Pricing the Internet

by

Jeffrey K. MacKie-Mason
Hal R. Varian
University of Michigan

April 1993
Current version: February 10, 1994

Abstract. This paper was prepared for the conference "Public Access to the Internet," JFK School of Government, May 26–27 , 1993. We describe the technology and cost structure of the NSFNET backbone of the Internet, and discuss how one might price Internet access and use. We argue that usage-based pricing is likely to be necessary to control congestion on the Internet and propose a particular implementating of usage-based pricing using a "smart market".

Keywords. Networks, Internet, NREN, NII.

Address. Hal R. Varian, Jeffrey K. MacKie-Mason, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220. E-mail: `jmm@umich.edu`, `halv@umich.edu`.

Pricing the Internet

Jeffrey K. MacKie-Mason
Hal R. Varian

On December 23, 1992 the National Science Foundation (NSF) announced that it will cease funding the ANS T3 Internet backbone in the near future. This is a major step in the transition from a government-funded to a commercial Internet. This movement has been welcomed by private providers of telecommunication services and businesses seeking access to the Internet.

No one is quite sure about how this privatization will work; in particular, it is far from clear how use of the privatized Internet will be priced. Currently, the several Internet backbone networks are public goods with exclusion: usage is essentially free to all authorized users. Most users are connected to a backbone through a "pipe" for which a fixed access fee is charged, but the user's organization nearly always covers the access fee as overhead without any direct charge to the user.[1] None of the backbones charge fees that depend at the margin on the volume of data transmitted. The result is that the Internet is characterized by "the problem of the commons," and without instituting new mechanisms for congestion control it is likely to soon suffer from server "overgrazing." We shall propose an efficient pricing structure to manage congestion, encourage network growth, and guide resources to their most valuable uses.

We first describe the Internet's technology and cost structure, since a feasible and efficient pricing scheme must reflect both technology and costs. We then describe congestion problems in the network, and some past proposals to control it. We turn to pricing by first describing in general terms the advantages and disadvantages of using pricing to control congestion, followed by the details of our proposed pricing structure. We devote particular attention to a novel feature of our proposal: the use of a "smart market" to price congestion in real time.

We wish to thank Guy Almes, Eric Aupperle, Hans-Werner Braun, Paul Green, Dave Katz, Mark Knopper, Ken Latta, Dave McQueeny, Jeff Ogden, Chris Parkin, Scott Shenker and Paul Southworth for helpful discussions, advice and data. We are also grateful to James Keller and Miriam Avins for extensive, helpful editorial advice. MacKie-Mason was visiting the Department of Economics, University of Oslo when this paper was completed.

[1] Most users of the NSFNET backbone do not pay a pipeline fee to ANS, the service provider, but instead pay for a connection to their "regional" or mid-level network, which then is granted a connection to the NSFNET.

1. Internet Technology and Costs

The Internet is a network of networks. We shall focus on backbone networks, although most of our pricing ideas apply equally well to mid-level and local area networks. There are essentially four competing backbones for the Internet: ANSnet, PSInet, Alternet, and SprintLINK.[2] ANS is a non-profit that was formed in 1990 to manage the publicly-funded NSFNET for research and educational users. ANSnet now provides the backbone service for NSFNET, as well as backbone service for commercial users through its subsidiary, ANS CO+RE, Inc. PSInet and Alternet are independent commercial providers of backbone Internet services to commercial and non-commercial users. Sprint, of course, is a major telecommunications provider as well as a provider of Internet transport services.

The Internet networks use packet-switching communications technology based on the TCP/IP protocols. While much of the traffic moves across lines leased from telephone common carriers, packet-switching technology is quite different from the circuit-switching used for voice telephony. When a telephone user dials a number, a dedicated path is set up between the caller and the called number. This path, with a fixed amount of network resources, is held open; no other caller can use those resources until the call is terminated.[3] A packet-switching network, by contrast, uses "statistical multiplexing": each circuit is shared by many users, and no open connection is maintained for a particular communications session. A data stream is broken up into small chunks called "packets." When a packet is ready, the computer sends it onto the network. When one computer is not sending a packet, the network line is available for packets from other computers. The TCP (Transmission Control Protocol) specifies how to break up a datastream into packets and reassemble it; the IP (Internet Protocol) provides the necessary information for various computers on the Internet (the routers) to move the packet to the next link on the way to its final destination.

The data in a packet may be 1500 bytes or so. Recently the average packet on NSFNET carries about 200 bytes of data (packet size has been steadily increasing). On top of these 200 bytes the

[2] In addition, a new alliance called CoREN has been formed between eight regional networks and MCI. This represents a move away from the traditional backbone structure towards a mesh-structured set of overlapping interconnections.

[3] Some telephone lines are multiplexed, but they are synchronous: $1/N$th of the line is dedicated to each open circuit no matter how lightly used is that circuit.

TCP/IP headers add about 40; thus about 17% of the traffic carried on the Internet is simply header information.

Packetization allows for the efficient use of expensive communications lines. Consider a typical interactive terminal session to a remote computer: most of the time the user is thinking. The network is needed only after a key is struck or when a reply is returned.[4] Holding an open connection would waste most of the capacity of the network link. Instead, the computer waits until after a key is struck, at which point it puts the keystroke information in a packet which is sent across the network. The rest of the time the network links are free to be used for transporting packets from other users.

The other distinguishing feature of Internet technology is that it is "connectionless."[5] This means that there is no end-to-end setup for a session; each packet is independently routed to its destination. When a packet is ready, the host computer sends it on to another computer, known as a router (or switch). The router examines the destination address in the header and passes the packet along to another router, chosen by a route-finding algorithm. A packet may go through 30 or more routers in its travels from one host computer to another. Because routing is dynamically calculated, it is entirely possible for different packets from a single session to take different routes to the destination.[6]

The postal service is a good metaphor for the technology of the Internet (Krol (1992), pp. 20–23). A sender puts a message into an envelope (packet), and that envelope is routed through a series of postal stations, each determining where to send the envelope on its next hop. No dedicated pipeline is opened end-to-end, and thus there is no guarantee that envelopes will arrive in the

---

[4] Some interactive terminal programs collect keystrokes until an `Enter` or `Transmit` key is struck, then sends the entire "line" off in a packet. However, most Internet terminal sessions use the `telnet` program, which sends each keystroke immediately in a separate packet.

[5] Some packet-switching networks are "connection-oriented" (notably, X.25 networks, such as Tymnet and frame-relay networks). In such a network a connection is set up before transmission begins, just as in a circuit-switched network. A fixed route is defined, and information necessary to match packets to their session and defined route is stored in memory tables in the routers. Thus, connectionless networks economize on router memory and connection set-up time, while connection-oriented networks economize on routing calculations (which have to be redone for every packet in a connectionless network).

[6] Dynamic routing contributes to the efficient use of the communications lines, because routing can be adjusted to balance load across the network. The other main justification for dynamic routing is network reliability, since it gives each packet alternative routes to their destination should some links fail. This was especially important to the military, which funded most of the early TCP/IP research to improve the ARPANET.

sequence they were sent, or follow exactly the same route.

The TCP protocol enables packets to be identified and reassembled in the correct order. TCP prefaces the data in a packet with a header containing the source and destination ports, the sequence number of the packet, an acknowledgment flag, and so on. The header takes up 20 or more bytes. TCP sends the packet to a router, a computer that is in charge of forwarding packets to their next destination. At the routers, IP adds another header (another 20 or more bytes) containing source and destination addresses and other information needed for routing the packet. The router then calculates the best next link for the packet to traverse, and sends it on. The best link may change minute by minute, as the network configuration changes.[7] Routes can be recalculated immediately from the routing table if a route fails. The routing table in a switch is updated nearly continuously.

Over the past five years, the speed of the NSFNET backbone has increased from 56 Kbps to 45 Mbps ("T3" service).[8] The newer backbones have also upgraded to 45 Mbps. These lines can move about 1,400 pages of text per second; a 20-volume encyclopedia can be sent across the Internet in half a minute. Many regional networks still provide T1 (1.5Mbps) service, but these too are being upgraded.

The transmission speed of the Internet is remarkably high. We recently tested the transmission delay at various times of day and night for sending a packet to Norway from Ann Arbor, Michigan. Each packet traversed 16 links: the IP header was read and modified 16 times, and 16 different routers calculated the best next link. Despite the many hops and substantial packetization and routing, the longest delay on one representative weekday was only 0.333 seconds (at 1:10 PM EST); the shortest delay was 0.174 seconds (at 5:13 PM EST).[9]

Current backbone network costs

The postal service is a good metaphor for packet-switching technology, not for the cost structure of Internet services. Most of the costs of providing the Internet are more-or-less independent of the

[7] Routing is based on a dynamic knowledge of which links are up and a static "cost" assigned to each link. Currently routing does not take congestion into account. Routes can change when hosts are added or deleted from the network (including failures), which happens often with about 2 million hosts and over 21,000 subnetworks.

[8] "Kbps" is thousand (kilo) bits per second; "Mbps" is million (mega) bits per second.

[9] While preparing the final manuscript we repeated our delay experiment for 20 days in October–November, 1993. The range in delay times between Ann Arbor and Norway was then 0.153 seconds and 0.303 seconds.

level of usage of the network; i.e., most of the costs are fixed costs. If the network is not saturated the incremental cost of sending additional packets is essentially zero.[10]

The NSF in 1993 spent about $11.5 million to operate the NSFNET and provided $7 million per year in grants to help operate the regional networks.[11] NSF grants also help colleges and universities connect to the NSFNET. Using the conservative estimate of 2 million hosts and 20 million users, this implies that the 1993 NSF Internet subsidy was less than $10 per year per host, or less than $1 per user.[12]

Total salaries and wages for NSFNET have increased by a little more than one-half (about 68% nominal) over 1988–1991, a time when the number of packets delivered has increased by a factor of 128.[13] It is hard to calculate total costs because of large in-kind contributions by IBM and MCI during the initial years of the NSFNET project, but it appears that total costs for the 128-fold increase in packets have increased by a factor of about 3.2.

