Some FAQs about Usage-Based Pricing

by

Jeffrey K. MacKie-Mason
Hal R. Varian

University of Michigan

August 1994
Current version: September 14, 1994

Abstract. This is a list of Frequently Asked Questions about usage-based pricing of the Internet. We argue that usage-based pricing is likely to come sooner or later and that some serious thought should be devoted to devising a sensible system of usage-based pricing.

Keywords. Internet, FAQs, usage-based pricing

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

Some FAQs about Usage-Based Pricing

Jeffrey K. MacKie-Mason
Hal R. Varian

1. Debunking some myths

Is usage-based pricing a "threat" or a "menace"?

Neither we hope. We think that some form of usage-based pricing for the Internet is likely (eventually) for reasons that we outline below. If some thought goes into designing a good pricing system, it should not cause problems for the vast majority of users. If a bad pricing system is adopted, it could cause a lot of problems. It is important to think carefully about how a reasonable pricing system might be designed far in advance of when one might be implemented.

Why is usage-based pricing desirable?

A major role of prices is to present information to people about the true costs of their actions. If prices accurately reflect costs then individuals can compare the benefits of their actions to the costs of their actions and make informed decisions.

Usage-based prices can be used to prioritize usage of a congested resource like a WWW server so that those who value access the most get the highest priority. Prices can also be used to allocate service classes to different uses and to recover costs of providing services. A key aspect of pricing services efficiently is that the revenues raised by the prices can be used to guide investment decisions and expand capacity.

Why is usage-based pricing undesirable?

The major objection to well-designed usage-based prices is the accounting and transactions cost. (We discuss these costs below.) Poorly-designed prices could have other costs such as impeding technical innovation and network usage. Usage-based prices may or may not be a good idea; it depends on how well they are designed, and whether the benefits they provide exceed the accounting and transactions costs.

Isn't usage-based pricing just a way to raise provider profits?

No. If Internet transport is provided by competitive firms (as it is today) then profits are determined by the degree of competition, not by the pricing mechanism. Some people have raised concerns about the potential monopolization of Internet transport. We think that this is unlikely in the foreseeable future. But regardless of how monopolization of the industry is, it doesn't have much to do with usage-based pricing: high profits can be obtained with flat rates and subscription charges, too.

This research was supported by the National Science Foundation Grant SES-93-20481. The NSF is not responsible for the content of this document.

Doesn't usage-based pricing necessarily raise users' total expenditures?

No. Again, if the industry is competitive or effectively regulated, then revenues will approximately equal costs. Costs may increase because of the added cost of accounting and billing. On the other hand, costs may decrease because usage-based pricing increases the efficiency of the network's functioning. When faced with usage charges, frivolous and low-value uses are likely to decrease, lowering total costs. In any case, instead of all users paying their *average* share of costs (through a connection, or subscription charge), they will start to pay for something closer to their own *incremental* share of costs. This means that low-intensity users should see a *reduction* in their total payments; high-intensity users will pay more.

Won't small users be hurt?

No. With flat-rate pricing, all costs are recovered through connection fees. These fees are based on average usage of a connection. That means that small (below-average) users are actually *subsidizing* the big users! With usage-based pricing, the heaviest users pay most of the costs. As we argue below, the heavy users are apt to be consumers of images and video; traditional text-based uses of the Internet will be tiny by comparison to multimedia.

2. Rationale for usage-based pricing

Why start pricing the Internet at all?

Internet transport is *already* priced, though many users seem unaware of that. Pricing is on the basis of a fixed monthly subscription fee for a connection of a given bandwidth. In most cases in the U.S. the incremental usage of that bandwidth is priced at a flat rate of zero. The reasonable question is not whether the Internet should be priced at all, but what type of pricing should be used.

How should prices be set?

One of the fundamental principles of economics is that prices should reflect costs. More specifically, the price of something should reflect it's *incremental social cost,* meaning the total cost to society of providing an additional unit of the good.

