A modest proposal for preventing Internet congestion

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

September 3, 1997

Abstract: A simple approach, called PMP (Paris Metro Pricing), is suggested for dealing with congestion in packet networks such as the Internet. It is to partition a network into several logical networks, each of which would treat all packets equally on a best effort basis, just as the current Internet does. There would be no formal guarantees of quality of service. The separate networks would differ only in the prices paid for using them. Networks with higher prices would attract less traffic, and thereby provide better service. Price would be the primary tool of traffic management.

1. Introduction

The Internet is the great success story of the 1990s. However, endemic congestion has led to wide dissatisfaction, and there is general agreement that new applications, especially real time ones such as packet telephony, will require higher quality of service. Various solutions to data network congestion are being developed, typically involving bandwidth reservation or priority setting. (See [Huitema, JordanJ, Keshav, Shenker2] for a discussion of some proposals and references.) Many of the proposed schemes are complicated, and involve substantial costs in both development and operations. Furthermore, since the basic problem is that of allocating a limited resource, any solutions will surely have to involve a pricing mechanism. This is felt by some to be a blemish, going against the tradition of the "free" Internet. Still, an explicit charging mechanism does appear inevitable to prevent the "tragedy of the commons" in which every packet is sent with the highest possible priority. Following in the footsteps of Jonathan Swift [Swift], I propose to turn a perceived burden into a solution, and rely on usage-sensitive pricing to control congestion, bypassing most of the complexity of other solutions. This should allow for simpler networks that are easier to design and deploy and operate faster.

The proposal (called PMP, an abbreviation of Paris Metro Pricing, for reasons explained below) is to partition a network into several logically separate networks. Each would have a fixed fraction of the capacity of the entire network. (Some variations on this design are possible and are discussed in Section 2.) All networks would route packets using protocols similar to the current TCP and UDP, with each packet treated equally. The only difference between the networks would be that they would charge different prices. Customers would choose the network to send their packets on (on a packet-by-packet basis, if they wished), and would pay accordingly. There would be no formal guarantees of quality of service, with packets handled on a "best effort" basis. The expectation is that the networks with higher prices would be less congested than those with lower prices.

All pricing mechanisms affect user demand, and thus can modify traffic loads. For example, the discount for evening calls on the voice telephone network shifts demand into the off-peak hours, and evens out the load on the network. The PMP proposal is to go further and use pricing as the main method of traffic management.

The PMP proposal was inspired by the Paris Metro system. Until about 15 years ago, when the rules were modified, the Paris Metro operated in a simple fashion, with 1st and 2nd class cars that were identical in number and quality of seats. The only difference was in the price of 1st and 2nd class tickets. (The Paris regional RER lines still operate on this basis.) The result was that 1st class cars were less congested, since only people who cared about being able to get a seat, not have to put up with noisy teenagers, etc., paid for 1st class. The system was self-regulating, in that whenever 1st class cars became too popular, some people decided they were not worth the extra cost, and traveled 2nd class, reducing congestion in 1st class and restoring the differential in quality of service between 1st and 2nd class cars.

The analogy of PMP with the Paris Metro should not be overdrawn. On the Paris Metro, both 1st and 2nd class passengers arrived at the destination at the same time. Different prices paid only for the expected differential in discomfort caused by congestion. In PMP, differences in service quality would be more complicated. For example, packets on lower-priced networks would have a higher chance of being dropped. The main point of the analogy is to show that a simple pricing scheme can induce users to separate themselves into classes that provide different quality of service, and that the division can be self-stabilizing.

Pricing is a crude tool. Different applications vary in requirements for bandwidth, latency, and jitter, for example. PMP would not provide any specific Quality of Service (QoS) guarantees. Unlike ATM, say, it would provide only a few channels, which would have only expected levels of service, not guaranteed ones. Moreover, subdividing a network into several pieces (even when the subdivision is on the logical and not the physical level) loses some of the advantages of statistical multiplexing that large networks offer. The justification for PMP is that, for all its deficiencies, the Internet does work, and with less congestion, even real-time applications can be run. Furthermore, there is no simple characterization of what QoS is required by different applications. The quality of service perceived by users depends in complicated ways on quantitative measures of network performance, and has a large subjective component. Thus there is little hope of satisfying everyone's quality demands. The hope of PMP is that a few classes of service will be satisfactory for most applications, just as a few classes of airline service suffice for most travelers, even though they have varied preferences for leg room, food, air temperature, and other attributes of air travel.

There are experts in the data networking community who argue that instead of working on complicated network schemes, all resources should be devoted to improving capacity (the "fat dumb pipe" model). (See p. 138 of [Huitema] and [Steinberg].) Technology is changing rapidly, and so there is an advantage to simple systems, since they can be developed and deployed much faster. As an example, for all the vaunted flexibility of ATM, it is currently being used primarily to provide "fat dumb pipes" [Steinberg]. ATM happens to be the fastest technology available, so it is deployed even though hardly any of its features are used. However, "fat dumb pipes" by themselves are unlikely to provide a workable solution, since all the evidence shows that with zero marginal costs, traffic will always grow to exhaust capacity. The PMP proposal is close to the "fat dumb pipe" one in the spectrum of possible approaches to network management. However, it brings in economic incentives to provide uncongested pipes (and thus higher quality services) for those who need them.

PMP inverts the usual order in which networks are designed. Usually an attempt is made to determine the QoS required by various application, then the network is designed to provide that QoS, and finally the prices are set. PMP sets the prices, and allows users to determine, based on their requirements and budgets as well as the feedback they receive about the collective actions of other user, what QoS they will receive. The expectation is that the different logical networks would usually have predictable performance and would provide sufficient QoS variety to satisfy most needs.

The pricing mechanism of PMP is about as simple as that of any usage-sensitive pricing scheme that has been proposed for the Internet. Thus the additional complexity it would introduce is minimal, and appears inevitable, since usage-sensitive pricing appears inevitable. The advantage of PMP is that it would provide congestion control essentially for free, once the pricing mechanism is in place, with only minor changes to the network infrastucture being required to handle the traffic management tasks.

The success of the Internet comes to a large extent from its connectionless nature, which simplifies the tasks of both users (who do not need to know anything about how their packets are handled) and networks (which, aside from the issue of routing tables, only have to react to local conditions, and do not need central or even distributed end-to-end coordination). Lack of adequate QoS is now leading the Internet community to consider bandwidth reservation policies such as RSVP. These require network nodes along the path of a transmission to coordinate their actions, and are thus "reinventing telephone technologies" [Steinberg]. The hope is that PMP would permit dispensing with measures such as RSVP and their complexity, and go back to the simpler model of the traditional Internet.

PMP is also designed to be acceptable to users, who have a strong preference for flat-rate pricing. It appears that consumers are willing to tolerate substantial variation in quality of a service or a product, but strongly prefer simple and predictable pricing schemes.

Section 2 presents PMP in greater detail. Section 3 discusses some of the potential problems of PMP, and possible ways to overcome them. Section 4 deals with the transition to PMP. Section 5 sketches the arguments for usage-sensitive pricing of the Internet, and also describes the public's aversion to such schemes, and the way in which PMP might help reconcile the two. Finally, Section 6 briefly outlines some of the other proposals for pricing data networks.

Modeling proposals such as PMP is hard, since our knowledge of the Internet and of user requirements and responses to different pricing schemes is sketchy at best. Appendix 1 presents some simple economic models of the gains that one could obtain from schemes such as PMP. Since there are large economies of scale and a steep learning curve in networking, lower cost service can be secured for all users by providing premium channels that attract additional, QoS-sensitive users. Such users are currently crowded out by the traffic that is insensitive to congestion.

Many aspects of PMP would require extensive research before it could be considered for deployment. This note is only a sketchy initial proposal.