Two components account for most of the costs of providing a backbone network: communications lines and routers. Lease payments for lines and routers accounted for nearly 80% of the 1992 NSFNET costs. The only other significant cost is for the Network Operations Center (NOC), which accounts for roughly 7% of total cost.[14] Thus we focus on the costs of lines and routers.

We have estimated costs for the network backbone as of 1992–93.[15] A T3 (45 Mbps) trunk line running 300 miles between two metropolitan central stations could be leased for about $32,000 per month. The cost to purchase a router capable of managing a T3 line was approximately

---

[10] In a postal service most of the cost is in labor, which varies quite directly with the volume of the mail.

[11] The regional network providers generally set their charges to recover the remainder of their costs, but there is also some subsidization from state governments at the regional level.

[12] This, of course, represents only backbone costs for NSFNET users. Total costs, including LAN and regional network costs, are higher.

[13] Since packet size has been slowly increasing, the amount of data transported has increased even more.

[14] A NOC monitors traffic flow at all nodes in the network and troubleshoots problems.

[15] We estimated costs for the network backbone only, defined to be links between common carrier Points of Presence (POPs) and the routers that manage those links. We did not estimate the costs for the feeder lines to the mid-level or regional networks where the data packets usually enter and leave the backbone, nor for the terminal costs of setting up the packets or tearing them apart at the destination.

$100,000. Assuming another $100,000 for service and operation costs, and 50-month amortization at a nominal 10% rate yields a rental cost of about $4900 per month for the router.

The costs of both lines and switching have been dropping rapidly for over three decades. In the 1960s, digital computer switching was more expensive (per packet) than lines (Roberts (1974)), but switching has since become substantially cheaper. In Table 1 we show estimated 1992 costs for transporting 1 million bits of data through the NSFNET backbone and compare these to estimates for earlier years. As can be seen, in 1992 lines cost about eight times as much as routers.

Table 1.
Communications and Router Costs
(Nominal $ per million bits)[1]

Year	Lines	Routers	Transmission Speed
1960	1.00		2.4 kbps
1962		10.00	
1963	0.42		40.8 kbps
1964	0.34		50.0 kbps
1967	0.33		50.0 kbps
1970		0.168	
1971		0.102	
1974	0.11	0.026	56.0 kbps
1992	0.00094	0.00007	45 mbps

Notes: 1. Costs are based on sending one million bits of data approximately 1200 miles on a path that traverses five routers.
Sources: 1960–74 from Roberts (1974). 1992 calculated by the authors using data provided by Merit Network, Inc.

The structure of the NSFNET backbone directly reflects its costs: lots of cheap routers manage a limited number of expensive lines. We illustrate a portion of the network in Figure 1. Each numbered square is an RS6000 router; the numbers listed beside a router are links to regional networks. In general, each packet moves through two separate routers at the entry and exit nodes. For example, if we send a message from the University of Michigan to Bell Laboratories, it will traverse link 131 to Cleveland, where it passes through two routers (41 and 40). The packet goes to New York, where it moves through another two routers (32 and 33) before leaving the backbone on link 137 to the JVNCnet regional network to which Bell Labs. Two T3 lines are navigated using four routers.

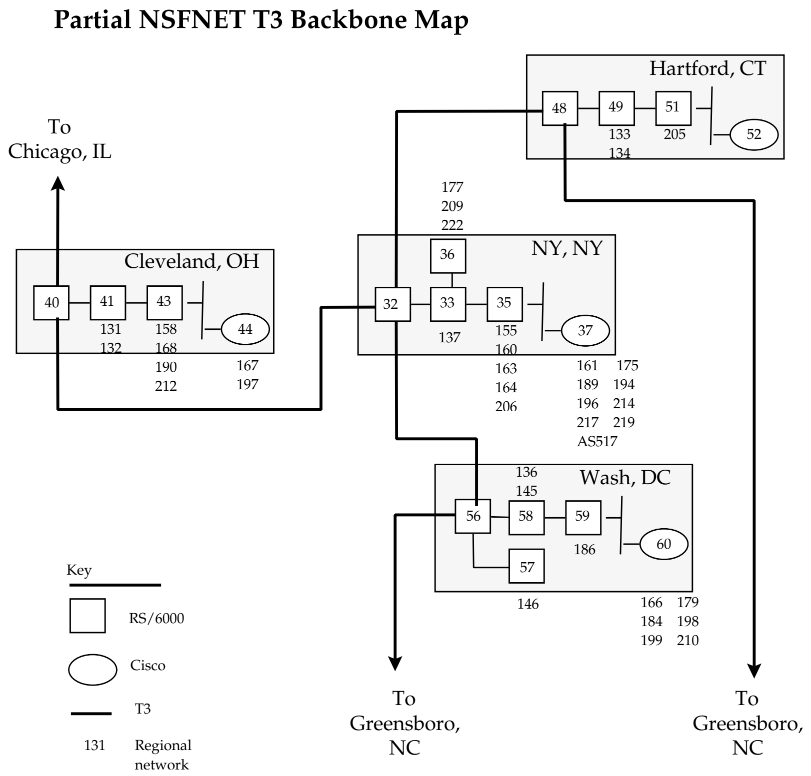


Figure 1. Network Map Fragment

Relation between technology and costs

Line and switching costs have been exponentially declining at about 30% per year (see the semi-log plot in Figure 2). But more interesting than the rapid decline is the change from expensive routers to expensive transmission links. Indeed, it was the crossover around 1970 (Figure 2) that created a role for packet-switching networks. When lines were cheap relative to switches it made sense to have many lines feed into relatively few switches, and to open an end-to-end circuit for each connection. In that way, each connection wastes transmission capacity (lines are held open whether data is flowing or not) but economizes on switching (one set-up per connection).

When switches become cheaper than lines the network is more efficient if data streams are broken into small packets and sent out piecemeal, allowing many users to share a single line. Each packet must be examined at each switch along the way to determine its type and destination, but this uses the relatively cheap switch capacity. The gain is that when one source is quiet, packets from other sources use the same (relatively expensive) lines.

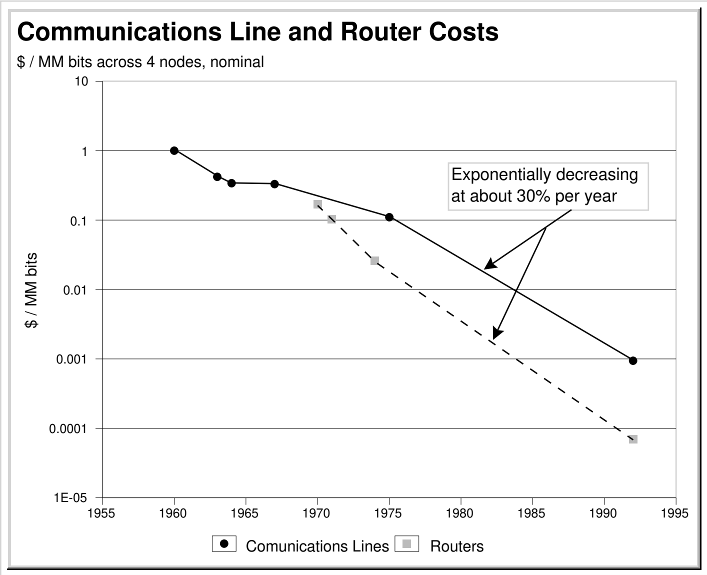


Figure 2. Trends in costs for communications links and routers.

2. Congestion Problems

The Internet is an extremely effective way to move information; for users, the Internet usually seems to work reliably and instantly. Sometimes, however, the Internet becomes congested, and there is simply too much traffic for the routers and lines to handle. At present, the only two ways the Internet can deal with congestion is to drop packets, so that some information must be resent by the application, or to delay traffic. These solutions impose external social costs: Sally sends a packet that crowds out Elena's packet; Elena suffers delay, but Sally does not for the cost she imposes on Elena.

In essence, this is the classic problem of the commons. When villagers have shared, unlimited access to a common grazing field, each will graze his cows without recognizing the costs imposed on the others. Without some mechanism for congestion control, the commons will be overgrazed. Likewise, as long as users have access to unlimited Internet usage, they will tend to "overgraze", creating congestion that results in delays and dropped packets for other users.

This section examines the extent of congestion, and explores some recent work on controlling congestion. Our proposal, which is based on charging per-packet prices that vary according to the degree of congestion, is explained later in the paper.

The Internet experienced severe congestion in 1987. Even now congestion problems are relatively common in parts of the Internet (although not yet on the T3 backbone). According to

Kahin (1992): "... problems arise when prolonged or simultaneous high-end uses start degrading service for thousands of ordinary users. In fact, the growth of high-end use strains the inherent adaptability of the network as a common channel" (page 11). Some contemplated uses, such as real-time video and audio transmission, will lead to substantial increases in the demand for bandwidth, and congestion problems will only get worse unless there is substantial increase in bandwidth.[16] For example, Smarr and Catlett write that:

> If a single remote visualization process were to produce 100 Mbps bursts, it would take only a handful of users on the national network to generate over 1Gbps load. As the remote visualization services move from three dimensions to [animation] the single-user bursts will increase to several hundred Mbps ... Only for periods of tens of minutes to several hours over a 24-hour period are the high-end requirements seen on the network. With these applications, however, network load can jump from average to peak instantaneously." Smarr and Catlett (1992), page 167.

