# The economics of the Internet: Utility, utilization, pricing, and Quality of Service

Andrew Odlyzko

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

July 7, 1998.

**Abstract.** Can high quality be provided economically for all transmissions on the Internet? Current work assumes that it cannot, and concentrates on providing differentiated service levels. However, an examination of patterns of use and economics of data networks suggests that providing enough bandwidth for uniformly high quality transmission may be practical. If this turns out not to be possible, only the simplest schemes that require minimal involvement by end users and network administrators are likely to be accepted. On the other hand, there are substantial inefficiencies in the current data networks, inefficiencies that can be alleviated even without complicated pricing or network engineering systems.

## 1. Introduction

The Internet has traditionally treated all packets equally, and charging has involved only a fixed monthly fee for the access link to the network. However, there are signs of an imminent change. There is extensive work on provision of Quality of Service (QoS), with some transmissions getting preferential treatment. (For a survey of this area and references, see the recent book [FergusonH].) Differential service will likely require more complicated pricing schemes, which will introduce yet more complexity.

The motivation behind the work on QoS is the expectation of continued or worsening congestion. As Ferguson and Huston say (p. 9 of [FergusonH])

> ... it sometimes is preferable to simply throw bandwidth at congestion problems. On a global scale, however, overengineering is considered an economically prohibitive luxury. Within a well-defined scope of deployment, overengineering can be a cost-effective alternative to QoS structures.

The argument of this paper is that overengineering (providing enough capacity to meet peak demands) on a global scale may turn out not to be prohibitively expensive. It may even turn out to be the cheapest approach when one considers the costs of QoS solutions for the entire information technologies (IT) industry.

Overengineering has been traditional in corporate networks. Yet much of the demand for QoS is coming from corporations. It appears to be based on the expectation that overengineering will not be feasible in the future. “There's going to come a time when more bandwidth is just not going to be available... and you'd better be able to manage the bandwidth you have,” according to one network services manager [JanahTD].

The abandonment of the simple traditional model of the Internet would be a vindication for many serious scholars who have long argued that usage-sensitive pricing schemes and differential service would provide for more efficient allocation of resources. (See [McKnightB] for references and surveys of this work.) The need for usage-sensitive pricing has seemed obvious to many on the general grounds of “tragedy of the commons”. As Gary Becker, a prominent economist, said recently (in advocating car tolls to alleviate traffic jams and the costs they impose on the economy [Becker]):

> An iron law of economics states that demand always expands beyond the supply of free goods to cause congestion and queues.

It may indeed be an iron law of economics that demand for free goods will always expand to exceed supply. The question is, will it do so anytime soon? An iron law of astrophysics states that the Sun will become a red giant and expand to incinerate the Earth, but we do not worry much about that event. Furthermore, the law of astrophysics is much better grounded in both observation and theoretical modeling than the law of economics. For example, consider Table 1 (based on data from tables 12.2 and 18.1 of [FCC]). It shows a dramatic increase in total length of toll calls per line. Such calls are paid by the minute of use, and their growth was presumably driven largely by decreasing prices, as standard economic theory predicts. On the other hand, local calls in the U.S. (which are almost universally not metered, but paid for by a fixed monthly charge, in contrast to many other countries) have stayed at about 40 minutes per day per line in the last two decades. The increase of over 62% in the total volume of local calls was accompanied by a corresponding increase in the number of lines. There is little evidence in this table of that “iron law of economics” that causes demand to exceed supply, and which, had it applied, surely should have led to continued growth in local calling per line. (There is also little evidence of the harm that Internet access calls are supposed to be causing to the local telephone companies. This is not to say there may not have been problems in some localities in California, for example, or that there won't be any in the future. However, at least through 1996 the increasing use of networked computers has not been a problem in aggregate.)

An obvious guess as to why we have stable patterns of voice calls is that people have limited time, and so, with flat-rate pricing, their demand for local calls had already been satisfied by 1980. However, that is not what the data in Table 1 shows. While the total volume of local calls went up almost 63% between 1980 and 1996, population increased only 16.5%, so minutes of local calls per person (including modem and fax calls) increased by 40%. Thus demand for local calls has been growing vigorously, but it was satisfied by a comparable increase in lines. Families and businesses decided, on average, to spend more on additional phone lines instead of using more of the "free good" that was already available. Somewhat analogous phenomena appear to operate in data networking, and may make it feasible to provide high quality undifferentiated service on the Internet.

In data networks, at first sight there does appear to be extensive evidence for that "iron law of economics." Comprehensive statistics are not available, but it appears that Internet traffic has been doubling each year for at least the last 15 years, with the exception of the two years of 1995 and 1996, when it appears to have grown by a factor of about 10 each year [CoffmanO]. Almost every data link that has ever been installed was saturated sooner or later, and usually it was sooner rather than later. An instructive example is that of the traffic between the University of Waterloo and the Internet, shown in Fig. 1. The Waterloo connection started out as a 56 Kbps link, was upgraded to 128 Kbps in July 1993, then to 1.5 Mbps in July 1994, and most recently to 5 Mbps in April 1997 [Waterloo]. Based on current usage trends, this link will be saturated by the end of 1998, and will need to be upgraded, or else some rationing scheme will have to be imposed. (A partial rationing scheme is already in effect, since the link is heavily utilized and often saturated during the day.)

The University of Waterloo statistics could be regarded as clinching the case for QoS and usage-sensitive pricing. They show a consistent pattern of demand expanding to exceed supply. However, I suggest a different view. The volume of data traffic in Fig. 1 grows at a regular pace, just about doubling each year. The 12-fold jump in network bandwidth from 128 Kbps to 1.5 Mbps in July 1994 did not cause traffic to jump suddenly by a factor of 12. Instead, it continued to grow at its usual pace. The students did not go wild, and saturate the link by downloading more pictures. Similarly, statistics for traffic on the Internet backbones show steady growth, aside from an anomalous period of extremely rapid increase in 1995 and 1996 [CoffmanO], and the NSFNet backbone in particular had traffic almost exactly doubling from the beginning to 1991 to the end of 1994. And should an increase in traffic be wrong? We are on the way to an Information Society, and so in principle we should expect growth in data traffic.

How much capacity to provide should depend on the value and the price of the service. To decide what is feasible or desirable, we have to consider the economics of the Internet. Unfortunately, the available sources (such as those in the book [McKnightB] or those currently available online through the links at [MacKieM, Varian]) are not adequate. The information they contain is often dated, and it usually covers only the Internet backbones. However, these backbones are a small part of the entire data networking universe. Sections 2 to 6 attempt to partially fill the gap in published information about the economics of the Internet.

Fig. 2 is a sketch of the Internet, with the label "Internet" attached just to the backbones (as the term is often used). As will be shown in Section 2 (based largely on the companion papers [CoffmanO, Odlyzko2]), these backbones are far smaller than the aggregate of corporate private line networks, whether measured in bandwidth or cost (although not necessarily in traffic). (See Table 2 for the sizes of data networks in the U.S. It is taken from [CoffmanO], and effective bandwidth, explained in that reference, compensates for most data packets traveling over more than a single link.) The private line networks, in turn, are dwarfed by the LANs (local area networks) and academic and corporate campus networks. Most of the pricing and differentiated service schemes that are being considered, though, are aimed at Internet backbones or private line WAN links. We need to consider how they would interact with the other data networks and the systems and people those networks serve.

Most of the effort on QoS schemes is based on the assumption of endemic congestion. However, when we examine the entire Internet, we find that most of it is uncongested. That the LANs are lightly used has been common knowledge. However, it appears to be widely believed that long distance data links are heavily utilized. The paper [Odlyzko2] (see Section 3 for a summary) shows that this belief is incorrect. Even the backbone links are not used all that intensively, and the corporate private line networks are very lightly utilized. There are some key choke points (primarily the public exchange points, the NAPs and MAEs, and the international links) that are widely regarded as major contributors to poor Internet performance, but there is even some dispute about their significance. (In general, while there have been numerous studies of the performance of the Internet, some very careful, such as [Paxson], there is still no consensus as to what causes the poor observed performance.)

What is not in dispute is that a large fraction of the problems that cause complaints from users are not caused by any deficiencies in transmission. Delays in delivery of email are frequent, but are almost always caused by mail server problems, as even trans-Atlantic messages do get through expeditiously. A large fraction of Web-surfing complaints are caused by server overloads or other problems. There are myriad other problems that arise, such as those concerned with DNS, firewalls, and route flapping.

A key question is whether QoS would help solve those other problems, or would aggravate them, by making the entire system more complicated, increasing the computational burden on the routers, and increasing the numbers and lengths of queues.

Many QoS schemes require end-to-end coordination in the network, giving up on the stateless nature of the Internet, which has been one of its greatest strengths. Essentially all QoS schemes have the defect that they require extensive involvement by network managers to make them work. However, it is already a major deficiency of the Internet that, instead of being the dumb network it is often portrayed as, it requires a huge number of network experts at the edges to make it work [Odlyzko3]. Instead of throwing hardware and bandwidth at the problem, QoS would require scarce human resources.

The evidence presented in this paper, combined with that of [Odlyzko2], shows that the current system, irrationally chaotic as it might seem, does work pretty well. There appear to be only a small number of choke points in the system, which should not be too expensive to eliminate. Further, there are some obvious inefficiencies in the system that can be exploited. By moving away from private lines to VPNs (Virtual Private Networks) over the public Internet, one could provide excellent service for everybody through better use of aggregation of traffic and complementarity of usage patterns. The bulk of the work on QoS may be unnecessary.

Anania and Solomon wrote a paper in 1988 (which was widely circulated and discussed at that time, but was only published recently in [AnaniaS]) that took the unorthodox approach of arguing for a flat-rate approach to broadband pricing. That paper was about pricing of what are now called ATM services, which have QoS built in, but many of Anania and Solomon's arguments also imply the desirability of a simple undifferentiated service. My work presents some additional arguments and extensive evidence of the extent to which the traditional undifferentiated service, flat-price system can work.

QoS does have a role to play. There will always be local bottlenecks as well as emergency situations that will require special treatment. Even when local network and server resources are ample, there will often be need to ration access to scarce human resources, such as technical support personnel. Even in the network, methods such as Fair Queueing [FergusonH] can be valuable in dealing with local traffic anomalies, for example. Implementing them would represent a departure from the totally undifferentiated service model, but a mild one, and one that can be implemented inside the network, invisible to the users, and without requiring end-to-end coordination in the network. My argument is that we need to make the network appear as simple as possible to the users, to minimize their costs.

Sections 2 through 12 describe the economics of the Internet. The conclusion is that with some exceptions, the system does work pretty well as is. There are bottlenecks, but there are also inefficiencies that can be exploited to eliminate the bottlenecks. Users in general behave sensibly, and although their demands for bandwidth are growing rapidly, these demands are reasonably regular and predictable. It appears likely that unit prices for transmission capacity will decline drastically (although total spending on high bandwidth connections will surely grow), which should make it economically feasible to meet the growing demand.