2. PMP

The main idea of PMP is simply to have several channels that differ in price. They would offer different expected quality of service through the action of users who select the channel to send their data on. This section presents some methods for implementing this idea, and also discusses some related issues.

The number of subnetworks in PMP should be small, possibly just two, but more likely three or four. Having few networks minimizes losses from not aggregating all the traffic, and also fits consumer preferences (discussed in Section 5) for simple schemes. Furthermore, it is known (cf. [Wilson]) that in many situations, most of the economic gains from subdivision into different classes of service can be gained with just a few classes. In other, somewhat similar settings, a small number of classes of service has worked satisfactorily. Some railroads in the 19th century had up to four classes of cars, whereas today they operate with one or two. Airlines mostly have either two or three classes of service.

The basic version of PMP mentioned in the Introduction assigns to each subnetwork a fixed fraction of the capacity of the entire network. One can also use priorities. In the proposals [BohnBCW, GuptaSW2], for example, packets with higher priorities would always be treated by a router before packets with lower priorities. The advantage of this approach is that the full gain from aggregating all traffic on one network would be obtained. However, allowing high priority packets to block completely lower priority ones violates the fairness criterion that appears to be important to consumers (see Section 5 for further discussion of this topic). A better approach might be to use weights in routing decisions, such as in the weighted round-robin technique [Keshav]. One could also use different approaches in different parts of the network. One can even mix these approaches on the same link. For example, one could assign 40% of the capacity of the network to class 1 traffic, and 60% to classes 2 and 3, with weighted priority queuing determining what packets in classes 2 and 3 are to be sent first. The fixed assignment of capacities to different classes of service would probably be best for the core of the network.

In general, assignments of capacities and prices to the subnetworks in PMP should stay constant for extended periods. This would fit consumer preferences for simplicity and also allow usage patterns to stabilize, and thus produce a predictable level of service on different networks. However, it might also be desirable to have different assignments of capacities and prices for nights and weekends, to encourage better utilization.

PMP is concerned primarily with the user interactions with the network. It does not specify how traffic management is to be carried out inside the network. Methods such as Fair Queuing can be used with PMP when appropriate, as can IP-switching and tag-routing. Just as some current Internet IP traffic is carried by ATM networks, PMP traffic can be sent over a variety of networks. The intention in PMP is to reduce the traffic management task by inducing users to separate themselves into classes with different requirements. This would eliminate or at least reduce the need for approaches such as RSVP [Huitema, Keshav], which violate the Internet's connectionless approach, and require complicated coordination across the network. However, PMP could be combined with RSVP, if that was felt to be necessary, by having a separate channel devoted to traffic with bandwidth reservations. (One could also carve out RSVP capacity out of the lowest-cost channel.)

PMP is concerned only with usage-sensitive charges for data sent over a network. Currently such charges are infrequent, but there strong arguments (summarized in Section 5) that such charges will be needed for efficient networks that provide the variety of new services that are emerging. Other charges are already common. Flat monthly fees based on the bandwidth of the access link currently pay for most of the Internet. There are also charges for connect time (common in Europe, for example, whereas in the U.S. they apply primarily for access through 800 numbers, or for some online services), which are appropriate when modems or telephone lines have to be paid for. (Such charges would be less appropriate if data splitting equipment is installed, so that data traffic does not use the switches of the voice phone network.) Some charges of these types would be expected to apply in addition to the usage-sensitive charges for PMP (but would be considerably lower than if there was no charge for data loads). (For a survey of different types of Internet access charges around the world, see [OECD].)

PMP would also not deal with some other problems where charging might be appropriate. The only effective way to deal with spam (massive junk email) may well be to impose charges for email delivery. However, a 200-byte spam message takes just as much effort to recognize and delete as a 200-kilobyte message, while the costs of handling a 200-byte message are extremely low. Therefore, to control spam, email charges would have to be considerably higher than the charges imposed by PMP, and should be considered separately.

PMP charges would be assessed on each packet, and would probably consist of a fixed charge per packet and a fee depending on the size of the packet. The combination of these two fees would depend on network costs. Application software would undoubtedly be written to generate packets of sizes that would minimize transmission costs, so the prices would have to be "incentive compatible," in economists' language.

3. PMP problems and solutions

Would users find the lack of guaranteed quality of service (QoS) of PMP acceptable? In voice telephony, experience has taught people to expect a uniform and high level of service. However, that is an exception. Most purchases (of books, cars, and so on) are made on the basis of expected, not guaranteed, quality. (Section 5 has further discussion of this topic.) Today's Internet provides extremely variable and mostly low quality of service. This is only because there is no alternative. Few people are happy with the service they get, and some applications are impossible to implement or perform poorly. This appears to be the driving force behind the numerous proposals to provide quality of service guarantees. (See [JordanJ] for an overview and references.) However, it seems likely that the main problem is not the variability in quality of service on the Internet but the generally low quality of that service. There are fewer complaints about QoS on various institutional LANs and WANs, which do not have any service guarantees, and even the Internet is generally regarded as good in the early morning hours when it is lightly loaded. This suggests that PMP, a best-effort system without guarantees, but with several channels of different congestion levels, might satisfy most needs.

Even though the concept of guaranteed QoS is attractive, it is largely a mirage. The only ironclad guarantees that can be made are for constant bandwidth. That is what voice phone users get, since 64 kbs of network capacity is devoted to each call. In addition, this voice call guarantee only applies to a connection that is established, as there are periods of congestion when call attempts fail. There are also occasional glitches, such as calls being dropped or noise on the line, but they are infrequent enough not to be a problem. In data networks, efficiency depends largely on statistical multiplexing of sources with varying and unpredictable bandwidth demands. However, it is clearly impossible to satisfy all user requirements and take advantage of the efficiency of multiplexing. A 100 Mbs channel can often handle 50 transmissions, each of which requires 1 Mbs on average, but occasionally has bursts of 5 Mbs. However, if many of the bursts occur at the same time, not all the demands can be accommodated. The current TCP forces all transmissions to slow down, which might be regarded as unfair to the sources that are transmitting at low rates. UDP does not slow down at all, which is unfair to TCP users. Many of the proposed schemes (and even existing ones, such as those in Frame Relay networks) guarantee each source 1.5 Mbs of capacity, say. Doing this, however, requires that the network have, if not centralized, then at least closely coordinated control, to set up end-to-end bandwidth reservations. Further, the network and the transmitters have to negotiate for each session. The result for the user, which, after all, should be the deciding factor, is that the perceived performance of the network can degrade suddenly as a result of unpredictable actions of others, when the bandwidth of a connection drops down to the minimal guaranteed level. In particular, applications have to be responsive to network conditions, just as they have to be in a best-effort system like PMP.

Guaranteed QoS is a mirage for another reason as well. For at least the next decade, it appears that ATM will not come to the desktop. Hence most applications (aside possibly from services such as packet telephony, which might use their own network infrastructure) will start out on Ethernet-like networks, which are inherently best-effort.

PMP would do away with the complexity of network control. There would be occasional service degradations, but if they are infrequent enough, this should be acceptable. In PMP, the higher-priced networks would be less congested, and would suffer less frequent service degradation. A service with a minimal bandwidth guarantee of 0.5 Mbs could be simulated by sending the most important 0.5 Mbs (the voice in a videoconference call as well as the high order bits of the picture, say) on a higher-priced channel, and the rest on a lower-priced one. There would be no latency or packet delivery guarantees, but with a sufficient differential in congestion on the two networks, the effect could be comparable to that of conventional networks.