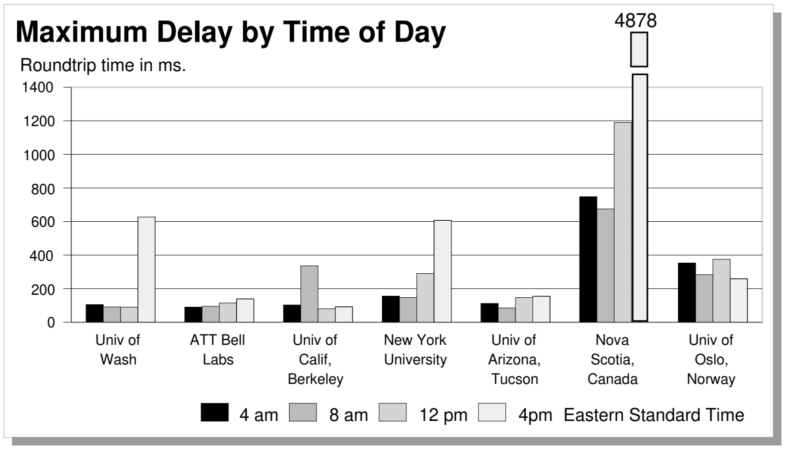
This has happened. For example, during the weeks of November 9 and 16, 1992, some packet audio/visual broadcasts caused severe delay problems, especially at heavily-used gateways to the NSFNET backbone and in several mid-level networks. Today even ordinary use is causing significant delays in many of the regional networks around the world as demand grows faster than capacity.

Of course, deliveries can be delayed for a number of other reasons. For example, if a router fails then packets must be resent by a different route. However, in a multiply-connected network, the speed of rerouting and delivery of failed packets measures one aspect of congestion, or the scarcity of the network's delivery bandwidth.

To characterize congestion on the Internet, we timed the delay in delivering packets to seven sites around the world. We ran our test hourly for 37 days during February and March 1993. Figure 3 and Figure 4 show our results from four of our hourly probes. Median and maximum delivery delays are not always proportional to distance: the delay from Michigan to New York was generally longer than to Berkeley, and delays from Michigan to Nova Scotia, Canada, were often longer than to Oslo, Norway. (Figure 3).

There is substantial variability in Internet delays. For example, the maximum and median delays vary greatly according to the time of day. There appears to be a large 4PM peak problem on

[16] We use the term bandwidth to refer to the overall capacity of the network to transport a data flow, usually measured in bits per second. The major bottleneck in backbone capacity today is in fact switch technology, not the bandwidth of the lines.



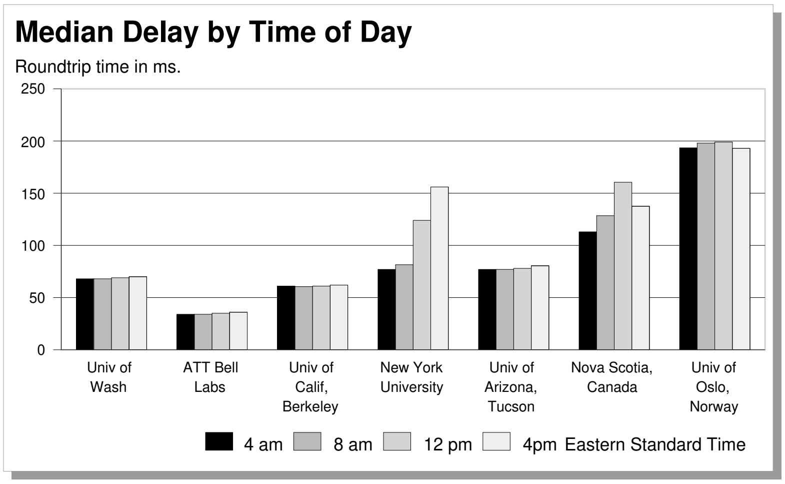


Figure 3. Maximum and Median Transmission Delays on the Internet

the east coast for packets to New York and Nova Scotia, but much less for ATT Bell Labs (in New Jersey).[17] The time-of-day variation is also evident in Figure 5(borrowed from Claffy, Polyzos, and Braun (1992)).[18]

In Figure 4 we measure delay variation by the standard deviation of delays by time of day for each destination. Delays to Nova Scotia, Canada were extraordinarily variable, yet delays to Oslo were no more variable than in transmission to New Jersey (ATT). Variability in delay fluctuates widely across times of day, as we would expect in a system with bursty traffic, but follows no

[17] The high maximum delay for the University of Washington at 4PM is correct, but appears to be aberrant. The maximum delay was 627 msec; the next two highest delays (in a sample of over 2400) were about 250 msecs each. After dropping this extreme outlier, the University of Washington looks just like UC Berkeley.

[18] Note that the Claffy et al. data were for the old, congested T1 network. We reproduce their figure to illustrate the time-of-day variation in usage; the actual levels of link utilization are generally much lower in the current T3 backbone. Braun and Claffy (1993) show time-of-day variations in T3 traffic between the US and three other countries.

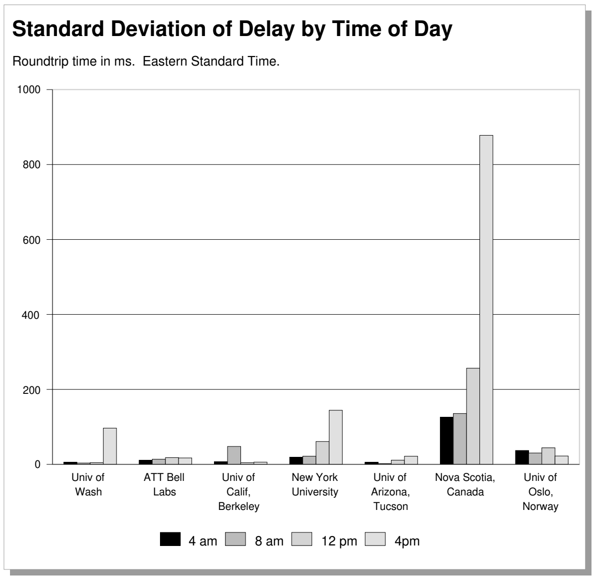


Figure 4. Variability in Internet Transmission Delays

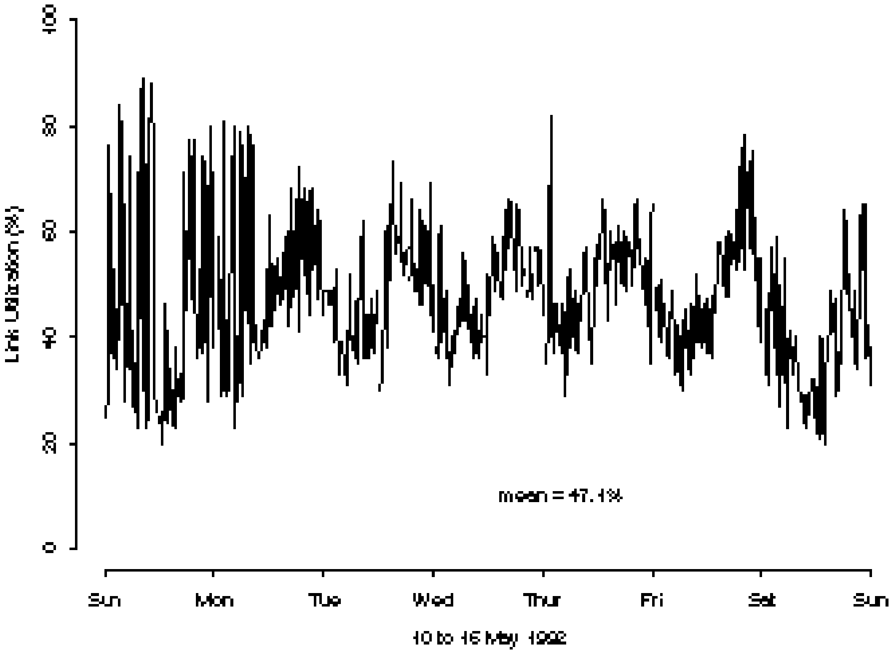


Figure 5. Utilization of Most Heavily Used Link in Each Fifteen Minute Interval (Claffy et al. (1992))

obvious pattern.

How much delay is involved, and who is inconvenienced? As seen in Figure 3, during our experiment we never experienced delays of more than 1 second in round trip time except to the site in Nova Scotia. Is that too trivial a delay to be concerned about? Probably "yes", for e-mail.

However, delays of that magnitude can be quite costly for some users and some applications.[19] For example, two-way voice communications must be limited to delays of 25 ms (500 ms with echo cancellors); likewise for interactive video.[20] Even a simple use like a terminal session can be very unpleasant if each character transmitted takes a second before echoing back to the screen. And of course, without a better mechanism for congestion control, we expect delays to increase in frequency and duration.

Controlling congestion

NSFNET usage has been growing at about 6% per month, or doubling every twelve months.[21] Although the cost of adding capacity is declining rapidly, we think it is very likely that congestion will continue to be a problem, especially as new very-high bandwidth uses (such as real-time broadcast video) become common. It is becoming increasingly important to consider how congestion in networks such as the Internet should be controlled, and much work is needed. As Kleinrock (1992) writes, "One of the least understood aspects of today's networking technology is that of network control, which entails congestion control, routing control, and bandwidth access and allocation."

There is some literature on network congestion control; see Gerla and Kleinrock (1988) for an overview. Most researchers have focused on schemes that offer different priorities and qualities of service, depending on users' needs. For example, users could send e-mail with a low priority, allowing it to be delayed during congested periods so that more time-critical traffic could get through.

In fact, IP packets contain fields called "Precedence" and "Type of Service" (TOS), but most commercial routers do not currently use these fields.[22] To facilitate the use of the TOS field,

[19] We should also note that our experiment underestimated the delay that many application might experience. We were sending probes consisting of a single packet. Some real data flows involve hundreds or thousands of packets, such as in terminal sessions, file transfers and multimedia transmissions. For these flows, periodic delays can be much longer due to the flow-control protocols implemented in the applications.

