Pricing the Internet

by

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

April 1993
Current version: November 26, 1993

Abstract. This paper was prepared for the conference ‘‘Public Access to the Internet,’’ JFK School of Government, May 26--27 , 1993. We describe some of the technology and costs relevant to pricing access to and usage of the Internet, and discuss the components of an efficient pricing structure. We suggest a possible smart-market mechanism for pricing traffic on the Internet.

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 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.

We think that it is safe to say that no one is quite sure about how this privatization effort will work. In particular, it is far from clear how access to 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] In any case, none of the backbones charge for actual usage in the sense of the volume of data transmitted.

In this paper we describe some of the technological, cost, and economic issues related to pricing access to and usage of the Internet. We strongly suspect that efficiency will require a combination of usage pricing for backbone services ane flat-fee pricing for connections Usage pricing in current networks will require new standards for TCP/IP packets in order to facilitate accounting and priority-based routing. We offer a proposal as to how access might be priced using a smart market.

1. Internet Technology and Costs

The Internet is a network of networks. In this paper we focus on network backbones, although most of our pricing ideas apply equally well to mid-level and local area networks. There are essentially

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. 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.

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 virtual 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 both commercial and non-commercial users. Sprint, of course, is a major telecommunications provider as well as a provider of Internet transport services.

The Internet is defined as those connected networks that use connectionless packet-switching communications technology based on the TCP/IP protocols. Even though much of the traffic moves across lines leased from telephone common carriers, the technology is quite different from the circuit-switching used for voice telephony. A telephone user dials a number and 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 and no other caller can use those resources until the call is terminated.[3] A packet-switching network, by contrast, uses statistical multiplexing to maximize use of the communications lines.[4] Each circuit is simultaneously shared by numerous users, and no single open connection is maintained for a particular communications session: some of the data may go by one route while the rest may take a different route. Because of the technology differences, pricing models appropriate for voice telephony will be inappropriate for data networks.

Packet-switching technology has two major components: packetization and dynamic routing. A data stream from a computer is broken up into small chunks called ‘‘packets.’’ The IP (Internet protocol) specifies how to break up a datastream into packets and reassemble it, and also 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.

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.

[2] 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.

[4] ‘‘Connection-oriented’’ packet-switching networks also exist: X.25 and Frame Relay are examples of such.

The network is needed only after a key is struck or when a reply is returned. 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.

With dynamic routing a packet's path across the network is determined anew for each packet transmitted. Because multiple paths exist between most pairs of network nodes, it is quite possible that different packets will take different paths through the network.[5]

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

So that packets can be identified and reassembled in the correct order, TCP packets consist of a header followed by data. The header contains the source and destination ports, the sequence number of the packet, an acknowledgment flag, and so on. The header comprises 20 (or more) bytes of the packet.

Once a packet is built TCP sends it to a router, a computer that is in charge of sending packets on to their next destination. At this point IP tacks on another header (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 towards its destination, and sends it on. The best link may change minute-by-minute, as the network configuration changes.[6] Routes can be recalculated immediately from the routing table if a route fails. The routing table in a switch is updated approximately continuously.

---

[5] 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.

[6] 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 1 million hosts and over 11,000 subnetworks.

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

Over the past 5 years, the speed of the NSFNET backbone has grown from 56 Kbps to 45 Mbps (''T-3'' service).[7] These lines can move data at a speed of 1,400 pages of text per second; a 20-volume encyclopedia can be sent across the net in half a minute. Many of the 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. Each packet traversed 16 links, and thus the IP header had to be read and modified 16 times, and 16 different routers had to calculate the best next link for the transmission. Despite the many hops and substantial packetization and routing overhead, 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).[8]

Current Backbone Network Costs

The postal service is a good metaphor for packet-switching technology, but a bad metaphor for the *cost structure* of Internet services. Most of the costs of providing the Internet are more-or-less independent of the 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.

The NSF currently spends about $11.5 million per year to operate the NSFNET and provides $7 million per year of grants to help operate the regional networks.[9] There is also an NSF grant program to help colleges and universities to connect to the NSFNET. Using the conservative

[7] In fact, although the communications lines can transport 45 Mbps, the current network routers can support only 22.5 Mbps service. ''Kbps'' is thousand (kilo) bits per second; ''Mbps'' is million (mega) bits per second.

[8] 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.

[9] 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.

estimate of 1 million hosts and 10 million users, this implies that the NSF subsidy of the Internet is less than $20 per year per host, and less than $2 per year per user.[10]

Total salaries and wages for NSFNET have increased by a little more than one-half (about 68% nominal) over 1988--1991, during a time when the number of packets delivered has increased 128 times.[11] 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 dominate 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.[12] In our discussion we focus only on the costs of lines and routers.

We have estimated costs for the network backbone as of 1992--93.[13] A T-3 (45 Mbps) trunk line running 300 miles between two metropolitan central stations can be leased for about $32,000 per month. The cost to purchase a router capable of managing a T-3 line is approximately $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 communications and switching have been dropping rapidly for over three decades. In the 1960s, digital computer switching was more expensive (on a per packet basis) than communications (Roberts (1974)), but switching has become substantially cheaper since then. We have estimated the 1992 costs for transporting 1 million bits of data through the NSFNET backbone and compare these to estimates for earlier years in Table 1. As can be seen, in 1992 the line cost is about eight times as large as the cost of routers.

---

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

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

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

[13] 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.

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

Year	Communications	Routers	Design Throughput
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 topology of the NSFNET backbone directly reflects the cost structure: lots of cheap routers are used to manage a limited number of expensive lines. We illustrate a portion of the network in Figure 1. Each of the numbered squares is an RS6000 router; the numbers listed beside a router are links to regional networks. Notice that in general any packet coming on to the backbone has to move through two separate routers at the entry and exit node. For example, a message we send from the University of Michigan to a scientist at Bell Laboratories will traverse link 131 to Cleveland, where it passes through two routers (41 and 40). The packet goes to New York, where it again moves through two routers (32 and 33) before leaving the backbone on link 137 to the JVNCnet regional network that Bell Labs is attached to. Two T-3 communications links are navigated using four routers.

