Center for Research on Economic and Social Theory

and

Department of Economics

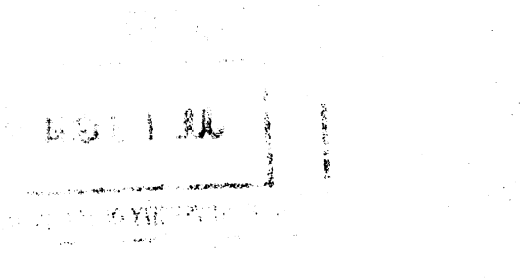
Working Paper Series

Economic FAQs About the Internet

Jeffrey K. MacKie-Mason and Hal R. Varian

May 1994
Number 94-05

DEPARTMENT OF ECONOMICS
University of Michigan
Ann Arbor, MI 48109-1220



Economic FAQs About the Internet

by

Jeffrey K. MacKie-Mason
Hal Varian
University of Michigan and NBER
University of Michigan

Current version: May 20, 1994

Abstract. This is a set of Frequently Asked Questions (and answers) about the economic, institutional, and technological structure of the Internet. We describe the current state of the Internet, discuss some of the pressing economic and regulatory problems, and speculate about future developments.

Keywords. Internet, computer networks

JEL Classification Numbers. L63, L86, O38

Address. Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220. E-mail: `Hal.Varian@umich.edu` and `jmm@umich.edu`. The most current version of this paper will be available for anonymous `ftp`, `gopher`, or `World Wide Web` access at `gopher.econ.lsa.umich.edu`.

1. What is a FAQ?

FAQ stands for Frequently Asked Questions. There are dozens of FAQ documents on diverse topics available on the Internet, ranging from physics to scuba diving to how to contact the White House. They are produced and maintained by volunteers. This FAQ answers questions about the economics of the Internet (and towards the end offers some opinions and forecasts). The companion paper in this Symposium, Goffe (1994), describes Internet resources of interest to economists, including how to find other FAQs.

2. Background

What is the Internet?

The Internet is a world-wide network of computer networks that use a common communications protocol, TCP/IP (Transmission Control Protocol/Internet Protocol). TCP/IP provides a common language for interoperation between networks that use a variety of local protocols (Netware, AppleTalk, DECnet and others).

Where did it come from?

In the late sixties, the Advanced Research Projects Administration (ARPA), a division of the U.S. Defense Department, developed the ARPAnet to link together universities and high-tech defense contractors. The TCP/IP technology was developed to provide a standard protocol for ARPAnet communications. In the mid-eighties the NSF created the NSFNET in order to provide connectivity to its supercomputer centers, and to provide other general services. The NSFNET adopted the TCP/IP protocol and provided a high-speed backbone for the developing Internet.

How big is the Internet?

From 1985 to April 1994, the Internet has grown from about 200 networks to well over 30,000 and from 1,000 hosts (end-user computers) to over two million. About 640,000 of these hosts are at educational sites, 520,000 are commercial sites, and about 220,000 are government/military sites, while most of the other 700,000 hosts are elsewhere in the world. NSFNET traffic has grown from 85 million packets in January 1988 to 6569 billion packets in March 1994. (A packet is about 200 bytes, and a byte corresponds to one ASCII character.) This is more than a six hundred-fold increase in only six years. The traffic on the network is currently increasing at a rate of 6% a month.[1]

The authors wish to acknowledge support from National Science Foundation grant SBR-9230481.

[1] Current NSFNET statistics are available by anonymous `ftp` from `nic.merit.edu`. All statistics we report are current as of March 1994 unless otherwise indicated.

What do people do on the Internet?

Probably the most frequent use is e-mail. After that are file transfer (moving data from one computer to another) and remote login (logging into a computer that is running somewhere else on the Internet). In terms of traffic, about 37% of total traffic is file transfer, 16% is e-mail and netnews, and 7% is from the information retrieval programs `gopher` and `World Wide Web`. People can search databases (including the catalogs of the Library of Congress and scores of university research libraries), download data and software, and ask (or answer) questions in discussion groups on numerous topics (including economics research). See Goffe (1994) for a catalog of network resources of interest to economists.

3. Organization

Who runs the Internet?

The short answer is "no one." The Internet is a loose amalgamation of computer networks run by many different organizations in over seventy countries. Most of the technological decisions are made by small committees of volunteers who set standards for interoperability.

What is the structure of the Internet?

The Internet is usually described as a three-level hierarchy. At the bottom are local area networks (LANs); for example, campus networks. Usually the local networks are connected to a regional, or mid-level network. The mid-levels connect to one or more backbones. The U.S. backbones connect to other backbone networks around the world. There are, however, numerous exceptions to this structure.

What is a regional net?

Regional networks provide connectivity between end users and the NSFNET backbone. Most universities and large organizations are connected by leased line to a regional provider. There are currently about a dozen regional networks.

Some of the regional networks receive subsidies from the NSF; many receive subsidies from state governments. A large share of their funds are collected through connection fees charged to organizations that attach their local networks to the mid-levels. For example, a large university will typically pay $60,000–$100,000 per year to connect to a regional.

Who runs the regionals?

The regionals are generally run by a state agency, or by a coalition of universities in a given geographic region. They are operated as nonprofit organizations.

What are the backbone networks?

As of January 1994 there are four public fiber-optic backbones in the U.S.: NSFNET, Alternet, PSInet, and SprintLink. The NSFNET is funded by the NSF, and is the oldest, having evolved directly out of ARPANET, the original TCP/IP network. The other backbones are private, for-profit enterprises.

Why is there more than one backbone?

Due to its public funding, the NSFNET has operated under an Acceptable Use Policy that limits use to traffic in support of research and education. When the Internet began to rapidly grow in the late 1980s, there was an increasing demand for commercial use. Since Internet services are unregulated[2] entry by new providers is easy, and the market for backbone services is becoming quite competitive.