It is impossible to predict with any certainty how the Internet will evolve, especially since its evolution depends on many factors, not only basic computing and networking technology and possible appearance of the proverbial "next killer app," but also on government regulation and sociology. Still, some conclusions can be drawn from the study of the current system. The complexity of the entire Internet is already so great, that the greatest imperative should be to keep the system as simple as possible. The costs of implementing involved QoS or pricing schemes are large and should be avoided. Section 13 outlines three scenarios that appear most likely. One is the continuation of the current flat rate pricing structure with almost uniform best-effort treatment of all packets, and enough bandwidth to provide high quality transmission. That scenario is likely to materialize if transmission prices decline rapidly enough. If they don't, the second scenario might arise, still with flat rate pricing and undifferentiated service, but with pricing reflecting expected usage of a customer. Finally, if even greater constraints are needed on traffic, ones that would provide congestion controls, approaches such as the Paris Metro Pricing (PMP) scheme of [Odlyzko1] might have to be used. PMP is the least intrusive possible usage-sensitive pricing scheme possible, and my prediction is that if any usage-sensitive pricing is introduced, it will eventually evolve towards (or degenerate into) PMP. None of these three scenarios would meet the conventional standards for economic optimality. However, the main conclusion of this paper is that optimality is unattainable, and we should seek the simplest scheme that works and provides necessary transmission quality.

## 2. The Internet and other networks

There are many excellent technical books and journal articles describing the technologies of the Internet (cf. [Keshav]). There is also a huge literature on how the Internet will change our economy and society (cf. [Gates]). On the other hand, practically nothing has been published on how the Internet is used, and how much it costs. It is as if we had shelves full of books telling us how to build internal combustion engines, and a comparable set of books on the effects of the automobile on suburban sprawl, income inequality, and other socioeconomic issues, but nothing about how many cars there were, or how much they cost to operate.

This section attempts to partially fill this gap in the knowledge of economics of data networks, but the picture it presents can only be a sketchy one. Still, it should help illuminate the major economic factors that are driving the evolution of the Internet.

The Internet (sometimes called the global Internet) refers to the entire collection of interconnected networks around the world that share a common addressing scheme. As such, it includes all of the elements shown in Fig. 2, which is a grossly simplified sketch of the data networking universe. The element called "Internet" in Fig. 2 is really just the public Internet, the core of the network consisting of the backbones and associated lines that are accessible to general users. WANs (Wide Area Networks) consist of some of the clouds in that figure (which are made up of LANs and campus networks) connected via either private line networks, or via public Frame Relay and ATM data networks provided by telecommunications carriers, or else via the public Internet. Fig. 2 omits many important elements of the data networking universe, such as regional ISPs.

This paper will concentrate on data networks in North America, primarily in the United States. Just as in the companion papers [CoffmanO, Odlyzko2], the justification is that most of the spending on data traffic is in the U.S. [DataComm]. Further, U.S. usage, influenced by lower prices than in most of the world [ITU], foreshadows what the rest of the world will be doing within a few years, as prices are reduced.

Data networks do not operate in isolation. To see them in the proper perspective, let us note that total spending on information technologies (IT) in the U.S. was about \$600 billion in 1997, approximately 8% of gross domestic product. The IT sector of the economy is credited with stimulating the high growth rate of the economy of the last few years, low unemployment, and low inflation [DOC].

Data communications cost about \$80 billion in the U.S. in 1997, or 13% of total IT spending, according to [DataComm]. Table 2 is a brief summary of the statistics on where this spending was directed, based on the more detailed information in [DataComm] (which also covers the rest of the world). These statistics show that transmission accounted for only \$16 billion, 20% of total for data communications, and 2.6% of total for all of IT. Thus data lines are a small part of the entire IT picture, and any scheme that attempts to improve their performance has to be weighed against costs that it might impose on the rest of the system. It is better to double the spending on transmission than to increase the average cost of all other IT systems by 3%.

Let us also note that U.S. spending on phone services from telecommunications carriers was around \$200 billion in 1997. Of this total, about \$80 billion was for long distance calls, but about \$30 billion was for access charges, paid to the local carriers. Thus it is more appropriate to say that $50 billion was for long distance services, and $150 billion for local services. In any event, today much more is being spent on voice phone services than on data. In the future, as broadband services grow, we can expect the balance to shift towards data. In particular, looking at total communications spending and how it is still dominated by voice, it is reasonable to expect substantial growth in spending on data transmission.

The core of the Internet, namely the backbones and their access links, is surprisingly inexpensive. There are many large estimates for total Internet spending, but those are misleading. There were about 20 million residential accounts with online services such as AOL at the end of 1997. At $20/month, they generated revenues of around $5 billion per year. However, most of that revenue is used to cover local access costs (the modems, customer service, and marketing expenses of the ISPs). The backbones are only a small part of the cost picture for residential customers (cf. [Leida]). In the statistics of [DataComm] (and of Table 2) they apparently are included in the "Commercial Internet services" category, which came to $1.5 billion in 1997. We now derive two other estimates that are both in that range. According to industry analysts [IDC], MCI's Internet revenues (which include only a small contribution from residential customers, and are dominated by corporate and regional ISP links to the MCI network) came to $251 million in 1997 (a 103% increase over 1996), and were running at an annual rate of $328 million in the last quarter of 1997. Since MCI is estimated to carry between 20 and 30 percent of the backbone traffic, we can estimate total revenues from all backbone operations between $1.1 and $1.6 billion at an annual rate at the end of 1997. (With revenues doubling each year, it is not adequate to look at annual statistics.) Yet another, rough way to estimate the costs of Internet backbones is to take their size, around 2,100 T3 equivalents at the end of 1997 [CoffmanO], and apply to that the $20,000 per month average cost of a T3 line [VS]. This produces an estimate of about $500 million per year for the main backbone links. When we add some additional costs for the access lines from carriers' Points of Presence to their backbones (cf. [Leida]), and apply the general estimate that for large carriers, transmission costs are about half of total costs, we arrive at an estimate of about $1.5 billion for the costs of the core of the Internet. (Costs and revenues are not the same, especially in the Internet arena, where red ink is plentiful as various players attempt to build market share, but within the huge uncertainty bounds we are working with, that should not matter much.)

Compared to the Internet backbones, the total cost of private line networks is at least 6 times as large. Furthermore, the aggregate bandwidth of leased lines is also much greater than of the public backbones, although the traffic they carry appears to be comparable in volume. (See tables 2 and 5, taken from [CoffmanO].) This helps explain why the evolution of the Internet is increasingly dominated by corporate networks.

Just as with switched voice networks, data network costs are dominated by the local part. However, there is much more heterogeneity in data than in voice. In the foreseeable future, large academic and corporate networks are likely to have 2.4 Kbps wireless links along with 14.4 and 28.8 Kbps modems, megabit xDSL and cable modem links, and gigabit fiber optic cables. In the local campus wired environment, overengineering with Ethernet, Fast Ethernet, Gigabit Ethernet, and similar tools appears certain to be the preferred solution. However, there will still be challenges of interconnecting the other transmission components (whether slower or faster), as well as all the servers and other equipment that require the bandwidth. Network managers will have a hard time making everything interoperate satisfactorily even without worrying about QoS.

The Internet backbones are small and inexpensive compared to the rest of the Internet. However, they are the heart of the Internet, just like a human heart that is small but crucial for the life of the body. The role of the backbones will likely become even more important in the future as a result of several related developments. One is that they are being traversed by an increasing fraction of data traffic. The traditional 80/20 rule, which said that 80% of the traffic stayed inside a local or campus network is breaking down. Ferguson and Huston [FergusonH] even mention some networks where as much as 75% of the traffic goes over long distance links. (We do not know how far that traffic goes, and in particular whether there will continue to be a strong distance dependence in the future. See [CoffmanO] for a more detailed discussion.) Another reason for the increasingly important role of the Internet backbones is that they are supplanting private line networks as corporate WAN links, and with the development of extranets, will be playing a crucial role in the functioning of the whole economy. Thus there is a reason to worry about the costs of the backbones, as they might become a larger fraction of the total networking pie. On the other hand, if one can overengineer only one part of the Internet, then it is best to do it to the core, as without high quality transmission at the core, other parts of the network will be only be able to offer poor service.

## 3. Network utilization

Network utilization rates are seldom discussed, yet they are the main factor determining costs of data services. A line that is used at 5% of capacity costs twice as much per byte of transmitted data as one whose average utilization rate is 10%.

Although there is no simple relation between the quality perceived by customers and how heavily their networks are used, the less heavily loaded the network, the better the service. Even the notoriously congested trans-Atlantic links do appear to provide good performance for applications as demanding as packet telephony in the early hours of Sunday morning. What this says is that even without any new QoS technologies, one can provide excellent quality by lowering utilization. Thus the main problem of the Internet is not a technical one, but an economic one, whether one can afford to have lightly utilized networks. As an example, the experimental vBNS network, discussed in [Odlyzko2], provides quality sufficient for even the most demanding applications, but it is expensive (or would be expensive, were it operated on a commercial basis), running at an average utilization of its links of around 3%.

The main question for the future of the Internet is whether customers are willing to pay for high quality by having low utilization rates, or whether many links will be congested, with QoS providing high quality for some select fraction of data transfers. We can find much about the likely evolution of data networks from observation of usage patterns of existing networks. When we consistently see lightly utilized links where customers can obtain higher utilization rates and lower costs by switching to lower capacity lines, we can deduce that they do want high quality data transport, and are willing to pay for it.

Table 4 shows utilization rates (averaged over a full week) for various networks. It is based on [Odlyzko2], except for the entry for local phone lines, which is derived from the data in Table 1 (which is based on [FCC]). A surprising result is that the long distance switched network is by far the most efficient in terms of utilizing transmission capacity. For most people, an even more surprising feature of the data is the low utilization rate of private line networks.

The paper [Odlyzko2] discusses the reasons data networks are lightly utilized. Lumpy capacity is a major one. Rapid and unpredictable growth is another. Small private networks are yet another. Perhaps the main reason, though, is the bursty nature of data traffic. This traffic is bursty on both the short and long time scales, and customers do value such bursty transmission. This means that we cannot reasonably expect data networks to approach the efficiency with which the switched voice network uses transmission capacity.

Utilization rates can also provide guides to the extent that QoS measures might improve perceived quality of networking. Practically all Internet users find service much better 5 in the morning than at noon. However, traffic on the backbones in the early hours of the morning is still about half that during the peak hours, as is seen in Fig. 10 of this paper and several figures in [Odlyzko2]. Further, traffic mix does not seem to vary much between trough and peak periods [ThompsonMW]. Therefore we can conclude that during peak periods, no QoS measure is likely to provide transmissions that have priorities around the median of all traffic with better service than the current undifferentiated service provides for all traffic early in the morning. (There could be some improvements in jitter, for example, but algorithms that do provide such improvements could be used inside the network, invisibly to the users, to provide similar improvements for undifferentiated service.)

An important point in considering the low utilization rates of data networks is that it is not caused primarily by inefficiency or incompetence. Customers choose the capacities of their lines, and their choices tell us what they want and what they are willing to pay for. This point will be treated at greater length in the following sections.

## 4. Inefficiency is good (if you can afford it)

Efficiency in utilization of transmission lines or switches should not be the main criterion for evaluating how good a network is. The crucial question is how well customers' needs are satisfied.