What are the costs of providing the Internet?

The easiest data to gather is data on the NSFNET, since it must report financial figures to the NSF. In recent years the NSF paid Merit about $12 million per year to maintain the NSFNET backbone. In addition they paid about $7 million to subsidize regional networks, and helped to subsidize various universities who wanted to connect to the Internet. In round terms the NSF support amounted to about $20 million per year.

The costs of Internet provision are dominated by the fixed costs. About 80% of the budget for the NSFNET goes to pay for line rental and routers. About 7% of the budget goes for the Network Operations Center. The incremental operating costs to servicing additional traffic are negligible at least up to the capacity of the network.

What about incremental social costs?

If a network resource is operating near capacity, other users who want to use the resource may be inconvenienced or delayed. If our file transfer delays your work by a minute, then the cost of our usage includes the value of that minute of your time. If our usage breaks up your interactive video conference, then our cost includes the value of your lost conference. Such *congestion costs* should be counted as part of the social costs of increasing network traffic.

What is the history of congestion on the Internet?

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 higher priority than file transfers. Subsequently, the NSFNET was upgraded to higher transmission speeds.

What is the current level of congestion on the Internet?

More recently, many services on the Internet have experienced significant congestion problems. Large `ftp` archives, `Web` servers at the NCSA, the original `Archie` site at McGill University and many services have had serious problems with overuse. The Mosaic home page at NCSA receives 1.3 million accesses a week. See Markoff (1993) for some anecdotes about “traffic jams on the Information superhighway.”. Afternoon delays in `telnet` sessions are observable every weekday in the Bay Area. As the use of multimedia increases, we expect to see significantly more congestion.

The *average* utilization of the NSFNET backbone is only about 5% of total capacity. But this is very misleading: IP traffic is very bursty and peak usage can be 10 times the average. With network traffic growing at 100% a year, we may easily face serious problems in the near future.

How much bandwidth does multimedia use?

The difference between plain old ASCII and multimedia is dramatic. Ordinary ASCII text uses about 44 bits per word. Telephone-quality voice uses 21,000 bits per word, and stereo CD uses 466,000 bits per word. Network quality video without compression is about 100 *megabits* per second. With compression, it’s about 45 Mbs—which is the entire capacity of the NSFNET backbone![1] (Some new compression schemes look like they will cut the bandwidth demands of video about in half.) Present-day video conferencing systems require about 400 Kbps.

In terms of file sizes, a 700 page book in ASCII is about 1 megabyte. On the other hand, a single non-compressed GIF image is about half a megabyte and a compressed JPG image is about a tenth of a megabyte. A 13 second compressed movie is over 4 megabytes.

[1] Figures taken from Lucky (1989).

What about new users?

The NSFNET backbone carries about 56 billion packets a month. If there are (conservatively) 10 million users on the backbone, the average user is sending 11,200 packets, or about 1 megabyte per month. This is about 1 ASCII book a month, or around 25 pages of text a day. This may seem like a lot—but one megabyte per month is only 2-4 GIF images per month, or about 4 seconds of a compressed movie. Existing users shifting to more bandwidth-intensive applications will put serious pressure on Internet bandwidth.

It also appears that there will be many new users of the Internet in the near-term future. The next release of MS-Windows and OS-2 are said to be "Internet ready". New users are likely to be attracted by high-bandwidth applications, which could also contribute significantly to an Internet crunch.

What might the future level of congestion look like?

If everyone just stuck to ASCII email congestion would not likely become a problem for many years, if ever. But even now, email only accounts for 15% of network usage; new ways to use the Internet are consuming ever increasing amounts of bandwidth.[2] The largest use of network bandwidth (37%) is `ftp` transfers. According to Ewing, Hall, and Schwartz (1992), around 9% were of these files were images.[3] WWW traffic, which makes heavy use of images, is one of the fastest growing components of network traffic. Hence, even today a significant component of network traffic is multimedia.