[20] An "ms" or millisecond is one one-thousandth of a second.

[21] The compound growth rate in bytes transported has been 5.8% per month from March 1991 to September 1993, and 6.4% per month from September 1992 to September 1993. This probably underestimates growth in Internet usage because traffic on other backbone routes has been growing faster.

[22] In 1986 the NSFNET experienced severe congestion and the there was some experimentation with routing based on the IP precedence field and the type of application. When the NSFNET was upgraded to T1 capacity, priority queuing was abandoned for end-user traffic.

its interpretation will probably be changed to the form described in Almquist (1992). Almquist proposes that the user be able to request that the network minimize delay, maximize total flow, maximize reliability, or minimize monetary cost. Prototype algorithms to provide such service are described in Prue and Postel (1988); a related proposal to ease congestion is in Bohn, Braun, Claffy, and Wolff (1993). In this scheme a router looks up the destination address and examines the possible routes. Each route has a TOS number; the router looks for one that matches the TOS number of the packet.

To an economist, this is too inflexible. In particular, the TOS value “minimize monetary cost” seems strange. Of course senders would want to minimize monetary cost for a given quality of service: that is an objective, not a constraint. Also, it is unfortunate that TOS numbers do not allow for inequality relations. Normally, one would think of specifying the maximum amount that one would be willing to pay for delivery, with the assumption that less expensive service (other things being equal) would be better.

As Almquist (1992) explains, “There was considerable debate over what exactly this value [minimize monetary cost] should mean.” However, he goes on to say:

> “It seems likely that in the future users may need some mechanism to express the maximum amount they are willing to pay to have a packet delivered. However, an IP option would be a more appropriate mechanism, since there are precedents for having IP options that all routers are required to honor, and an IP option could include parameters such as the maximum amount the user was willing to pay. Thus, the TOS value defined in this memo merely requests that the network ‘minimize monetary cost.”’ Almquist (1992)

Almquist’s remarks reflect the limited attention to pricing in most research to date, especially to control congestion. But without pricing it is hard to imagine how priority schemes could be implemented. What is to stop an e-mail user from setting the highest priority if it costs nothing? What political or organizational authority should be allowed to dictate the relative priority to give college student real-time multimedia rap sessions versus elementary school interactive classrooms?[23] Cocchi, Estrin, Shenker, and Zhang (1992) and Shenker (1993) make the important point that if applications require different combinations of network characteristics (responsiveness, reliability, throughput, etc.), then some sort of pricing will be needed to sort out users’ demands for these characteristics.

[23] Enforcing externally determined priorities may be impossible anyway since bytes are bytes and it is difficult to monitor anything about the content of a data stream.

Faulhaber (1992) has considered some of the economic issues related to pricing access to the Internet. He suggests that "transactions among *institutions* are most efficiently based on *capacity per unit time.* We would expect the ANS to charge mid-level networks or institutions a monthly or annual fee that varied with the size of the electronic pipe provided to them. If the cost of providing the pipe to an institution were higher than to a mid-level network . . . the fee would be higher."

Faulhaber's suggestion makes sense for recovering the cost of a dedicated line, one that connects an institution to the Internet backbone. But we don't think that it is appropriate for charging for backbone traffic itself because the bandwidth on the backbone is inherently a shared resource—many packets "compete" for the same bandwidth. There is an overall constraint on capacity, but there is no such thing as an individual's capacity on the backbone.

Although it is appropriate to charge a flat fee to cover the costs of a network connection, it is important to charge for network usage when the network is congested. During times of congestion bandwidth is a scarce resource. Conversely, when the network is not congested the marginal cost of transporting additional packets is essentially zero; it is therefore appropriate to charge users a very low or no price for packets when the system is not congested.

One problem with usage-sensitive pricing is the cost of accounting and billing. The cost would be astronomical if network providers were required to keep detailed accounts for every packet sent (comparable to call accounting by phone companies), because packets are very small units.[24] However, the accounting load could be greatly reduced. First, given the huge number of packets traversing backbones (currently over one billion per day on the NSFNET), charges based on a statistical *sample* of packets sent might be acceptable. Second, if usage is priced only during congested periods, most packets need no accounting. Third, traditional phone company accounting systems, which seem like the natural comparison, may not be a good model. They are centralized and off-line; we think that breakthroughs are likely in the area of in-line, distributed accounting, which will substantially lower costs.[25]

[24] A vigorous, one-minute phone call on a digital network today utilizes about $60 \times 64\mathrm{K}/8$ bytes of network throughput capacity, but only 1 accounting record. This much information would require roughly 2500 average-sized IP packets, each potentially with its own accounting record if full packet accounting were required.

[25] The most obvious example is to have the billing information transmitted, and the bank account debited, through the network rather than through off-line printed bills and checks written several weeks later.

There has been some recent work to design mechanisms for usage accounting on the Internet. As a first attempt, ANS developed a usage sampling and reporting system, called COMBits, which collected aggregate measures of packets and bytes using a statistical sampling technique.[26] Unfortunately, COMBits collects data only down to the network-to-network level of source and destination; the resulting data can only be used to charge at the level of the subnetwork, and the local network administrator must split up the bill (Ruth and Mills (1992)).[27] In 1992, the a committee of the Internet standards body published a draft architecture for Internet usage reporting (Group (1992)). Braun and Claffy (1993) describe measurement of Internet traffic patterns by type of application and by international data flows, and discuss some of the accounting issues that must be solved. We are also undertaking research on methods for reducing accounting costs.

For the remainder of this paper, we assume that some amount of usage-level accounting will be economically feasible in the future, and focus on the problem of efficiently pricing network resources.

3. Should Prices Be Used?

Congestion is likely to be a serious problem in the future Internet, and past proposals to control it are unsatisfactory. We think an economic approach to allocating scarce Internet resources is warranted. Telecommunications lines, computer equipment, and labor are not free; if not employed by the Internet, they could be put to productive use in other activities. Bandwidth is also scarce: when the backbone is congested, one user's packet crowds out another's, resulting in dropped or delayed transmissions. Allocating scarce resources among competing uses is the central focus of economics. In this section we discuss the benefits and costs of using economic methods to control congestion.

Our objective is not to raise profits above a normal rate of return by pricing backbone usage. Rather, our goal is to find a pricing mechanism that will lead to the most efficient use of existing resources, and will guide investment decisions appropriately. Of course, a network need not be private to be priced; governments are perfectly capable of setting prices.[28]

---

[26] See Claffy, Braun, and Polyzos (1993) for a detailed study of sampling techniques for measuring network usage.

[27] COMBits has been plagued by problems and resistance and currently is used by almost none of the mid-level networks.

[28] In fact, many of the mid-level regional networks are government agencies, and they charge prices to connect organizations to their networks.

Currently, the Internet uses a mix of two non-price resource allocation mechanisms: randomization and first-come, first-served (FIFO). With randomization, each packet has an equal chance of getting through (or being dropped). With FIFO, all packets are queued as they arrive; if the network is congested, every packet's delay is based on its arrival time in the queue. It is easy to see why these schemes are not efficient–delay is surely more costly for some packets than for others. For example, a real-time video transmission of a heart operation to a remote expert may be more valuable than a file transfer of a recreational game or picture. Economic efficiency is enhanced if the mechanism allocating scarce bandwidth gives higher priority to uses that are more socially valuable.

We do not feel that the service provider—government or otherwise—should decide which packets are more socially valuable; Soviet experience shows that allowing bureaucrats to decide whether work shoes or designer jeans are more valuable is a deeply flawed mechanism. A price mechanism works quite differently. The provider informs users of the cost of providing services; users decide for themselves whether their packets are more or less valuable than the cost of the packet transport service. When the backbone is congested, the cost of service will be high due to the the cost of crowding out or delaying the packets of other users; if prices reflect costs only those packets with high value will be sent until congestion diminishes.

Furthermore, if network congestion is properly priced, the revenues collected from the congestion surcharges can be used to fund further capacity expansion. Under certain conditions, the fees collected from the congestion charges turn out to be just the "right" amount to spend on expanding capacity.

One common concern about pricing the Internet is that "poor" users will be deprived of access. This is not a problem with pricing itself, but with the distribution of wealth; we could ensure that certain users have sufficient resources to purchase a base level of services by redistributing initial resources through vouchers or lump sum grants.[29] Indeed, total costs will be lower in an efficient network so it will be less costly to meet distributional objectives than in an unpriced network.

[29] Food stamps are an example of such a scheme. The federal government more or less ensures that everyone has sufficient resources to purchase a certain amount of food. But food is priced, so that given one's wealth plus food stamps, the consumer still must decide how to allocate scarce resources relative to the costliness of providing those resources. The government does not guarantee unlimited access to foodstuffs, nor to all varieties of caloric substances (alcoholic beverages are not eligible).

Highways are often suggested as an analogy for the future of Internet. Many people argue that publicly provided interstate highways without tolls work well and should be the model. But this analogy is flawed. First, not all democratic governments agree that toll-free roads are the best allocation of social resources; most European countries have extensive toll systems, and even some U.S. interstates have tolls. More important, an interstate offers a single, undifferentiated service. Users who need different services pay for access to rail lines, canals, or airports. No one argues that use of *all* transportation networks should be free. The interstate highway system might be viewed as the one-size-fits-all universal access option (for those who can afford cars), with the option to pay for using a mode with a different combination of service characteristics. Likewise, a government might want to provide universal, free access to a baseline set of Internet transport services, and allow charges for usage of other services above a threshold. Appropriate free services might include plain-text e-mail (with lower priority when the network is congested) but not guaranteed, zero-delay multimedia broadcast.