Consider figures 3 and 4, which show utilizations of dial-in modems at Columbia University and the University of Toronto. (These figures are based on detailed data supplied by those institutions. Graphs for more recent periods, separated out further by 14.4 and 28.8 Kbps modem pools, can be found at [Columbia, Toronto].) The average utilizations (over the periods shown in the figures) were 52% at the University of Toronto and 78% at Columbia University. Clearly Columbia was utilizing its modems more efficiently. Was it providing better service, though? Its modems were completely busy for more than 12 hours a day. (The slight drops below 100% in the utilization in Fig. 3 are misleading, since they represent only a little idle time, and are largely the resetting of modems after a session is terminated.) Clearly much demand is unsatisfied, and there are many frustrated potential users who do not accomplish their work. Further, the high 78% utilization rate is misleading, since many users are probably staying online for longer periods than they would if they had assurance they could get a new connection when they wanted it. Toronto, with a lower utilization rate, managed to accommodate all demands except for a brief period on Monday night.

The University of Toronto managed to satisfy essentially all demands of its students and faculty for modem connections and still achieve a 52% utilization rate. This rate is extremely high. There are many examples of low utilization rates. The family car is typically used around 5% of the time. The fax in our office or home, the PC on the desktop, and the road we drive on are all designed for peak usage, and are idle most of the time. We are willing to pay for this inefficiency because the costs are low compared to the benefits we receive. As costs decrease, we usually accept lower efficiency. For example, in the early days of computing, programs were written in assembly language. Later, as the industry advanced, there was a shift towards compiled programs. They typically run at half the speed of assembly coded versions, but they make software easier to write and more portable. With further advances in computing power, industry was willing to jump on the Java bandwagon, even though the early versions of Java typically ran a hundred times slower than compiled programs. With over 99% of computing cycles devoted to running screen savers, this was a worthwhile tradeoff.

When capital costs are high, it makes sense to maximize the utilization of facilities. For example, latest semiconductor fab lines, approaching $2 billion in cost, are run around the clock. Among services that do face wildly varying demands, cannot stockpile their outputs, and have high fixed costs and lumpy facilities problems, the airlines manage to operate at 60-70% of capacity. (This is measured as a fraction of filled seats. Political opposition to night landings and difficulty in persuading people to fly at 3 am from Boston to Chicago keep airplanes on the ground more than half the time.) They do this through intensive use of pricing. What is noteworthy about this example is that the yield management techniques used by the airlines are not liked by the public, but they work to keep the planes filled, while at the same time allowing people to fly at a moment's notice (provided they are willing to pay).

Airlines are an example of extremely expensive equipment that needs to be operated intensively. In contrast, LAN equipment is relatively inexpensive. Most of data traffic is on LANs, and it appears that utilization rates (measured over a week) are almost universally low, around 1%. Users notice degradation in service when traffic grows, and respond by pressuring network administrators to increase capacity (and, in the short run, pressure their colleagues to avoid things like large file transfers during busy hours). In many organizations, LAN spending is controlled by small units, so that money not spent on LANs can be devoted to other purposes. When those units do overengineer, they must feel they are getting their money's worth. Spending on LAN equipment (see Table 3) is in the tens of billions of dollars, but that is only a fraction of the spending on PCs, say, and a tiny fraction of the salaries of the people relying on those LANs. Low utilization rates of LANs are a necessary penalty for good performance.

The decisions on utilization rates of LANs are made in rather convoluted ways, with network administrators taking into account user complaints, plans for the future, available budgets, and other factors. Thus it is hard to say that the existing arrangement with 1% utilization factors is the outcome of a rational process. Let us instead consider a simpler case. Consider the growth in phone lines shown in Table 1. They increased by 63% between 1980 and 1996. Only a fraction (under 30%) of this increase was caused by growth in the number of households. Detailed statistics on the sources of this growth between 1988 and 1996 are available in Table 18.3 of [FCC]. During those 8 years, households with phone service increased by 9.7 million, additional residential lines (second, third, ...) increased by 13.5 million, and business lines went up by 13.7 million. Even though there was plenty of spare capacity on their phone lines (with average utilization of 4%), both households and businesses decided it was better to pay for additional lines than miss important calls, or bother switching between voice and fax. In earlier decades, similar considerations led to the effective elimination of party lines. In all these cases, individuals and businesses voted with their pocketbooks for lower utilizations and higher quality service.

Although the connection is less clear than for phone lines, corporate WANs are lightly utilized because users either pay directly or else pressure their centralized networking organization for high quality data services. Mission-critical applications cannot be neglected. What is important to realize is that data services customers decide on their utilization levels. The business customers of the Zocalo ISP listed in Table 2 of [Odlyzko2] who use their T1 lines at less than 0.5% of capacity on average could transmit the same amount of data over a 56 Kbps line, and still stay under 14% in average utilization. This would reduce their costs by a factor of 4. Since they are paying for the higher bandwidth, we have to conclude that they do find it worthwhile.

Almost all applications benefit from high bandwidth and low latency. Further, high bandwidth can often substitute for low latency, since (as is explained in detail in [Cavanagh], for example), it is the transmit time of a whole file or screen image that needs to be minimized. Being able to send an X-ray to a physician, authorize a credit card transaction in a few seconds or find out a customer's previous purchase history right away have obvious value, one that cannot be determined a priori by any arbitrary rule that average utilization has to be at least 10%. Decisions have to be made by those responsible for the application.

Since corporate IT organizations are always under pressure to cut costs, one might conclude that corporate LANs and WANs are properly provisioned for their tasks. If that is so, though, we are left with the puzzle of the Internet. Why does it not provide better service? As we have seen above, spending on the backbones of the Internet is a tiny fraction of the spending on LANs and WANs. Diversion of a small portion of LAN and WAN budgets might seem to suffice to expand bandwidth of the backbones and solve the congestion problem.

The solution to the Internet performance puzzle seems to have three parts. One is that given the utilization levels discussed in Section 3, lowering backbone congestion to that of a corporate WAN would involve about a 3 to 5-fold increase in capacity, with corresponding increases in costs. Second is that the Internet lacks a method for users to signal providers effectively what quality of service they want and are willing to pay for. Third, most of the users of the public Internet are very price sensitive, and are (almost by definition, since they continuing to pay) comfortable with the present service. This is largely a legacy of the Internet's origins as an academic research network.

The public Internet is perfectly satisfactory for most applications, such as email. Even the notoriously congested trans-oceanic links do handle email expeditiously. (Most of the complaints about email one hears are caused not by backbone network performance, but by mail servers, a different subject.) Large database updates can also be handled, provided they can tolerate delays. (Large preprint servers report few difficulties in mirroring overnight, for example.) It is only real-time applications, such as packet telephony, transaction processing, telnet, or video conferencing, and to a lesser extent Web surfing, that are negatively affected by congestion. The Internet thus serves well what we might call the lower tier of the data networking market.

On the other hand, corporate networks serve well the upper tier. What this means, though, is that they provide unnecessarily high quality of service for the lower tier demands. If a single level of service is all that can be provided, this is probably the optimal strategy for corporate networks. However, it is expensive, and often involves using a Rolls-Royce to transport gravel. QoS could remedy that problem, but would have its own costs. I will discuss those later, but first I will present a survey of the costs of current networks.

## 5. Costs and prices

One observation that led to this paper was that although packet networks are universally extolled as much more efficient than switched voice networks, they are surprisingly expensive. At most corporations, sending data over internal packet networks is more expensive than by using modems over the public circuit switched network. This is especially striking for two reasons. One is that modem calls use transmission capacity of the switched voice network extremely inefficiently. Although 128 Kbps of network bandwidth is dedicated to each such call, usually only 28.8 or 33.6 Kbps worth of data is transmitted. The other reason is that only a small part (estimated by industry analysts at around $0.015 of the approximately $0.12 that the major carriers collect for a minute of a voice call in the continental U.S.) goes towards running their networks. (By far the largest expense, one that data lines are not burdened with, is access charges to the local telephone companies.) Thus for a data transfer using a modem to cost less than over a data network goes counter to assertions such as those about a "fundamental reason ... the telephone network is more expensive to operate than the Internet" (pp. 5-6 of [Keshav]).

Table 7 shows costs of sending a megabyte (MB) of data over various networks. Most of this section is devoted to an explanation of how the figures in this table were derived. First let me remark that costs in this section refer to the costs incurred by users, and so reflect prices charges by service providers. They are not necessarily the costs of those providers.

The cost of transporting a megabyte of data over a network is just one measure of performance, and it is seldom quoted. Customers usually do not pay per megabyte. Further, for large file transfers that can tolerate an overnight delay, sending a bunch of magnetic tapes via Federal Express will produce costs orders of magnitude lower than those calculated below. (That is indeed how many data bases are mirrored.) However, the cost per MB is fundamental to understanding the economics of data networks. After all, the main function of these networks is to move bits from one point to another. Further, there is a clear distinction between transporting data overnight on magnetic tapes and sending it on a data network right away. The distinction between transmission via a modem and the Internet is often immaterial to users.

The costs in Table 7 might seem high. At the other end of the scale, the high value of reliable and low delay data communication is shown by the example of services such as Advantis, CompuServe, and Infonet, which as recently as 1993 were charging between $20 and $40 per MB (in addition to the per-minute charges for connect time), far higher than any of the costs of Table 7. This reinforces the claim made earlier that the Internet has grown not because of the low cost of packet data transmission, but because of the added functionality that such transmission made possible.

The comparisons below involve apples, oranges, and pears. Further, all cost estimates are rough ones, due to lack of precise statistics. Still, it is worth looking at the numbers, especially since the disparities in costs are so large that they swamp any errors in estimates, and help explain the migration of data traffic from private lines to Frame Relay and the Internet.

First let us consider sending data via a modem over the switched voice network, paying retail rates. If we achieve a sustained speed of 28.8 Kbps (and, to be conservative, assume that we are only transferring data in one direction, with the other channel idle), we will send 200 KB per minute. Several carriers, including AT&T, offer their residential customers calling plans with a flat rate of $0.10 per minute for calls in the continental US at any time. With such a rate, the cost of transmitting a megabyte becomes $0.50.

The computation above assumes that we use a modem to transfer data in just one direction. If we have two file transfers of equal size in opposite directions, cost drops to $0.25 per megabyte. Use of a 33.6 Kbps or a 56 Kbps modem lowers the cost further.

Frame Relay is offered primarily in the form of PVCs (Permanent Virtual Circuits). There is a charge for a port, which imposes an absolute limit on the rate data can be sent into the Frame Relay network. There is also a charge for CIR (Committed Information Rate), which is the rate that the service provider guarantees to carry successfully to the destination. However, there is usually no charge for the amount of data transmitted. Taking the total traffic carried for a variety of businesses, and dividing it into the prices paid for the service, it appears that on average Frame Relay providers collect revenues between $0.30 and $0.40 per MB.

There are occasional references to a cost per MB of $0.03 or $0.05 for use of SVCs (switched virtual circuits) on Frame Relay or ATM networks. Although accurate, those figures are misleading, since they cover only the usage-sensitive component of the price. With the low utilization levels that are observed, fixed costs dominate, and average $0.30 and $0.40 per MB.

Private line costs per MB are even harder to estimate than those of Frame Relay. However, industry sources (cf. [Cavanagh] and reports in *Data Communications* and other magazines) report that Frame Relay is usually 20% to 50% cheaper than private line for equivalent service. (It is impossible to be precise, because this is an oranges-to-apples comparison. Among other factors, Frame Relay pricing does not depend on distance in the continental U.S., while that of private lines does.) Hence I conclude that private line transmission costs between $0.50 and $1.00 per MB. This is confirmed by some sample data points, such as that of a large corporation in which network services are charged to departments according to their traffic, with total charges of $0.80 per MB.

