reasonably well defined bandwidth, latency, and jitter requirements. However, that is not what we have on our networks today. Most of the traffic consists of Web page downloads that are small, and what matters is how quickly the entire page is delivered. Therefore ATM is irrelevant from users' perspective. It finds its greatest applications in core networks, where aggregate traffic flows do resemble the traffic conditions for which ATM was designed.

Most QoS measures (see [FergusonH] for a survey) are also of doubtful utility in the current environment. In an environment such as that of figures 2, 4, and 5, it is intuitively appealing to create a special lane for high priority traffic. In the environment of Fig. 3, though, which is much more representative of the universe of data networks today than figures 2, 4, and 5, that is much more questionable. High priority and low priority traffic do go through just about whenever they need to. Even when two demands coincide, it is just as likely that they will both be of high priority, so no prioritization scheme would help. Applications would still have to cope with occasional congestion.

Bandwidth reservations are especially questionable in the environment of Fig. 3. In a recent work on the guaranteed service features being developed for the vBNS high speed network, the authors say that [SongCW]

> ... we give a relatively firm commitment of bandwidth. The word "relatively" suggests

that we have not excluded the use of bandwidth overbooking for the benefit of statistical sharing. However, we must make sure that an equivalent throughput is not compromised.

For traffic like that in Fig. 3, it is impractical to avoid bandwidth overbooking, as firm guarantees would involve tiny utilization rates and astronomical costs. If we allow overbooking, though, then we are basically dealing with a best-effort network for the high priority traffic, with low utilization providing an expected high quality of transmission. That, however, can be accomplished by much simpler methods, such as the Paris Metro Pricing scheme of [Odlyzko1], to be discussed later.

So far I have been arguing that ATM and QoS are inappropriate for today's Internet. However, there are limits to these arguments. Both ATM and QoS might become much more relevant if the Internet changes (as will be discussed in Section 7). ATM already plays a big role in the core of the network, as the basic transport mechanism used by the backbones to carry IP traffic. The arguments in this paper do not say anything about the advantages of ATM in that context as compared to packet-over-SONET or other technologies. The focus here is on how the Internet appears to the users and system administrators at the edges of the network, and from their point of view ATM does not offer serious advantages, and does have serious shortcomings.

I am also not suggesting that QoS is useless. It will be vital in the many situations where there are stringent bandwidth constraints, such as at the edges of the network, especially in the wireless arena. Further, techniques such as Weighted Fair Queueing (WFQ) and Random Early Detection (RED) can be invaluable in controlling congestion, and can be implemented inside the network without destroying the exceedingly valuable stateless nature of the Internet, and without complicating the lives of users or even the lives of users' system administrators. Such simple techniques might suffice to provide congestion controls for networks that offer uniformly high quality of service to all traffic.

Some congestion responsive techniques are essential. Even the highest priority transmissions will occasionally have to compete for limited bandwidth with other transmission of similar priority. Therefore essentially all applications will have to possess some mechanism for limiting their bandwidth demands in the presence of congestion. (The few exception that do require absolute bandwidth guarantees are likely to stay with private lines or some special channels on the public network, as has historically been the case.) Since the human perceptual system is insensitive to small changes, very simple, although suboptimal, algorithms like the ones in TCP, WFQ, and RED should suffice, especially if the network is not heavily loaded.

Why do I keep insisting on keeping the Internet simple? It is not just that the Internet is too complicated; it is that all the other things that rely on the Internet are too complicated! To make the applications that people care about interact effectively with most of the QoS schemes that are being proposed would be an undesirable additional burden.

5. Lessons to be learned from switched voice networks

The Internet community appears to be learning precisely the wrong lessons from the switched voice networks. The ideas that data moves in flows and that precise guarantees of service quality are required appear to be inappropriate for the Internet, and are leading to development efforts that are likely to be of little use.