Universal access and a base endowment of usage for all citizens could be provided through vouchers or other redistribution schemes. But for any given distribution of resources, how should backbone services be allocated? They are currently allocated (among paid-up subscribers) on the basis of randomization and first-come, first-served. In other words, users now pay the costs of congestion through delays and lost packets. A pricing mechanism will convert delay and queuing costs into dollar costs. If prices reflect the costs of providing the services, they will force the user to compare the value of her packets to the costs she is imposing on the system. Allocation will then be based on the value of the packets, and the total value of service provided by the backbones will be greater than under a non-price allocation scheme.[30]

In the rest of the paper we discuss how one might implement pricing that reflects the cost, including congestion costs, of providing backbone services. We begin with a review of some current pricing schemes and their relationship to costs.

[30] Furthermore, in the pricing scheme we propose, users willing to tolerate delay when the network is congested would face a usage price close to or equal to zero.

4. Current Pricing Mechanisms

Most organizations do not connect directly to the NSFNET. For example, a university typically connects to its regional network; the regional connects to the NSFNET. The regional networks (and the private backbone networks) charge their customers for access, but not actual usage. Regionals, private backbones and users are not charged for connections to or usage of the NSFNET, which has been the primary backbone of the Internet. The full costs of NSFNET have been paid by NSF, IBM, MCI and the State of Michigan through 1994.

Table 3 summarizes the prices offered to large universities by ten major providers for T1 access (1.5 Mbps).[31] There are three major components: an annual access fee, an initial connection fee and in some cases a separate charge for the equipment on the customer's premises (a router to serve as a gateway between the customer network and the Internet provider's network).[32] The current annual total cost per T1 connection is about $25,000 to $35,000.

---

[31] The fees for some providers are dramatically lower due to public subsidies.

[32] Customers generally also pay a monthly "local loop" charge to a telephone company for the line between the customer's site and the Internet provider's "point of presence" (POP), but this charge depends on mileage and is generally set by the telephone company, not the Internet provider.

Table 2.
Representative Prices for T-1 Connection*

		Fee Components		
		Annual Fee	Initial Connection Cost	Customer Premises Equipment
Service Provider	ALTERnet	24,000	8,900	incl.
	ANS	32,000	incl.	incl.
	CERFnet	20,100	3,750	incl.
	CICnet	10,000	15,000	incl.
	JvNCnet	33,165	13,850	incl.
	Michnet	24,000	14,250	incl.
	MIDnet	6,000	15,000	incl.
	NEARnet	30,000	13,500	incl.
	PREPnet	3,720	1,900	not incl.
	SURAnet	25,000	3,500	3,300

Notes:
* Prices as reported by the vendors. These are prices for a large university. There are some variations in the bundle of services provided, so the prices are not strictly comparable.

Source: Complied by Bill Yurcik, NASA/Goddard Space Flight Center, 11/13/92, with corrections by the authors.

All of the providers use the same type of pricing: an annual fee for unlimited access, based on the bandwidth of the connection. These pricing schemes provide no incentives to flatten peak demands, nor any mechanism for allocating network bandwidth during periods of congestion. It would be relatively simple for a provider to monitor a customer's usage and bill by the packet or byte. Monitoring requires that outgoing packets be counted at a single point: the customer's gateway router. However, pricing every packet would not necessarily increase efficiency, because the marginal cost of a packet is nearly zero. Since it is bandwidth that is scarce, efficient prices must reflect the current availability of bandwidth. Neither a flat price per packet nor time-of-day prices would closely approximate efficient pricing.

5. Matching Prices to Costs

As a general rule, users should face prices that reflect the resource costs that they generate so that they can make informed decisions about resource utilization. In this section we explain our approach to how users should pay for each resource. We consider:

- *The incremental costs of sending extra packets.* If the network is not congested, this is essentially zero.
- *The social costs of delaying other users' packets when the network is congested.* This is not directly a resource cost, but should be considered part of the social cost of a packet. Users bear this cost through delay and dropped packets, and would often be willing to pay to reduce congestion.
- *The fixed costs of providing the network infrastructure.* This is the rent for the line, the cost of the routers, and the salary for the support staff.
- *The incremental costs of connecting to the network.* Each new connection to the Internet involves costs for access lines and switching equipment.
- *The cost to expand network capacity.* This normally consists of adding new routers, new lines, and new staff.

We first consider how *ideal* prices would reflect these costs; then we consider how market-based prices might work.

The incremental costs of sending extra packets.

The price of sending a packet in an uncongested network should be close to zero; a higher price is socially inefficient since it does not reflect the true incremental costs. If the incremental cost is high enough to justify the cost of monitoring and billing, it should be charged as a per-packet cost. Much of the needed monitoring and billing system would be needed to implement our other pricing proposals.

The social costs of delaying other users' packets when the network is congested.

The price for sending a packet when the network is congested should be positive: if my packet precludes or delays another user's packet, then I should pay the cost I impose on the other user. If my packet is more valuable than hers, then it should be sent; if hers is more valuable than mine, then hers should be sent.

We can depict the logic of this argument graphically using demand and supply curves. Suppose the packet price were very high: only a few users would want to send packets. As the packet price decreases, more users would be willing to send packets.[33] We show this relationship between price and the demand for network access in Figure 6. If the network capacity is fixed at K, then the optimal price for admitting the packets is where the demand curve crosses the capacity supply. If demand is small relative to capacity, the efficient price is zero—all packets are admitted. If demand is high, users that are willing to pay at least the price of admission to the network are admitted; others are not.

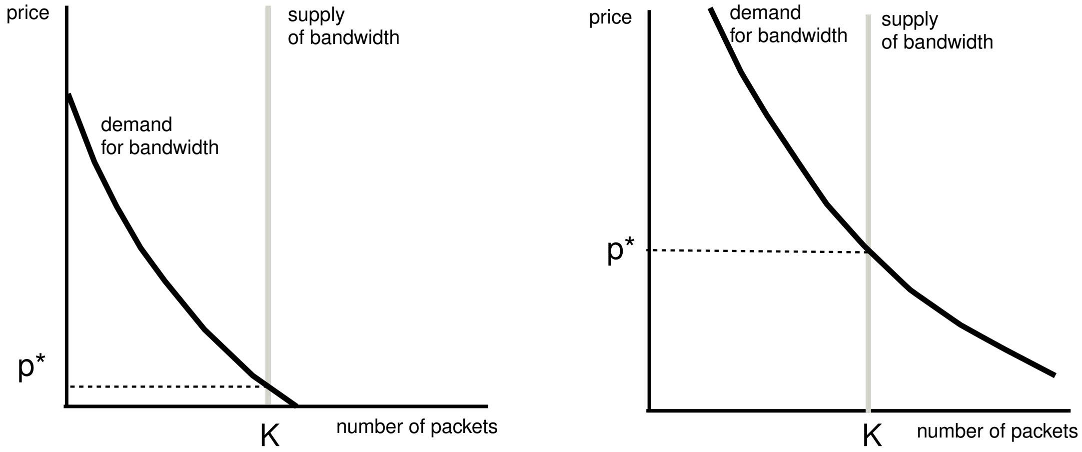


Figure 6. Demand for network access with fixed capacity. When demand is low, the packet price is low. When demand is high, the packet price is high.

This analysis applies to the extreme case where capacity is fixed. If an increase in packets from some users imposes delay on other users, but not outright exclusion, the analysis is slightly different. Suppose that we know how delay varies with the number of packets, and that we have some idea of the costs imposed on users by a given amount of delay. Then we can calculate a relationship between number of packets sent and delay costs. The relevant magnitude for determining the optimal number of users is the *marginal* cost of delay, the cost added by the next single packet (see Figure 7).

The efficient price is where the user's willingness to pay for an additional packet equals the marginal increase in delay costs generated by that packet. If a potential user faces this price, she

[33] One complication in implementing packet pricing is dealing with the difference between packets *sent* and packets *received*. The former will be greater than or equal to the latter due to dropped packets, which becomes important especially during periods of congestion.

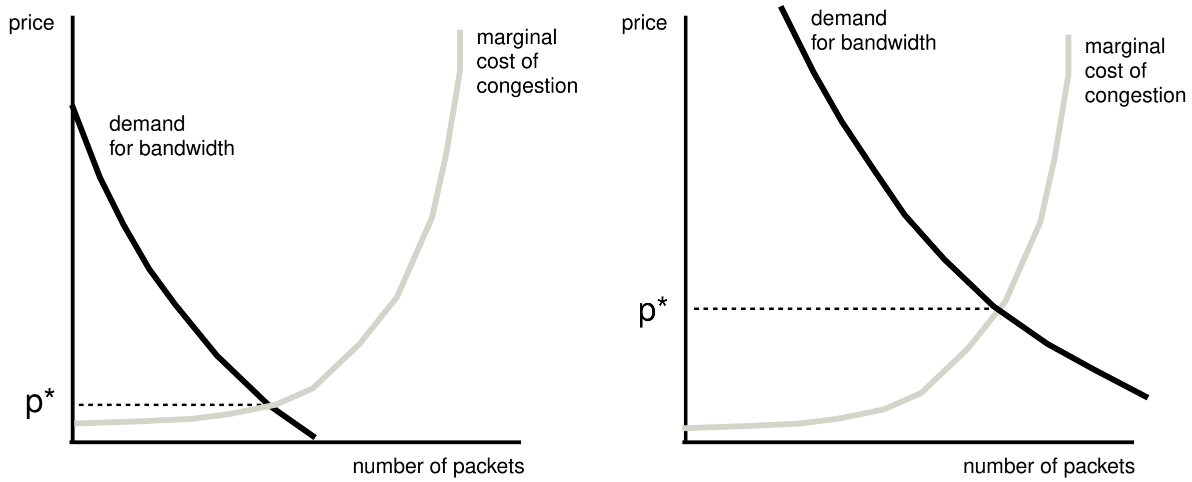


Figure 7. Demand for network access with a marginal cost of delay. When demand is low, the packet price is low. When demand is high, and congestion is high, the packet price is high.

can be able to compare her own benefit from sending a packet to the marginal delay costs she imposes on other users.