Won't technological progress increase the supply of bandwidth?

It is true that the supply of bandwidth is increasing, but so is the demand. The routers that handle network traffic are just computers. Improvements in the technology that increase their switch speed of the routers will also increase the speed of the computers that generate traffic on the net. Furthermore, the number of users with ever-faster computers connected to the Internet is exploding. There is no reason to think that bandwidth supply will fall in cost faster than demand increases. At some times there will be periods where the supply of bandwidth exceeds the demand. Between now and the time when every household on the planet is wired with a fiber optic connection (and every citizen has a broadband wireless device) we anticipate that congestion will be an increasingly serious problem.

What is the relationship between pricing and type-of-service?

Different kinds of traffic requires different treatment from the network. E-mail can be delayed without much loss; real-time video needs very rapid service. In order to provide appropriate treatment for different kinds of service, the person who generates the data has to indicated what

[2] NSFNET backbone usage has been increasing at 6–10% *per month* for the past 6 years, so total traffic has more than doubled each year.

[3] This estimate is based on the file name; this is certainly an underestimate since it does not count images that are compressed, tarred, or transferred using a non-standard naming convention.

type of stream it is. But if some sorts of data get "better treatment" than others, and all data costs the same to send, what is to prevent users from misrepresenting the type of their data?[4]

In order for a pricing system to "incentive compatible" it is necessary that use of higher quality service incurs a higher cost to the user. We have stated this principle for "priority", but it also holds for any type of special handling. This argument is laid out in detail in Shenker (1993).

What about recovery of fixed costs?

We don't think that usage prices are a very good way to recover the cost of providing network capacity. Since the network costs are primarily the fixed cost of capacity, it makes more sense to charge users a fixed fee depending on the capacity of their connection to the net. This is essentially the scheme that is used now. In general you want to recover costs that aren't sensitive to usage with non-usage-sensitive prices, and costs that are sensitive to usage with usage-sensitive prices.

How large would usage prices be for the current Internet?

The current NSFNET backbone costs about $\$10^6$ per month and carries $60,000 \times 10^6$ packets per month. This implies a cost per packet of about $1/600$ cents. If there are 10 million users of the NSFNET backbone then full cost recovery of the NSFNET subsidy would imply an average monthly bill of about 10 cents per person. If we accept the guesstimate that the total cost of the U.S. portion of the Internet is about 10 times the NSFNET subsidy, we come up with one dollar per person per month for *full* cost recovery. The revenue from congestion fees would presumably be significantly less than this amount, since if revenue from congestion fees exceeded the cost of the network, it would be profitable to expand the size of the network.

3. Different approaches to allocating network usage

What are "smart" markets?

In MacKie-Mason and Varian (1994a) we proposed a way to price network usage that we called "smart markets." Much of the time the network is uncongested; at such times the price for usage should be zero. However, when the network is congested, packets are queued, delayed, and dropped. 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 *highest* priority packet that is not admitted to the network. It can be shown that this mechanism provides the right incentives for users to reveal their true priority.

[4] For example, some advocate identifying traffic types by TCP port number in a TCP/IP network. As WWW users are well aware, however, it is straightforward to configure applications to use different port numbers.

Why are smart markets incentive compatible?

The basic idea is that the price a user pays is not determined by the priority he sets, but by the bid of the first packet that is rejected from the network. This means that the user has no incentive to misrepresent his true valuation of his packets. This is a form of a “Vickrey auction” or “second-price auction”, which economists have shown to be incentive compatible. See, for example, Vickrey (1961).

Why do smart markets send the right signals for capacity expansion?

The price for network usage set by the smart market is effectively the value of the packets that are *not* admitted to the network. If the total value of those packets exceeds the cost of expanding the network in order to handle them, then it is appropriate to expand capacity. Investing the revenues from congestion fees in capacity expansion is just the right rule to follow.

What are other proposals?