Nowadays the commercial backbones and the NSFNET backbone interconnect so that traffic can flow from one to the other. Given the fact that both research and commercial traffic is now flowing on the same fiber, the NSF's Acceptable Use Policy has become pretty much of a dead letter. The charges for these interconnections are currently relatively small lump-sum payments, but there has been considerable debate about whether usage-based "settlement charges" will have to be put in place in the future.

Who runs the NSFNET?

Currently the NSF pays Merit Network, Inc. (Michigan Educational Research Information Triad) to run the NSFNET. Merit in turn subcontracts the day-to-day operation of the network to Advanced Network Services (ANS), which is is a nonprofit firm founded in 1990 to provide network backbone services. The initial funding for ANS was provided by IBM and MCI.

How much does NSFNET cost?

It is difficult to say how much the Internet as a whole costs, since it consists of thousands of different networks, many of which are privately owned. However, it is possible to estimate how much the NSFNET backbone costs, since it is publicly supported. As of 1993, NSF pays Merit about $11.5 million per year to run the backbone. Approximately 80% of this is spent on lease payments for the fiber optic lines and routers (computer-based switches). About 7% of the budget is spent on the Network Operations Center, which monitors traffic flows and troubleshoots problems.

To give some sense of the scale of this subsidy, add to it the approximately $7 million per year that NSF pays to subsidize various regional networks, for a total of about $20 million. With current estimates that there are approximately 20 million Internet users (most of whom are connected to the NSFNET in one way or another) the NSF subsidy amounts to about $1 per user per year. Of course, this is significantly less than the total cost of the Internet; indeed, it does not even include all of the public funds, which come from state governments, state-supported universities, and other national governments as well. No one really knows how much all this adds up to, although there are some research projects underway to try to estimate the total U.S. expenditures on the Internet. It has been estimated—read "guessed"— that the NSF subsidy of $20 million per year is less than 10% of the total U.S. expenditure on the Internet.

[2] Transport of TCP/IP packets is considered to be a "value-added service" and as such is not regulated by the FCC or state public utility commissions.

What is the future for a federally-funded backbone?

The NSFNET backbone will likely be gone by the time this article is published, or soon thereafter. With the proliferation of commercial backbones and regional network interconnections, a general-purpose federally subsidized backbone is no longer needed. In the new NSF awards just announced, the NSF will only fund a set of Network Access Points (NAPs), which will be hubs to connect the many private backbones and regional networks. The NSF will also fund a service that will provide fair and efficient routing among the various backbones and regionals. Finally, the NSF will fund a very-high speed backbone network service (vBNS) connecting the six supercomputer sites, with restrictions on the users and traffic that it can carry. Its emphasis will be on developing capabilities for high-definition remote visualization and video transmission. The new U.S. network structure will be less hierarchical and more interconnected. The separation between the backbone and regional network layers of the current structure will become blurred, as more regionals are connected directly to each other through NAPs, and traffic passes through a chain of regionals without any backbone transport.

What are independent providers?

Most users access the Internet through their employer's organizational network, which is connected to a regional. However, in the past few years a number of for-profit independent providers of Internet access have emerged. These typically provide connections between small organizations or individuals and a regional, using either leased lines or dial-up access. Starting in 1993 some of the private computer networks (e.g., Delphi and World) have begun to offer full Internet access to their customers (Compuserve and the other private networks have offered e-mail exchange to the Internet for several years).

Who provides access outside of the U.S.?

There are now a large number of backbone and mid-level networks in other countries. For example, most western European countries have national networks that are attached to EBone, the European backbone. The infrastructure is still immature, and quite inefficient in some places. For example, the connections between other countries often are slow or of low quality, so it is common to see traffic between two countries that is routed through the NSFNET in the U.S. (Braun and Claffy (1993)).

4. Technology

Is the Internet different from telephone networks?

Yes and no. Most backbone and regional network traffic moves over leased phone lines, so at a low level the technology is the same. However, there is a fundamental distinction in how the lines are used by the Internet and the phone companies. The Internet provides connectionless packet-switched service whereas telephone service is circuit-switched. (We define these terms below.) The difference may sound arcane, but it has vastly important implications for pricing and the efficient use of network resources.

What is circuit-switching?

Phone networks use circuit switching: an end-to-end circuit must be set up before the call can begin. A fixed share of network resources is reserved for the call, and no other call can use those resources until the original connection is closed. This means that a long silence between two teenagers uses the same resources as an active negotiation between two fast-talking lawyers. One advantage of circuit-switching is that it enables performance guarantees such as guaranteed maximum delay, which is essential for real-time applications like voice conversations. It is also much easier to do detailed accounting for circuit-switched network usage.

How is packet-switching technology different from circuit-switching?

The Internet uses "packet-switching" technology. The term "packets" refers to the fact that the data stream from your computer is broken up into packets of about 200 bytes (on average), which are then sent out onto the network.[3] Each packet contains a "header" with information necessary for routing the packet from origination to destination. Thus each packet in a data stream is independent.

The main advantage of packet-switching is that it permits "statistical multiplexing" on the communications lines. That is, the packets from many different sources can share a line, allowing for very efficient use of the fixed capacity. With current technology, packets are generally accepted onto the network on a first-come, first-served basis. If the network becomes overloaded, packets are delayed or discard ("dropped").

How are packets routed to their destination?

The Internet technology is *connectionless*. 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*. 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 routes are dynamically updated, it is possible for different packets from a single session to take different routes to the destination.

Along the way packets may be broken up into smaller packets, or reassembled into bigger ones. When the packets reach their final destination, they are reassembled at the host computer. The instructions for doing this reassembly are part of the TCP/IP protocol.