Technological and Cost Trends

The decline in both communications link and switching costs has been exponential at about 30% per year (see the semi-log plot in Figure 2). But more interesting than the rapid decline in costs 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

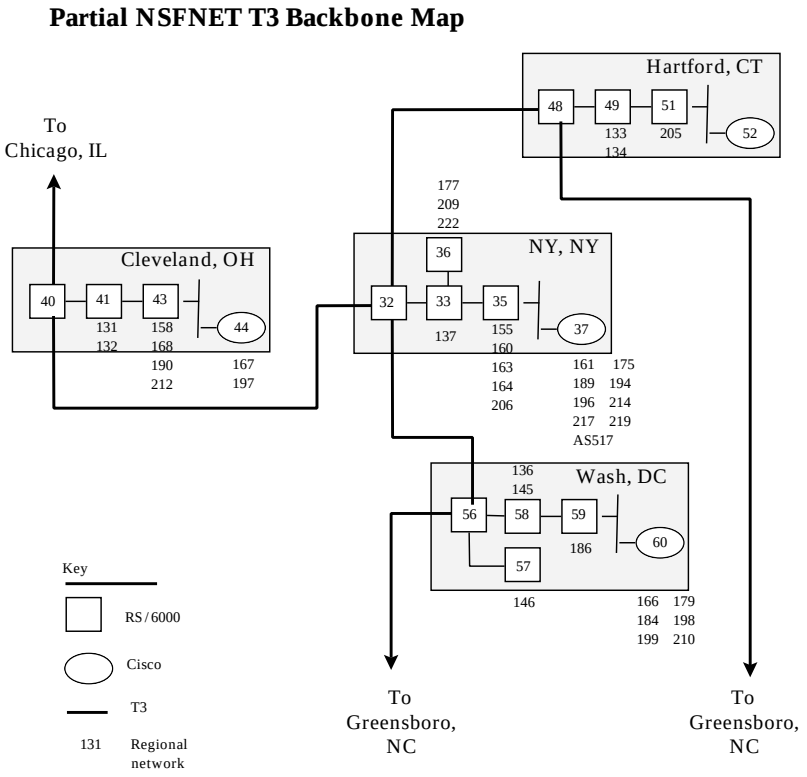

Figure 1. Network Topology Fragment

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 became cheaper than lines the network is more efficient if data streams are broken into small packets and sent out piecemeal, allowing the packets of 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.

Although the same reversal in switch and line costs occurred for voice networks, circuit-switching is still the norm for voice. Voice is not well-suited for packetization using current protocols because of variation in delivery delays, packet loss, and packet ordering.[14] Voice cus-

[14] Our tests found packet delays ranging between 156 msec and 425 msec on a trans-Atlantic route (N=2487 traces, standard deviation = 24.6 msec). Delays were far more variable to a Nova Scotia site: the standard deviation was 340.5 msec when the mean delay was only 226.2 msec (N=2467); the maximum delay was 4878 msec.

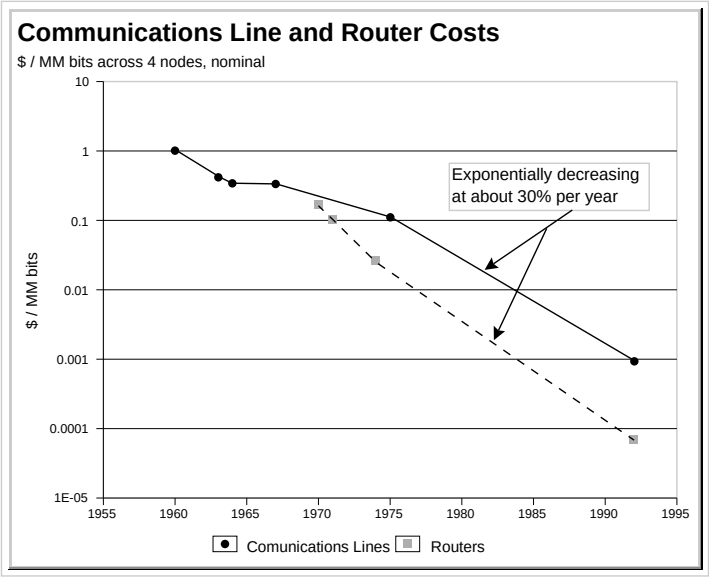

Figure 2. Trends in costs for communications links and routers.

tomers will not tolerate these delays in transmission (although some packetized voice applications are beginning to emerge as transmission speed and reliability increases, see Anonymous (1986).[15] [16]

2. Congestion problems

Another aspect of cost of the Internet is congestion cost. Although congestion costs are not paid for by the *providers* of network services, they are paid for by the *users* of the service. Time spent by users waiting for a file transfer is a social cost, and should be recognized as such in any economic accounting.

The Internet experienced severe congestion problems in 1987. Even now congestion problems are relatively common in parts of the Internet (although not currently on the T-3 backbone). According to Kahin (1992): ‘‘However, problems arise when prolonged or simultaneous high-end

[15] The reversal in link and switch costs *has* had a profound effect on voice networks. Indeed, Peter Huber has argued that this reversal made inevitable the breakup of ATT (Huber (1987)). He describes the transformation of the network from one with long lines all going into a few central offices into a web of many switches with short lines interconnecting them so that each call could follow the best path to its destination.

[16] The economies and flexibility of packet switching with the need for circuit-switched quality guarantees has stimulated the development of new connection-oriented packet-switching technologies, of which asynchronous transfer mode (ATM) is the apparent leader. Telecommunications researchers are working towards using ATM or a related technology to carry voice as well as other data in a packet-type network (using fixed-length ‘‘cells’’ rather than variable length packets, however).

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). It is apparent that contemplated uses, such as real-time video and audio transmission, would lead to substantial increases in the demand for bandwidth and that congestion problems will only get worse in the future unless there is substantial increase in bandwidth:

> 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.

There are cases where 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.

To investigate the nature of congestion on the Internet we timed the delay in delivering packets to seven different sites around the world. We ran our test hourly for 37 days during February and March, 1993. Deliveries can be delayed for a number of reasons other than congestion-induced bottlenecks. 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.