The main potential PMP problem is inefficiency. To provide a higher QoS than the current Internet, the premium networks would need to be less heavily loaded. Would capacity utilization have to be so low as to make the scheme uneconomic? Unfortunately we do not have enough information to answer this question. We do not even know how efficiently the Internet is operating. There have been careful studies of Internet performance (see [MonkC, NLANR, Paxson2, YajnikKT] and especially [Paxson3]), but the difficulties of collecting the data and analyzing it are substantial. There is not even comprehensive and widely accepted data on packet loss rates [Metcalfe]. There are regular workshops on Internet traffic measurement (see [NLANR] for pointers to these and other information sources), but the state of our ignorance about the Internet is astounding. The large network providers do not provide basic data on their total traffic or capacity utilization, and apparently many do not collect careful statistics. It appears that every part of the Internet is a bottleneck, and that the most serious choke points move around. Even such notoriously congested links as the one across the Atlantic do have periods when traffic moves smoothly.

The difficulties in deciding how efficiently the Internet is operating substantial. For the voice phone network, the problem is much simpler. Calls are discrete items, and are either completed or not. The fraction of calls that are blocked provides a precise measure of congestion. The AT&T voice phone network routinely operates at over 80% of its maximal capacity during the peak business hours, and few calls are blocked. However, the total capacity utilization is in the 15-20% range, since there are few calls in the slack periods. (It is worth mentioning that although these are precise figures, they are based on the idea that a phone call is 64 kbs. With compression, much lower transmission capacity would suffice. Thus even in the phone network the measurement of capacity utilization is not easy.)

On the Internet, capacity utilization is much harder to define. The statistics for the NSFNET compiled by Merit (and available through the links at [NLANR]) show that this backbone, towards the end of its existence, when it consisted exclusively of T3 lines, transmitted data at a rate that was only about 5% of the link capacity. What this presumably means is that the bottlenecks were elsewhere, most likely at the routers, or at the links connecting to the backbone. We do not know what the true capacity utilization was.

There are also problems in interpreting current Internet statistics. For example, consider the 15-minute average throughput data for the PacBell NAP for the period Aug. 3, 1997 to Aug. 27, 1997, available through the links at [NLANR]. After removing the data for Aug. 14 and 15 (when apparently the entire NAP was down for a few hours), we find that the minimum transmission rate was 50.3 Mbs, maximum was 309.4 Mbs, average was 222.3 Mbs, and the standard deviation was 48.0 Mbs. Thus the average transmission rate was 72% of the maximal one, much higher than for the phone network. Remarkably enough, the statistics for just Saturdays and Sundays during that period show figures of 107.8, 273.3, 205.5, and 41.3, respectively. This is again in contrast to the voice phone network, where there is little traffic on weekends. Similar utilization profiles apply to the other major switching points for which data is available at [NLANR].

One conclusion that could be drawn from these statistics is that the Internet is much more efficient than the voice phone network, with capacity utilization of over 70% as against 15-20%. However, such an argument is easy to question. For one thing, the maximal transmission rate through a node on the Internet under normal conditions is much less than the theoretical throughput. This is because data traffic is fractal [LelandTWW] (an observation that was first made in LANs, and has now been confirmed in many other data networks). This suggests that all data networks with heterogeneous sources will use only a fraction of their capacity, a considerably smaller fraction than the phone network does. There are further complications. Taking the ratio of observed average traffic to observed maximal traffic is a misleading utilization statistic, since the observed maximum is small compared to capacity. Further, observed traffic is not the same as useful traffic. When packets get lost, they are retransmitted (when using TCP, for example), which inflates traffic counts. The retransmission problem gets worse precisely when congestion increases. (The TCP acknowledgement packets appear to pose less of an overhead, but the routing information that is constantly being transmitted is another burden.)

Probably the main conclusion that can be drawn from available traffic statistics is that the Internet is terribly congested, and that it is extremely inefficient in the social and economic sense by repressing demand. Some of the apparently high utilization rate that is observed for the PacBell NAP, for example, is caused by the desirable shift of large data transfers (such as in mirroring databases) to the slack night hours. Most of it, though, appears to be caused by not satisfying existing and potential demand for data service. There is some data available through [NLANR] on packet loss rates, which are one indication of congestion. That is not the full story, though. Most of the Internet traffic is TCP, which uses variants of Van Jacobson's backoff algorithm (introduced in 1988 to prevent another collapse of the type that the Internet had suffered then). This algorithm slows down individual transmissions in the presence of congestion. Further, this algorithm has the effect of slowing demand from users, who, as a result of slow transmission, do less work on the Net than they could otherwise. Therefore the actual demand for data transmission is probably much higher during peak hours than is apparent from the statistics.

One indication of the repressed demand for Internet service is provided by comparing modem usage with traffic statistics. Data from an ISP show that the average number of modems in use is about 30% of the maximum number, a pattern of usage closer to that of the voice phone network than of the transmission pattern through the NAPs, say. This shows that it is not that traffic demand is more even on the Internet than on the voice phone network. Instead, what we are seeming is the result of severe congestion and rationing.

We are all familiar with highway traffic, when cars are moving smoothly, and then a sudden perturbation leads to a jam. It is a general phenomenon of queuing systems that close to a critical point, small increases in utilization can yield dramatic deterioration in service quality. Consider the simplest system, the $M/M/1$ queue. If the average throughput is increased from 90% of maximum feasible to 99%, the average queue size will grow from 9 to 99, and therefore the average time spent waiting in the queue will grow by the same factor of 11. Conversely, if we decrease utilization from 99% to 90%, by less than 10%, queue size and the time waiting in the queue will both decrease by a factor of 11. Note that such dramatic increases in service quality at small costs of efficiency are possible only near the critical point. If the $M/M/1$ queue is operating at 50% of maximum throughput, an 11-fold decrease in average queue size is possible only by going down to a utilization rate of 1/11, a 5.5-fold decrease.

The analogy should not be overdrawn, but the data cited earlier suggest that the Internet is operating closer to the 99% utilization level than to the 50% level in a queue. Looking at queue sizes in routers may be misleading, since most of the demand reduction is probably coming from the automatic action of TCP and users' reactions. Real congestion may be much worse. If this is true, small decreases in network utilization might lead to dramatic improvements in perceived QoS. The problem is how to achieve and maintain such a reduction. Usage-sensitive pricing would provide an incentive for users to keep their traffic demands from clogging the network, and also for service providers to build the capacity that there is demand for.

Much better data on the perceived quality of service as a function of capacity utilization is needed to determine how well PMP would perform. The hope that PMP would not require extreme overengineering of the network is supported by the observation that during the night and early morning, the Internet provides much better perceived service than during the busy hours. However, the load carried by the Internet does not vary much, as is shown by data cited before for the NAPs.

Various additional aspects of PMP that are important for its operation will not be dealt with here, as they would require further study, but do not seem to be crucial. For example, how does a network that implements PMP interoperate with one that does not? (A simple rule might be to send all traffic from a network that does not use PMP on the lowest priority subnetwork, but other rules could be more appropriate.) How would revenues be split among different service providers? Also, one would need to provide facilities for either the sender or the receiver to pay for the transmission, a problem that also occurs in other schemes. Both these problems have been considered in the literature for other pricing schemes. How frequently would the capacities and prices of different subnetworks in PMP vary? (In particular, should there be off-peak discounts, given that the Internet is a global network, and peak hours might occur at different times in different regions?)

The remainder of this section concentrates on a few aspects of PMP. One crucial problem is how to set prices and capacities of the separate networks. This is a difficult problem in general. However, it should not be too difficult to get nearly optimal solutions. Aside from relying on customer surveys and user complaints, one could obtain the necessary data from time of day variations in traffic patterns. I suggest that prices and capacities of the networks should stay constant for extended periods, to provide the predictability of price and service quality that consumers like. (However, one might allow for some time of day price variations, such as the evening discount on long distance phone calls). Since consumers could choose for each packet the network to send it on, I expect that some would go by some general expectation of quality of service for different networks, while others would hunt (using software on their computers) for the cheapest way to satisfy their requirements. The latter class would serve a role similar to that of speculators in commodity markets, who provide liquidity. The natural variation in total demand for transmission with time of day would lead these users to shift their demand among different channels. This should allow network operators to deduce what the distribution of consumer demands and valuations is.