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).

[3] Recall that a byte is equivalent to one ASCII character.

What is the physical technology of the Internet?

Most of the network hardware in the Internet consists of communications lines and switches or routers. In the regional and backbone networks, the lines are mostly leased telephone trunk lines, which are increasingly fiber optic. Routers are computers; indeed, the routers used on the NSFNET are modified commercial IBM RS6000 workstations, although custom-designed routers by other companies such as Cisco, Wellfleet, 3-Com and DEC probably have the majority share of the market.

What does "speed" mean?

"Faster" networks do not move electrons or photons at faster than the speed of light; a single bit travels at essentially the same speed in all networks. Rather, "faster" refers to sending more bits of information simultaneously in a single data stream (usually over a single communications line), thus delivering n bits faster. Phone modem users are familiar with recent speed increases from 300 bps (bits per second) to 2400, 9600 and now 19,200 bps. Leased-line network speeds have advanced from 56 Kbps (kilo, or 10^3 bps) to 1.5 Mbps (mega, or 10^6 bps, known as T-1 lines) in the late 80s, and then to 45 Mbps (T-3) in the early 90s. Lines of 155 Mbps are now available, though not yet widely used. The U.S. Congress has called for a 1 Gbps (giga, or 10^9 bps) backbone by 1995.

The current T-3 45 Mbps lines can move data at a speed of 1,400 pages of text per second; a 20-volume encyclopedia can be sent coast to coast on the NSFNET backbone in half a minute. However, it is important to remember that this is the speed on the superhighway—the access roads via the regional networks usually use the much slower T-1 connections.

Why do data networks use packet-switching?

Economics can explain most of the preference for packet-switching over circuit-switching in the Internet and other public networks. Circuit networks use lots of lines in order to economize on switching and routing. That is, once a call is set up, a line is dedicated to its use regardless of its rate of data flow, and no further routing calculations are needed. This network design makes sense when lines are cheap relative to switches.

The costs of both communications lines and computers have been declining exponentially for decades. However, since about 1970, switches (computers) have become relatively cheaper than lines. At that point packet switching became economic: lines are shared by multiple connections at the cost of many more routing calculations by the switches. This preference for using many relatively cheap routers to manage few expensive lives is evident in the topology of the backbone networks. For example, in the NSFNET any packet coming on to the backbone has to pass through two routers at its entry point and again at its exit point. A packet entering at Cleveland and exiting at New York traverses four NSFNET routers but only one leased T-3 communications line.

What changes are likely in network technology?

At present there are many overlapping information networks (e.g., telephone, telegraph, data, cable TV), and new networks are emerging rapidly (paging, personal communications services, etc.). Each of the current information networks is engineered to provide a particular type of service and

the added value provided by each of the different types was sufficient to overcome the fixed costs of building overlapping physical networks.

However, given the high fixed costs of providing a network, the economic incentive to develop an "integrated services" network is strong. Furthermore, now that all information can be easily digitized separate networks for separate types of traffic are no longer necessary. Convergence toward a unified, *integrated services network* is a basic feature in most visions of the much publicized "information superhighway." The migration to integrated services networks will have important implications for market structure and competition.

The international telephone community has committed to a future network design that combines elements of both circuit and packet switching to enable the provision of integrated services. The ITU (formerly CCITT, an international standards body for telecommunications) has adopted a "cell-switching" technology called ATM (asynchronous transfer mode) for future high-speed networks. Cell switching closely resembles packet switching in that it breaks a data stream into packets which are then placed on lines that are shared by several streams. One major difference is that cells have a fixed size while packets can have different sizes. This makes it possible in principle to offer bounded delay guarantees (since a cell will not get stuck for a surprisingly long time behind an unusually large packet).

An ATM network also resembles a circuit-switched network in that it provides connection-oriented service. Each connection has set-up phase, during which a "virtual circuit" is created. The fact that the circuit is virtual, not physical, provides two major advantages. First, it is not necessary to reserve network resources for a given connection; the economic efficiencies of statistical multiplexing can be realized. Second, once a virtual circuit path is established switching time is minimized, which allows much higher network throughput. Initial ATM networks are already being operated at 155 Mbps, while the non-ATM Internet backbones operate at no more than 45 Mbps. The path to 1000 Mbps (gigabit) networks seems much clearer for ATM than for traditional packet switching.

When will the "information superhighway" arrive?

The federal High Performance Computing Act of 1991 aimed for a gigabit per second (Gbps) national backbone by 1995. Six federally-funded testbed networks are currently demonstrating various gigabit approaches. To get a feel for how fast a gigabit per second is, note that most small colleges or universities today have 56 Kbps Internet connections. At 56 Kbps it takes about five hours to transmit one gigabit!

Efforts to develop integrated services networks also have exploded. Several cable companies have already started offering Internet connections to their customers.[4] ATT, MCI and all of the "Baby Bell" operating companies are involved in mergers and joint ventures with cable TV and other specialized network providers to deliver new integrated services such as video-on-demand. ATM-based networks, although initially developed for phone systems, ironically have been first implemented for data networks within corporations and by some regional and backbone providers.

[4] Because most cable networks are one-way, these connections usually use an "asymmetric" network connector that brings the input in through the TV cable at 10 Mbps, but sends the output out through a regular phone line at about 14.4 Kbps. This scheme may be popular since most users tend to download more information than they upload.

5. How is Internet access priced?

What types of pricing schemes are used?

Until recently, nearly all users faced the same pricing structure for Internet usage. A fixed-bandwidth connection was charged an annual fee, which allowed for unlimited usage up to the physical maximum flow rate (bandwidth). We call this "connection pricing". Most connection fees were paid by organizations (universities, government agencies, etc.) and the users paid nothing themselves.