Our results are summarized in Figure 3 and Figure 4; we present the results only from four of the 24 hourly probes. Figure 3 shows the median and maximum delivery delays by time of day. Average delays are not always proportional to distance: the delay from Michigan to New York University was generally longer than to Berkeley, and delays from Michigan to Nova Scotia, Canada, were often longer than to Oslo, Norway.

There is substantial variability in Internet delays. For example, the maximum and median delays in Figure 3 are quite different by time of day. There appears to be a large 4PM peak problem on 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,

[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.

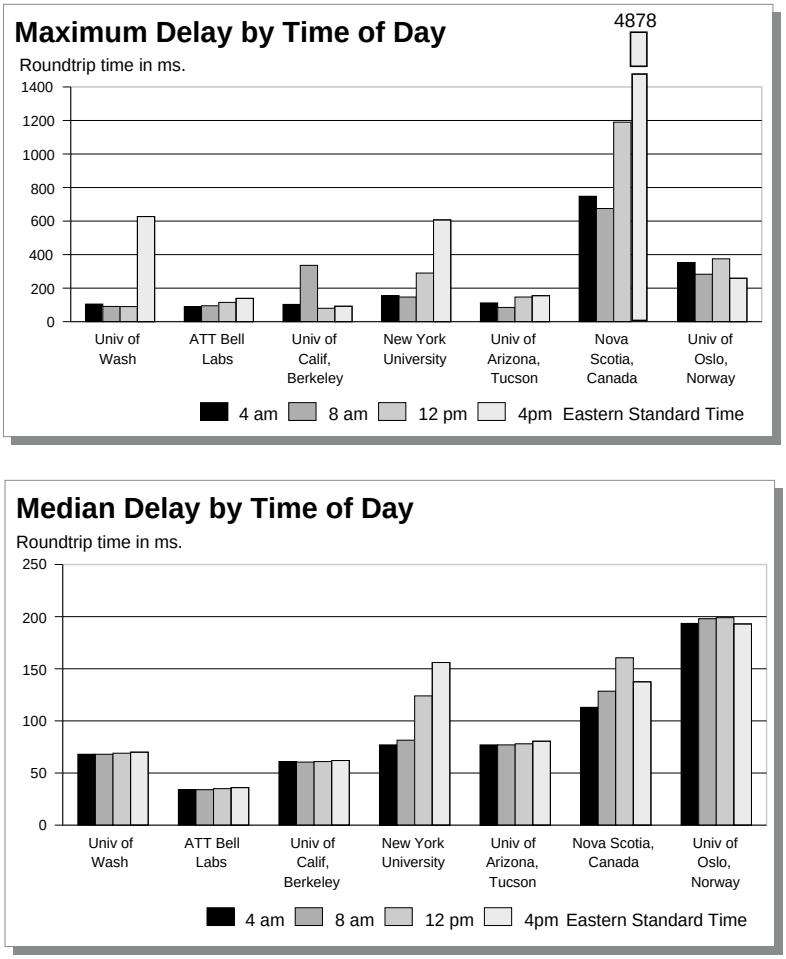

Figure 3. Maximum and Median Transmission Delays on the Internet

Polyzos, and Braun (1992).[18]

Figure 4 shows the standard deviation of delays by time of day for each destination. The delays to Nova Scotia, Canada were extraordinarily variable, yet the delays to Oslo had no more variability than does 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 obvious pattern.

[18] Note that the Claffy et al. data were for the old, congested T-1 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 T-3 backbone. Braun and Claffy (1993) show time-of-day variations in T-3 traffic between the US and three other countries in their Figure 5.

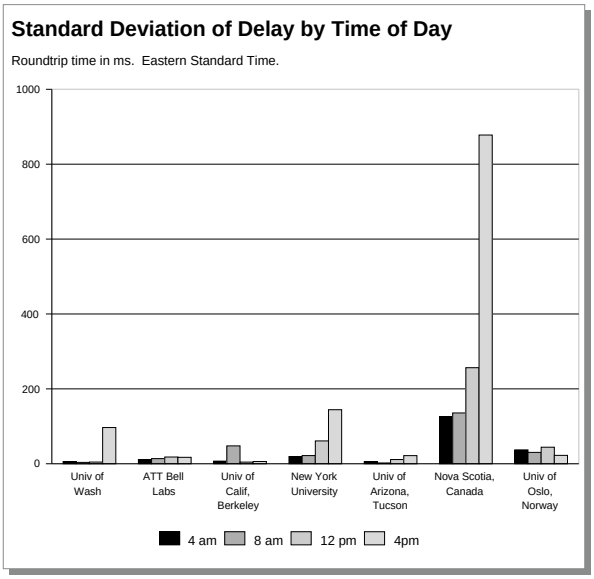

Figure 4. Variability in Internet Transmission Delays

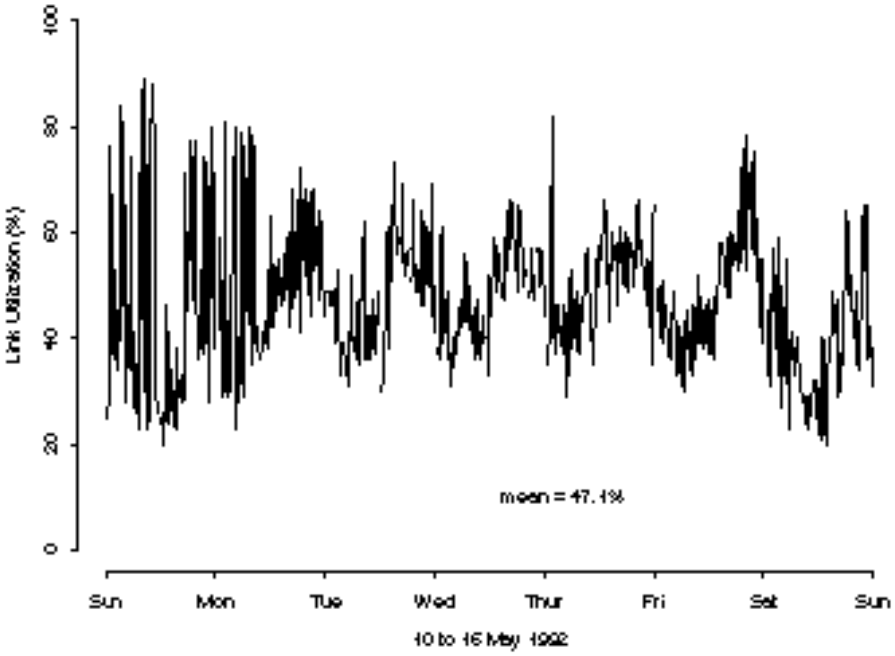

Figure 5. Utilization of Most Heavily Used Link in Each Fifteen Minute Interval (Claffy et al. (1992))