On the other hand, there are valuable lessons that can be learned from the traditional phone networks. One of them is simplicity. The Internet is just too complicated. Not only is it expensive (as the figures in the first table in Section 2 show), but it requires many experts at the edges to keep it running. A telling sign is that only about two thirds of U.S. households that have PCs also have Internet accounts. In contrast, the phone system has developed simple user interfaces and standards that allow billions of people to use it easily. There is an unavoidable conflict between simplicity and flexibility, and the phone system is not flexible enough to survive in its present form. On the other hand, the Internet will surely have to become simpler to attract more users. Furthermore, there should be ways to do that without sacrificing much flexibility, by moving towards standards such as IPv6, providing more security inside the network, and so on.

Another important lesson that the Internet can learn from the switched voice network is about economics. It is worth remembering that initially the phone was an extremely expensive technology. A hundred years ago, monthly charges for phone service in New York City amounted to half of the average monthly wage. Yet with diligent effort, phone service has become affordable for the masses. As is discussed in [Odlyzko2], some factors that went into the high utilization of the switched voice network (which was a substantial, although not the dominant contributor to the lowering of costs) are the slow and predictable rate of growth of voice traffic, factors that do not apply to the Internet. However, there are other factors that can also be taken advantage of in data networks, such as complementary traffic patterns and statistical aggregation.

Fig. 7 shows the traffic patterns on the switched voice network separated into the residential and business customers. By carrying both types of calls on the same network, over 40% of capacity is saved. (See [Odlyzko4] for a more detailed discussion.) Yet in data networks, we have corporate private line networks that are used mostly just during the business day. We also have ISP networks used primarily by residential customers, who use them largely in the evenings and on weekends. (See [Odlyzko2, Odlyzko4] for graphs of usage patterns.) Those networks are disjoint. If they were combined, they could provide a uniformly high quality of service to all current traffic at all times and do so with lower total capacity than the separated networks have now.

Aggregation of traffic is another great opportunity for the Internet. If data networks were as congested as is widely believed (supposedly with 70% peak hour utilizations on most circuits, cf. [Odlyzko2]) there would be little that could be done. However, corporate users purchase lines for their burst capacity, so utilizations are low, and the situation is vastly different. A dozen lines like the one in Fig. 3 can be aggregated onto a single T1 with all dozen users obtaining essentially the same service as with their own lines, since their traffic peaks are uncorrelated. The line in Fig. 3 has lower utilization than the average for private line networks, but experiments with combining traffic traces for different lines show that statistical aggregation has great promise for producing higher utilization. It appears that by aggregating traffic from different private lines onto larger common links, one should be able to produce average utilizations at least two or three times as high as the 3-5% range that is typical of corporate networks. (Note that this is just about the operating range of current Internet backbones, which do appear to provide high quality services.) Furthermore, this would still be corporate traffic only, which is significant only during the regular business day. By mixing in the complementary household traffic patterns, it might be possible to raise utilization levels by factors of four compared to the current levels, and provide uniformly high quality service for all traffic.

A large common network would also serve to reduce the enormous administrative costs of running separate corporate networks with point-to-point connections. It might also make it easier to provide some form of the dynamic routing that has been so important in reducing costs and increasing utilization rates in switched networks [Ash]. In general, the cost advantages of a single common network have been demonstrated overwhelmingly with electric and other utilities, but are still waiting to be realized in data networks.

The costs advantages of a large common data network are already being partially realized by the Internet backbones. A large part of the reason for the huge differential in transport costs shown in the first table in Section 2 is precisely because the public Internet aggregates many sources of traffic, both business and residential, does operate large links, and centralizes many network management functions. However, to be a convincing substitute for private line networks, the public Internet will have to provide higher quality and security. Quality on many backbones already appears to be sufficient, and VPN (virtual private network) technology exists to assure necessary security.

6. Differential service, usage sensitive charging, and Paris Metro Pricing

The main objection to QoS is that it would complicate the Internet. As a simple example, if traffic priority were to be set according to application, then either encryption technologies such as IPSec, which conceal packet payloads, would have to be banned, or else an elaborate additional signaling scheme would be required. The ideal solution is to keep the Internet as close to a dumb network as possible, one that just accepts packets and delivers then to the destination. As in the switched voice network, this might require building in more intelligence inside the network (to deal with security issues, for example), but it should be intelligence that is invisible to the end users.