For the PMP proposal to work, the performance of the different networks has to be predictable, at least on average. Unfortunately, the fractal nature of data traffic [LelandTWW] means that we have to expect that all PMP channels will experience sporadic congestion. All we can expect is that the higher-priced channels will experience this service degradation less frequently. This could lead to network instability, with degradation on one channel propagating to other channels. For example, an extended congestion episode on the lowest-priced channel might lead a large fraction of users of that channel to decide to pay extra and send their packets to the higher-priced networks, which would then become intolerably congested. There are several ways to overcome this problem (should it turn out to be a serious one). One is by modifying the charging mechanism. Access to the premium channels might be not on a packet-by-packet basis, but instead the user would pay for the right to send 1,000 packets on that channel in the next second. This would increase the financial barrier to upgrading channels.

Another way to lessen the instability problem is to promote segregation of different types of services on different networks. For example, the lowest-priced network (where the price per packet might be zero, as mentioned before) could have artificial delays and packet losses induced by the network operators, to make it unusable for videoconferencing, say. (For example, the capacity of the lowest-priced channel could be lowered in slack times by requiring that packets in that channel spend some time in the buffer before being transmitted.) This would be analogous to the policies of various companies. For example, Federal Express has next-day delivery and "next-day-by-10am" delivery. Regular next-day delivery packages that are available for delivery at 10 am are not delivered then, but in a separate trip in the afternoon. This type of approach, referred to as "damaged goods," has been studied by Deneckere and McAfee [DeneckereM], who show that it is common in high-tech industries, and that it often serves to promote social welfare. (This approach appears to be especially suited for trade in information goods. See [Odlyzko, Varian2].) Methods of this type could be used to induce a more even load on the separate networks, and thus compensate for some of the potential difficulties.

4. PMP implementation

The PMP proposal can be regarded as a logical development of some current trends. Various Internet service providers (ISPs) are planning to distinguish their networks through higher quality of service (QoS), and plan to charge extra for that. Customers with connections to several ISPs would then have a choice similar to that in PMP. S. Keshav has pointed out that MCI is planning a network for business customers that would be physically separate from MCI's regular network for individuals. MCI customers who sign up for both networks will then have a limited version of PMP available to them. The PMP proposal would simply let each ISP offer its customers an array of choices that they might have available through different ISPs anyway, and should therefore be more efficient.

PMP would be easy to introduce. It would not be necessary to wait for the deployment of IPv6 [Huitema] or other protocols. The current IPv4 packets already have a 3-bit priority field that is unused. (It was used for only a brief period a decade ago [BohnBCW, Bailey].) Since the number of networks in PMP is likely not to exceed 4, this is more than sufficient. Interoperability would be easy, as all packets that do not contain any bits indicating class of service could be sent on the lowest cost (and lowest priority) network.

At least initially, the cost per packet on the lowest cost network would undoubtedly be zero. There are strong arguments (see Section 5) for usage-sensitive prices even on this network, but zero prices would make this network look like the current Internet, and so make the transition easier. It might also be possible to have zero prices on this network in the long run during slack periods.

Eventually applications, such as videoconferencing software, would be rewritten to give users the choice of network (and thus of quality of their transmission channel) from within each application. Since that would take time, initially one would need to write "wrapper" software that would handle all IP traffic on a user's machine and set the priority bits to the level specified by the user. Network administrators would have a chance to police users' behavior at the firewall. For example, a university might reset priorities of packets coming from students' computers to that of the lowest class.

Inside the network, changes would only have to be done in the router software. It would be necessary to maintain logically separate queues or to give appropriate priority to packets from different channels.

The major change required in a network by PMP is the same one as that needed for any usage-sensitive pricing scheme. It would be necessary to install hardware or software to count the packets and bytes for each user. Essentially all of this accounting could be done at the edges of the network, although there would probably have to be some measurement at the inter-ISP gateways. This task could be simplified by using sampling. Unlike some other pricing schemes, PMP would not require any detailed accounting or pricing decision to be made in the core of the network, where speed of operations is the greatest requirement, and so simplicity is desirable.

There is often a chicken and egg problem with introduction of new network services. They require users to justify introducing the service, but there are no users until the service is widely deployed. PMP could be implemented within a single ISP initially, and used to provide substitutes for private line and Frame Relay services for large organizations that have several facilities in areas covered by that ISP.

As with most other pricing schemes, there are still areas requiring further research. For example, how should one charge for multicasting? (Cf. [HerzogSE].) It would also be necessary to arrange for 800-like services, in which the receiver pays. These have already been considered in the literature, and the authenticated transactions required for them can also be carried out just by the service providers at the edges of the network.

5. The irresistible force runs into the immovable object

The need for usage-sensitive pricing appears to be irresistible. It has impeccable economic logic as well as increasing practical evidence behind it. Unfortunately, it collides with users' unshakeable preference for flat-rate pricing. The problem is how to reconcile the two.

The case for usage-sensitive pricing of the Internet has been ably made many times already, for example in [Clark2, GuptaSW4, MacKieMV1, MacKieMV2, Shenker1, Shenker2, ShenkerCEH]. The basic problem is that the demand for transmission capacity appears to be practically unlimited, especially as high bandwidth services are developed. This guarantees a continuation of the nearly constant congestion we are experiencing right now. This congestion will make many novel services, such as teleconferencing, impossible, as data transfers that are insensitive to delay continue to crowd out all other traffic.

The basic argument in favor of usage-sensitive charges is magnified by the many incentives for Internet users to behave in ways detrimental to other users, an example of the "tragedy of the commons." When America Online switched to flat-rate pricing at the end of 1996, its system could not cope with increased demand. As it became harder for users to get a new connection, they started leaving their connections open even when they were not doing anything, seriously aggravating the problem. For America Online, the problem was a shortage of modems, which would not have been alleviated by charging for packets sent. However, similar perverse incentives exist on the Internet to increase data transfers. To get better performance from the "World Wide Wait" while Web surfing, tools such as PeakJet use the time that a user spends looking at a Web page to download all pages linked to it, so that if the user decides to read one of them, it can be fetched quickly from a local disk. The worse the congestion, the greater the incentive for individuals to employ such tools, and many servers have experienced overloads as a result. Similarly, there is an increasing temptation to use systems such as WebWhacker. This program can spend a whole night downloading Web pages to the hard disk of a PC, just in case that PC's owner wants to spend a few minutes looking at a small selection of those pages the next day. While so far only local congestion problems have been documented that are caused by PeakJet, WebWhacker, and similar systems, their widespread use would overwhelm the current Internet. Most computers on average use only a small fraction of the capacity of their link to the Internet. PeakJet and WebWhacker exploit the full available bandwidth. It would take fewer than 200,000 PCs (under 1of all networked PCs) connected at 28.8 Kbs to saturate the current Internet (which is estimated to have a capacity of about 5 Gbs) if all were downloading Web pages at a steady 28.8 Kbs rate. There is nothing wrong with PeakJet and WebWhacker per se, as they can be useful, especially when the user has urgent tasks and the local connection to the Internet is slow. The main problem with such tools is that with flat-rate pricing, they create incentives for users to rely on them indiscriminately, even when the benefit to those users is minor.

In general, the survival of the Internet owes much to altruism and ignorance. There are all too many incentives for users to abuse the system. These incentives are growing, and as the user population becomes increasingly heterogeneous, less inclined to cooperate. The official TCP standard requires transmissions to slow down when congestion occurs, and the Internet would collapse without this feature. However, there are many faulty implementations of TCP that are already deployed, and if they were used more widely, the Internet would almost surely suffer congestion collapses [Paxson1]. Further, there is no effective method to prevent the creation and use of rogue versions of TCP, which would speed up transmission of packets when they encounter delays or packet losses. Such versions would provide better service to their users (as long as not too many others follow the same strategy and cause a collapse), and in a flat-rate pricing environment would not cost those users anything extra.