Let us next consider the Internet. There are several ways to do this. Residential customers currently stay connected about 45 minutes per day, according to public statements from AOL and other carriers, and while they are online, they transfer data at the rate of 5 Kbps (overwhelmingly to their computers). Thus they average 50 MB per month of data transfer, and at the standard rate of $20 per month, are in effect paying $0.40 per MB.

The more interesting case is that of large business and academic customers who have dedicated high speed data lines to the Internet. They usually pay for the local connection to the nearest Point of Presence (PoP) of their ISP, and then for the link to the ISP. The estimate of 3,000 TB per month of traffic on Internet backbones, together with the estimate of $1.5 billion per year for the revenue of those backbones (from Section 2 above) yield an estimated cost of $0.04 per MB for the use of the backbones. This is almost an order of magnitude less than the cost of Frame Relay or private line transport.

As another confirmation that $0.04 per MB is a reasonable cost estimate for the Internet backbones, let us note that Web hosting services usually do have usage-sensitive pricing. For large data transfers, they typically charge between $0.03 and $0.05 per MB.

As an interesting historical note, according to [MacKieMV], in 1993 NSF was paying $11.5 million for the NSFNet backbone, and about $7 million as subsidies for some of the regional networks. To be comparable to the current backbone providers, which pay for the links from their PoPs to their backbones, it seems reasonable to say that the total cost of the NSFNet backbone was around $18 million per year. Since this network carried about 6 TB per month in 1993 [NSFNet], the cost per MB was $0.25. Towards the end of 1994, traffic grew to 15 TB per month, but costs probably did not (since the network was very lightly utilized, see [Odlyzko2]), so the cost per MB most likely dropped to $0.10 or so.

The computations above suggest $0.04 per MB as the current cost of data transfer over the Internet backbones. However, that is not a full picture. Although comprehensive statistics are not available, it appears that the ISPs with national backbones (the NSPs) deal directly with only a small fraction of businesses that have Internet connections, primarily the big organizations with large capacity links. Most of the small corporations go through regional ISPs, a crucial part of the Internet picture that is not represented in Fig. 2. In terms of revenues, low bandwidth connections still dominate the private line market (see Table 6 and the more detailed discussion in [CoffmanO]) and the same appears to be true for connections to the public Internet (cf. Table 2 of [Odlyzko2]). Low bandwidth links are much more expensive (relative to their capacity) than high bandwidth ones, as will be discussed in greater detail in the next section. Hence the cost of sending data for the bulk of business customers is likely to be considerably higher than the $0.04 per MB incurred by the typically larger customers that connect directly to the backbone providers. That is why the Internet entry in Table 7 has a range from $0.04 to $0.15 per MB.

A crucial point in considering the costs of transporting data over various networks is that these costs are largely under user control. Since most utilization rates are low, as is shown in Table 7, and fees for private line, Frame Relay, and Internet connection consist of fixed monthly charges, most users can save money by switching to lower bandwidth links. They have a choice, and by selecting the capacities and prices they do, they signal their interest in high quality transmission. As an example, those companies in Table 2 of [Odlyzko2] that use their T1 links to the Zocalo ISP at 0.5% of capacity are paying $0.34 per MB.

The cost estimates of Table 7 are approximate. However, the cost advantages of Frame Relay over private line networks, and of the public Internet over Frame Relay are widely accepted. For example, a recent article [Cray] estimated annual costs of a small corporate network to be $133,272 using leased lines, $89,998 using Frame Relay, and $38,400 over the public Internet. “Your mileage may vary” is still good advice, but there is a definite economic advantage to moving away from private lines and Frame Relay solutions and towards the Internet. The major barrier to wide use of the public Internet by corporations is security, but that one is on the way to being overcome through the use of VPN (Virtual Private Networks). The other major barrier is performance. However, even there progress is taking place, with an increasing number of ISPs providing high quality transmission (within their networks) and even service guarantees [Makris].

## 6. Future of costs and prices

Preceding sections showed that costs of data transfers over private line and Frame Relay networks are surprisingly high. Although packet networks are reputed to be extremely efficient, on average they cost more per MB than the switched voice network does. However, those costs are under customer control, in that it is the customers who decide how heavily to utilize their links, and therefore what their costs will be. Given the prices that have been available for transmission in the past, they have chosen low utilization rates and relatively high prices. Will that change in the future?

While the technical press is full of stories about progress in fiber optic technologies, network managers have had to face rising prices for data transmission capacity over the last half a dozen years. Fig. 2 of [CoffmanO] shows the historical record of T1 prices, which decreased by a factor of 5 from 1983 to 1992, but have gone up by about 50% since 1992 (in nominal dollars). The rate of increase has accelerated recently. The article [Rendleman] presents data about MCI's rates for leased lines. Those for T1s increased by between 19% and 22% between June 1996 and December 1997, while those for T3s went up by 43%.

In the past, network managers coped with increased data traffic by moving to higher bandwidth links, which are much less expensive per unit of capacity. As an example, Table 6 (based on [VS], and copied from [CoffmanO]) shows that T3 leased lines accounted for about 41% of the bandwidth of all private lines in the U.S. at the end of 1997, but only about 7% of the revenue. (For concrete examples of prices for varying speeds, see [FishburnO].) However, with growth rates in volume of data increasing, and leased line prices rising at an accelerating pace, it is natural that network managers are uncertain whether it will be possible to continue providing large pipes that are lightly utilized, and are asking for QoS measures to provide high quality transmission for at least the most important part of their traffic.

Although transmission prices have been rising, it seems likely that they will soon start decreasing. The main reason is that new technologies, primarily WDM (Wavelength Division Multiplexing) are leading to dramatic cost reductions, and those will eventually show up in lower prices. A more detailed discussion is contained in the companion papers [CoffmanO, FishburnO]. What we see now is the transition from a network that was sized for voice traffic, in which rising data volumes were causing a squeeze on capacity, to one in which the underlying fiber is not a constraint, and the capacity of the entire network depends on the deployment of electronics that will be decreasing in price at a steady pace. What we are likely to see is what has been observed in semiconductors (see Table 1 in [FishburnO] for the history of Intel), with total spending on data transmission rising, but unit prices plummeting.

## 7. Aggregation of traffic

The greatest inefficiency in data networking today is that thousands of corporations are running their own private networks. (See Table 6, based on data from [VS], for statistics on leased lines.) Not only are the direct transmission costs higher than over the public Internet or the Frame Relay networks (Table 7), but the indirect costs of network management are high as well. A grave defect of the Internet is that in pushing intelligence to the edges of the network, it also pushed network administration and maintenance to the edges, where it is wastefully duplicated [Odlyzko3].

Data traffic does not smooth out as much as voice traffic when it is aggregated (cf. [FeldmannGWK, LelandTWW]). However, it does smooth out, and that is a part of the economic foundations for ISPs. The national backbones that reportedly do provide good service manage to operate at considerably higher utilization levels than private line networks. Why don't large corporations manage to achieve similar gains? One reason is the concentration of traffic during the business day, discussed more fully in Section 9. Another possible reason is that as the size of a corporation increases, the high bandwidth sources also increase, so that the bursts on large pipes are about as large, relative to capacity, as on small ones. This is a very tentative hypothesis for which I do not have much evidence, but it is supported by the statistics on utilization and by pictures such as that of Fig. 7, showing traffic on some corporate T3s. The traffic patterns of Fig. 7 is not much smoother than that in Fig. 6, which comes from a 128 Kbps link. Would aggregating traffic from many corporations into a large public network avoid this problem? It should. Estimates of traffic on the public Internet and on private line networks suggest [CoffmanO] that no single corporation carries even 1% of the total traffic. Thus even large sources, that do produce the spikes seen in Fig. 7, would be merged into much larger streams of data that should produce smoother traffic.

Even without smoothing gains, aggregation of traffic does offer huge savings. The reason is that price per unit of transmission capacity decreases dramatically as the bandwidth of a connection increases. This is discussed at greater length in [CoffmanO, FishburnO], but generally, increasing the size of a link by a factor of 25 causes price to rise by a factor of only 6 to 10. This relation is currently valid only in the 56 Kbps to T3 range, but in the future is likely to extend to higher capacities [FishburnO]. These economies of scale reflect lower selling, administrative, and maintenance costs, as well as avoidance of many of the costs of the multiplexing hierarchy that is required to provide low bandwidth links on a fast fiber network.

**8. Stability of usage and growth patterns**

A major reason for the high utilization of the long distance switched networks is that voice traffic can be predicted accurately and grows slowly. There is a large literature on teletraffic engineering that was developed over most of the 20-th century (cf. [Ash] for information and references), and the prediction and traffic engineering tools that have been developed are surprisingly effective. A few calls fail to get through as a result of congestion, but there are so few of them, even on the busiest days, which typically are Mother's Day and the Monday after the Thanksgiving holiday, that people don't notice them. As a result of the high predictability, capacity can be sized correctly. However, constant vigilance by network operators and network designers is required to keep the switched voice network operating at high fraction of its capacity and yet accommodate all demands, since those demands to change. Special events, such as airline ticket wars, or call-in promotions, do cause strains.

The switched phone network does block calls on a large scale in emergency situations, such as an earthquake. However, an early decision was made in the Bell System to satisfy all normal demands. (As network costs decreased, though, the percentage of calls that were allowed to be blocked was reduced dramatically, see [Ash].) In studies of the economics of voice phone systems, there have been proposals for lowering costs by providing less capacity, so that many calls would be blocked during peak hours, and it would be the persistent callers, who kept redialing (and thus presumably valued their calls the most) who would get through [MitchellV]. However, this idea was never taken seriously. The switched voice system has attracted an enviable reputation for quality and usability by meeting even peak demands.

Data traffic grows much faster than voice traffic, and is inherently much more bursty. A good example is that of a dedicated 128 Kbps link to the Internet, profiled in Fig. 6 here and in Fig. 4 of [Odlyzko2], and the traffic from the Internet to a set of mostly residential customers shown in Fig. 6 of [Odlyzko2]. Such usage patterns have convinced many that data traffic is simply too chaotic to accommodate fully, and that rationing by queue will be essential for most traffic, with only a select portion getting high quality transmission using QoS tools.

This section provides evidence against the basic assumption of unpredictable data traffic. Yes, data does not behave like voice, and in particular is fractal, so does not smooth as much when more sources are aggregated (cf. [FeldmannGWK, LelandTWW]). However, all that this might mean is that we won't be able to use as high a fraction of the bandwidth as we can with voice. There is much predictability in data traffic. The regular growth rate in the volume of transmissions at the University of Waterloo (Fig. 1) was already mentioned in the Introduction. Other examples showing rather regular increase (typically a doubling each year) are cited in [CoffmanO]. There is also regularity on smaller time scales. First of all, there is regular time-of-day and day-of-week variation, shown in many graphs in this paper and in [Odlyzko2]. (These were studied from a more technical perspective in [Mukherjee].) How predictable those patterns are can be seen in Fig. 9, which shows data traffic from AltaVista during several days, each a Tuesday, Wednesday, or Thursday, in January 1998. (The reason for restricting to the middle of the week is that, just as with voice traffic, there is a characteristic pattern with data, with Fridays showing less traffic than other workdays, and weekends even less traffic, and distributed differently across the day.)