Although the ideal of a simple network is very attractive, there are advantages to more complicated systems. In particular, several levels of service quality would lead to more efficient use of network capacity. Today, all corporate traffic, including delay-insensitive file transfers, receive the same high quality service, and in addition, networks are idle most of the time. On the other hand, on the public Internet, there are choke points that lead to frustration with the "World Wide Wait," yet there is no way for users to obtain better service. A network offering different levels of service to different types of traffic therefore has attractions. All the standard economic arguments argue in favor of such a solution [MacKieM, McKnightB, Varian]. Charges do not have to be high to have a noticeable effect on human behavior. Still, the question is whether the gains are worth the cost.

Any universal differential service scheme will almost inevitably involve a usage sensitive pricing system. So far such pricing has been applied only in limited cases, most notably in countries such as Chile, New Zealand, and Australia, where communication costs to the U.S. (which is where most of their international and often of their total Internet traffic comes from) are very high. However, with traffic increasing and transmission costs still growing, charging per byte is spreading. Even corporate IT managers are implementing it, in order to allocate costs to divisions. So far, though, all such charges are simply for each byte sent or received (although sometimes it is just for bytes received from international links). This has obvious attractions all by itself, as it would promote fairness. As it is, flat rate pricing means that corporations that use their Internet links lightly pay the same as ISPs and universities that transmit at high fractions of the capacity of their connections. This is already leading to modifications of the standard flat rate approach, with ISPs often facing higher charges than other users.

Even with simple per-byte charging, there are problems caused by the lack of a reliable measurement infrastructure. (As a simple example, in the two years' of data I have received for the link of Fig. 3, about 15% of the values are missing, and there are some obviously erroneous entries, such as some indicating data transfer rates of over 100 Mbps.) There are also questions of fairness, since packets that are dropped further on in their journey get counted and charged for, and this overcharge gets worse precisely when the network is congested and provides the worst service.

With differential services, the accounting difficulties increase. Traffic counts would have to become more robust, and there would surely have to be a substantial infrastructure to allow either the sender or the receiver to pay. If it turns out that it is worth modifying the Internet to that extent, then, in a compromise with the overwhelming need for simplicity, I propose using the Paris Metro Pricing (PMP) scheme of [Odlyzko1]. In PMP, the backbones would be divided into several logically separate channels, each with a different price per byte. Users would be free to select for each packet which channel to send it on. The expectation is that the more expensive channels would attract less traffic, and therefore would be much less congested. The details of PMP and in particular further justifications for it are contained in [Odlyzko1]. The basic intuition of PMP is tohave a scheme that is as simple as possible. If there are going to be different service levels on the Internet, there will have to be differential pricing. In that case, though, why not take advantage of that pricing to deal with congestion control, and preserve the stateless nature of the Internet? PMP keeps the pricing part, which seems unavoidable, and dispenses with everything else. My expectation is that if a differentiated service system is introduced on the Internet, it will eventually evolve towards PMP (or degenerate towards it, depending on one's view).

7. The rapidly evolving Internet

The preceding sections dealt with the current Internet, and explained the utilization rates and patterns that dominate on it. They suggested why ATM is not a solution to the bulk of the problems on the current Internet, and that most of the QoS measures are also of questionable applicability. Will this also be true on the future Internet? That will depend on prices of transmission and the nature of data traffic.

Prices: In Section 2, we saw the crucial role that pricing plays in utilization patterns of data links. If prices of transmission continue to go up, as they have been doing recently, then most data networks might be used like the corporate trans-Pacific link of Fig. 4, and then differentiated services might become necessary.

Predicting data transport prices is hazardous. As an example, in 1993 Irvin [Irvin] published a study of private line prices in the U.S.. He constructed two plausible models that fit the historical record well up to that point, and used them to predict a continuation of the declining trend in prices. Unfortunately, that happened to be just the time when prices hit their absolute minimum. Since 1992, prices have increased over 50%, and in 1999 were three times as high as Irvin's models predicted. (See [CoffmanO] for a graph of historical private line prices.) However, as is discussed in [FishburnO, Odlyzko4], we are entering a new era, with new technologies and new competitors. Until recently the Internet was so small, that even its 100% per year growth rate did not affect the much smaller growth rate of the underlying telecommunications network. Very soon, though, that growth rate for the Internet will mean a similar growth rate for the entire network, which is likely to lead to rapid introduction of new equipment and a rapid decline in prices. (It is also likely to lead to an increase in total revenues from data transport, in analogy to what has been happening in microprocessors, hard disks, and other high tech areas.) The paper [FishburnO] presents some simple economic models which demonstrate that when prices fall rapidly, a lightly loaded network with uniformly high quality of service can often be economically optimal.

