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

Andrew Odlyzko

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

July 7, 1998.

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

1. Introduction

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

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

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

The argument of this paper is that overengineering (providing enough capacity to meet peak demands) on a global scale may turn out not to be prohibitively expensive. It may even turn out to be the

cheapest approach when one considers the costs of QoS solutions for the entire information technologies (IT) industry.

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

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

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

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

networked computers has not been a problem in aggregate.)

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

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

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

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

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

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

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

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

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

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

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

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

Sections 2 through 12 describe the economics of the Internet. The conclusion is that with some ex-

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.