Fig. 10 shows traffic on an OC3 Internet link. It is the same MCI link profiled in figures 3 and 5 of [Odlyzko2], and more extensively in [ThompsonMW], except that this time traffic is shown to the north, the flow rate is in 5-minute averages, and the patterns for the two workdays are overlaid. Comparing this graph to that of Fig. 3 of [Odlyzko2] shows how much smoothing hourly averages introduce. On a 5-minute scale, there is much more burstiness, and presumably if could obtain measurements on millisecond scales, the oscillations would be even greater. On the other hand, there is a remarkable similarity in the traffic over the two days, a similarity that carries over to a comparison of each of Monday through Thursday of that week.

Fig. 11 shows the distribution of hourly traffic averages from the Library of Congress server during the four Sunday-Monday pairs of Dec. 7-8, 1997, and then Jan. 11-12, March 1-2, and May 17-18, 1998. Over this period traffic grew by about 50% (from 7.1 GB on Dec. 8. 1997, to 10.9 GB on March 2, 1998), but the traffic pattern has been remarkably stable. However, there are occasional spikes (such as the one on Sunday, May 17, 1998, the highest one in Fig. 9). Also, some days have unusually heavy traffic. The largest volume sent out (through the end of April, 1998) was on Tuesday, February 17, after a long weekend that included Valentine's Day and ended with the Presidents' Day holiday on Monday, when 16.2 GB was sent out. (Note that the cumulative statistics available at [LOC] are not fully reliable, since they show heavy traffic on Monday, Feb. 16, whereas the detailed hourly statistics for that day show extremely light traffic, as is typical on holidays.)

What is one to do when faced with demands such as those on the Library of Congress, which grow rapidly but steadily at over 100% a year, and are reasonably predictable, but with occasional surges that are 50% higher than normal for that day? Given rapidly dropping server prices, it does not seem unreasonable to adopt the policy that guided the building of the voice phone network, namely put in just enough capacity to meet the foreseeable increases in demand, like that of Feb. 17, 1998. If even higher demand materializes, let some simple queuing mechanism take care of rationing access during the overload episode. If that is done, though, then on most days there will be no congestion at the servers, and no need for any complicated QoS mechanisms. A similar approach appears to be workable for transmission capacity, especially if bandwidth prices do start dropping rapidly.

A factor that makes traffic on backbone links smoother than it might be is that it gets naturally smoothed at its sources or at intermediate links on the way to the backbones. A workstation might be capable of putting out 50 Mbps, but if it is attached to an Ethernet, it will be lucky to transmit (or receive) 5 Mbps. We again have to keep in mind the picture of the entire Internet in Fig. 2. What happens at the core results from a combination of events at the edges, and those edges act as natural controllers. Even if the capacity of one link is drastically increased, it requires improvements in the rest of the infrastructure to be able to utilize the added bandwidth sensibly. Admittedly, a malicious user can cause harm. However, in general people act sensibly, not maliciously.

## 9. Complementary traffic patterns

Table 4 shows that the long distance switched voice network is utilized much more efficiently than any of the data networks. A key factor behind this efficiency in usage of transmission facilities is that they are shared among several classes of users with complementary traffic pattern. Fig. 1 of [Odlyzko2] shows the aggregate traffic pattern on the switched voice network. Fig. 5 of this paper shows this same traffic broken down into residential and business components. (This figure is based on Fig. 1 of [Odlyzko2] and Fig. 30.2 of [Clark3].) The distinction is not clear, in that much of what is officially called residential calling comes from small offices and home offices of business customers who are not identified as such. Similarly, much of 800-number calling is by residential users, but it is classified in the business category, since it is paid for by corporate customers. Still, even though it is imprecise, the division into these two classes is enlightening. Their behavior is strikingly different. What's most important, their demands are largely complementary. Residential customers' heaviest calling period is on Sunday, when there are practically no business calls, and during the business day, they concentrate their calls in the evening, again when there are few business calls. If one had to build two separate networks for these two classes of customers, then, even ignoring the loss of efficiency coming from less aggregation, the total capacity of the two networks would have to be over 60% higher than that of the single network.

Data traffic shows comparable differences in traffic patterns among different classes of users, differences that are similar to those observed in voice phone statistics. Business customers generate almost all of their traffic during the business day, in the pattern shown in Fig. 5 and to some extent also in figures 4 and 7 of [Odlyzko2]. Residential customers, whose behavior dominates Fig. 4 of this paper and Fig. 6 of [Odlyzko2], generate their traffic on weekends and late in the day (even later than voice calls, with the peak for modem calls around midnight). What this means is that if the private line networks were absorbed into the public Internet's backbones, they could provide for everybody service as good as the corporate users enjoy right now and total costs would be far lower.

## 10. Pricing and demand shifting

Often natural demands are not sufficiently complementary to provide an even demand profile. In such cases, one can modify customer demands through pricing. In the Bell System, long distance rates depended on distance from the beginning. This pricing originated in the high marginal costs of setting up and carrying such calls. On the other hand, prices were uniform around the clock until 1919. At that time a three-tier price scheme was introduced (on top of the complicated distance-sensitive structure), with evening discounts for calls placed between 8:30 pm and midnight, and a larger night discount from midnight until 4:30 am. The number of tiers and hours they were effective kept changing, but it is noteworthy that Sunday discounts were introduced only in 1936, and Saturday ones a couple of decades later. The AT&T archives appear not to contain any detailed records justifying the evening and night discounts, but the likely rationale is easy to deduce. Long distance phone service was extremely expensive for several decades after Alexander Graham Bell's invention. In 1919, four years after the introduction of transcontinental service, a three-minute phone call from New York City to San Francisco (the minimal one could buy) cost $16.50, more than most families spent on housing in a month. Telephone service was for businesses and the very rich. As late as 1930, the average number of long distance calls was only 160,000 during the week, 136,000 on Saturday (which was largely still a working day, and helps explain the long delay in introduction of Saturday discounts), and 67,000 on Sundays. The network was largely idle in the evening and at night, yet it had to be kept operational (and this involved substantial costs in the 1910s, since operators were required to set up calls), since much of its value was in the ability to provide service at unusual times. It was sensible therefore to encourage use in off-peak hours through discounts. The late start at 8:30 pm and early phaseout at 4:30 am were presumably designed to avoid diversion of regular business calls into lower-cost periods.

Extensive study of economics of telecommunications did not start until the 1960s (see [MitchellV] and the references there). It is interesting to consider for comparison the electric power industry. There economic concerns were at the forefront of business planning, and an extensive literature was generated at the end of the 19th and in early 20th centuries. For a brief overview of the subject, see the excellent short summary in [Friedlander]. For more detailed accounts, see the famous comprehensive survey [Hughes] and the papers [HausmanN1, HausmanN2, Platt]. Interestingly, the literature on pricing of electricity was created primarily by engineers and businessmen, without noticeable input from economists. (This was several decades before Ramsey, for example.) It did not have the quantitative analyses of modern economics, but it was sophisticated. In particular, careful attention was paid to diversity in patterns of use through the day. The problem in launching the electric power industry was that initial demand was for lighting homes, and that basically lasted for a few hours each evening. Discounts were introduced for industrial use in order to utilize the available capacity during the day. (Note the interesting reversal, with commercial users getting lower rates on electricity than residential ones, but paying more for telephone service.) A remarkable 1914 address by Samuel Insull, a pivotal figure in the development of the electric power in the U.S., enumerated eleven different classes of consumers (such as residential homes, industry, street cars, and street lighting), their diverse usage patterns, and how they contributed to leveling of total demand. For example, during the day of maximum use in 1914, had each of those eleven categories made its peak demand at the same time, Commonwealth Edison would have had to supply 26,640 kw. Instead, the peak demand on the system was only 9,770 kw. (For details, see Chapter 9 of [Hughes], especially pp. 217-226.) The electric power industry even took deliberate steps to stimulate development of new applications that would generate usage patterns that would complement those of other sources of demand.

Flat rate pricing for Internet access may turn out not to be sustainable, and the variations in traffic patterns among different classes of customers may not provide enough smoothing. In that case, it might be possible to use pricing to induce better utilization. Pricing would not necessarily have to be usage-sensitive, in that one can have different prices for different classes of consumers, if one can discriminate among them, and prevent arbitrage.

In the time-sharing arena, there have been successful examples of evening out the load through time-of-day pricing [GaleK]. However, the rapid growth in processing power that has provided more cycles than people know what to do with, and this has put an end to just about the entire business of pricing processing power. It seems likely that in data transmission a similar phenomenon will apply. If it does not, though, then schemes like those in [GaleK] might provide a workable alternative.

## 11. Quality of Service

For a survey of QoS techniques, see [FergusonH]. This section will not deal with the technologies, and will concentrate on raising global questions about the place of QoS in networking.

QoS approaches would be of greatest value if networks were seriously congested, and the traffic requiring high quality transmission were a small fraction of the total. Under those conditions, creating an express lane for the priority traffic is sensible. However, neither of these conditions applies today, and it is questionable whether either will apply in the future. As was shown in sections 3 and 5, most data networks are very lightly used, and the costs appear to be acceptable, since it is the users who choose the tradeoff between price and performance. Further, there are reasons to expect that prices of broadband transmission capacity will start a rapid and sustained decline, so that it will continue to be feasible to provide uncongested pipes.

When utilization rates are low, users are clearly paying for the ability to burst at high speed. As is discussed in the Introduction to [Odlyzko2], in that environment the main constraint is not competition with other demands, but just the bandwidth of the connection. QoS does not provide any help with that.

Many QoS schemes prioritize traffic depending on what application it comes from. This approach has many problems. Currently the majority of Internet traffic is http (around 70%, [ThompsonMW]), and it is tempting to say that this represents "Web surfing" that should travel with lower priority than packetized voice, say. However, much of this "Web surfing" may represent (in the future, if not now) customers making product inquiries or purchases, and it is questionable whether it should be consigned to a congested channel. In general, it is not at all clear whether high priority traffic will be a small fraction of the total. Further, as IPSec and other encryption methods are applied to an increasing fraction of the traffic, it will be hard to decide based on packets alone what application originated them. It will then be necessary to resort to elaborate signaling schemes to make prioritization work, further complicating an already complex system.

Advocates of QoS appear not to have produced any quantitative analyses of the advantages of their schemes, nor of the costs of implementing them, although there are some admissions that those costs are not going to be negligible [FergusonH]. Most of the costs are likely to be hard to quantify ones, those involving application developers, end users, and especially network administrators. All those costs are already high.

Many QoS schemes would require deployment throughout the Internet to be effective. That is a serious defect, given the tremendous heterogeneity and lack of centralized control, since there are plenty of ancient systems around, and no signs that changes will be any faster in the future. Further, one of the most pressing issues for the future of the Internet appears to be that of ISP interconnections. It is already complicated, with no clear models as to what service agreements and settlements will be common. Throwing in QoS questions would only make this issue harder to resolve.

In conclusion, total system costs suggest that it would be best to confine most QoS measures to the edges of the network, for those infrequent cases where it is absolutely necessary. The core of the network would be best run at a uniformly high standard for all traffic, with no end-to-end coordination. This would make it easier to reach and monitor agreements among customers and service providers, and would allow application developers and users a simple view of the network. Some QoS methods could still be used in the core of the network, provided they were invisible to the end users. (Fair Queueing is an example of such a technique.)