Even without unintentionally defective or rogue implementations of TCP, the Internet is already threatened by the growth of services that use protocols such as UDP, which do not slow down transmission in the presence of congestion [BradenFM]. Usage-sensitive pricing could provide incentives for cooperative behavior. If every packet incurred a charge, sending two copies of a packet to cope with network losses would double the cost of the transmission, and induce marginal users to postpone or abandon their transmissions.

Usage-sensitive pricing would also play a useful role in providing incentives to service providers to build adequate capacity in the core of the Internet. It is estimated that ISPs currently spend only about a third of their budget buying bandwidth. Gains in market share appear to be the highest priority, and providing good connectivity to existing customers is secondary. This is only to be expected with the current flat-rate scheme, since revenues depend only on the number of users.

Consumer usage as well as satisfaction with good or services depend in large part on their subjective reactions to pricing schemes (cf. [Brittan]). In particular, while the arguments for usage-sensitive pricing seem to be irresistible, they run into users' seemingly immovable preference for flat rates. This preference has attracted considerable attention recently, especially when America Online was forced to offer such a plan. However, there are many earlier examples in the online world, as when services such as Prodigy and CompuServe were forced to stop charging for individual email messages. Large organizations also show a strong preference for flat rates. The introduction by the U. S. Defense Data Network of usage-sensitive pricing resulted in the different branches of the U. S. armed forces building their own networks [Bailey]. This preference for flat rates is not unique to data networking. It is a general phenomenon that was probably first explored and documented in the context of pricing of local telephone calls in the Bell System in the 1970s [CosgroveL]. In practice, what it means is that consumers are willing to pay more for a flat-rate plan than they would under a per-user pricing scheme. This preference is being exploited by various businesses, to the extent that there is even a utility that offers an annual supply of natural gas for heating for a flat fee. (The fee is based on the previous year's usage, with surcharges or refunds if consumption deviates by more than 20% from the expected level.) As was already recognized in [CosgroveL], there appear to be three main reasons for the preference for flat rates:

(i) Predictability: Users know ahead of time how much the service will cost, and do not have to worry about sudden large bills. (A recent study showed that a large fraction of those households in the United States that do not have telephone service could afford it, since they have cable TV and other services. However, they do not install phones since they are concerned about family and friends generating large bills [MuellerS].)

(ii) Overestimate of usage: Customers typically overestimate how much they use a service, with the ratio of their estimate to actual usage following a log-normal distribution.

(iii) Hassle factor: With per-use pricing, consumers keep worrying whether each call is worth the money it costs, and it has been observed that their usage goes down. Charges for local calls in the United States had the effect of shortening the lengths of calls, even when the charges were on a per-call basis.

Flat rates are preferred by consumers, but they also have major advantages for service providers. They were already advocated for broadband services by Anania and Solomon in [AnaniaS], a paper that was first presented almost a decade ago. On the Internet, they eliminate the need for a traffic measurement and charging infrastructure, which, even for a system such as PMP, where almost all the work would be done at the edges of the network, would be costly to implement. (Flat rates often have socially desirable effects, as well. In pricing of household garbage disposal, they decrease dumping of garbage, for example [FullertonK].)

Flat rate pricing often allows service providers to collect more revenue. This is often true even when the user preferences mentioned above (which are hard to incorporate into conventional utility maximization arguments) are ignored. In general, flat-rate (or subscription) pricing is likely to be dominant in sales of information goods [BB, FishburnOS, Odlyzko, Varian1]. The conventional economic utility maximization arguments show that the advantages of bundling strategies (selling combinations of goods for a single price) increase as marginal costs decrease (cf. [BakosB]). Even sales of software are likely to be more profitable in the conventional arrangement of a fixed fee for unlimited use than on a per-use basis [FishburnOS]. However, all those predictions are for goods and services with negligible marginal costs. Moreover, there are often positive network externalities that strengthen the case for subscription or site licensing plans. For example, a software producer benefits from users recruiting other users, generating enhancements to the basic package, and so on.

While there are strong arguments, such as those mentioned above, that flat-rate pricing will be increasing as electronic commerce grows, those arguments have limited applicability to data network pricing. What makes flat-rate local calling plans feasible is that making a call requires time from the consumer. As a result, most households make only about 5 local calls per day (of about 4 minutes each). There would be no monetary cost for making more, but there would be a cost in time. Similarly, demand for natural gas for heating does not vary too much with price, since for most homes, the price of gas is not a huge part of the budget, and maintaining a temperature of 90 degrees Fahrenheit appeals to few people. Internet access is different, since the usable bandwidth is growing, and computers can keep a link occupied even in the absence of human intervention. Marginal costs are still negligible. However, there are substantial negative network externalities. If Alice uses a word processor, that does not preclude anyone else using it, and her usage is likely to benefit others, who will be able to share files with her and consult her about bugs. When she sends a packet, though, it can only impede other users' transmissions. Therefore usage-sensitive pricing does seem to be necessary. The problem is how to make it palatable to users.

Consumers have long accepted a variety of usage-sensitive rates. In the United States, long distance phone calls have largely been paid for on a per-use basis, and in most of the rest of the world even local calls have traditionally incurred charges. While there is a tendency towards flat rates in general, as marginal costs diminish and it becomes easier to satisfy consumer preferences, this trend is not universal. For example, Federal Express and United Parcel Service are moving towards charging for delivery of express mail according to distance, instead of using a flat fee. Even in Internet transmissions, there have been many instances of charging for the amount of transmitted data [Brownlee, OECD]. Such usage-sensitive pricing appears to be spreading for large customers in the U.S., with UUnet, MCI, and other carriers offering them. Many ISPs have declared that they intend to move away from flat rate pricing for individuals. It seems it should be possible to persuade users to accept usage-sensitive pricing, especially if the benefits are made clear. PMP should make the transition easier than with most other schemes, since the lowest-priced channel could be offered initially at zero cost per packet, and would thus behave just like today's Internet.

In PMP, the preference for flat-rate pricing can be partially accommodated by selling large blocks of transmission capacity (giving the user the right to send or receive 100 MB of data over a week through the lowest priced channel, or 60 MB through the next most expensive channel, say). Such pricing has worked well in long distance telephony in the United States, with consumers typically paying for more capacity than they used [MitchellV].

PMP offers a simple pricing plan with constant and easily understood pricing, which is an advantage, as it fits consumer desires. It does not offer any service guarantees, however. Such guarantees are popular. L. L. Bean has developed an enviable reputation, partially as a result of its no-questions-asked return policy. Cable TV companies are trying to improve their notoriously bad customer relations by offering days of free service when interruptions occur. Marketing of telecommunications services to large corporate users also increasingly relies on guarantees of features such as availability and data delivery delays. However, few guarantees are absolute, and most purchases are made on the basis of expectations. The restaurant meals and books we buy, the movies we go to, even the clothes we purchase after trying them on in a store, all involve large elements of uncertainty about the quality we experience. When we subscribe to a newspaper or a magazine, neither we nor the editors know in advance precisely what we will get. Expectations, based on our own experience, word of mouth recommendations, and other sources, is what we rely on. Moreover, consumers are willing to accept occasional large deviations from the expected quality of service. An airplane passenger in first class may have an uncomfortable trip, if there is a sick and crying child in the seat behind. On the other hand, a coach passenger may have three seats to herself, enough to stretch out and get a good night's sleep on a trans-oceanic flight, and have a much better experience than those in first class. On average, though, a first class ticket does provide superior service, and that is enough to maintain a huge price differential. It seems likely that consumers could accept the lack of guarantees of QoS in PMP, especially if the average quality of different channels were predictable enough.