Bohn, Braun, Claffy, and Wolff (1993) have suggested using a mixture of altruism and quotas to implement priority-based routing. In their framework, users voluntarily declare a priority for their packets, and these priorities are subsequently charged against a usage quota. In their scheme, user quotas are charged for the requested priority whether or not the network is congested, unlike the smart market in which the price reflects the current state of congestion.

There are also several other interesting suggestions worth exploring. Hardy and Tribble (1993) describe a low-cost, bilateral transfer arrangement for routing packets based on “trust” and repeated interaction. Cocchi, Estrin, Shenker, and Zhang (1992) provide motivation for pricing, describe a general framework for dealing with pricing issues and conduct some simulations. Murphy and Murphy (1994) describe some simulations with pricing ATM traffic.

What about pricing multiple qualities of service?

Most studies to date have considered pricing for a single service dimension (priority in a first-come, first-served network, or call admission to an ATM network). However, with the growing demand for multimedia, we need to think about how to allocate multiple service qualities in an integrated network. For example, file transfers tolerate zero errors, but can tolerate substantial delay. Interactive video can tolerate some packet loss, but requires tight bounds on maximum delay and variation in delay. It is possible to use some generalizations of a smart market for pricing multiple qualities of service, but the computational burdens are significant. It may be that responsive pricing will be feasible in the near term only for reserving bandwidth and service qualities in advance.

How would prices be set in a free market?

We are not advocating *imposing* usage prices on the Internet. However we suspect that usage pricing will emerge for the reasons that we’ve outlined above. One scenario might go like this. In 5 or 6 years there could be a half-a-dozen competing backbone providers in the U.S. who interconnect at NAPS, or something like NAPS.[5] If high-volume users disrupt network traffic seriously, one or

[5] “NAPS” (Network Access Points) are switching centers where several independent networks interconnect.

more of the backbone providers might institute usage-based fees of the sort we described above (e.g., on the order of a thousandth of a cent per packet). If this happens, the high-volume users would move off of the charging network and on to one of the non-charging networks, immediately congesting it. The non-charging networks would be forced to follow suit quickly to avoid being swamped with traffic.

MacKie-Mason and Varian (1994b) sketch out an economic model of how a competitive market for network services might function.

What is the role of standards setting?

We don't think that the private market *alone* can provide a complete solution to the pricing problem. Although we think some form of usage pricing is likely, the form it takes could improve or worsen the quality of the Internet. To implement *good* usage pricing there clearly must be coordination among different carriers about what *forms* of pricing they will implement: for example, they have to interpret header information that sets "bids" or "priorities" in the same way. Furthermore, they will likely have to make "settlements" to compensate each other for carrying large amounts of each other's traffic. If we want an effective and beneficial pricing system, we must think now about standards for the necessary infrastructure to support such a system.

4. Accounting costs

How much does it cost to keep accounts for telephone calls?

Lots of numbers are tossed around about accounting costs in the telephone system. It is important to distinguish two categories of cost: *incremental cost,* (also known as *marginal cost*) and *average cost.* Average cost is just total cost of providing some level of service, divided by the total amount of that service provided. Marginal cost is the cost of providing *additional* units of service, given that some level of service is already being provided. The difference between the two boils down to the fixed costs.

In the telecommunications industry, as with the Internet, almost all costs are fixed costs: once the lines and switching equipment has been provided, it costs very little to use it, up to its capacity. When capacity has been reached, you have to pay to increase capacity if you want to increase usage.

This means that the marginal cost of a phone call is essentially zero if it is made in an off-peak period. The question is, how much does it cost to make an additional phone call during a *peak* period? Mitchell (1990) estimates that the the incremental capacity cost of a call during peak usage is about :5 1:0 cents per call.