## 12. Usage-sensitive pricing: Panacea or boondoggle?

There are strong arguments for usage-sensitive pricing even in a network with a single class of service (see [FishburnO, Odlyzko1], for example, and the references there). Even in today's undifferentiated service model, there is a strong trend by corporate networks to charge business units according to their usage of the network, to promote greater awareness of costs and more rational use of resources. These arguments will become even more compelling as we move into the bandwidth explosion period. For example, in the residential market, we will want to offer Internet access to neighbors, one of whom has just a POTS line with a 28.8 Kbps modem, the other a cable modem capable of 1.5 Mbps. The prices we charge them cannot differ by much, since otherwise the cable modem customer will not buy our service. On the other hand, we do not want that cable modem customer to engage in arbitrage, selling 28.8 Kbps service to his entire neighborhood, since we won't be able to afford to carry a constant 1.5 Mbps traffic stream for the price of a single residential account.

If we offer a network with different classes of service, there will be no way to avoid the necessity of usage-sensitive pricing. Customers will have to have a clear incentive to send only the traffic that requires high quality service on the better channels. Pricing does not have to be onerous to have a large effect. Even a small charge can have a noticeable effect on human behavior. For example, New York City has recently started installing water meters. They were led to this step by statistics that showed municipalities with water metering have per capita water consumption up to a half smaller than New

York City. As another example, usage of courier services is not explicitly restricted in corporations, but knowledge of the extra costs and management spot checks lead employees to limit their use to important cases. In Internet access as well, small prices have often had large effects. For example, in Europe, residential Internet access is largely on a flat rate basis, but users have to pay the phone company for each minute of line use. The result is a huge surge in Internet use in the evening, when the relatively modest per-minute phone charges drop.

Although there are strong arguments for usage-sensitive pricing, the burden of introducing it would be heavy. Careful accounting of traffic volumes would be required, as well as techniques for either sender or receiver paying. The switched voice phone system is often derided for its expensive billing system. However, that system is designed to be accurate, reliable, and auditable. Existing traffic monitoring schemes for packet networks are none of these things. When one examines current traffic statistics, there are many missing or clearly incorrect data points. To introduce a robust usage-sensitive pricing mechanism would be a long and expensive undertaking.

If there are to be per-byte charges on the Internet, they should be based on edge pricing; i.e., charging at the entrance and exit from the network, not based on what happens at internal nodes. Congestion pricing, in which tolls are levied only when there is too much traffic, have been advocated by MacKie-Mason and Varian and others [MacKieMV] and do have nice optimality properties. However, they involve complex operations in the heart of the network, where resources are most scarce, and also go against user preferences [Clark1, Clark2, Odlyzko1, ShenkerCEH].

## 13. Alternatives to Quality of Service

The simplest alternative to QoS is to simply provide the big pipes that will accommodate user demands, and continue to treat all packets equally as well as to charge fixed monthly fees only. As was shown earlier, this strategy would work with current networks and users. Evidence shows that users would not go wild and instantly saturate the newly available bandwidth. What would be required is to replace existing private networks by public ones of the same aggregate bandwidth. Such networks would exploit the complementarity of traffic demands and the smoothing effect of aggregation of many sources to provide universally high quality transmission. This would minimize the cost to application developers, users, and network managers, and would let them concentrate on other serious issues.

There are reasons (outlined in Section 6) to expect that this strategy might work in the future as well, as prices of broadband capacity decline. However, it would be a race between demand and supply, and one could argue that the balance would be unstable, or else would settle into a situation like that of the current public Internet, where quality is mostly inadequate. We already see the traditional model breaking down. Backbone providers are beginning to levy extra charges on ISPs, and Web hosting services do have usage-sensitive pricing. When ISPs load their lines 10 times more heavily than corporate customers, the arbitrage opportunities and the disparities in costs imposed by different customers are just too large to ignore. Since these factors will continue in the future, it might be necessary to have a slightly more intrusive pricing scheme that does provide correct incentives for service providers and users. I propose using a uniformly high transmission quality with all packets treated equally on the backbones, and with "expected usage profile" to price access to the backbones. It resembles Clark's "expected capacity allocation" [Clark1, Clark2], except that it would not be used on a short time scale for congestion control, as Clark's scheme would, and would not have his "in" and "out" bits. Users would enter into contracts with service providers that would specify within broad ranges the volumes of traffic they would be transmitting. Such contracts could be renegotiated quarterly or annually, in a pattern used in the insurance industry. Contracts might even specify what fraction of traffic would be have to be "cooperative" (like TCP, which slows down in the presence of congestion), and what fraction would be sent at different times of day.

It is possible that even the expected usage profile approach would not be sufficient to provide good service, and that some method that would provide congestion control on short time scales would still be required. In that case, as a slightly more intrusive scheme, and the strongest that is likely to be justified I propose 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]. While that paper was written when I thought the Internet was much more congested than it is, the basic intuition of PMP still appears to be valid, namely to have 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).

PMP would not optimize anything except simplicity of the system (subject to having usage-sensitive pricing). There are other proposals, described or referenced in [McKnightB], that are provably better in terms of optimizing some economic criteria. However, all the evidence collected in this paper shows that in the rush towards the Information Society, where just keeping the Internet going in the face of rapid change is an overwhelming challenge, it is simplicity that is most desirable.

## 14. Conclusions

Almost all of the Internet is overengineered, designed to accommodate even peak loads, and likely to remain that way. Only a few key parts are congested. One option is to implement a variety of Quality of Service measures, to attempt to provide adequate service for the crucial applications. However, it is not clear how efficient that would be. Further, the effectiveness of QoS would depend on substantial modifications throughout the whole computing and communications infrastructure. Considering the economics of the whole system, and the likely declines in costs of high bandwidth connections, it appears preferable to overengineer the few parts that are now bottlenecks. Should that not be feasible, only the simplest possible schemes should be implemented, to minimize the burden on the users and network administrators. Most of the QoS tools are useful, but should be implemented either only locally, or else deep inside the network, invisible to the users.

**Acknowledgements:** I thank the many people who have already been acknowledged in [Odlyzko2], as well as Roderick Beck, Arthur Berger, Douglas Carson, Costas Courcoubetis, John Denker, Amy Friedlander, Bill Gale, Frank Kelly, Jeff Lagarias, Rodolfo Milito, Louis Monier, Philip Murton, Craig Partridge, Vern Paxson, Richard Steinberg, David Reed, and Roger Watt for comments and information that were invaluable in the preparation of this paper.

## References


[ActonP] J. P. Acton and R. E. Park, Response to time-of-day electricity rates by large business customers: Reconciling conflicting evidence, Report R-3477-NSF, RAND Corp., 1987.

[AnaniaS] L. Anania and R. J. Solomon, Flat–the minimalist price, pp. 91-118 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

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

[Becker] G. S. Becker, Good-bye, tollbooths and traffic jams?, *Business Week,* May 18, 1998.

[Boardwatch] *Boardwatch* magazine, ⟨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/⟩.

[Clark1] D. D. Clark, Adding service discrimination to the Internet, *Telecommunications Policy,* 20 (1996), pp. 169-181.

[Clark2] D. D. Clark, Internet cost allocation and pricing, pp. 215-252 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Clark3] M. P. Clark, *Networks and Telecommunications: Design and Operations,* 2nd ed., Wiley, 1998.

[CoffmanO] K. G. Coffman and A. M. Odlyzko, The size and growth rate of the Internet. 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⟩.

[Cray] A. Cray, Secure VPNs: Lock the data, unlock the savings, *Data Communications,* May 21, 1997. Available at ⟨http://www.data.com/roundups/vpn_servers.html⟩ and ⟨http://www.data.com/roundups/vpn_servers_save.html⟩.

[DataComm] The 1998 Data Comm Market Forecast, *Data Communications,* Dec. 1997, pp. 54ff. Available at ⟨http://www.data.com/whats_hot/forecast98.html⟩.

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

[Egan] B. L. Egan, *Information Superhighway Revisited: The Economics of Multimedia,* Artech House, 1996.

[FCC] U.S. Federal Communications Commission, *Trends in Telephone Service,* Feb. 1998. 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,* April 1998 (to appear).

[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. Available at ⟨http://www.research.att.com/∼amo⟩.

[FishburnOS] P. C. Fishburn, A. M. Odlyzko, and R. C. Siders, Fixed fee versus unit pricing for information goods: competition, equilibria, and price wars, *First Monday,* vol. 2, no. 7 (July 1997), ⟨http://www.firstmonday.dk/⟩. Also to appear in *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property,* D. Hurley, B. Kahin, and H. Varian, eds., MIT Press. Available at ⟨http://www.research.att.com/∼amo⟩.

[Friedlander] A. Friedlander, *Power and Light: Electricity in the U.S. Energy Infrastructure, 1870-1940,* Corp. for National Research Initiatives, 1996.

[GaleK] W. A. Gale and R. Koenker, Pricing interactive computer services, *Computer Journal,* vol. 27, no. 1 (1984), pp. 8-17.

[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⟩.

[Gates] B. Gates, with N. Myhrvold and P. Rinearson, *The Road Ahead,* Viking, 1995.

[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⟩.

[HausmanN1] W. J. Hausman and J. L. Neufeld, Engineers and economists: Historical perspectives on the pricing of electricity, *Technology and Culture* 30 (1989), pp. 83-104.

[HausmanN2] W. J. Hausman and J. L. Neufeld, Time-of-day pricing in the U.S. electronic power industry at the turn of the century, *Rand J. Economics* 15 (1984), pp. 116-124.

[Hughes] T. P. Hughes, *Networks of Power: Electrification in Western Society, 1880-1930,* Johns Hopkins Univ. Press, 1983.

[IDC] International Data Corporation, Report IDC #15062, Business Network Services Bulletin, WorldCom: Putting it all together. Jan. 1998. Corrected in Feb. '98, in IDCFlash, IDC #15549.

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

[JanahTD] M. Janah with M. E. Thyfault and B. Davis, Customized bandwidth, *Information Week,* May 11, 1998, pp. 18-20.

[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.

[LOC] Library of Congress, data traffic statistics, available at ⟨http://lcweb.loc.gov/stats/⟩.

[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⟩.

[MacKieMV] J. K. MacKie-Mason and H. R. Varian, Economic FAQs about the Internet, pp. 27–62 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. A version is available in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Makris] J. Makris, Guerrilla ISPs: Young guns threaten the ISP regime–but can they really give the people what they want?, *Data Communications,* May 1998, pp. 90–102. Available at ⟨http://www.data.com/roundups/guerrilla.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/⟩.

[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⟩.

[MitchellV] B. M. Mitchell and I. Vogelsang, *Telecommunications Pricing: Theory and Practice,* Cambridge Univ. Press, 1991.

[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⟩.

[Mukherjee] A. Mukherjee, On the dynamics and significance of low frequency components of Internet load, *Internetworking: Research and Experience*, 5 (1994), pp. 163-205.

[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, A modest proposal for preventing Internet congestion. 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. 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.

[ParkM] R. E. Park and B. M. Mitchell, Optimal peak-load pricing for local telephone calls, Report R-3401-1-RC, RAND Corp., 1987.

[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/⟩.

[Platt] H. L. Platt, The cost of energy: Technological change, rate structures, and public policy in Chicago, 1880-1920, *Urban Studies* 26 (1989), pp. 32-44.

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

[Rendleman] J. Rendleman, Connectivity crunch stymies IT access to high-speed lines, *PCWeek,* April 13, 1998, pp. 1 and 20. Available at ⟨http://www.zdnet.com/pcweek/news/0413/13t1.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⟩.