Unfortunately, it is impossible to predict how soon transmission prices will start declining. Even when they do, it is not clear how fast they will do so.

Nature of Internet traffic: ATM and many QoS schemes were dismissed in an earlier section as irrelevant for today's Internet on the grounds that current traffic does not consist of extended flows with well-defined rate requirements. However, the Internet can change extremely rapidly. After all, hardly anyone had heard of the Web half a dozen years ago, and yet it is now the dominant application on the Internet. What if multimedia traffic begins to dominate? Under those conditions ATM might be the right answer, but this appears an unlikely scenario. St Arnaud [StArnaud, StArnaudCFM] has already argued convincingly that multimedia is not the future of the Internet, and instead computer-to-computer communication will dominate. To his arguments I would add another one, namely that there are limits on how much multimedia material people will want to be consume. On the other hand, as long as computers keep growing in numbers and power, their potential communication demands will grow, and are likely to fill available bandwidth. However, my views differ from those of St Arnaud in an important respect. He expects that the dominant computer-to-computer traffic will be insensitive to delay and jitter. My prediction is that while that is true in principle, it will not be so in practice. I expect that future computer-to-computer traffic will in some important respects be similar to today's Web surfing. It is likely to be overwhelmingly generated in response to human demands. An example might be a surgeon sending data from the operating room to a specialist for a consultation, who in turn sends out software agents to scour databases for similar data. Another example might be a salesman trying to generate a quote for a prospective customer, and kicking off a flurry of communications between the ERP (enterprise resource planning) systems of his company and its suppliers. All such communications will have the same feature that we see on the Web today, features that corporations spend much for, namely delivering results as quickly as possible. Networks engineered to provide that level of service should be able to provide low latency and jitter as automatic byproducts with only a few simple mechanisms such as Fair Queueing that are completely invisible to the users.

8. Conclusions

The current state of the entire Internet, and the utilization patterns on it, show that ATM and QoS are of limited utility. On the other hand, there are huge inefficiencies in the Internet that can be alleviated without intrusive measures that affect users. There is enough data transport capacity to provide high quality transmission for all current traffic as well as allow for substantial growth, if only all data networks were combined into a common network. There is also overwhelming evidence of users' desire and willingness to pay for high quality services. How the Internet will evolve is likely to be determined by the trends in transmission prices and in the nature of traffic. Given the huge potential costs to the entire IT system of any modifications to the Internet, though, simplicity will surely be at a premium.

If prices do decrease sufficiently rapidly compared to traffic growth, then it might be economically optimal to continue with the present system of flat rate pricing, and to provide high quality service to all packets. If prices do not decline sufficiently, then something like the "expected usage profile" proposal of [Odlyzko4] might be appropriate. In this scheme, all traffic would still get the same high quality transmission. However, users would pay ISPs according to their past usage (based on sampling, say), with lowered rates for sending their traffic at night, say, or for making sure most of the traffic is congestion-sensitive (such as TCP). Finally, if traffic growth outpaces price declines, then some version of Paris Metro Pricing might be called for as a last resort.

Acknowledgements: I thank my coauthors, Kerry Coffman and Peter Fishburn, the many people who have already been acknowledged in previous papers [CoffmanO, FishburnO, Odlyzko2, Odlyzko3, Odlyzko4] on which this one is based, as well as David Charlton, Jon Crowcroft, Tim Dorcey, Vince Fedele, Bob Frankston, Jim Gray, Eric Grosse, Fotios Harmantzis, Alan Kotok, Jacek Kowalski, Michael Lesk, Chris Ramming, Frank Schmidt, Bill St Arnaud, Ed Vielmetti, Ivan Vukovic, Damon Wischik, and Ed Zajac for their comments.

References

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

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

[CoffmanO] 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⟩.