Controlling Congestion

According to Kleinrock (1992), ‘‘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.’’ NSFNET usage has been growing at about 6% per month, or

doubling every twelve months.[19] Although costs of providing additional capacity are declining rapidly, we think it is very likely that congestion will continu problem, especially as new very-high bandwidth uses (such as real-time broadcast video) become common. Essentially, this is the classic problem of the commons: without a mechanism for congestion control, there will inevitably be inefficient use of the common resource. As long as users have access to flat-rate usage, they will continue to ''overgraze.'' It is becoming increasingly important to consider how congestion in networks such as the Internet should be controlled.

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

Indeed, IP packets contain fields known Precedence and Type of Service (TOS). Currently, most commercial routers do not use these fields.[20] However, it is widely anticipated that this must change due to increased congestion on the Internet: ''An obvious application would be to allow router and host configuration to limit traffic entering the internet to be above some specific precedence. Such a mechanism could be used to reduce traffic on an internet as often as needed under crisis conditions'' (Cerf (1993)).

The current interpretations of these fields described in Postel (1981) will probably be changed to the more flexible form described in Almquist (1992). Almquist discusses only the TOS fields, and proposes that the user be able to request that the network should minimize delay, maximize throughput, maximize reliability, or minimize monetary cost when delivering the packet. Prototype algorithms to provide such service are described in Prue and Postel (1988). In this proposed protocol a router looks up the destination address and examines the possible routes. Each route has a TOS number. If the TOS number of the route matches the TOS number of the datagram, then that route is chosen. Note that the TOS numbers must *match;* inequality relationships are not allowed.

[19] 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 alternative backbone routes has probably been growing faster.

[20] 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.

To an economist's eye, this specification seems somewhat inflexible. In particular, the TOS value ''minimize monetary cost'' seems somewhat strange. Of course senders would want to minimize monetary cost for a given quality of service: minimizing monetary cost is an objective, not a constraint. Also, the fact that TOS numbers do not allow for inequality relations is strange. Normally, one would think of specifying the amount that one would be willing to pay for delivery, with the implicit assumption that any 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. In particular, there is very little work on using pricing for congestion control. But without pricing it is hard to imagine how priority schemes can be implemented. What is to stop an e-mail user from setting the highest priority on her packets if it costs her nothing to do so? 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?[21] 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 it will be necessary to have some sort of pricing to sort out users' demands for these characteristics. Their papers lay out the problem in general and describe how it might be solved.

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.''

[21] 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's suggestion makes sense for a dedicated line---e.g., a line connecting an institution to the Internet backbone. But we don't think that it is necessarily appropriate for charging for backbone traffic itself. The reason is that 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 individual capacity level on the backbone.[22]

Although it is appropriate to charge a flat fee to cover the one-time costs of a network connection, we also think that it is important to charge for network usage when the network is congested. After all, during times of congestion the scarce resource is bandwidth for additional packets.[23] Conversely when the network is not congested the marginal cost of transporting additional packets is essentially zero. It would therefore be appropriate to charge users a very low or zero price for packets when the system is not congested.

One problem with usage-sensitive pricing is the accounting and billing overhead. The cost would be astromical 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 of data.[24] However, a number of factors could reduce the accounting load by orders of magnitude. First, given the huge number of packets traversing backbones (currently over one billion per day on the NSFNET), basing charges on a statistical *sample* of packets sent may be acceptable. Second, if usage is priced only during congested periods, most packets sent need no accounting. Third, traditional phone company accounting systems, which seem like the natural comparison, are centralized and off-line. We think that dramatic breakthroughs are likely in the area of in-line, distributed accounting, which will substantially lower the costs.[25]

---

[22] It may be true that an organization's use of the backbone bandwidth is more-or-less proportional to the bandwidth of its connection to the backbone. For such an organization, the size of the dedicated line to the backbone may be a good signal of its intended usage of the common backbone.

[23] As we have already pointed out the major bottleneck in backbone capacity is not the bandwidth of the medium itself, but the switch technology. We use the term bandwidth to refer to the overall capacity of the network.

[24] A vigorous, one-minute phone call on a digital network today utilizes about $60 \times 64\text{K}/8$ 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 and banal 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 on designing mechanisms for usage accounting on the Internet. As a first attempt, ANS developed a usage sampling and reporting system it called COMBits. COMBits collects aggregate measures of packets and bytes, using a statistical sampling technique.[26] However, COMBits only collects data down to the network-to-network level of source and destination. Thus, the resulting data can only be used to charge at the level of the subnetwork; the local network administrator is responsible for splitting up the bill (Ruth and Mills (1992)).[27] More recently, the Internet Accounting Working Group has published a draft architecture for Internet usage reporting (Internet Accounting: Usage Reporting Architecture, July 9, 1992 draft). 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 need to 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. We thus turn to a more thorough discussion of pricing.

3. General observations on pricing

The Internet uses scarce resources. 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. Economics is concerned with ways to allocate scarce resources among competing uses, and it is our belief that economics will be useful in allocating Internet resources as well.

We are not concerned with pricing the Internet to generate profits from selling backbone services. Indeed, a network need not be private to be priced; governments are perfectly capable of charging prices.[28] Rather, our goal is to find pricing mechanisms that lead to the most efficient use of existing resources, and that guide investment decisions in an appropriate manner.

---

[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.