[Shenker] S. Shenker, Fundamental design issues for the future Internet, *IEEE J. Selected Areas Comm.,* 13 (1995), pp. 1176-1188.

[ShenkerCEH] S. Shenker, D. Clark, D. Estrin, and S. Herzog, Pricing in computer networks: reshaping the research agenda, *Telecommunications Policy,* 20 (1996), pp. 183-201.

[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⟩.

[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⟩.

[UUNet] UUNet Access Services, available at ⟨http://www.us.uu.net/html/access_services.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/⟩.

[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: Number and usage of U.S. telephone lines.

| year | lines (millions) | local calls (minutes per day per line) | intrastate toll calls (minutes per day per line) | interstate toll calls (minutes per day per line) |
|---|---|---|---|---|
| 1980 | 102.2 | 39 | 4 | 4 |
| 1984 | 112.6 | 40 | 5 | 5 |
| 1988 | 127.1 | 39 | 6 | 7 |
| 1992 | 143.3 | 37 | 6 | 7 |
| 1996 | 166.3 | 40 | 6 | 8 |

Table 2: Effective bandwidth of long distance networks, year-end 1997.

| network | bandwidth (Gbps) |
|---|---|
| US voice | 350 |
| Internet | 75 |
| other public data networks | 40 |
| private line | 330 |

Table 3: U.S. data communications market (derived from Dec. 1997 issue of *Data Comunications*).

| | 1997 revenues (millions) | 1997 growth rate | 1998 projected growth rate |
|---|---|---|---|
| PRODUCTS: | | | |
| LAN switches | $4,297 | 75% | 38% |
| Routers | 3,095 | 12 | 11 |
| Hubs | 1,789 | −20 | −12 |
| NICs | 2,115 | −20 | −14 |
| Servers | 18,370 | 15 | 16 |
| Wiring | 2,685 | 22 | 15 |
| Network operating systems | 1,921 | 13 | 15 |
| Remote access devices | 1,604 | 35 | 27 |
| Modems | 3,077 | −7 | 13 |
| Frame Relay switches and access devices | 1,231 | 36 | 24 |
| PBXs | 4,805 | −3 | 27 |
| Videoconferencing equipment | 930 | 30 | 25 |
| Network and systems management | 3,094 | 25 | 25 |
| Other | 6,713 | 29 | 30 |
| PRODUCTS TOTAL | $55,726 | 14% | 19% |
| NETWORK SUPPORT SERVICES | $7,880 | 14% | 17% |
| DATA AND NETWORK SERVICES: | | | |
| Leased lines | $9,750 | 16% | 10% |
| ISDN | 1,036 | 35 | 60 |
| Frame Relay | 2,330 | 106 | 103 |
| Commercial Internet services | 1,517 | 42 | 103 |
| Other | 1,378 | 22 | 40 |
| DATA AND NETWORK SERVICES TOTAL | $16,011 | 28% | 38% |
| PRODUCTS AND SERVICES TOTAL | $79,617 | 17% | 23% |

Table 4: Average utilization levels

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

Table 5: Traffic on long distance networks, year-end 1997.

| network | traffic (TB/month) |
|---|---|
| US voice | 40,000 |
| Internet | 2,500 - 4,000 |
| other public data networks | 500 |
| private line | 3,000 - 5,000 |

Table 6: Retail leased line market in the U.S., end of year 1997. Bandwidth in Gbps, revenue in billions of dollars.

| line speed | no. lines | bandwidth Gbps | projected 98 growth | revenue $ billions |
|---|---|---|---|---|
| 56/64 & lower | 447,530 | 57 | -1% | 4.87 |
| fractional T1 | 19,880 | 10 | 2% | 0.26 |
| T1 | 98,850 | 304 | 7% | 4.58 |
| T3 & higher | 3,010 | 259 | 34% | 0.72 |

Table 7: Costs of transmitting a megabyte of data over various networks.

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

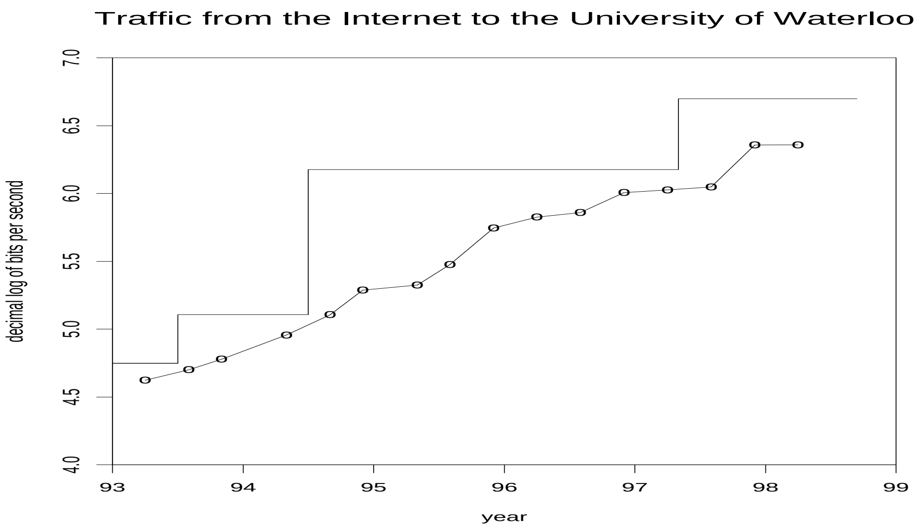


Figure 1: Traffic on the link from the public Internet to the University of Waterloo. The line with circles shows average traffic during the month of heaviest traffic in each school term. The step function is the full capacity of the link. By permission of University of Waterloo.

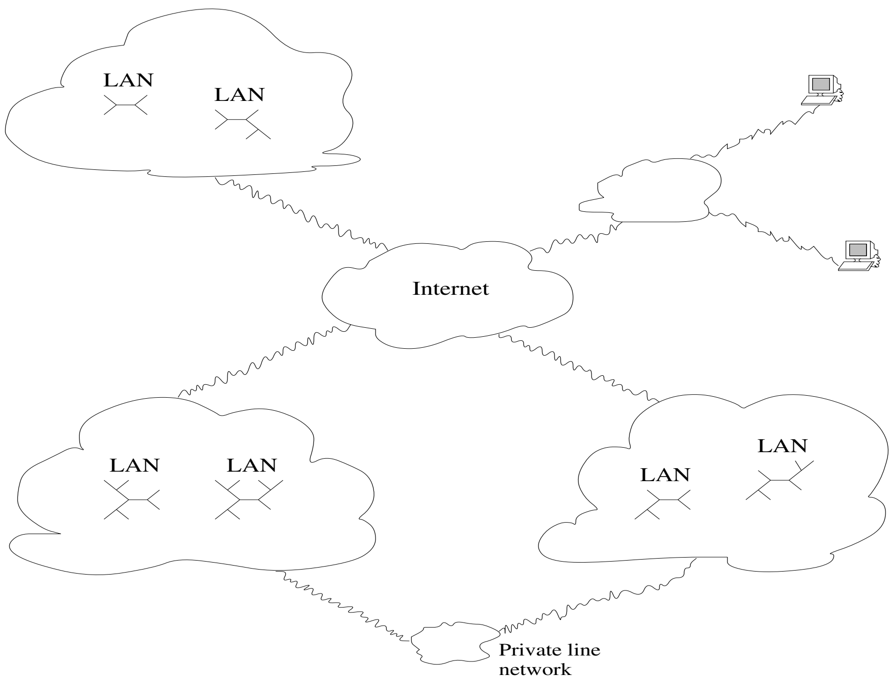


Figure 2: The core Internet and other data networks. Artwork by Thelma Pickell and Sue Pope.

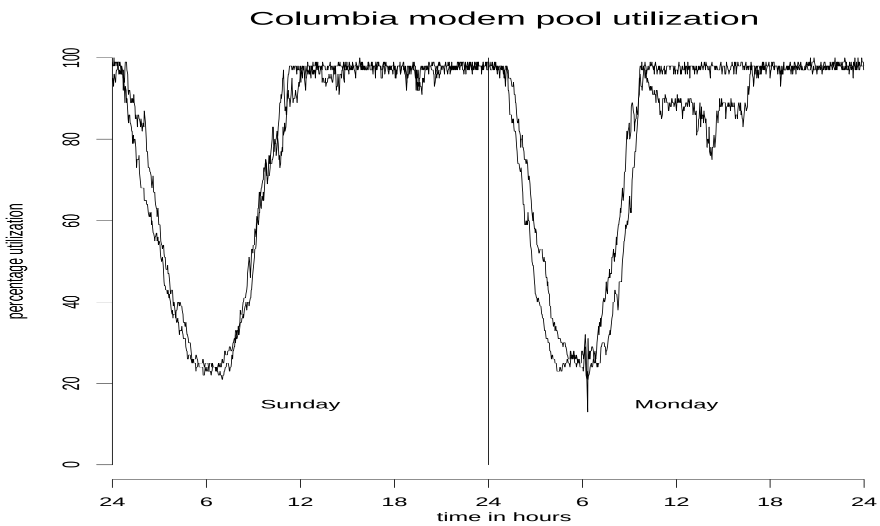


Figure 3: Columbia University modem pool utilization during the Sunday and Monday pairs of Jan. 18-19 and Jan. 25-26, 1998. 2-minute averages. By permission of Columbia University.

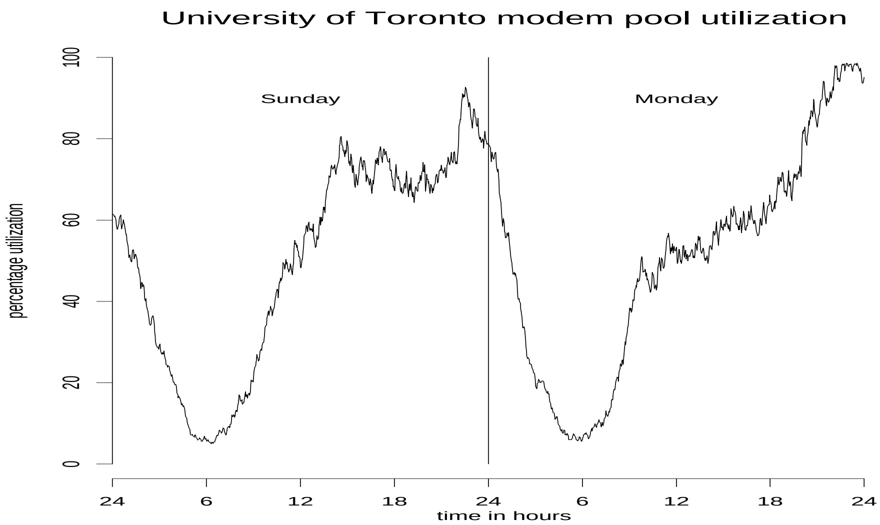


Figure 4: University of Toronto modem pool utilization during Mar. 7-8, 1998. 2-minute averages. By permission of University of Toronto.