Simple connection pricing still dominates the market, but a number of variants have emerged. The most notable is "committed information rate" pricing. In this scheme, an organization is charged a two-part fee. One fee is based on the bandwidth of the connection, which is the maximum *feasible* flow rate; the second fee is based on the maximum *guaranteed* flow to the customer. The network provider installs sufficient capacity to simultaneously transport the committed rate for all of its customers, and installs flow regulators on each connection. When some customers operate below that rate, the excess network capacity is available on a first-come, first-served basis for the other customers. This type of pricing is more common in private networks than in the Internet because a TCP/IP flow rate can be guaranteed only network by network, greatly limiting its value unless a large number of the 20,000 Internet networks coordinate on offering this type of guarantee.

Networks that offer committed information pricing generally have enough capacity to meet the entire guaranteed bandwidth. This is a bit like a bank holding 100% reserves, but is necessary with existing technology since there is no commonly used way to prioritize packets.

For most usage, the marginal packet placed on the Internet is priced at zero. At the outer fringes there are a few exceptions. For example, several private networks (such as Compuserve) provide email connections to the Internet. Several of these charge per message above a low threshold. The public networks in Chile and New Zealand charge their customers by the packet for all international traffic. We discuss some implications of this kind of pricing below.

6. What economic problems does the Internet face?

If you have read this far in the article, you should have a good basic understanding of the current state of the Internet—we hope that most of the questions you have had about the how the Internet works have been answered. Starting here we will move from FAQs and "facts" towards conjectures, FEOs (firmly expressed opinions), and PBIs (partially baked ideas).

How can the Internet deal with increasing congestion?

Nearly all usage of the Internet backbones is unpriced at the margin. Organizations pay a fixed fee in exchange for unlimited access up to the maximum throughput of their particular connection. This is a classic problem of the commons. The externality exists because a packet-switched network is a shared-media technology: each extra packet that Sue User sends imposes a cost on all other users because the resources Sue is using are not available to them. This cost can come in form of delay or lost (dropped) packets.

Without an incentive to economize on usage, congestion can become quite serious. Indeed, the problem is more serious for data networks than for many other congestible resources because of the tremendously wide range of usage rates. On a highway, for example, at a given moment a

single user is more or less limited to putting either one or zero cars on the road. In a data network, however, single user at a modern workstation can send a few bytes of e-mail or put a load of hundreds of Mbps on the network. Within a year any undergraduate with a new Macintosh will be able to plug in a video camera and transmit live videos home to mom, demanding as much as 1 Mbps. Since the maximum throughput on current backbones is only 45 Mbps, it is clear that even a few users with relatively inexpensive equipment could bring the network to its knees.

Congestion problems are not just hypothetical. For example, congestion was quite severe in 1987 when the NSFNET backbone was running at much slower transmission speeds (56 Kbps). Users running interactive remote terminal sessions were experiencing unacceptable delays. As a temporary fix, the NSFNET programmed the routers to give terminal sessions (using the `telnet` program) higher priority than file transfers (using the `ftp` program). (See Goffe (1994) paper for a description of `telnet` and `ftp`.)

More recently, many services on the Internet have experienced severe congestion problems. Large `ftp` archives, `Web` servers at the National Center for Supercomputer Applications, the original `Archie` site at McGill University and many services have had serious problems with overuse. See Markoff (1993) for more detailed descriptions.

If everyone just stuck to ASCII email congestion would not likely become a problem for many years, if ever. However, the demand for multi-media services is growing dramatically. New services such as Mosaic and Internet Talk Radio are consuming ever-increasing amounts of bandwidth. The supply of bandwidth is increasing dramatically, but so is the demand. If usage remains unpriced is is likely that there will be periods when the demand for bandwidth exceeds the supply in the foreseeable future.

What non-price mechanisms can be used for congestion control?

Administratively assigning different priorities to different types of traffic is appealing, but impractical as a long-run solution to congestion costs due to the usual inefficiencies of rationing. However, there is an even more severe technological problem: it is impossible to enforce. From the network's perspective, bits are bits and there is no certain way to distinguish between different types of uses. By convention, most standard programs use a unique identifier that is included in the TCP header (called the "port" number); this is what NSFNET used for its priority scheme in 1987. However, it is a trivial matter to put a different port number into the packet headers; for example to assign the `telnet` number to `ftp` packets to defeat the 1987 priority scheme. To avoid this problem, NSFNET kept its prioritization mechanism secret, but that is hardly a long-run solution.

What other mechanisms can be used to control congestion? The most obvious approach for economists is to charge some sort of usage price. However, to date, there has been almost no serious consideration of usage pricing for backbone services, and even tentative proposals for usage pricing have been met with strong opposition. We will discuss pricing below but first we examine some non-price mechanisms that have been proposed.

Many proposals rely on voluntary efforts to control congestion. Numerous participants in congestion discussions suggest that peer pressure and user ethics will be sufficient to control congestion costs. For example, recently a single user started broadcasting a 350–450Kbps audio-video test pattern to hosts around the world, blocking the network's ability to handle a scheduled audio broadcast from a Finnish university. A leading network engineers sent a strongly-worded e-mail message to the user's site administrator, and the offending workstation was disconnected from

the network. However, this example also illustrates the problem with relying on peer pressure: the inefficient use was not terminated until after it had caused serious disruption. Further, it apparently was caused by a novice user who did not understand the impact of what he had done; as network access becomes ubiquitous there will be an ever-increasing number of unsophisticated users who have access to applications that can cause severe congestion if not properly used. And of course, peer pressure may be quite ineffective against malicious users who want to intentionally cause network congestion.