Consumer and business behavior is often hard to fit into the standard economic framework. A puzzle of modern economics is the reluctance of businesses to use price overtly as a method of rationing popular goods or services. With some minor exceptions, ski-lift ticket prices do not depend on the quality of the snow, nor on whether it is the peak vacation season. Opera tickets usually do not depend on who the lead singers are, and admission prices to first-run movies do not depend on the length of ticket lines. For some reason, free enterprise companies prefer the socialist method of rationing by queue to that of rationing by price. This appears to reflect a general public aversion to the auction mechanism. During the oil crises of the 1970s, bizarre gasoline rationing rules that were (correctly) derided by economists as ineffective and inefficient were popular with the public. Laws against ticket scalping are common, and are widely supported. Yet, to most economists, scalpers fulfill a socially useful role of getting tickets into the hands of those who are willing to pay the most for them. The main puzzle for most economists in this area seems to be that scalpers can make a living. Why don't theaters and sports arenas simply adjust ticket prices to clear the market and appropriate to themselves some of the gain that the public or the scalpers obtain? However, that is simply not done, except in unusual circumstances. There have been attempts to explain this phenomenon using conventional economic utility maximization arguments (cf. [BarroR]), but they are not entirely convincing. It seems likely that the cause lies more in the realm of consumers' seemingly irrational economic behavior, whose study was pioneered by Kahneman and Tversky. The challenge is to design pricing schemes that approach the goal of efficiency that can be achieved by auction mechanisms, and yet do respect consumer aversion to the auction.

A particularly important role in consumer behavior in the economic and political arenas is played by the notion of fairness [Odlyzko, Zajac]. Fairness is likely to play an increasing role in electronic commerce. Decreasing marginal costs are increasing the incentives for sellers to impose artificial barriers, and at the same time the nature of electronic commerce makes it much more apparent to consumers that the barriers are artificial. Therefore it will be increasingly important to convince consumers of the fairness of pricing schemes. In the design of PMP, assigning fixed capacity to different subnetworks is likely to appeal to consumers more than some of the priority schemes mentioned in Section 2. It avoids the appearance of an auction, in which users willing to pay higher prices hog all the bandwidth. It also throws the onus for congestion on other users, and not on the network provider, which again seems to be more palatable.

6. Other pricing proposals

Several proposals have been made for usage-sensitive pricing. Extensive information can be found on the Web site [Varian0] and in the collection of paper edited by McKnight and Bailey (of which the reference [AnaniaS] below is one). Further references, short summaries, and criticisms can be found in [Clark1, Shenker1, ShenkerCEH]. Here I only make a few remarks on the main features of some of these proposals, and how they compare to PMP.

Among the earliest and most influential pricing proposals is that of MacKie-Mason and

Varian [MacKieMV1, MacKieMV2]. (A preprint with their scheme had circulated much earlier. For extensions of their work, see also [LehrW].) They propose imposing charges on packets when those packets contribute to congestion. In particular, charges would be zero when the network is not fully utilized. Their Vickrey auction mechanism has some desirable properties. However, as is pointed out in [Clark1, ShenkerCEH], for example, it requires complicated systems to conduct an auction among individual packets (which, moreover, would be most involved in the core of the network, where simplicity is of highest value to obtain high speed of operation). In addition, this proposal does not deal with the problem that delay or loss of an individual packet at a single node is not a good measure of network performance for most applications. Further, since a packet typically goes over a dozen or more routers, in the absence of global information about all routers on the path, how could the user decide how much to bid to get through the first router on the path? Finally, in terms of meeting customer preferences, the MacKie-Mason and Varian proposal is likely to be unsatisfactory, since it is impossible to predict how much it will cost to transmit any single packet.

The Gupta et al. proposal [GuptaSW1, GuptaSW2, GuptaSW3, GuptaSW4] is (oversimplifying a lot) to have a set of service classes and priorities. As is pointed out in [Clark1, ShenkerCEH], there are problems with this approach, among them that low priority classes could fail to get any bandwidth at all if enough traffic from higher priority classes show up. The scheme also has substantial overhead. It requires collecting and processing extensive information about the network.

The schemes that are closest to PMP are those advocated by Clark [Clark1, Clark2] and Shenker et al. [ShenkerCEH]. These authors point out that quality as perceived by consumers is not just a matter of minimizing packet delays or losses, but depends on the application, and is hard to quantify. It is also highly unlikely that an optimal policy can be found that would deal with the varied requirements of a heterogeneous user population and many different services. Those authors argue for edge pricing (i.e., charging at the entrance and exit from the network, not based on what happens at internal nodes, as is required by the MacKie-Mason and Varian proposal), which is a feature of PMP. They also argue for at least some variant of Clark's proposal of charging for expected usage, with those portions of a consumer's offered load that deviate from negotiated statistics being treated at lower priorities. (This part is similar to another proposal of Kelly [Kelly].) The problem with charging for negotiated usage profiles, just as with applications of future markets to networks, is that they do not deal with the inevitable short-term fluctuations in traffic. It is desirable to provide incentives for users to either lower their load on the network or else switch to a higher-priced network when congestion occurs.

Feng, Kandlur, Saha, and Shin [FengKSS1, FengKSS2, FengKSS3] have proposed implementing services such as controlled-load and guaranteed service (cf. [BradenCS]) without end-to-end network coordination. They use adaptive packet marking with two classes, with higher priority packets treated preferentially at the routers, to provide soft guarantees of QoS. They are not concerned with pricing as such, but appear to assume a variant of Clark's scheme of charging for expected usage. In many ways their proposal is similar to PMP in lack of hard QoS guarantees and having separate classes of packets. However, they assume more intelligence in the network (changing marking of packets, for example) and have just two classes of packets. Their main concern is with modifying TCP to accomplish their goals.

Acknowledgements: I thank Jerry Ash, Vijay Bhagavath, Steve Bellovin, Kim Claffy, Kerry Coffman, John Denker, Nick Duffield, Bruce Emerson, Anja Feldmann, Philippe Flajolet, John Friedman, Paul Ginsparg, Albert Greenberg, Paul Henry, Andrew Hume, Chuck Kalmanek, S. Keshav, Chuck McCallum, Nick Maxemchuk, Rodolfo Milito, Deborah Mills-Scofield, Gerry Ramage, Jennifer Rexford, Paul Resnick, Don Towsley, Greg Wetzel, Walter Willinger, and Pat Wirth for comments on an earlier draft or providing useful information.

References

[AnaniaS] L. Anania and R. J. Solomon, Flat–the minimalist price, pp. 91-118 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Bailey] J. Bailey, Internet economics, available at ⟨http://far.mit.edu/Pubs/inet_econ/abstract.html⟩.

[BakosB] Y. Bakos and E. Brynjolfsson, Aggregation and disaggregation of information goods: Implications for bundling, site licensing and micropayment systems, in *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property,* D. Hurley, B. Kahin, and H. Varian, eds., MIT Press (1997). To appear. Available at ⟨http://www.gsm.uci.edu/∼bakos⟩.

[BarroR] R. J. Barro and P. M. Romer, Ski-lift pricing, with applications to labor and other markets, *Am. Econ. Rev.* 77 (1987), 875-90.

[BohnBCW] R. Bohn, H.-W. Braun, K. C. Claffy, and S. Wolff, Mitigating the coming Internet crunch: multiple service levels via Precedence, March 22, 1994 report, available at ⟨ftp://ftp.sdsc.edu/pub/sdsc/anr/papers/precedence.ps.Z⟩.

[BradenCS] R. Braden, D. Clark, and S. Shenker, Integrated services in the Internet architecture: an overview, RFC1633, available at ⟨ftp://ds.internic.net⟩.

[BradenFM] B. Braden, S. Floyd, and G. Minshall, White paper on mechanisms for unresponsive traffic, available at ⟨ftp://ftp.ee.lbl.gov/floyd/NGI97.txt⟩.