One common resource allocation mechanism is randomization: each packet has an equal chance of getting through (or being dropped). Another allocation scheme is first-come, first-served: all packets are queued as they arrive and if the network is congested, every packet suffers a delay based on its arrival time in the queue. It is easy to see why these schemes are not good ways to achieve efficiency.[29] However one measures the social value of expeditious delivery for a packet, it will surely be true that some packets are worth more than 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 will be enhanced if the mechanism allocating scarce bandwidth gives higher priority to uses that are more socially valuable.

We do not want the service provider---government or otherwise---to decide which packets are more socially valuable and allocate scarce bandwidth accordingly. We know from the Soviet experience 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 knows the costs of providing services and can announce these to the users; users then can decide for themselves whether their packets are more or less valuable than the cost of providing 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. The users themselves decide how valuable each packet is, and sort out for themselves which packets are serviced (or in a multiple service quality network, receive which quality of service; see Shenker (1993)).

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. We return to this point below.

One commonly expressed concern about pricing the Internet is that ‘‘poor’’ users will be deprived of access. This is not a problem with pricing, but with the distribution of wealth. A pricing mechanism determines how the scarce bandwidth will be allocated given the preferences and resources of the users. If we wish to ensure that certain users have sufficient resources to

[29] Current backbones use a mix of queuing and random dropping as their mechanisms for allocating congested capacity.

purchase a base level of services then we can redistribute initial resources, say by providing vouchers or lump sum grants.[30]

Highways are often suggested as an analogy for the future of Internet. Many people argue that publicly provided interstate highways free of tolls work well and should be the model. But the analogy is flawed. First, many democratic governments do not 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 importantly, an interstate offers only a single, undifferentiated class of service. Users who need different service characteristics pay for access to rail lines, canals, or airports. No ones has seriously argued that usage of *all* transportation networks should be free. The interstate highway system might be viewed as the one-size-fits-all universal access option (universal for those who can afford cars!), with usage charges imposed for those who want a different combination of service characteristics. Likewise, a government might want to provide universal access to a baseline set of free Internet transport services, yet allow all other services be provided with usage charges. Appropriate universal access to free services might include e-mail on a best-efforts basis (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---if desired---can be provided through vouchers or other redistribution schemes. But for a 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 are already paying the costs of congestion through delays and lost packets. A pricing mechanism will convert delay and queuing costs into dollar costs. If prices are designed to 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 on the basis of 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] 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).

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.

4. Current Pricing Mechanisms

NSFNET, the primary backbone network of the Internet, has been paid for by the NSF, IBM, MCI and the State of Michigan until the present.[31] However, most organizations do not connect directly to the NSFNET. A typical university will connect to its regional mid-level network; the mid-level maintains a connection to the NSFNET. The mid-level networks (and a few alternative backbone networks) charge their customers for access, but not actual usage.

There are dozens of companies that offer connections to the Internet. Most large organizations obtain direct connections, which use a leased line that permits unlimited usage subject to the bandwidth of the line. Some customers purchase ''dial-up'' service which provides an intermittent connection, usually at much lower speeds. We will discuss only direct connections below.

Table 3 summarizes the prices offered to large universities by ten of the major providers for T-1 access (1.5 mbps).[32] There are three major components: an annual access fee, an initial connection fee and in some cases a separate charge for the customer premises equipment (a router to serve as a gateway between the customer network and the Internet provider's network).[33] The current annual total cost per T-1 connection is about $30--35,000.

[31] NSF restricts the use of the backbone to traffic with a research or educational purpose, as defined in the Acceptable Use Policies.

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

[33] Customers will generally also have to 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 will generally be 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: annual fee for unlimited access, based on the bandwidth of the connection. This is the type of pricing recommended by Faulhaber (1992). However, 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 only that the outgoing packets be counted at a single point: the customer's gateway router.

However, pricing every packet would not necessarily increase the efficiency of network service provision, because the marginal cost of a packet is nearly zero. As we have shown, the important scarce resource is bandwidth, and thus efficient prices need to reflect the current state of the network. Neither a flat price per packet nor even time-of-day prices would come very close to efficient pricing.

5. Matching prices to costs

In general we want the prices that users face to reflect the resource costs that they generate so that they can make intelligent decisions about resource utilization. In the case of the Internet, there are several costs that might be considered:

- *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 certainly 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.* As we have seen this is basically 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 user (or network of users) connection to Internet involves costs for access lines and switching equipment.
- *The cost of expanding capacity of the network.* This will normally consist of adding new routers, new lines, and new staff.

We first consider how *ideal* prices would incorporate this cost information, then consider how market-based prices might work.

The incremental costs of sending extra packets.

The price of sending a packet in a non-congested network should be close to zero; any 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.[34]

[34] Note that much of the necessary monitoring and billing cost may already be incurred 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 in a congested state should be positive: if my packet precludes (or delays) another user's packet, then I should face the cost that 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; then only a few users would want to send packets. As the packet price decreases, more users would be willing to send more packets.[35] We depict this relationship between price and the demand for network access in Figure 6. If the network capacity is some fixed amount 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 users are admitted. If demand is high, users that are willing to pay more than the price of admission to the network are admitted; the others are not.

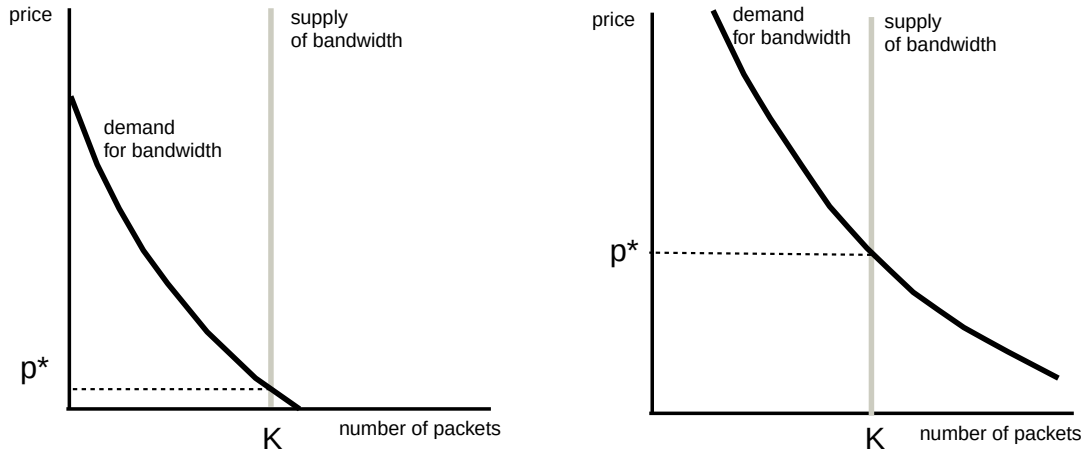

Figure 6. Demand for network access with fixed capacity. When demand is low, the packet price is low. When demand is high, the packet price is high.