How does this compare with billing costs? As of July 1990 he estimates:

$$\text{incremental itemized billing cost} = 0.7 – 1.2 \text{ cents per call}$$
$$\text{incremental summary billing cost} = 0.1 – 0.2 \text{ cents per call}$$
$$\text{account maintenance and collection costs} = 50 – 75 \text{ cents per month}:$$

Itemized billing costs are therefore more than 50% of the cost of an incremental call. But this is because the cost of an incremental call is so small, not because billing costs are so large. Since it costs almost nothing to make a call during non-peak periods, accounting costs are almost 100% of the incremental cost of a non-peak call!

It is also worth observing that *summary* billing costs are only about 10% of itemized billing costs: just counting message units is a lot cheaper than itemizing every call.

This gives us a picture of the *marginal* costs of billing; what do the *total* costs of billing look like? In 1984 AT&T paid the regional Bell operating companies to do their billing. The RBOCs charged them about 10 cents a transactions for billing, which was about 6% of their revenue from long-distance calls.[6] When AT&T was regulated, its revenues were supposed to be about the same as its costs. This suggests that billing and accounting were probably no more than 10% of *total* costs.

The difference between the 100% and 10% figure is that they use a different denominator: accounting costs are a large fraction of incremental costs, but a small fraction of total costs. This is because total costs are dominated by fixed costs and incremental costs are almost zero.

Are there differences between telephone and Internet accounting?

Unfortunately yes. Telephony is a connection-oriented service: each call has a setup phase, during which a connection is established. The connection is maintained for the length of the call. A single call generates only a single accounting record, no matter how long the connection lasts. The Internet, on the other hand, is a connectionless packet service. A given session is broken into many small packets, each of which traverses the network independently of the others.[7] If telephone-style accounting were implemented, the equivalent of a one-minute local phone call would generate about 2500 accounting records, and a ten-minute call would require 25,000 records! If usage-based pricing is to be feasible in a connectionless network, it may be essential to devise more efficient methods of accounting.

Another problem for Internet accounting is the pervasiveness of client-server applications, like `www`, `gopher`, and anonymous `ftp`. With telephony, the person placing the call is billed.[8] With a client-server application, most of the traffic may be sent by the server, on behalf of the client. A one-packet user request to a `www` server may generate a few hundred thousand packets of return traffic to download a file. To the network it appears that the server originated most of the traffic, but naturally it is the client who should be charged.

[6] By way of comparison, Hansell (1994) reports that transactions costs amount to 4% of banks' *cash* deposits.

[7] The problem of accounting for a connectionless service arises for applications as well as for network transport. For example, `www` and `gopher` are connectionless: each client request is a separate session. Accounting for `www` server usage requires a separate record for every "hit", even if a single user makes 30 hits during what she perceives to be a single continuous session.

[8] There are well-known ways to take advantage of this. For example, people on the East Coast of the U.S. will generally place late evening calls to people on the West Coast, where it is still early evening and the rates are higher. Similarly, there are redialing services that re-establish international calls so that they seemingly originate in the U.S. because U.S. rates are much lower than those in many other countries.

Are their other ways to do accounting?

Since organizations that use the Internet can be assumed to have ready access to computers, there may be ways to automate the billing process. One idea that we have been thinking about is *distributed accounting*. In this model, consumers purchase “digital stamps”—essentially numbers. These numbers are then sent along with the information that is being transferred. The routers can examine the numbers to make sure that they haven’t been used already. If they are valid, the packet will be transported. There are a variety of encryption techniques that can be used to ensure security in such a system.

This idea is closely related to Chaum (1985)’s idea of digital cash and Hardy and Tribble (1993)’s “digital silk road”. The advantage of such a system is that there is no centralized billing overhead. The burden of network accounting is distributed among the users, just like the burden of keeping track of postage is distributed among the customers of the Post Office.

Have other countries tried usage-based pricing?

Chile and New Zealand both have several years of experience with usage-based pricing. Baeza-Yates, Piqnera, and Poblete (1993) describe the situation in Chile and Brownlee (1994) describes the situation in New Zealand. According to these authors, the Chilean experience has not been very positive, but the New Zealand experience has been much more successful. The accounting software developed by the New Zealanders, NetTraMet, is available for anonymous `ftp` at `ftp.funet.fi` in `/pub/networking/management/NeTraMet`.