The fixed costs of providing the network infrastructure.

The initial investment in network infrastructure is a discrete decision: a certain amount of money can buy a usable network of minimal size. What criterion can be used to decide whether the initial investment is warranted? The natural principle to apply is that total benefits should exceed costs. The existence of an uncongested network is a public good that provides benefits for all users without exclusion; that is, Bob's use doesn't preclude Peter's use. Therefore we should add up how much all potential users would be willing to pay for the network infrastructure, and see if this total "willingness-to-pay" exceeds the cost of provision.

In the case of a computer network like the Internet, it is natural to think of paying for the network infrastructure with a flat access fee. Each party who connects to the network pays a flat price for network access distinct from the usage based fee described earlier. In general, these connection fees will vary, since different people and institutions value connection to the net differently. The infrastructure cost recovery will be efficient if each customer connects for a fee less than or equal to the amount she is her willing to pay, because then no customers would be needlessly excluded. If the total amount that users are willing to pay exceeds the infrastructure cost, the fees could be assigned in a variety of ways, depending on market conditions and the network providers' objectives. For example, a public sector network might want to charge a higher proportion of the willingness-to-pay of large or commercial users, and a lower fraction of the willingness-to-pay of poor or residential customers. This scheme would resemble the cross-subsidization for universal access historically regulated in the Bell System's telephone service.

The incremental costs of connecting to the network.

Each new user requires a connection to the network. In some cases, this connection may share an existing facility, for instance using a home phone to make a dial-up connection. Such a connection imposes no new costs and should be priced at zero. Other connections may require new cables, a router, and other investments. Each user should be charged the cost of installing a connection to the backbone as a single, one-time connection fee.

It may be that the public at large benefits from having more users connected, so that it would be efficient to provide a connection subsidy to ensure that some users who would not otherwise connect, do so. This does not mean that when there are network externalities all connections should be free, but that it would be efficient to have some subsidy per connection that is related to the *public* gain from an additional connection. Theoretically, an even more efficient scheme would target those users who are most likely to abstain *without* a subsidy, but targeted subsidies are difficult to implement.

The cost of expanding capacity of the network.

If network usage never reaches capacity, even at no cost for packets, then clearly there is no need to expand capacity. Usage prices that are based on congestion provide guidance about when to expand capacity. Consider the model with fixed capacity: Packet prices measure the marginal value of the last admitted packet. If the cost of expanding capacity to accommodate one more packet is less than the marginal value of that packet, then it makes economic sense to expand capacity. If expansion costs more, it is not economically worthwhile.

Hence optimal congestion pricing plays two roles—it efficiently rations access to the network in times of congestion, *and* it sends the correct signals about capacity expansion. In this framework, all the revenues generated by congestion prices should be used to expand capacity.

One advantage of this scheme is that only the users who want to use the network when it is at capacity pay for expansion. Users who are willing to wait do not pay anything toward expanding network capacity. We think that this point is important politically. The largest constituency on the Internet is apparently e-mail users;[34] a proposal to charge high prices for e-mail is likely to be politically infeasible. However, e-mail can usually tolerate moderate delays. Under congestion pricing of the sort we are describing, e-mail users could put a low or zero bid price on their traffic, and would continue to face a very low cost.

The situation is only slightly different in the case of delay costs. Here the price measures the marginal benefit of an additional packet, which is equal to the marginal cost of delay. If additional investment would reduce the marginal cost of delay by more than the amount users are willing to pay for reduced delay, then it should be undertaken, and otherwise not. (We examine the analytics of pricing a congested network in the Appendix.) If the packet price accurately reflects delay and congestion costs, it is the appropriate guidance to determine whether capacity should be expanded.

Pricing summary

An efficient pricing mechanism would have the following structure: (1) a packet charge close to zero when the network is not congested; (2) a positive packet charge when the network is congested; (3) a fixed connection charge that differs from institution to institution. Current pricing is almost always limited to a fixed connection charge. The main difference in what we propose is the addition of a usage-sensitive charge when the network is congested.

6. Implementing Congestion Prices

We now describe one method to implement efficient congestion prices. The connection charges are simplest: the current method needs no alterations. Each customer pays a flat fee for connection; this fee often depends on the bandwidth of the connection. Presumably the bandwidth of the connection purchased by an organization is correlated to some degree with the organization's willingness to pay, so this should serve as a reasonable characteristic upon which to base connection charges.[35]

[34] More traffic is generated by file transfers, but this reflects fewer users sending bigger data streams (files vs. e-mail messages).

[35] In future work we will investigate how a profit-maximizing or welfare-maximizing provider of network access might use price discrimination in connection fees.

No charges for sending packets when the network is not congested is also easy to arrange—that's what we have now. The novel part of the pricing mechanism we propose is the per-packet charge when the network is congested. We have discussed how one might implement such a fee in MacKie-Mason and Varian (1993), and we briefly review that proposal here.

If congestion has a regular pattern with respect to time of day, or day of week, then prices could vary in a predictable way over time. However, this is relatively inflexible. We think that it would be better to use a "smart market": the price to send a packet would vary minute-by-minute to reflect the current degree of network congestion.

A smart market would not be terribly difficult to implement, at least conceptually. Each packet would have a "bid" field in its header to indicate how much its sender is willing to pay to send it. Users would typically set default bids for various applications, and override the defaults in special circumstances. For example, a user might assign a low bid to e-mail packets. Real-time audio or visual data might be assigned a high bid price. The network would admit all packets with bid prices that exceed the current cutoff amount, determined by the marginal congestion costs imposed by the next additional packet.

This mechanism guarantees only relative priority, and is not an absolute promise of service. A packet with a high bid gains access sooner than one with a low bid, but delivery time cannot be guaranteed.[36] Rejected packets could be bounced back to the users, or be routed to a slower network, possibly after being stored for a period in a buffer in case the congestion falls sufficiently a short time later.

A novel feature of such a smart market is that users do *not* pay the price they actually bid; rather, they pay the market-clearing price, which is always lower than the bids of all admitted packets. This is different from priority-pricing by say, the post office, where you pay for first-class mail even if there is enough excess capacity that second-class mail is moving at the same speed.

The smart market has many other desirable features. Its outcome is the classic supply-equals-demand level of service of which economists are so fond. The equilibrium price, at any one point, is the bid of the marginal user. Each infra-marginal user is charged this price, so each infra-

[36] It is hard to see how absolute guarantees *can* be made on a connectionless network. However, there have been proposals to provide hybrid networks, with some connection-oriented services in parallel to the connectionless services. Connection-oriented services are well-suited for delivery guarantees.

marginal user gets a consumer surplus from the purchase. Further, as we show in the Appendix, the congestion revenues equal the optimal investment in capacity expansion.

The major differences from the textbook demand-and-supply story is that no iteration is needed to determine the market-clearing price—the market is cleared as soon as the users have submitted their bids for access.[37] Waldspurger, Hogg, Huberman, Kephart, and Stornetta (1992) describe some (generally positive) experiences in using this kind of "second-bid" auction to allocate network resources. However, they do not examine network access itself, as we are proposing here.[38]

We have assumed that the bid-price set by the users accurately reflects users' willingness to pay. Does our scheme provide the correct incentives for users to reveal this value—is there anything to be gained by trying to "fool" the smart market? Fortunately not; the dominant strategy in the second-bid auction is for users to bid their true values. By the nature of the auction, users are assured that they always get access when their value is higher than the current price, yet they will never be charged more than this amount, and normally less.

7. Other Concerns about the Smart Market Mechanism

Our smart market proposal is preliminary and tentative. It is only one theoretically appealing way to implement efficient congestion control. In this section we discuss a number of issues that must be studied and resolved before the smart market can be successfully implemented.

Who sets the bids?

We expect that bids would be set by three agents: the local administrator who controls access to the net, the user of the computer, and the computer software. Organizations with limited resources, for example, might choose low bid prices for all sorts of traffic. This would mean that they may not have access during peak times, but still would have access during off-peak periods.[39]

---

[37] Of course, in real time operation, one would presumably cumulate demand over some time interval. It is an interesting research issue to consider how often the market price should be adjusted. The bursty nature of Internet activity suggests a fairly short time interval. However, if users were charged for the congestion cost of their usage, it is possible that the bursts would be dampened.

[38] In technical terms, our mechanism can be viewed as an auction where the n highest bidders gain access at the $n + 1^{st}$ highest price bid, otherwise known as a Vickrey auction.

[39] With bursty traffic, low-priority packets at "peak time" might experience only moderate delays before getting through. This is likely to be quite different from the telephone analogue of making customers wait until after 11PM to obtain

An organization might not pass usage-sensitive prices through to its users. If not, then all traffic types would normally be sent with the same, organization-wide priority bid. If organizations do provide price or other incentives to their users, then users can assign different bids to various data flows. Normally, users would set default values in their software for different services. For example, file transfers might have lower priority than e-mail, e-mail would be lower than `telnet` terminal sessions, `telnet` would be lower than audio, and so on. The user could override default values in special cases, for example when a particular e-mail message is especially urgent, if he is willing to pay an increased price during congested periods.

Offline accounting

If the smart market system uses the sampling system suggested above, accounting overhead need not slow traffic much, since it can be done in parallel. All the router must do is compare the bid of a packet with the current value of the cutoff. The accounting information on every 1000^{th} packet, say, would be sent to a dedicated accounting machine that determines the equilibrium access price and records the usage for later billing.[40] Such sampling would require changes in current router technology and might well prove expensive. For example, NSFNET modified routers to collect sampled usage data; they found that the cost of the monitoring system was significant.