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

[FCC] U.S. Federal Communications Commission, *Trends in Telephone Service,* Sept. 1999. Available from ⟨http://www.fcc.gov/ccb/stats⟩.

[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⟩.

[FergusonH] P. Ferguson and G. Huston, *Quality of Service: Delivering QoS on the Internet and in Corporate Networks,* Wiley, 1998.

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

[Gibbs] W. W. Gibbs, Transportation's perennial problems, *Scientific American,* 277, no. 4 (Oct. 1997), pp. 54-57. Available at ⟨http://www.sciam.com/1097issue/1097gibbs.html⟩.

[Huitema] C. Huitema, The required steps towards high quality Internet services, unpublished Bellcore report, 1997.

[Irvin] D. R. Irvin, Modeling the cost of data communication for multi-node computer networks operating in the United States, *IBM J. Res. Develop.* 37 (1993), pp. 537-546.

[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.

[MacKieM] J. MacKie-Mason, Telecom Information Resources on the Internet, Web site with links to online sources, ⟨http://china.si.umich.edu/telecom/telecom-info.html⟩.

[McKnightB] L. W. McKnight and J. P. Bailey, eds., *Internet Economics,* MIT Press, 1997. Preliminary version of many papers available in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[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⟩.

[Odlyzko1] A. M. Odlyzko, Paris Metro Pricing for the Internet, in *Proc. ACM Conference on Electronic Commerce (EC-99),* ACM, 1999, pp. 140-147. Based on a 1997 unpublished manuscript, A modest proposal for preventing Internet congestion. Both available at ⟨http://www.research.att.com/∼amo⟩.

[Odlyzko2] A. M. Odlyzko, Data networks are lightly utilized, and will stay that way. Available at ⟨http://www.research.att.com/∼amo⟩.

[Odlyzko3] A. M. Odlyzko, Smart and stupid networks: Why the Internet is like Microsoft. *ACM netWorke* 2(5) (Dec. 1998), 38-46. Available at ⟨http://www.research.att.com/∼amo⟩.

[Odlyzko4] A. M. Odlyzko, The economics of the Internet: Utility, utilization, pricing, and Quality of Service. Available at ⟨http://www.research.att.com/∼amo⟩.

[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/⟩.

[SchaferV] A. Schafer and D. Victor, The past and future of global moobility, *Scientific American*, vol. 277, no. 4 (Oct. 1997), pp. 58-61. Available at ⟨http://www.sciam.com/1097issue/1097schafer.html⟩.

[SongCW] C. Song, L. Cunningham, and R. Wilder, Quality of Service development in the vBNS, *IEEE Communications Magazine* 36 (May 1998). Available at ⟨http://www.vbns.net/presentations/papers/⟩.

[StArnaud] B. St Arnaud, The future of the Internet is NOT multimedia, *Network World*, Nov. 1997. Available at ⟨http://tweetie.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://wwww.canet2.net⟩.

[Steinberg] S. G. Steinberg, Netheads vs. Bellheads, *Wired,* 4, no. 10 (Oct. 1996), pp. 144-147, 206-213. Available at ⟨http://wwww.wired.com/wired/4.10/features/atm.html⟩.

[Varian] H. R. Varian, The economics of the Internet, information goods, intellectual property and related issues, reference Web pages with links, ⟨http://www.sims.berkeley.edu/resources/infoecon/⟩.

[WiltziusBD] D. Wiltzius, L. Berc, and S. Devadhar, BAGNet: Experiences with an ATM metropolitan-area network, *ConneXions,* vol. 10, no. 3, (March 1996). Available at ⟨http://www-itg.lbl.gov/BAGNet.html⟩.

Figure 1: The analogy between the Internet and the switched voice network is weak and often misleading.