This analysis applies for the extreme case where there is a fixed capacity. If increase in use by some agents imposes delay on other agents, but not outright exclusion, the analysis is slightly

[35] 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.

different. Suppose that we know the amount of delay as a function of 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, as depicted in Figure 7.

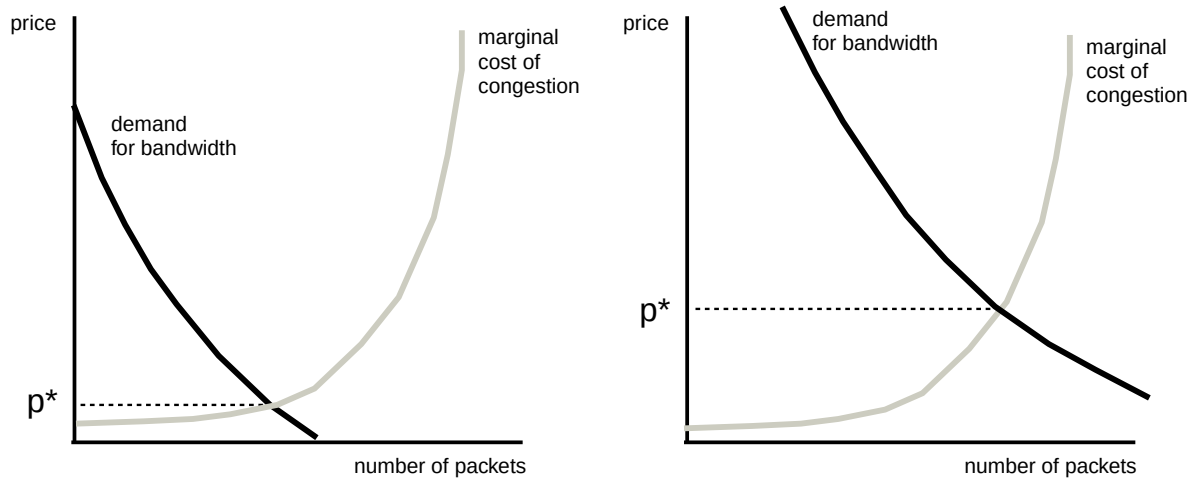

Figure 7. Demand for network access with a marginal cost of delay. When demand is low, the packet price is low. When demand is high, and congestion is high, the packet price is high.

The efficient price is where the marginal willingness to pay for an additional packet just covers the marginal increase in delay costs generated by that packet. If a potential user faces this price he will be able to compare his own benefit from sending a packet to the marginal delay costs that this imposes on other users.

The fixed costs of providing the network infrastructure.

Think of the initial investment in network infrastructure as a discrete decision: if you pay a certain amount of money you can create a usable network of minimal size. What criterion can be used to decide whether the initial investment is warranted? The simple answer is that the investment should be undertaken if total benefits exceed costs. The existence of an uncongested network is a public good that provides benefits for all users without exclusion. Therefore we should add up all potential users' willingnesses-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 via 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 be different for different people, since different people and institutions will value connection to the net differently. The infrastructure cost recovery will be efficient if each customer is allowed to connect for a fee less than or equal to her willingness-to-pay, because then no customers would be needlessly excluded. If the sum of total willingness-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 their willingness-to-pay to large or commercial users, and a lower fraction of their willingness-to-pay to poor or residential customers. This scheme would resemble the universal access cross-subsidization historically regulated in the Bell System for 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 (e.g., 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 cabling, a router, and other investments. Each user should be charged the cost of making their 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. Then it would be efficient to provide a connection subsidy to ensure that some users who would not otherwise connect end up doing 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. An even more efficient scheme would be to 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 the network usage never reaches capacity, even at a zero price of packets, then clearly there is no need for expanding capacity. It is only appropriate to expand capacity when the network is sometimes congested. Consider first the model with fixed capacity. If the packet prices are set correctly, we have seen that they measure the marginal value of the last admitted packet. If the cost

of expanding capacity enough to accommodate one more packet is less than the marginal value of that packet, then it makes economic sense to expand capacity. If this condition is not satisfied, then capacity expansion is not economically worthwhile.

Hence the optimal congestion prices play a two roles---they serve to efficiently ration access to the network in times of congestion *and* they send the correct signals with respect to capacity expansion. In this framework, all the revenues generated by congestion prices should be plowed back into capacity expansion.

Note that only the users who want to use the network when it is at capacity pay for capacity expansion. Users who are willing to wait until after the demand peak do not pay anything towards expanding network capacity. We think that this point is important from a political perspective. The largest constituency of the Internet apparently is e-mail users.[36] 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 can reduce the marginal cost of delay by more than the willingness-to-pay for reduced delay then it should be undertaken, otherwise it should not. We examine the analytics of pricing a congested network in the Appendix. It turns out that essentially the same result holds: if the packet price is chosen to be optimal with respect to delay and congestion costs it will be the appropriate price to use for determining whether capacity should be expanded.

Pricing Summary.

There are five types of costs associated with providing a broad-based computer network: (1) the marginal costs of sending packets when the network is not congested; (2) the congestion costs of sending packets when the network is congested; (3) the fixed costs of providing initial infrastructure; (4) the costs of connecting new users; and, (5) the costs of expanding backbone capacity. An efficient pricing mechanism will have a structure that is parallel to this cost structure:

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

(1) a packet charge close to zero when the network is not congested; (2) a positive packet charge when the network is congested; (3&4) a fixed connection charge that differs from institution to institution. The congestion charge revenues can then be used to guide capacity expansion decisions (5).