Fluctuations in the spot market price

Many colleagues are uncomfortable with the idea of fluctuating prices for bandwidth. Some feel that predictable prices, and hence budgets, are important to users. We have several responses. First, if prices and uses of the network turn out to be relatively predictable, expenditures would fluctuate very little. Enterprises have little difficulty now dealing with fluctuations in postage, electricity, and telephone bills from month to month; there is no reason to expect that network usage would be different.

---

low-priority, low-rate service. The average length of delays for low-priority traffic will depend on the average level of excess capacity in the system. One advantage of our scheme is that it correctly signals the efficient level of capacity to maintain.

[40] We don't discuss the mechanics of the billing system here. Obviously, there is a need for COD, third-party pricing, and other similar services.

Second, it is important to remember that in the smart market, prices only fluctuate *down.* The user sets the maximum he or she is willing to pay; the actual cost will never be higher. Furthermore, the user should have virtually instantaneous feedback about expenditures, so there should be little difficulty in budgetary control. As an extreme example, a user's employer might simply set a single bid price on all traffic: the unit price would only fluctuate down, and some cost would be borne in delay—when the market price was higher—rather than in expenditure jumps.

Finally, and most important, the price set by the smart market is a "wholesale" price, not necessarily a "retail" price. If a user does not wish to bear the risk of price fluctuations, he or she can always contract with another party who is willing to bear that risk; either the network service provider or a third party. The third party could offer different levels of service at different prices: "premium" service would send all packets immediately regardless of cost (to the third party); "economy" service would send packets only when the congestion price was below a certain level, and delay them when congestion was higher. The user would face an incentive to reduce traffic through a congested network because she would choose different priced fixed-budget plans.

For example, consider an extreme case where the network price has significant fluctuations: the price for an hour of teleconferencing at a particular time of day might be \$200 or \$50. A third party could offer to sell bandwidth to anyone who demands it at \$100 an hour. If the price turned out to be \$50, the bandwidth reseller would make a profit; if it turned out to be \$200, the bandwidth reseller would take a loss. The purchaser would pay \$100 no matter what.

If the price fluctuations are large, most retail customers might prefer to contract for bandwidth at a fixed price. But the existence of a spot market would be very important; it allows "wholesalers" to buy bandwidth on an "as available" basis, thereby encouraging efficient use of bandwidth.

In the end, cost fluctuations—either in the form of price or delay, or both—are unavoidable in a congestible network. In a well-functioning market users can choose to bear the mix of delay and price fluctuation directly, or they can pay a third party an "insurance premium" in exchange for a reduction in price or delay fluctuations.

Burstiness

Traffic on the network fluctuates quite significantly over periods as short as a few seconds; packet transfers are "bursty". Can the smart market keep up with such change?

We have two answers to this question. First, it is simple to buffer packets for short periods. During a burst of high-priority bids, packets with low-priority bids are buffered. After the high-priority packets are admitted, the low-priority packets move onto the network. In network engineering this is known as priority-based routing, and is reasonably well understood.

The second answer is a bit deeper. We conjecture that if usage were priced according to our scheme, network traffic would be much less bursty; the bursts exist because there is no charge for them. If bursts were costly to users there would be fewer. Users would have an incentive to use applications that smooth the network traffic flow. For example, in countries where electricity is priced by time of day, water heaters heat water in the middle of the night, when rates are low. If a refrigerator can be that smart, think what a workstation could do—if it faced the right prices.

Routing

As mentioned above, the Internet is a connectionless network. Each router knows the final destination of a packet, and uses its routing tables to determine the best way to get the packet from the current location to its next hop. These routing tables are updated continuously to indicate the current state of the network. They reflect failed links and new nodes, but not congestion on the links of the network. Indeed, there is no standard measurement for congestion available on current T3 networks.

Currently, all packets follow the same route at a given time; however, if each packet carried a bid price, this information could be used to facilitate routing. for example, packets with higher bids could take faster routes, while packets with lower bids could be routed through slower links. Obviously this description is very incomplete, but it seems likely that having packets bid for access will help to distribute traffic more efficiently.

Distributional aspects

Charging prices for usage during congested times may be politically acceptable, because it would largely preserve the cost structure for the many current users who can live with some delay and unreliability. In a smart market, low-priority access to the Internet (such as e-mail) would continue to cost very little. Indeed, with relatively minor public subsidies to cover the marginal *resource* costs, it would be possible to have efficient pricing with a price of close *zero* most of the time, since the network is usually not congested.

If there are several competing carriers, the usual logic of competitive bidding suggests that the price for low-priority packets should approach the marginal cost—which, as we have argued, is essentially zero. In the plan that we have outlined the high priority users would pay most of the costs of expanding the Internet.

Interruptible service

Implementing the smart market mechanism for pricing congestion on the Internet would require adding new information to the TCP/IP headers, which will take considerable discussion and debate. However, there is an interim way to handle congestion pricing that requires very little change in existing protocols. Suppose that providers of Internet services offer two classes of service: full service and interruptible service. Users would pay a flat fee based on bandwidth of their connection and the type of service they prefer. Full service would cost more than interruptible service.

When the load on the routers used by the Internet provider reached a certain level, users who purchase interruptible service would be denied access until the congestion subsided. All that is needed to implement this rationing mechanism is a simple change to the routing algorithms.

The defect of interruptible service is that it is inflexible compared to the smart market solution: it applies to all participants in a single administrative billing unit and cannot be overridden by individual users. On the other hand it would be very simple to implement. See Wilson (1989) for a detailed study of the analytics of interruptible service.

8. The Roles of the Public and Private Sectors

The technical problems associated with a usage-pricing scheme, including our proposed smart market, are enormous. The current Internet has developed through a collaboration between the private sector and governments; we think the development of the future broadband Internet with mechanisms for accounting and usage-sensitive pricing will also require government involvement.

The NSF is moving the Internet backbone away from the "interstate" model toward the "turnpike" model, as evidenced by the emergence of private-sector backbone competitors. The "Interstate" approach is for the government to develop the "electronic superhighways of the future" as part of an investment in infrastructure. The "turnpike" approach relies on the private sector to develop the network infrastructure for Internet-like operations, with the government providing subsidies to offset the cost of access to the private networks.

We believe an intermediate solution is necessary. The private sector is probably more flexible and responsive than a government bureaucracy; however, competing network standards could lead to an electronic Tower of Babel. A publicly imposed standard is important—turnpikes have the same traffic regulations as Interstates. For example, customer demand for low-delay, uncongested networks will give providers an incentive to implement some form of congestion control pricing, but individual network providers will probably not choose to implement such methods unless there is coordination in standards and widespread adoption of the mechanism. We think that there is an important role for public and quasi-public bodies in designing coordinated policies and protocols for congestion control, accounting and usage-sensitive pricing. As Estrin (1989) explains: "The Internet community developed its original protocol suite with only minimal provision for resource control . . . This time it would be inexcusable to ignore resource control requirements and not to pay careful attention to their specification."

One role for government is to insure interconnectivity between competing network providers. It may also be important for governments to provide the regulatory framework for in-line accounting and billing.[41] Whether protocols for actually implementing accounting and billing should be defined by a public body or an industry consortium is not immediately obvious.[42]

The history of standards for voice networks offers an interesting lesson. U.S. voice communications are now provided by a mesh of overlapping and connected networks operated by competing providers (ATT, MCI and Sprint being the largest). This is similar to the situation we expect to emerge for data networks. However, during the decades when switching and billing standards were being designed and refined, the only significant provider was ATT, so it could develop a single, coordinated standard that later providers adopted. International voice networks, in contrast, require interconnection and traffic handoff between various (mostly national) providers. These standards were designed and imposed by a public body, the CCITT.

A pricing standard must contain enough information to encourage efficient use of network bandwidth, and contain information for accounting and billing. A privatized network is simply not

[41] A recent Congressional bills submitted by Representative Boucher to begin implementing the NREN requires uniform protocols for interconnection between providers. It is not clear whether Congress will also mandate uniform standards for providing management information like accounting data.

[42] The current standards body for the Internet is the Internet Engineering Task Force (IETF), which is a voluntary, loosely-knit organization run by network specialists from industry, academia and other interested groups.

viable without such standards: work should start immediately on developing them.

The other important task for government is to estimate the public benefit from access and usage by users who might not be willing to pay their own costs, and then to design subsidies to encourage those users. We think the growth and development of the Internet will be best served if network services are priced according to cost (including congestion costs), and subsidies should be distributed so that users can pay those charges. Implementing subsidies instead by continuing to charge zero prices would give the biggest subsidies to the wrong users, would not provide useful signals to guide the use of costly resources, and would not guide investments in network expansion and upgrading. Using an efficient pricing scheme instead will encourage growth in network use and capacity, and guide resources to the highest-value uses.

Appendix: Some analytics of pricing a congestible resource

The classic "problem of the commons" describes a situation where property that is held in common will tend to be overexploited. Each user is aware of his private costs incurred by accessing the common property but neglects the costs he imposes on others. In the context of the Internet we have seen that the scarce resource is the switching capacity of the routers. When the network is highly congested, an additional user imposes costs on other users to the extent that his use of switching capacity prevents, or at least slows down, the use of the same capacity by other users.

Efficient use of the switch capacity requires that users that are willing to pay more for access should be admitted before users with lower willingness-to-pay. The price for admission to the switches should be that price that reflects the social cost of an additional packet.

Here we briefly examine some of the analytics of a standard (static) congestion model.[43] Arnott, de Palma, and Lindsey (1990) have argued strongly that congestion models should examine dynamic microbehavior in a more detailed way than the standard model does. Although we agree with this point, and think that modeling congestion behavior for computer networks is a promising avenue for future research, we here consider only the simplest textbook case of congestion.