One recent proposal for voluntary control is closely related to the 1987 method used by the NSFNET (Bohn, Braun, Claffy, and Wolff (1993)). This proposal would require users to indicate the priority they want each of their sessions to receive, and for routers to be programmed to maintain multiple queues for each priority class. Obviously, the success of this scheme would depend on users' willingness to assign lower priorities to some of their traffic. In any case, as long as it is possible for just one or a few abusive users to create crippling congestion, voluntary priority schemes that are not robust to forgetfulness, ignorance, or malice may be largely ineffective.

In fact, a number of voluntary mechanisms are in place today. They are somewhat helpful in part because most users are unaware of them, or because they require some programming expertise to defeat. For example, most implementations of the TCP protocols use a "slow start" algorithm which controls the rate of transmission based on the current state of delay in the network. Nothing prevents users from modifying their TCP implementation to send full throttle if they do not want to behave "nicely."

A completely different approach to reducing congestion is purely technological: overprovisioning. Overprovisioning means maintaining sufficient network capacity to support the peak demands without noticeable service degradation.[5] This has been the most important mechanism used to date in the Internet. However, overprovisioning is costly, and with both very-high-bandwidth applications and near-universal access fast approaching, it may become too costly. In simple terms, will the cost of capacity decline faster than the growth in capacity demand?

Given the explosive growth in demand and the long lead time needed to introduce new network protocols, the Internet may face serious problems very soon if productivity increases do not keep up. Therefore, we believe it is time to seriously examine incentive-compatible allocation mechanisms, such as various forms of usage pricing.

How can users be induced to choose the right level of service?

The current Internet offers a single service quality: "best efforts packet service." Packets are transported first-come, first-served with no guarantee of success. Some packets may experience severe delays, while others may be dropped and never arrive.

However, different kinds of data place different demands on network services. E-mail and file transfers requires 100% accuracy, but can easily tolerate delay. Real-time voice broadcasts require much higher bandwidth than file transfers, and can only tolerate minor delays, but they can tolerate significant distortion. Real time *video* broadcasts have very low tolerance for delay *and* distortion.

Because of these different requirements, network routing algorithms will want to treat different types of traffic differently—giving higher priority to, say, real-time video than to e-mail or file transfer. But in order to do this, the user must truthfully indicate what type of traffic he or she is

[5] The effects of network congestion are usually negligible until usage is very close to capacity.

sending. If real-time video bit streams get the highest quality service, why not claim that all of your bit streams are real-time video?

Cocchi, Estrin, Shenker, and Zhang (1992) point out that it is useful to look at network pricing as mechanism design problem. The user can indicate the "type" of his transmission, and the workstation in turn reports this type to the network. In order to ensure truthful revelation of preferences, the reporting and billing mechanism must be incentive compatible. The field of mechanism design has been criticized for ignoring bounded rationality of human subjects. However, in this context, the workstation is doing most of the computation, so that quite complex mechanisms may be feasible.

What are the problems associated with Internet accounting?

One of the first necessary steps for implementing usage-based pricing (either for congestion control or multiple service class allocation) is to measure and account for usage. Accounting poses some serious problems. For one thing, packet service is inherently ill-suited to detailed usage accounting, because every packet is independent. As an example, a one-minute phone call in a circuit-switched network requires one accounting entry in the usage database. But in a packet network that one-minute phone call would require around 2500 average-sized packets; complete accounting for every packet would then require about 2500 entries in the database. On the NSFNET alone nearly 60 billion packets are being delivered *each month*. Maintaining detailed accounting by the packet similar to phone company accounting may be too expensive.

Another accounting problem concerns the granularity of the records. Presumably accounting detail is most useful when it traces traffic to the user. Certainly if the purpose of accounting is to charge prices as incentives, those incentives will be most effective if they affect the person actually making the usage decisions. But the network is at best capable of reliably identifying the originating host computer (just as phone networks only identify the phone number that placed a call, not the caller). Another layer of expensive and complex authorization and accounting software will be required on the host computer in order to track which user accounts are responsible for which packets.[6] Imagine, for instance, trying to account for student e-mail usage at a large public computer cluster.

Accounting is more practical and less costly the higher the level of aggregation. For example, the NSFNET already collects some information on usage by each of the subnetworks that connect to its backbone (although these data are based on a sample, not an exhaustive accounting for every packet). Whether accounting at lower levels of aggregation is worthwhile is a different question that depends importantly on cost-saving innovations in internetwork accounting methods.

Does network usage need to be priced?

Network resources are scarce, and thus some allocation scheme is required. We explained above why voluntary and technological allocation mechanisms are unlikely to remain satisfactory. Various forms of usage pricing have desirable features for congestion control, and are likely to be equally desirable for allocating multiple service classes in an integrated services network.

[6] Statistical sampling could lower costs substantially, but its acceptability depends on the level at which usage is measured—e.g., user or organization—and on the statistical distribution of demand. For example, strong serial correlation can cause problems.

In any case, voluntary schemes will require substantial overprovisioning to handle the burstiness of demand, and the wide range of bandwidths required by different applications. Excess capacity has been subsidized heavily—directly or indirectly—through public funding. While providing network services as a zero marginal price public good probably made sense during the research, development and deployment phases of the Internet, it is harder to rationalize as the network matures and becomes widely used by commercial interests. Why should data network usage be free even to universities, when telephone and postal usage are not?[7]

Indeed, the Congress required that the federally-developed gigabit network technology must accommodate usage accounting and pricing. Further, the NSF will no longer provide backbone services, leaving the general purpose public network to commercial and state agency providers. As the net increasingly becomes privatized, competitive forces may necessitate the use of more efficient allocation mechanisms. Thus, it appears that there are both public and private pressures for serious consideration of pricing. The trick is to design a pricing system that minimizes transactions costs.

What should be priced?