6. Implementing prices

We have argued that prices should reflect costs. But we have not yet considered how these efficient prices should be implemented. We turn now to that task.

The connection charges are the easiest to deal with, since that essentially the current method of charging for provision. Each customer pays a flat fee for connection; often this fee will depend on the characteristics of the customer (educational, commercial) and on the size of the bandwidth of the connection. Presumably the bandwidth of the connection purchased by a user is correlated to some degree with the user's willingness to pay, so this should serve as a reasonable characteristic upon which to base connection charges.[37]

A zero cost of packet charges when the network is not congested is not hard to arrange either---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). 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 a relatively inflexible form for pricing. We think that it would be better to use a ''smart market'': a price for packet access to the net that varies minute-by-minute to reflect the current state of the network congestion.

This would not be terribly difficult to implement, at least conceptually. Each packet would have a ''bid'' field in the header that would indicate the willingness-to-pay for that packet. Users would typically set default bids for various applications, then override these defaults in special circumstances. For example, a user might assign a low bid to e-mail packets, for which immediate access to the net is usually not required. Real-time audio or visual data might be assigned a high bid price. The network would then admit all packets whose bid exceeded some cutoff amount. The

[37] We intend to investigate how a profit-maximizing or welfare-maximizing provider of network access might price discriminate in connection fees in future work.

cutoff amount is determined by the condition that the marginal willingness-to-pay for an additional packet has to equal the marginal congestion costs imposed by that packet.

A novel feature of this kind of smart market is that users do *not* pay the price that they actually bid; rather they pay for their packets at the market-clearing price, which by construction will be lower than the bids of all admitted packets. Note how this is different from priority-pricing by say, the post office. In the post-office model you pay for first-class mail even if there is enough excess capacity that second-class mail could move at the same speed. In the smart market described here, a user pays *at most* their willingness-to-pay for an additional packet.

The smart market has many desirable features. By construction the outcome is the classic supply-equals-demand level of service of which economists are so fond.[38] The equilibrium price, at any point in time, is the bid of the marginal user. Each infra-marginal user is charged this price, so each infra-marginal user gets positive consumer surplus from his or her purchase.

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.[39] This mechanism can be viewed as a Vickrey auction where the n highest bidders gain access at the n + 1^{st} highest price bid.[40]

We have assumed that the bid-price set by the users accurately reflects the true willingness-to-pay. One might well ask whether users have the correct incentives to reveal this value: is there anything to be gained by trying to ‘‘fool’’ the smart market? It turns out that the answer is ‘‘no.’’ It can be shown that it is a dominant strategy in the Vickrey auction to bid your true value, so users have no incentive to misprepresent their bids for network access. By the nature of the auction, you are assured that you will never be charged more than this amount and normally you will be charged much less.

---

[38] For good reason, we might add.

[39] 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.

[40] Waldspurger, Hogg, Huberman, Kephart, and Stornetta (1992) describes 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.

7. Remarks about the smart market solution

Who sets the bids?

We expect that choice of bids would be done by three parties: the local administrator who controls access to the net, the user of the computer, and the computer software itself. An organization with limited resources, for example, might choose low bid prices for all sorts of access. This would mean that they may not have access during peak times, but still would have access during off-peak periods.[41]

Within any limits imposed by institution policies, the users could then set priority values for their own usage. 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 these default values in special cases, for example when a particular e-mail message is especially urgent, if he was willing to pay for any increase in congestion during peak periods.

Note that this access control mechanism only guarantees relative priority, not absolute priority. A packet with a high bid is guaranteed access sooner than a low bid, but no absolute guarantees of delivery time can be made.[42] 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 permitted priority level falls sufficiently a short time later.

Offline accounting

If the smart market system is used with the sampling system suggested earlier the accounting overhead doesn't have to slow things down much since it can be done in parallel. All the router has to do is to compare the bid of a packet with the current value of the cutoff. The accounting

[41] 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 10PM to obtain 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.

[42] 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.

information on every 1000^{th} packet, say, is sent to a dedicated accounting machine that determines the equilibrium access price and records the usage for later billing.[43] However, such sampling would require changes in current router technology. Such accounting may well prove expensive. NSFNET has modified routers to collect sampled usage data; they found that the cost of the monitoring system is significant.

Fluctuations in the spot market price

Many readers have been unhappy with the idea that the price of bandwidth would fluctuate in the smart market system. It is felt by some that having predictable prices and budgets is important to users. We have several responses to this set of issues. First, everything depends on how much expenditures fluctuate. 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, and there is no reason to expect that network usage would be different.

Second, it is important to remember that in the smart market, prices only fluctuate *down.* The user (or the user's application) sets the maximum he or she is willing to pay for network access; the actual price paid will almost always be less than this. Furthermore, the user should have virtually instantaneous feedback about the current state of his or her expenditures, so there should be little difficulty in budgetary control.

Finally, the most important point that we need to make is that the price set by the smart market is a ''wholesale'' price, not necessarily a ''retail'' price. If a particular user doesn't want to bear the risk of price fluctuations, he or she can always contract with another party who is willing to bear that risk. This party may be the supplier of the network service, or it may be a third party.

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 could be $200 or could be $50. A third party could offer to sell bandwidth to anyone demanding it at, say, $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 make a loss. But the purchaser would pay a flat $100 no matter what.

[43] 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.

If the price fluctuations are large, it may well happen that most retail customers buy bandwidth on a contract basis at a fixed price. But the fact that the spot market is available is very important since it allows ‘‘wholesale’’ customers to buy bandwidth on an ‘‘as available’’ basis, thereby encouraging efficient use of bandwidth.

It is also worthwhile to note that the fluctuations in price represent a real resource cost---congestion costs. If the user doesn't bear that cost, then someone else will have to: the other users who find their packets delayed or dropped. Of course, there is no reason why the risk of price fluctuations couldn't be borne by third-parties. One could imagine a futures market for bandwidth in which third-parties offer to absorb the risk of price fluctuations for a fee.

Short term price fluctuations

Another problem arises at the other end of the time scale. It is widely observed that packet transfers are ‘‘bursty.’’ Traffic on the network fluctuations quite significantly over short time periods. Can a market price keep up with this kind of fluctuation?