[Brittan] D. Brittan, Spending more and enjoying it less?, *Tech. Rev.* 100, no. 5 (July 1997), pp. 11-12. Available at ⟨http://web.mit.edu/afs/athena/org/t/techreview/www/articles/july97/brittan.html⟩.

[Brownlee] N. Brownlee, Internet pricing in practice, pp. 77-90 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Clark1] D. D. Clark, Adding service discrimination to the Internet, *Telecommunications Policy,* 20 (1996), 169-181.

[Clark2] D. D. Clark, Internet cost allocation and pricing, pp. 215-252 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[CocchiSEZ] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, Pricing in computer networks: motivation, formulation, and example, draft of Nov. 18, 1993, available at ⟨ftp://parcftp.xerox.com/pub/net-research/pricing2.ps.Z⟩.

[CosgroveL] J. G. Cosgrove and P. B. Linhart, Customer choices under local measured telephone service, *Public Utilities Fortnightly,* August 30, 1979, 27-31.

[Crawford] D. W. Crawford, Internet services: A market for bandwidth or communication, pp. 379-400 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[CrawleyNRS] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, A Framework for QoS-based routing in the Internet, draft of March 21, 1966, available at ⟨ftp://ietf.org/internet-drafts/draft-ietf-qosr-framework-00.txt⟩.

[DanielsonW] K. Danielson and M. Weiss, User control modes and IP allocation, pp. 305-322 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[DeneckereM] R. J. Deneckere and R. P. McAfee, Damaged goods, *J. Economics and Management Strategy,* 5, no. 2 (1966), 149-174.

[EdellMV] R. J. Edell, N. McKeown, and P. P. Varaiya, Billing users and pricing for TCP, *IEEE J. Selected Areas Comm.,* 13 (1995), 1162-1175.

[FengKSS1] W.-C. Feng, D. D. Kandlur, D. Saha, and K. G. Shin, TCP enhancements for an Integrated Services Internet, available at ⟨http://www.eecs.umich.edu/~wuchang/index2.html⟩.

[FengKSS2] W.-C. Feng, D. D. Kandlur, D. Saha, and K. G. Shin, Understanding TCP dymanics in an Integrated Services Internet, available at ⟨http://www.eecs.umich.edu/~wuchang/index2.html⟩.

[FengKSS3] W.-C. Feng, D. D. Kandlur, D. Saha, and K. G. Shin, daptive packet marking for providing differentiated services in the Internet, available at ⟨http://www.eecs.umich.edu/~wuchang/index2.html⟩.

[FishburnOS] P. C. Fishburn, A. M. Odlyzko, and R. C. Siders, Fixed fee versus unit pricing for information goods: competition, equilibria, and price wars, *First Monday,* vol. 2, no. 7 (July 1997), ⟨http://www.firstmonday.dk/⟩. Also to appear in *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property,* D. Hurley, B. Kahin, and H. Varian, eds., MIT Press. Available at ⟨http://www.research.att.com/~amo⟩.

[FloydJ] S. Floyd and V. Jacobson, Link-sharing and resource management models for packet networks, *IEEE/ACM Trans. Networking,* 3 (1995), 365-386. Available at ⟨ftp://ftp.ee.lbl.gov/papers/link.ps.Z⟩.

[FullertonK] D. Fullerton and T. Kinnaman, Household responses to pricing garbage by the bag, *Am. Econ. Rev.* 86, no. 4 (Sept. 1996), 971-984.

[GongS] J. Gong and P. Srinagesh, The economics of layered networks, pp. 63-76 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in J. Electronic Publishing, special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[GuptaSW1] A. Gupta, D. O. Stahl, and A. B. Whinston, Pricing of services on the Internet, available at ⟨http://cism.bus.utexas.edu/res/wp.html⟩.

[GuptaSW2] A. Gupta, D. O. Stahl, and A. B. Whinston, Priority pricing of integrated services networks, pp. 323-352 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[GuptaSW3] A. Gupta, D. O. Stahl, and A. B. Whinston, A stochastic equilibrium model of Internet pricing, available at ⟨http://cism.bus.utexas.edu/res/wp.html⟩.

[GuptaSW4] A. Gupta, D. O. Stahl, and A. B. Whinston, The Internet: A future tragedy of the commons?, available at ⟨http://cism.bus.utexas.edu/res/wp.html⟩.

[HerzogSE] S. Herzog, S. Shenker, and D. Estrin, Sharing multicast costs, pp. 169-212 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Huitema] C. Huitema, *IPv6: The New Internet Protocol,* Prentice Hall, 1996.

[Ipsilon] Ipsilon IP switching applications, available at ⟨http://www.ipsilon.com/products/applications.htm⟩.

[JordanJ] S. Jordan and H. Jiang, Connection establishment in high-speed networks, *IEEE J. Selected Areas Comm.,* 13 (1995), 1150-1161.

[Kelly] F. P. Kelly, Charging and accounting for bursty connections, pp. 253-278 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in J. Electronic Publishing, special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Keshav] S. Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network,* Addison-Wesley, 1997.

[LehrW] W. H. Lehr and M. B. H. Weiss, The political economy of congestion charges and settlements in packet networks, *Telecommunications Policy,* 20 (1996), 219-231.

[LelandTWW] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Trans. Networking* 2 (1994), 1-15.

[MacKieMMM] J. K. MacKie-Mason, L. Murphy, and J. Murphy, The role of responsive pricing in the Internet, pp. 279-304 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in J. Electronic Publishing, special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[MacKieMV1] J. K. MacKie-Mason and H. R. Varian, Pricing the Internet, in *Public Access to the Internet,* B. Kahin and J. Keller, eds., MIT Press, 1995, pp. 269-314. Available at ⟨http://www.sims.berkeley.edu/∼hal/people/hal/papers.html⟩.

[MacKieMV2] J. K. MacKie-Mason and H. R. Varian, Pricing congestible network resources, *IEEE J. Selected Areas Comm.,* 13 (1995), 1141-1149. Available at ⟨http://www.sims.berkeley.edu/∼hal/people/hal/papers.html⟩.

[Metcalfe] B. Metcalfe, NetNow's statistics trigger defensive responses from some corners of the 'net, *InfoWorld,* Feb. 3, 1997. Available at ⟨http://www.infoworld.com/cgi-bin/displayNew.pl?/metcalfe/bm020397.htm⟩.

[MitchellV] B. M. Mitchell and I. Vogelsang, *Telecommunications Pricing: Theory and Practice,* Cambridge Univ. Press, 1991.

[MonkC] T. Monk and K. C. Claffy, A survey of Internet statistics / metrics activities. Available at ⟨http://www.tomco.net/~tmonk/metrics.htm⟩.

[MuellerS] M. L. Mueller and J. R. Schement, Universal service from the bottom up: A study of telephone penetration in Camden, New Jersey, *The Information Society* 12, no. 3 (1996), 273-292.

[NLANR] National Laboratory for Applied Network Research, ⟨http://www.nlanr.net/⟩.

[Odlyzko] A. M. Odlyzko, The bumpy road of electronic commerce, in *WebNet 96 - World Conf. Web Soc. Proc.,* H. Maurer, ed., AACE, 1996, pp. 378-389. Available at ⟨http://www.research.att.com/~amo⟩.

[OECD] OECD, Information infrastructure convergence and pricing: The Internet, report available at ⟨http://www.oecd.org/dsti/gd_docs/s96_xxe.html⟩.

[Paxson1] V. Paxson, Automated packet trace analysis of TCP implementations, *Proc. SIGCOMM '97,* to be published. See also [Paxson3].

[Paxson2] V. Paxson, End-to-end Internet packet dynamics, *Proc. SIGCOMM '97,* to be published. See also [Paxson3].