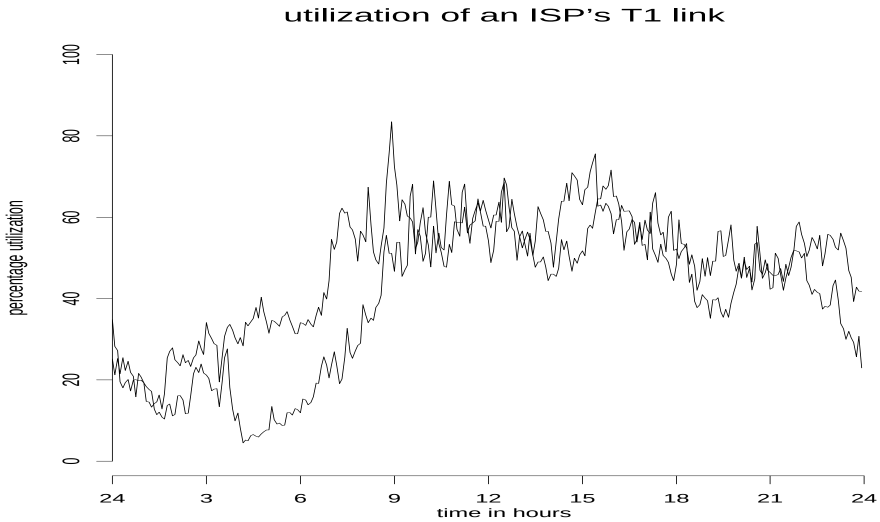

Figure 2: Traffic on an ISP's T1 line on Tuesdays of April 14 and 21, 1998. 5-minute averages.

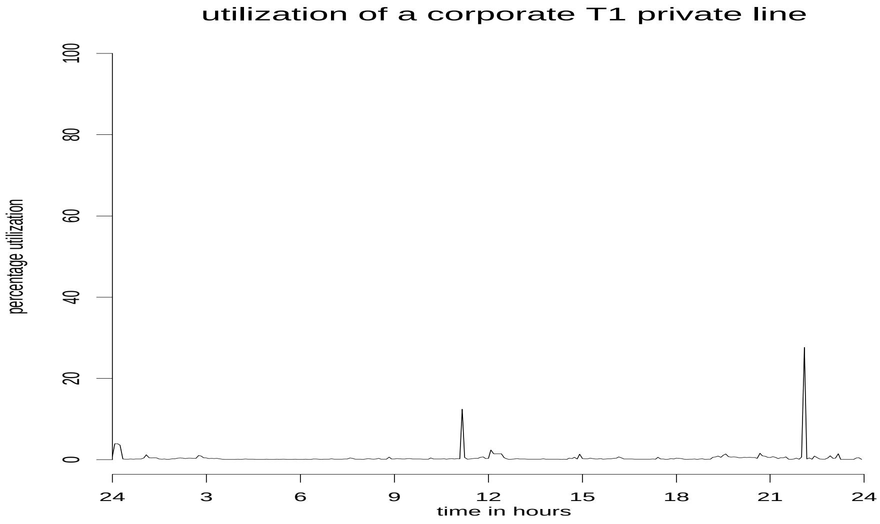

Figure 3: Traffic on a corporate T1 line in the continental U.S. during Thursday, May 28, 1998. 5-minute averages.

Figure 4: Traffic from the U.S. to the Far East on a corporate 128 Kbps line during a weekday. The peak traffic hours between 1800 and 2400 coincide with the busy hours in Far East location. Hourly averages.