Standard economic theory suggests that prices should be matched to costs. There are three main elements of network costs: the cost of connecting to the net, the cost of providing additional network capacity, and the social cost of congestion. Once capacity is in place, direct usage cost is negligible, and by itself is almost surely is not worth charging for given the accounting and billing costs.[8]

Charging for connections is conceptually straightforward: a connection requires a line, a router, and some labor effort. The line and the router are reversible investments and thus are reasonably charged for on annual lease basis (though many organizations buy their own routers). Indeed, this is essentially the current scheme for Internet connection fees.

Charging for incremental capacity requires usage information. Ideally, we need a measure of the organization's demand during the expected peak period of usage over some period, to determine its share of the incremental capacity requirement. In practice, it might seem that a reasonable approximation would be to charge a premium price for usage during pre-determined peak periods (a positive price if the base usage price is zero), as is routinely done for electricity. However, casual evidence suggests that peak demand periods are much less predictable than for other utility services. One reason is that it is very easy to use the computer to schedule some activities for off-peak hours, leading to a shifting peaks problem.[9] In addition, so much traffic traverses long distances around the globe that time zone differences are important. Network statistics reveal very irregular time-of-day usage patterns (MacKie-Mason and Varian (1994)).

[7] Many university employees routinely use email rather than the phone to communicate with friends and family at other Internet-connected sites. Likewise, a service is now being offered to transmit faxes between cities over the Internet for free, then paying only the local phone call charges to deliver them to the intended fax machine.

[8] See MacKie-Mason and Varian (1993).

[9] The single largest current use of network capacity is file transfer, much of which is distribution of files from central archives to distributed local archives. The timing for a large fraction of file transfer is likely to be flexible. Just as most fax machines allow faxes to be transmitted at off-peak times, large data files could easily be transferred at off-peak times—if users had appropriate incentives to adopt such practices.

How might congestion be priced?

We have elsewhere described a scheme for efficient pricing of the congestion costs (1994a,b). The basic problem is that when the network is near capacity, a user's incremental packet imposes costs on other users in the form of delay or dropped packets. Our scheme for internalizing this cost is to impose a congestion price on usage that is determined by a real-time Vickrey auction. Following the terminology of Vernon Smith and Charles Plott, we call this a "smart market."

The basic idea is simple. Much of the time the network is uncongested, and the price for usage should be zero. When the network is congested, packets are queued and delayed. The current queuing scheme is FIFO. We propose instead that packets should be prioritized based on the value that the user puts on getting the packet through quickly. To do this, each user assigns her packets a bid measuring her willingness-to-pay for immediate servicing. At congested routers, packets are prioritized based on bids. In order to make the scheme incentive-compatible, users are not charged the price they bid, but rather are charged the bid of the *lowest* priority packet that is admitted to the network. It is well-known that this mechanism provides the right incentives for truthful revelation.

This scheme has a number of nice features. In particular, not only do those with the highest cost of delay get served first, but the prices also send the right signals for capacity expansion in a competitive market for network services. If all of the congestion revenues are reinvested in new capacity, then capacity will be expanded to the point where its marginal value is equal to its marginal cost.

What are some problems with a smart market?

Prices in a real-world smart market cannot be updated continuously. The efficient price is determined by comparing a list of user bids to the available capacity and determining the cutoff price. In fact, packets arrive not all at once but over time, and thus it would be necessary to clear the market periodically based on a time-slice of bids. The efficiency of this scheme, then, depends on how costly it is to frequently clear the market and on how persistent the periods of congestion are. If congestion is exceedingly transient then by the time the market price is updated the state of congestion may have changed.

A number of network specialists have suggested that many customers—particularly not-for-profit agencies and schools—will object because they do not know in advance how much network utilization will cost them. We believe that this argument is partially a red herring, since the user's bid always controls the *maximum* network usage costs. Indeed, since we expect that for most traffic the congestion price will be zero, it should be possible for most users to avoid ever paying a usage charge by simply setting all packet bids to zero.[10] When the network is congested enough to have a positive congestion price, these users will pay the cost in units of delay rather than cash, as they do today.

We also expect that in a competitive market for network services, fluctuating congestion prices would usually be a "wholesale" phenomenon, and that intermediaries would repackage the services and offer them at a guaranteed price to end-users. Essentially this would create a futures market for network services.

[10] Since most users are willing to tolerate some delay for email, file transfer and so forth, most traffic should be able to go through with acceptable delays at a zero congestion price, but time-critical traffic will typically pay a positive price.

There are also auction-theoretic problems that have to be solved. Our proposal specifies a single network entry point with auctioned access. In practice, networks have multiple gateways, each subject to differing states of congestion. Should a smart market be located in a single, central hub, with current prices continuously transmitted to the many gateways? Or should a set of simultaneous auctions operate at each gateway? How much coordination should there be between the separate auctions? All of these questions need not only theoretical models, but also empirical work to determine the optimal rate of market-clearing and inter-auction information sharing, given the costs and delays of real-time communication.

Another serious problem for almost any usage pricing scheme is how to correctly determine whether sender or receiver should be billed. With telephone calls it is clear that in most cases the originator of a call should pay. However, in a packet network, both "sides" originate their own packets, and in a connectionless network there is no mechanism for identifying party B's packets that were solicited as responses to a session initiated by party A. Consider a simple example: A major use of the Internet is for file retrieval from public archives. If the originator of each packet were charged for that packet's congestion cost, then the providers of free public goods (the file archives) would pay nearly all of the congestion charges induced by a user's file request.[11] Either the public archive provider would need a billing mechanism to charge requesters for the (ex post) congestion charges, or the network would need to be engineered so that it could bill the correct party. In principle this problem can be solved by schemes like "800", "900" and collect phone calls, but the added complexity in a packetized network may make these schemes too costly.