We suppose that a representative user has a utility function $u(x_i) - D$, where x_i is the number of packets sent by user i and D is the total delay experienced by the user. The delay depends on the total utilization of the network, $Y = X/K$ where $X = \sum_{i=1}^{n} x_i$ is the total usage and K is network capacity.[44] This specification implies that if usage X is doubled and capacity K is doubled, then network utilization $Y = X/K$ and delay $D(Y)$ remain the same.

If there is no congestion-based pricing, user i will choose x_i to satisfy the first-order condition[45]

$$u'(x_i) = 0.$$

[43] The treatment is intended for economists; it is probably too terse for non-economists.

[44] We could also make the utility of packets depend on the delay by writing utility as $u(x_i, D)$. We choose the additively separable specification only for simplicity.

[45] We assume that the user ignores the fact that his own packets impose delay on his own packets; we can think of this effect as being built into the utility function already. There is no problem in relaxing this assumption; the calculations just become messier.

The *efficient* utilization of the network maximizes the sum of all users' utilities, $\sum_{i=1}^{n} u(x_i) - nD(X/K)$. This yields the n first-order conditions

$$u'(x_i) - \frac{n}{K} D'(Y) = 0.$$

One way to achieve this efficient outcome is to set a congestion price per packet of

$$p = \frac{n}{K} D'(Y), \tag{1}$$

so that user i faces the maximization problem

$$\max_{x_i} \; u(x_i) - D(Y) - px_i.$$

The first-order condition to this problem is

$$u'(x_i) = p = \frac{n}{K} D'(Y) \tag{2}$$

which is easily seen to lead to the optimal choice of x_i. The price has been chosen to measure the congestion costs that i's packets impose on the other users.

Optimal capacity expansion

Suppose now that it costs $c(K)$ for capacity K and that we currently have some historically given capacity. Should the capacity be expanded? The welfare problem is

$$W(K) = \max_{K} \sum_{i=1}^{n} u(x_i) - nD(Y) - c(K).$$

Since x_i is already chosen so as to maximize this expression, the envelope theorem implies that

$$W'(K) = nD'(Y) \frac{X}{K^2} - c'(K).$$

Substituting from equation (1)

$$W'(K) = p \frac{X}{K} - c'(K). \tag{3}$$

Suppose that the marginal cost of capacity expansion is a constant, $c_K = c'(K)$. Then we see that $W'(K)$ is positive if and only if $pX - c_K K > 0$. That is, *capacity should expanded when the revenues from congestion fees exceed the cost of providing the capacity.*

A competitive market for network services

Suppose that there are several competing firms providing network access. A typical producer has a network with capacity K and carries X packets, each of which pays a packet charge of p. The producer's operating profits are $pX - c(K)$.

Let $p(D)$ be the price charged by a provider that offers delay D. In general, if the delay on one network is different than on another the price will have to reflect this quality difference. The utility maximization problem for consumer i is to choose which network to use and how much to use it:

$$\max_{x_i, D} \; u(x_i) - D - p(D)x_i$$

which has first-order conditions

$$\begin{aligned} u'(x_i) - p(D) &= 0 \\ -1 - p'(D)x_i &= 0. \end{aligned}$$

The first equation says that each user will send packets until the value of an additional packet equals its price. The second equation says that the user will choose a network with a level of delay such that the marginal value to the user of additional delay equals the marginal cost of paying for the delay (by switching suppliers). Adding up this last first-order condition over the consumers yields

$$n = -p'(D)X. \tag{4}$$

A competitive producer offering delay $D(Y)$ wants to choose capacity and price so as to maximize profits, recognizing that if it changes its delay the price that it can charge for access will change. The profit maximization problem is

$$\max_{X,K} \; p(D(Y))X - c(K),$$

which gives us first-order conditions

$$\begin{aligned} p'(D)D'(Y)Y + p(D) &= 0 \\ -p'(D)D'(Y)Y^2 - c'(K) &= 0. \end{aligned} \tag{5}$$

Combining these two conditions and using equation (4) gives us two useful expressions for $p(D)$:

$$\begin{aligned} p(D) &= \frac{n}{K}D'(Y) \\ &= c'(K)\frac{K}{X} \end{aligned}.$$

Comparing the first equation to (2) we see that the competitive price will result in the optimal degree of congestion. Comparing the second equation to equation (3) we see that competitive behavior will also result in optimal capacity.

Adding capacity

Suppose now that a competitive firm is trying to decide whether to add additional capacity ΔK. We consider two scenarios. In the first scenario, the firm contemplates keeping X fixed and simply charging more for the reduction in delay. The amount extra it can charge for each packet is

$$\frac{dp}{dK}\Delta K = -p'(D)D'(Y)\frac{X}{K^2}\Delta K.$$

Using equation (5) this becomes

$$\frac{p}{K}\Delta K.$$

Since the firm can charge this amount for each packet sent, the total additional revenue from this capacity expansion is

$$p\frac{X}{K}\Delta K.$$

This revenue will cover the costs of expansion if

$$p\frac{X}{K}\Delta K - c'(K)\Delta K = \left[p\frac{X}{K} - c'(K)\right]\Delta K \geq 0,$$

which is precisely the condition for social optimality as given in equation (3).

Consider now the second scenario. The firm expands its capacity and keeps its price fixed. In a competitive market it will attract new customers due to the reduction in delay. In equilibrium this firm must have the same delay as other firms charging the same price. Suppose that in the initial equilibrium $X/K = Y$. Then the additional number of packets sent must satisfy $\Delta X = Y\Delta K$. It follows that the increase in in profit for this firm is given by

$$pY\Delta K - c'(K)\Delta K = \left[p\frac{X}{K} - c'(K)\right]\Delta K.$$

Again we see that capacity expansion is optimal if and only if it increases profits.

The relationship between capacity expansion and congestion pricing was first recognized by Mohring and Hartwize (1962) and Strotz (1978). Some recent general results can be found in Arnott and Kraus (1992b, 1992a).

References

Almquist, P. (1992). Type of service in the internet protocol suite. Tech. rep. RFC 1349, Network Working Group.

Arnott, R., de Palma, A., and Lindsey, R. (1990). Economics of a bottleneck. *Journal of Urban Economics*, *27*, 111–130.

Arnott, R., and Kraus, M. (1992a). Financing capacity on the bottleneck model. Tech. rep., Department of Economics, Boston College.

Arnott, R., and Kraus, M. (1992b). Self-financing of congestible facilities in a dynamic environment. Tech. rep., Economics Department, Boston College.

Bohn, R., Braun, H.-W., Claffy, K., and Wolff, S. (1993). Mitigating the coming Internet crunch: Multiple service levels via precedence. Tech. rep., UCSD, San Diego Supercomputer Center, and NSF.

Braun, H.-W., and Claffy, K. (1993). Network analysis in support of internet policy requirements. Tech. rep., San Diego Supercomputer Center.

Claffy, K., Braun, H.-W., and Polyzos, G. (1993). Application of sampling methodologies to wide-area network traffic characterization. Tech. rep. Technical Report CS93-275, UCSD.

Claffy, K. C., Polyzos, G. C., and Braun, H.-W. (1992). Traffic characteristics of the T1 NSFNET backbone. Tech. rep. CS92-252, UCSD. Available via Merit gopher in Introducing the Internet directory.

Cocchi, R., Estrin, D., Shenker, S., and Zhang, L. (1992). Pricing in computer networks: Motivation, formulation, and example. Tech. rep., University of Southern California.

Estrin, D. (1989). Policy requirements for inter administrative domain routing. Tech. rep. RFC1125, USC Computer Science Department.

Faulhaber, G. R. (1992). Pricing Internet: The efficient subsidy. In Kahin, B. (Ed.), *Building Information Infrastructure*. McGraw-Hill Primis.

Gerla, M., and Kleinrock, L. (1988). Congestion control in interconnected LANs. *IEEE Network*, *2*(1), 72–76.

Group, I. A. W. (1992). Usage reporting architecture. Tech. rep., Internet Engineering Task Force. Draft.

Kahin, B. (1992). Overview: Understanding the NREN. In Kahin, B. (Ed.), *Building Information Infrastructure*. McGraw-Hill Primis, NY.

Kleinrock, L. (1992). Technology issues in the design of NREN. In Kahin, B. (Ed.), *Building Information Infrastructure*. McGraw-Hill Primis.

Krol, E. (1992). *The Whole Internet*. O'Reilly & Associates, Inc., Sebastopol, CA.

MacKie-Mason, J. K., and Varian, H. (1993). Some economics of the internet. Tech. rep., University of Michigan.

Mohring, H., and Hartwize, M. (1962). *Highway Benefits: An Analytical Approach*. Northwestern University Press, Evanston.

Prue, W., and Postel, J. (1988). A queuing algorithm to provide type-of-service for IP links. Tech. rep. RFC 1046, USC Information Sciences Institute.

Roberts, L. G. (1974). Data by the packet. *IEEE Spectrum*, *XX*, 46–51.

Ruth, G., and Mills, C. (1992). Usage-based cost recovery in internetworks. *Business Communications Review*, *xx*, 38–42.

Shenker, S. (1993). Service models and pricing policies for an integrated services internet. Tech. rep., Palo Alto Research Center, Xerox Corporation.

Smarr, L. L., and Catlett, C. E. (1992). Life after Internet: Making room for new applications. In Kahin, B. (Ed.), *Building Information Infrastructure*. McGraw-Hill Primis.

Strotz, R. (1978). Urban transportation parables. In Margolis, J. (Ed.), *The Public Economy of Urban Communities*. Resources for the Future, Washington, D.C.

Waldspurger, C. A., Hogg, T., Huberman, B. A., Kephart, J. O., and Stornetta, W. S. (1992). Spawn: A distributed computational economy. *IEEE Transactions on Software Engineering*, *18*(2), 103–117.

Wilson, R. (1989). Efficient and competitive rationing. *Econometrica*, *57*(1), 1–40.