5. How do we keep things the way they were?

We can’t. The multimedia genie is out of the bottle. The bad news is that this is going to mean that the Internet is going to have to find new ways to allocate bandwidth. The good news is that if the increases in capacity to handle multimedia are put in place, there should be plenty of room for plain old ASCII transactions.

Many people are forecasting movies-on-demand, real-time video conferencing and other bandwidth-consuming applications that require orders of magnitude more bandwidth than we use today. If people are going to use these new applications they will have be priced so they are affordable—but this means that traditional text-based uses of the Internet will end up being essentially free. The challenge facing the Internet will be how to make the transition to this new multimedia environment.

References

Baeza-Yates, R., Piqnera, J. M., and Poblete, P. V. (1993). The Chilean internet connection or I never promised you a rose garden. In *Proc. INET '93*.

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.

Brownlee, J. N. (1994). Kawaihiko charging workshop report. Tech. rep., Computer Centre, The University of Auckland, Auckland, New Zealand.

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

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

Ewing, D. J., Hall, R. S., and Schwartz, M. F. (1992). A measurement study of Internet file transfer traffic. Tech. rep., Department of Comptuter Science, University of Colorado. ftp://ftp.cs.colorado.edu/pub/techreports/schwartz/FTP.Meas.ps.Z.

Hansell, S. (1994). An end to the 'nightmare' of cash?. *New York Times, Tuesday*, C1.

Hardy, N., and Tribble, E. D. (1993). The digital silk road. Tech. rep., Agorics, Inc.

Lucky, R. W. (1989). *Silicon Dreams: Information, Man and Machine*. St. Martin's Pres, New York.

MacKie-Mason, J. K., and Varian, H. R. (1994a). Economic FAQs about the internet. *Journal of Economic Perspectives*, *8*(3).

MacKie-Mason, J. K., and Varian, H. R. (1994b). Pricing congestible network resources. Tech. rep., University of Michigan. http://gopher.econ.lsa.umich.edu.

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

Mitchell, B. (1990). Incremental costs of telephone access and local use. Tech. rep. R3909, RAND.

Murphy, J., and Murphy, L. (1994). Bandwidth allocation by pricing in ATM networks. Tech. rep., EECS Department, University of California, Berkeley.

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

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, *16*, 8–37.

Biographies

Jeffrey K. MacKie-Mason

Jeffrey K. MacKie-Mason is an Associate Professor of Economics and Public Policy at the University of Michigan, and a Research Associate at the National Bureau of Economic Research in Cambridge, MA. He works in a number of fields in economics, including utility pricing, industrial organization, taxation, and corporate finance. He received his Ph.D. from the Massachusetts Institute of Technology in 1986, and a Master's in Public Policy from the University of Michigan in 1982. He has been a National Fellow at the Hoover Institution (Stanford), and a Visiting Scholar at the University of California, Berkeley, and the University of Oslo, Norway. His email address is jmm@umich.edu.

Hal R. Varian

Hal R. Varian is the Reuben Kempf Professor of Economics and a Professor of Finance at the University of Michigan. He received his S.B. degree from MIT in 1969 and his MA (mathematics) and Ph.D. (economics) from the University of California at Berkeley in 1973. He has taught at MIT, Berkeley, Stanford, Oxford, and several other universities, and has been at the University of Michigan since 1977. Professor Varian was a Guggenheim Fellow in 1979–80 and was elected a Fellow of the Econometric Society in 1983. He has served as Co-Editor of the *American Economic Review,* and is currently on the editorial boards of several journals.

Professor Varian has published numerous papers in economic theory, industrial organization, public finance, and econometrics. He is the author of a graduate textbook, *Microeconomic Analysis* and an undergraduate textbook, *Intermediate Microeconomics.* His email address is Hal.Varian@umich.edu.