How large would congestion prices be?

Consider the *average* cost of the current NSFNET backbone: about $\$10^6$ per month, for about $60,000 \times 10^6$ packets per month. This implies a cost per packet (around 200 bytes) of about 1/600 cents. If there are 20 million users of the NSFNET backbone (10 per host computer), then full cost recovery of the NSFNET subsidy would imply an average monthly bill of about $0.08 per person. If we accept the estimate that the total cost of the U.S. portion of the Internet is about 10 times the NSFNET subsidy, we come up with 50 cents per person per month for *full* cost recovery. The revenue from congestion fees would presumably be significantly less than this amount.[12]

The average cost of the Internet is so small today because the technology is so efficient: the packet-switching technology allows for very cost-effective use of existing lines and switches. If everyone only sent ASCII email, there would probably never be congestion problems on the Internet. However, new applications are creating huge demands for additional bandwidth. A *video* e-mail message could easily use 10^4 more bits than a plain text ASCII e-mail with the "same" information content and providing this amount of incremental bandwidth could be quite expensive. Well-designed congestion prices would not charge everyone the average cost of this incremental bandwidth, but instead charge those users whose demands create the congestion and need for additional capacity.

[11] Public file servers in Chile and New Zealand already face this problem: any packets they send in response to requests from foreign hosts are charged by the network. Network administrators in New Zealand are concerned that this blind charging scheme is stifling the production of information public goods. For now, those public archives that do exist have a sign-on notice pleading with international users to be considerate of the costs they are imposing on the archive providers.

[12] If revenue from congestion fees exceeded the cost of the network, it would be profitable to expand the size of the network.

How should information services be priced?

Our focus thus far has been on the technology, costs and pricing of network transport. However, most of the value of the network is not in the transport, but in the value of the information being transported. For the full potential of the Internet to be realized it will be necessary to develop methods to charge for the value of information services available on the network.

There are vast troves of high-quality information (and probably equally large troves of dreck) currently available on the Internet, all available as free goods. Historically, there has been a strong base of volunteerism to collect and maintain data, software and other information archives. However, as usage explodes, volunteer providers are learning that they need revenues to cover their costs. And of course, careful researchers may be skeptical about the quality of any information provided for free.

Charging for information resources is quite a difficult problem. A service like Compuserve charges customers by establishing a billing account. This requires that users obtain a password, and that the information provider implement a sophisticated accounting and billing infrastructure. However, one of the advantages of the Internet is that it is so decentralized: information sources are located on thousands of different computers. It would simply be too costly for every information provider to set up an independent billing system and give out separate passwords to each of its registered users. Users could end up with dozens of different authentication mechanisms for different services.

A deeper problem for pricing information services is that our traditional pricing schemes are not appropriate. Most pricing is based on the measurement of replications: we pay for each copy of a book, each piece of furniture, and so forth. This usually works because the high cost of replication generally prevents us from avoiding payment. If you buy a table we like, we generally have to go to the manufacturer to buy one for ourselves; we can't just simply copy yours. With information goods the pricing-by-replication scheme breaks down. This has been a major problem for the software industry: once the sunk costs of software development are invested, replication costs essentially zero. The same is especially true for any form of information that can be transmitted over the network. Imagine, for example, that copy shops begin to make course packs available electronically. What is to stop a young entrepreneur from buying one copy and selling it at a lower price to everyone else in the class? This is a much greater problem even than that which publishers face from unauthorized photocopying, since the cost of replication is essentially zero.

There is a small literature on the economics of copying that examines some of these issues. However, the same network connections that exacerbate the problems of pricing "information goods" may also help to solve some of these problems. For example, Cox (1992) describes the idea of "superdistribution" of "information objects" in which accessing a piece of information automatically sends a payment to the provider via the network. However, there are several problems remaining to be solved before such schemes can become widely used.

What is required for electronic commerce over the Internet?

Some companies have already begun to advertise and sell products and services over the Internet. Home shopping is expected to be a major application for future integrated services networks that transport sound and video. Electronic commerce could substantially increase productivity by reducing the time and other transactions costs inherent in commerce, much as mail-order shopping

has already begun to do. One important requirement for a complete electronic commerce economy is an acceptable form of electronic payment.[13]

Bank debit cards and automatic teller cards work because they have reliable authentication procedures based on both a physical device and knowledge of a private code. Digital currency over the network is more difficult because it is not possible to install physical devices and protect them from tampering on every workstation.[14] Therefore, authentication and authorization most likely will be based solely on the use of private codes. Another objective is anonymity so individual buying histories cannot be collected and sold to marketing agencies (or Senate confirmation committees).

A number of recent computer science papers have proposed protocols for digital cash, checks and credit, each of which has some desirable features, yet none of which has been widely implemented thus far. The seminal paper is Chaum (1985) which proposed an anonymous form of digital cash, but one which required a single central bank to electronically verify the authenticity of each "coin" when it was used. Medvinsky and Neuman (1993) propose a form of digital check that is not completely anonymous, but is much more workable for widespread commerce with multiple banks. Low, Maxemchuk, and Paul (1994) suggest a protocol for anonymous credit cards.

What does the Internet mean for telecommunications regulation?

The growth of data networks like the Internet are an increasingly important motivation for regulatory reform of telecommunications. A primary principle of the current regulatory structure, for example, is that local phone service is a natural monopoly, and thus must be regulated. However, local phone companies face ever-increasing competition from data network services. For example, the fastest growing component of telephone demand has been for fax transmission, but fax technology is better suited to packet-switching networks than to voice networks, and faxes are increasingly transmitted over the Internet. As integrated services networks emerge, they will provide an alternative for voice calls and video conferencing, as well. This "bypass" is already occurring in the advanced private networks that many corporations, such as General Electric, are building.