We have two answers to this question. First, it is very easy to buffer packets for short time intervals. When a high-priority/high-bid burst comes along, packets with low priority and low bid, 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 a reasonably well-understood policy.

The second answer is a bit deeper. We conjecture that if usage were priced in the way we advocate, network traffic would be a lot less bursty. Said another way: bursts in network traffic are there now because there is no charge for bursts. If bursts were costly to the user there would be fewer of them.

Of course, this is not only because the user would change behavior---the bursts are at a much higher frequency than the users control. Rather, the users would have an incentive to use applications that smoothed the network traffic flow. In countries where electricity is priced by time of day, water heaters are smart enough to 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 knew the right prices.

Routing

As we have mentioned several times, the Internet is a connectionless network. Each router knows the final destination of a packet, and determines, from its routing tables, what the best way is to get from the current location to the next location along the way. These routing tables are updated continuously to indicate the current state of the network. Routing tables change to reflect failed links and new nodes, but they do not change to reflect congestion on various links of the network. Indeed, there is no standard measurement for congestion available on the current NSFNET T-3 network.

Currently, there is no prioritization of packets: all packets follow the same route at a given time. However, if each packet carried a bid price, as we have suggested, this information could be used to facilitate routing through the Internet. For example, packets with higher bids could take faster routes, while packets with lower bids could be routed through slower links.

The routers could assign access prices to each link in the net, so that only packets that were ‘‘willing to pay’’ for access to that link would be given access. Obviously this description is very incomplete, but it seems likely that having packets bid for access will help to distribute packets through the network in a more efficient way.

Distributional aspects

As we mentioned earlier, the issue of pricing the Internet is highly politicized. One nice feature of smart market pricing is that low-priority access to the Internet (such as e-mail) would continue to have a very low cost. 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 to *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 marginal cost---which, as we have argued, is essentially zero. In the plan that we have outlined the high priority users would end up paying most of the costs of expanding the Internet.

Interruptible service

Implementing the smart market mechanism for pricing congestion on the Internet would involve adding new information to the TCP/IP headers. It will take a considerable amount of discussion and debate to accomplish this. However, there is a partial way to handle congestion pricing that requires very little change in existing protocols.

Suppose that providers of Internet services had two classes of service: full service and interruptible service. Users would pay a flat fee based on the size of their pipeline for the type of service they preferred and full service would cost more than interruptible service.

When the load on the routers used by the Internet provider reached a certain level, the users who had purchased 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 rather 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 *is* very simple to implement. See Wilson (1989) for a detailed study of the analytics of interruptible service.

8. Role of public and private sector

Current private providers of access to the Internet generally charge for the ‘‘size of the pipe’’ connecting users to the network. We have argued that charges for usage, at least during congested periods, will also be necessary for efficient use and growth of the Internet. However, the technical problems associated with any 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 current NSF is moving the Internet backbone away from the ‘‘Interstate’’ model, and towards the ‘‘turnpike’’ model; indeed, with the emergence of private-sector backbone competitors this has already commenced. 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 is that the private sector should develop the network infrastructure for

Internet-like operations, with the government providing subsidies to offset the cost of access to the private networks.

We think that an intermediate solution is necessary. The private sector is probably more flexible and responsive than a government bureaucracy. However, the danger is that competing network standards would lead to an electronic Tower of Babel. It is important to remember that turnpikes have the same traffic regulations as the Interstates. We think that there is an important role for public and quasi-public bodies to play in designing coordinated policies and protocols for congestion control, accounting and usage-sensitive pricing. As Estrin (1989) puts it: ''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 governments is to insure interconnectivity between competing network providers.[44] It may also be important for governments to provide the necessary regulatory framework for the emergence of ''digital cash'' and digital credit to promote in-line accounting and billing. Whether the necessary protocols for actually implementing accounting and billing should be defined by a public body or an industry consortium is not immediately obvious.[45]

It is worth remarking on the history of standards for voice networks. U.S. voice communications are now provided by a mesh of overlapping and connected networks operated by multiple, competing providers (ATT, MCI and Sprint being the largest). This is quite a bit like the situation we expect to emerge for data networks. However, over the decades when switching and billing standards were being designed and refined, the only significant provider was ATT, so it could impose a single, coordinated standard that later providers accepted. International voice networks, by contrast, have always required interconnection and traffic handoff between various (mostly national) providers. Standards were designed and imposed by a public body, the CCITT.

A pricing standard has to be carefully designed to contain enough information to encourage efficient use of network bandwidth, as well as containing the necessary hooks for accounting and

[44] 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 the bill will also mandate uniform standards for providing management information like accounting data.

[45] 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.

rebilling information. A privatized network is simply not viable without such standards, and work should start immediately on developing them.

The other important task for government, as we alluded to above, is to estimate how much the public benefits 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 by a bifurcation of prices and the resources to pay those prices. Network services should be priced according to cost (including congestion costs), then subsidies should be distributed so that users can pay those charges. Implementing subsidies through zero prices instead would give the biggest subsidies to the wrong users. Zero prices also would not provide the useful signals to users that would guide the use of costly resources, and guide investments in network expansion and upgrading.

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.[46] 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.[47] 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[48]

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

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

[47] 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.

[48] 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

$$u'(x_i) - p(D) = 0$$

$$-1 - p'(D)x_i = 0.$$

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 that 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 K. 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

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

Using equation (5) this becomes

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

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

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

This revenue will cover the costs of expansion if

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

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 $X = YK$. It follows that the increase in in profit for this firm is given by

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

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.

Anonymous (1986). Stratacom, inc. introduces 'packetized voice system'. *Communications Week, 2*.

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.

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

Cerf, V. G. (1993). Core protocols. In *Internet System Handbook.* Addison Wesley, Reading, MA.

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.

Huber, P. W. (1987). *The Geodesic Network: 1987 Report on Competition in the Telephone Industry.* U.S. Gov't Printing Office, Washington, DC.

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.

Postel, J. (1981). Darpa internet program protocol specification. Tech. rep. RFC 791, Information Sciences Institute, University of Southern California.

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.