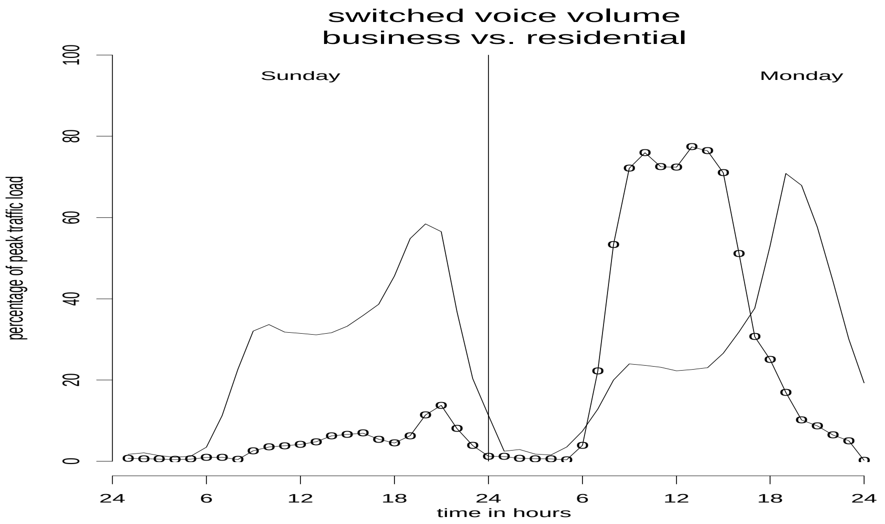


Figure 5: Residential (thin line) and business (line with circles) voice traffic on U.S. long distance switched voice networks, as percentage of peak traffic on those networks.

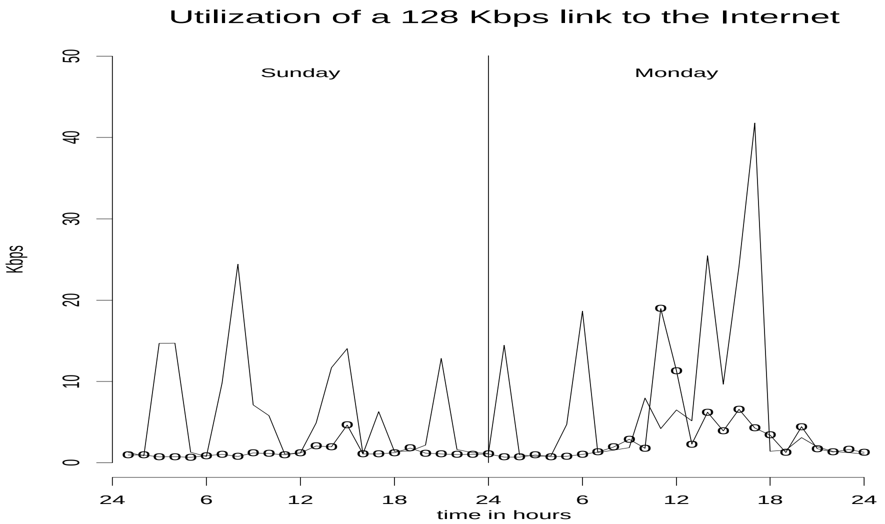


Figure 6: Utilization of a 128 Kbps dedicated business connection to the Internet during February 22 and 23 (thin line) and March 1 and 2 (line with circles), 1998. Hourly averages of traffic to the customer.

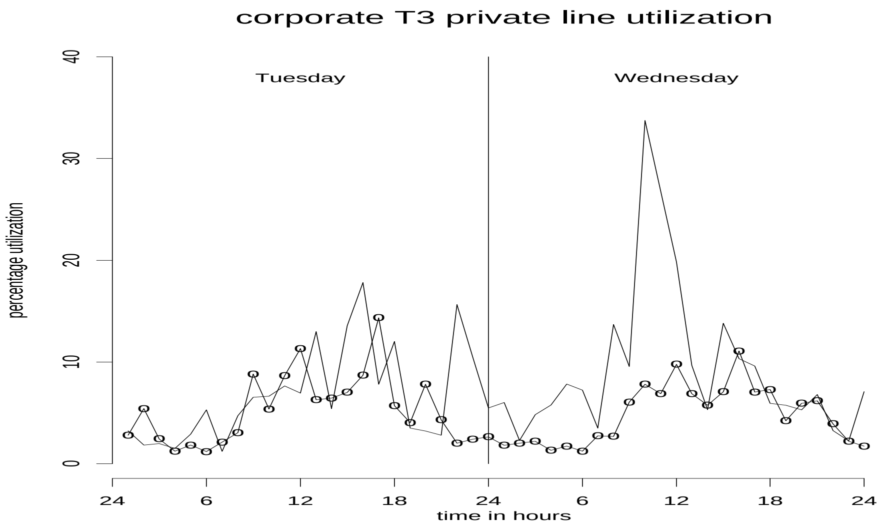


Figure 7: Traffic on two corporate T3 private lines. Hourly averages.

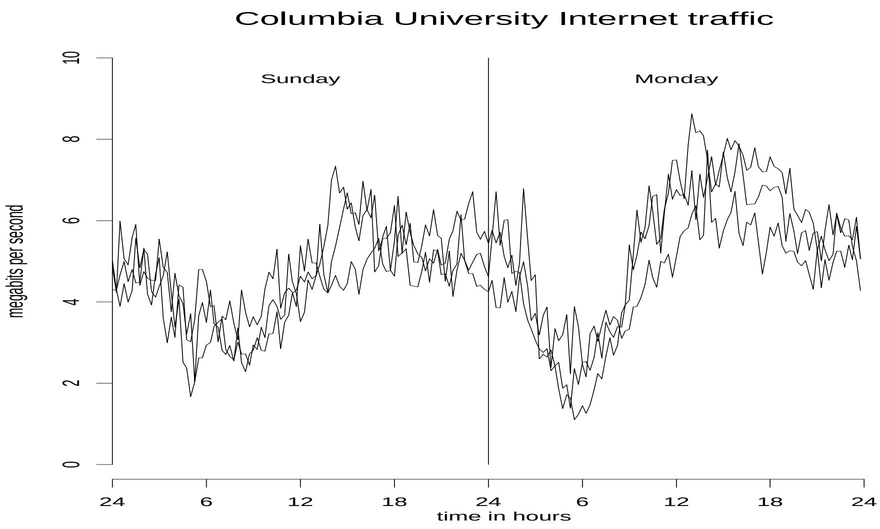


Figure 8: Traffic into Columbia University on Jan. 18, 19, 25, 16 and Feb. 1 and 2, 1998. 15-minute averages, Eastern Standard Time. By permission of Columbia University.

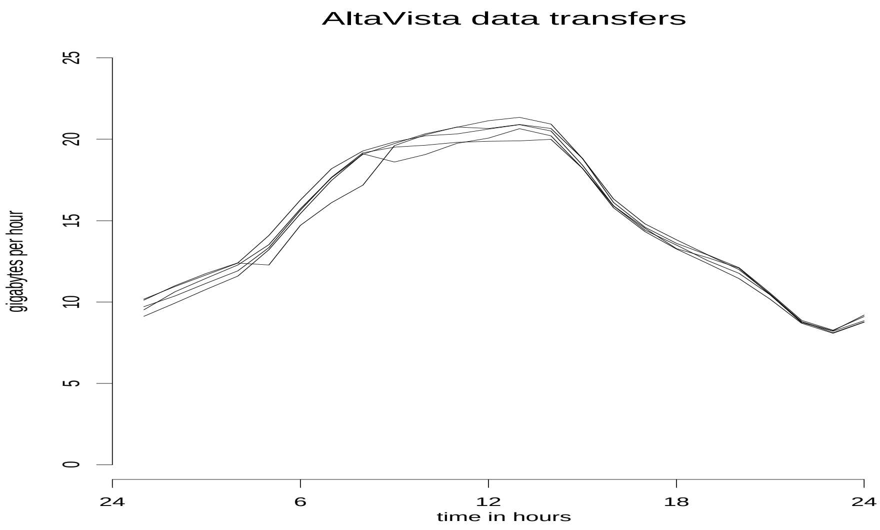


Figure 9: Data transfers by the AltaVista search engine, during several workdays in January 1998. Hourly averages, Pacific Standard Time. By permission of Digital Equipment Corp.

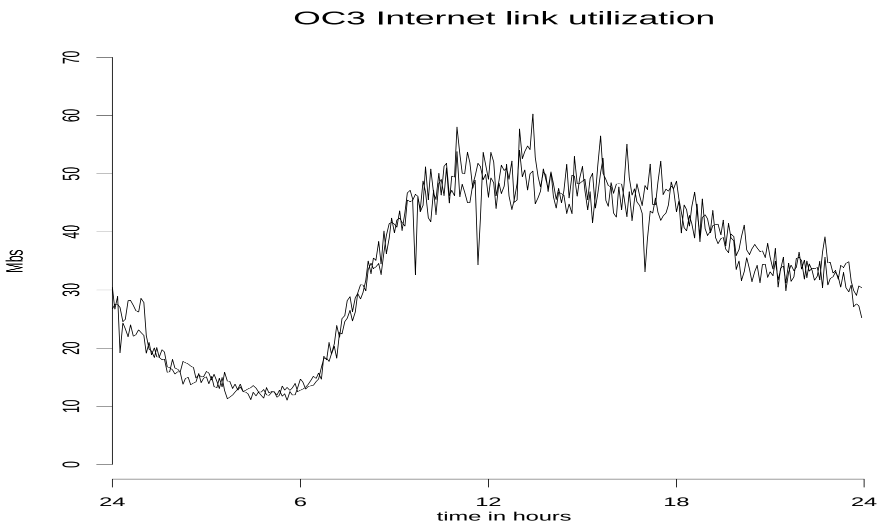


Figure 10: Traffic to the north on an MCI OC3 Internet trunk on Tuesday and Wednesday, August 26 and 27, 1997. 5-minute averages, Eastern Standard Time. By permission of MCI.

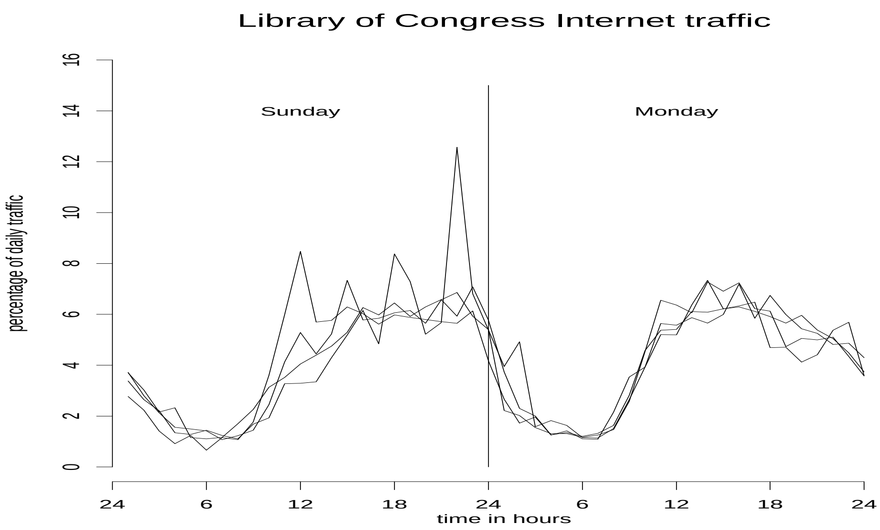


Figure 11: Library of Congress traffic over several Sunday-Monday pairs in late 1997 and first half of 1998. Shows fraction of daily traffic transmitted during each hour.