As a result, the trend seems to be toward removing of barriers against cross-ownership of local phone and cable TV companies. The regional Bell operating companies have filed a motion to remove the remaining restrictions of the Modified Final Judgement that created them (with the 1984 breakup of ATT). The White House, Congress, and the FCC are all developing new models of regulation, with a strong bias towards deregulation (for example, see the *New York Times,* 12 January 1994, p. 1).

Internet transport itself is currently unregulated. This is consistent with the principal that common carriers are natural monopolies, and must be regulated, but the services provided over those common carriers are not. However, this principal has never been consistently applied to phone companies: the services provided over the phone lines are also regulated. Many public interest groups are now arguing for similar regulatory requirements for the Internet.

[13] In our work on pricing for network transport (1994a, 1994b), we have found that some form of secure electronic currency is almost surely necessary if the transactions costs of accounting and billing are to be low enough to justify usage pricing.

[14] Traditional credit cards are unlikely to receive wide use over a data network, though there is some use currently. It is very easy to set up an untraceable computer account to fraudulently collect credit card numbers; fraudulent telephone mail order operations are more difficult to arrange.

One issue is "universal access," the assurance of basic service for all citizens at a very low price. But what is "basic service"? Is it merely a data line, or a multimedia integrated services connection? And in an increasingly competitive market for communications services, where should the money to subsidize universal access be raised? High-value uses which traditionally could be charged premium prices by monopoly providers are increasingly subject to competition and bypass.

A related question is whether the government should provide some data network services as public goods. Some initiatives are already underway. For instance, the Clinton administration has required that all published government documents be available in electronic form. Another current debate concerns the appropriate access subsidy for primary and secondary teachers and students.

What will be the market structure of the information highway?

If different components of local phone and cable TV networks are deregulated, what degree of competition is likely? Similar questions arise for data networks. For example, a number of observers believe that by ceding backbone transport to commercial providers, the federal government has endorsed above-cost pricing by a small oligopoly of providers. Looking ahead, equilibrium market structures may be quite different for the emerging integrated services networks than they are for the current specialized networks.

One interesting question is the interaction between pricing schemes and market structure. If competing backbones continue to offer only connection pricing, would an entrepreneur be able to skim off high-value users by charging usage prices, but offering more efficient congestion control? Alternatively, would a flat-rate connection price provider be able to undercut usage-price providers, by capturing a large share of low-value "baseload" customers who prefer to pay for congestion with delay rather than cash? The interaction between pricing and market structure may have important policy implications, because certain types of pricing may rely on compatibilities between competing networks that will enable efficient accounting and billing. Thus, compatibility regulation may be needed, similar to the interconnect rules imposed on regional Bell operating companies.

7. Further Reading

We have written two papers that provide further details on Internet technology, costs, and pricing problems (1994a, 1994b). In addition, a longer and more up-to-date version of this paper is available as a World Wide Web (WWW) document, with hypertext links to many related papers and data sources. These files can be found at `http://gopher.econ.lsa.umich.edu`.

Scott Shenker and his colleagues have written two papers dealing with pricing problems and the use of mechanism design to deal with them (Cocchi et al. (1992), Shenker (1993), Cocchi, Estin, Shenker, and Zhang (1991)). Huberman (1988) is a book that discusses computer networks as market economies.

Partridge (1993) has written an excellent book for a general audience interested in network technology now and in the near future. For a detailed discussion of computer networking theory and technologies, see Tanenbaum (1989). The best detailed treatment of the emerging ATM technology is de Prycker (1993).

References

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.

Chaum, D. (1985). Security without identification: Transaction systems to make big brother obsolete. *Communications of the ACM, 28*(10), 1030–1044.

Cocchi, R., Estin, D., Shenker, S., and Zhang, L. (1991). A study of priority pricing in multiple service class networks. In *Proceedings of Sigcomm '91*. Available at ftp: //ftp.parc.xerox.com/pub/net-research/pricing-sc.ps.

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

Cox, B. (1992). What if there is a silver bullet and the competition gets it first?. *Journal of Object-oriented Programming, xx.*

de Prycker, M. (1993). *Asynchronous Transfer Mode : Solution for ISDN* (2nd edition). Ellis Horwood, New York.

Goffe, W. (1994). Internet resources for economists. Tech. rep., University of Southern Mississippi. To appear in *Journal of Economic Perspectives*, Summer 1994. Available at gopher://niord.shsu.edu.

Huberman, B. (1988). *The Ecology of Computation*. North-Holland, New York.

Low, S., Maxemchuk, N. F., and Paul, S. (1994). Anonymous credit cards. Tech. rep., AT&T Bell Laboratories, Murray Hill, NJ. Available at ftp://research.att.com/dist/anoncc/anoncc.ps.Z.

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

MacKie-Mason, J. K., and Varian, H. (1994). Pricing the internet. In Kahin, B., and Keller, J. (Eds.), *Public Access to the Internet*. Prentice-Hall, Englewood Cliffs, New Jersey. Available from ftp://gopher.econ.lsa.umich.edu/pub/Papers.

Markoff, J. (1993). Traffic jams already on the information highway. *New York Times, November 3*, A1.

Medvinsky, G., and Neuman, B. C. (1993). Netcash: A design for practical electronic currency on the Internet. In *Proceedings of the First ACM Conference on Computer and Communications Security* New York. ACM Press. Available at ftp: //gopher.econ.lsa.umich.edu/pub/Archive/netcash.ps.Z.

Partridge, C. (1993). *Gigabit Networking*. Addison-Wesley, Reading, MA.

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

Tanenbaum, A. S. (1989). *Computer Networks*. Prentice Hall, Englewood Cliffs, NJ.