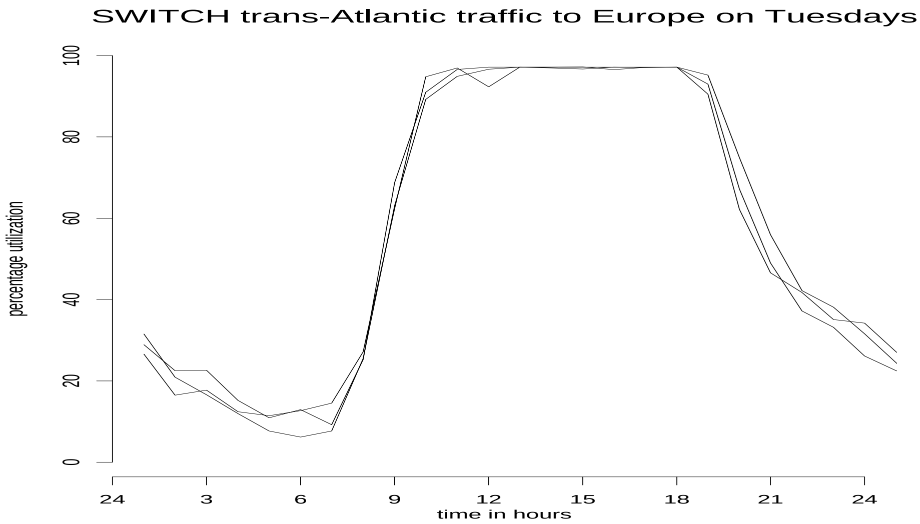

Figure 5: Traffic on the 8 Mbps link from the U.S. to SWITCH, the Swiss academic and research network, during Tuesdays of February 3, 10, and 17, 1998. Hourly averages, Swiss time. By permission of SWITCH.

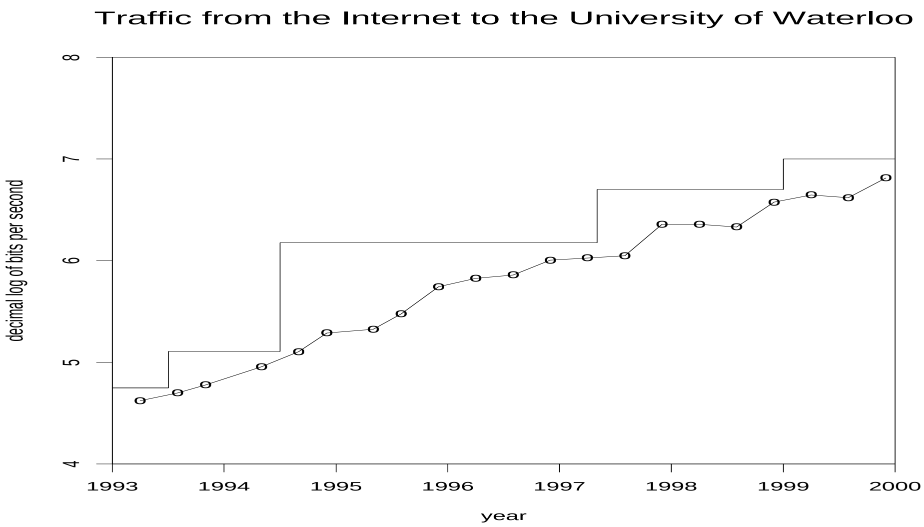

Figure 6: Traffic on the link from the public Internet to the University of Waterloo. The line with circles shows average traffic during the month of heaviest traffic in each school term. The step function is the full capacity of the link. By permission of University of Waterloo.

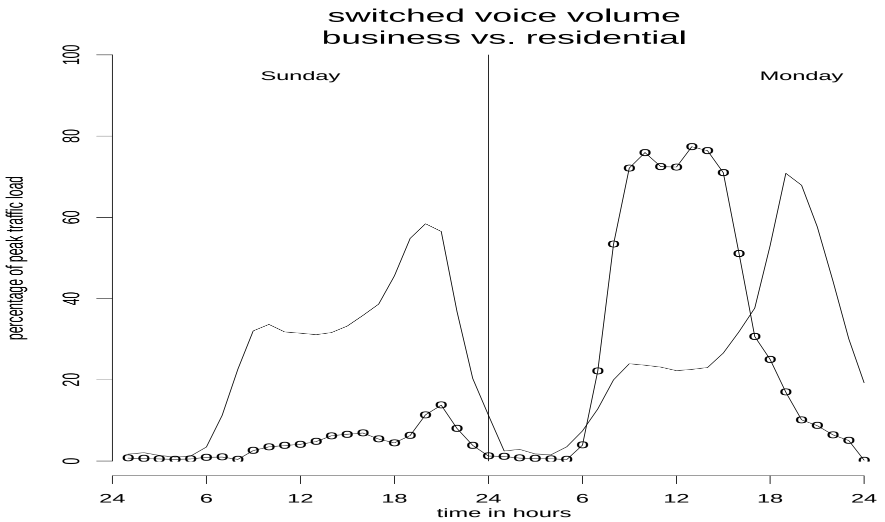

Figure 7: Residential (thin line) and business (line with circles) voice traffic on U.S. long distance switched voice networks, as percentage of peak traffic on those networks.