[Paxson3] V. Paxson, *Measurements and Dynamics of End-to-End Internet Dynamics,* Ph.D. thesis, Computer Science Division, Univ. Calif. Berkeley, April 1997. Available at ⟨ftp://ftp.ee.lbl.gov/papers/vp-thesis/⟩.

[Shenker1] S. Shenker, Service models and pricing policies for an integrated services Internet, in *Public Access to the Internet,* B. Kahin and J. Keller, eds., MIT Press, 1995, pp. 315-337.

[Shenker2] S. Shenker, Fundamental design issues for the future Internet, *IEEE J. Selected Areas Comm.,* 13 (1995), 1176-1188.

[ShenkerCEH] S. Shenker, D. Clark, D. Estrin, and S. Herzog, Pricing in computer networks: reshaping the research agenda, *Telecommunications Policy,* 20 (1996), 183-201.

[Steinberg] S. G. Steinberg, Netheads vs. Bellheads, *Wired,* 4, no. 10 (Oct. 1996), pp. 144-147, 206-213. Available at ⟨http://wwww.wired.com/wired/4.10/features/atm.html⟩.

[Srinagesh] P. Srinagesh, Internet cost structures and interconnection agreements, pp. 121-154 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Swift] J. Swift, *A Modest Proposal for Preventing the Children of Poor People in Ireland from Being A Burden to their Parents or Country, and for Making them Beneficial to the Public,* 1729.

[Varian0] H. R. Varian, The economics of the Internet, information goods, intellectual property and related issues, reference Web pages with links, ⟨http://www.sims.berkeley.edu/resources/infoecon/⟩.

[Varian1] H. R. Varian, Pricing information goods, available at ⟨http://www.sims.berkeley.edu/~hal/people/hal/papers.html⟩.

[Varian2] H. R. Varian, Versioning information goods, available at ⟨http://www.sims.berkeley.edu/~hal/people/hal/papers.html⟩.

[WangSP] Q. Wang, J. M. Peha, and M. A. Sirbu, Optimal pricing for integrated services networks, pp. 353-376 in *Internet Economics,* L. W. McKnight and J. P. Bailey, eds., MIT Press, 1997. Preliminary version in *J. Electronic Publishing,* special issue on Internet economics, ⟨http://www.press.umich.edu/jep/⟩.

[Wilson] R. Wilson, Efficient and competitive rationing, *Econometrica* 57 (1989), pp. 1-40.

[YajnikKT] M. Yajnik, J. Kurose and D. Towsley, Packet loss correlation in the MBone multicast network, to appear in *Proc. IEEE Global Internet Conf.* (London, Nov. 1996). Available at ⟨http://www-net.cs.umass.edu/mcast.html⟩.

[Zajac] E. E. Zajac, *Political Economy of Fairness,* MIT Press, 1995.

Appendix 1. Gains from network segmentation

Various aspects of PMP require additional study and modeling. Here we consider only some simple models of the gains that can be obtained by having logically separate networks that operate at different utilization levels. These models are crude and are not specific to PMP. Any other scheme that exploits the economies of scale of aggregating traffic with different utilization levels would provide comparable benefits in this model. For an example of other types of economic models dealing with pricing in data networks, see [CocchiSEZ], for example. Still, even these models may shed some light on how benefits of better data networks would be divided.

We will assume that there are two types of demands for data transport. Users (generally processes, and not individuals) will be assumed to fall into types A and B. Type A users might correspond to bulk file transfers that are not sensitive to delays. We will assume that when the price is x (per byte, say), type A users will wish to send

$$ax^{-1}e^{-x} \tag{A1}$$

bytes (per day, say). They will then generate network revenues of

$$ae^{-x} \ . \tag{A2}$$

This is an unconventional model, but might not be unreasonable for data traffic, with total demand limited primarily by general budget constraints at low prices. Note that historically, prices of data transmission have been dropping, but total spending has been climbing. We will assume that the cost (the ongoing operational cost, as well as depreciation and profit, which will be assumed to be limited by competition) of operating a network that carries w bytes is

$$cw^{3/4} \tag{A3}$$

for some constant $c > 0$. This is a conservative assumption, since it corresponds to less than a 16% reduction in costs when the network doubles in size ($2^{3/4} = 1.68179\ldots$). The economies of scale faced by a single ISP that moves from purchasing T1 lines to T3 lines or the learning curve experience faced by the network equipment manufacturers justify assumptions of even higher reductions in costs, which correspond to exponents even lower than the 3/4 assumed above.

With the above assumptions, if there are only type A users, we expect the cost of the network to equal the revenues, so that

$$ae^{-x} = c(ax^{-1}e^{-x})^{3/4} \ , \tag{A4}$$

which is equivalent to

$$x^3e^{-x} = a^{-1}c^4 \ . \tag{A5}$$

The unique maximum of x^3e^{-x} occurs at $x = 3$ and equals $27e^{-3} = 1.344250\ldots$. Hence for combinations of a and c with $c^4 > 27ae^{-3}$, (i.e., high costs of network compared to demand), there is no price x that will recover costs, and so the network will not be built. For $c^4 < 27ae^{-3}$, there will be two solutions for x, and it is the smaller one, call it x_A, that will be preferred, since it corresponds to higher revenue and higher traffic.

Suppose that there are also type B users, who will only use a network when its utilization rate is at most half of that acceptable to type A users. (This is a pessimistic assumption, since it seems likely that much smaller reductions in network loads would suffice to produce substantial improvements in service.) Suppose that at price x, they will generate traffic of

$$bx^{-1}e^{-x} \ . \tag{A6}$$

Constructing a separate network for these users will cost

$$c(2bx^{-1}e^{-x})^{3/4} \tag{A7}$$

(the 2 coming from lower utilization rate), and bring revenues of

$$be^{-x} \ . \tag{A8}$$

Thus in this case the price x that equalizes revenue and cost is a solution to

$$x^3e^{-x} = 8b^{-1}c^4 \tag{A9}$$

(provided it exists, which happens when $27b \geq 8c^4e^3$). We will use x_B to denote the minimal solution to (A9).

Suppose a single network with a single price were to be built for both type A and type B users. Then its average utilization would have to be half that of a network meant for type A users alone, and so at price x would have revenue

$$(a+b)e^{-x} \tag{A10}$$

but cost

$$c(2(a+b)x^{-1}e^{-x})^{3/4} \ . \tag{A11}$$

Hence the price x that equalizes cost and revenue would have to satisfy

$$x^3e^{-x} = 8(a+b)^{-1}c^4 \ . \tag{A12}$$

We let x_{AB} denote the minimal solution to (A12) (when one exists, which happens precisely for $27(a+b) \geq 8c^4e^3$). We note that if $b > 7a$, so demand from type B users is large compared to that of type A users, type A users will benefit by having lower prices than if they had their own network, since $x_{AB} < x_A$. If b is small compared to a, though, then even if x_{AB} exists, x_{AB} will be larger than x_A, so type A users will be paying more than if they had their own network. They will also get better service, but the assumption is that they do not need it. (Note that type B users will always benefit from having type A users on their network, as prices will be lower, reflecting greater economies of scale.)

Suppose finally that we can have two networks for type A and type B users that are logically separate but physically part of the same network. We also assume that the provision of the logical separation imposes negligible additional costs. Then, if the price for type A users is set at y and those of type B at z, revenue will be

$$ae^{-y} + be^{-z} \tag{A13}$$

and the cost of the network will be

$$c(ay^{-1}e^{-y} + 2bz^{-1}e^{-z})^{3/4} \ . \tag{A14}$$

Prices y and z now need to satisfy

$$ae^{-y} + be^{-z} = c(ay^{-1}e^{-y} + 2bz^{-1}e^{-z})^{3/4} \ . \tag{A15}$$

Since we have two prices to select, we have more freedom of choice. By letting $y \to \infty$ or $z \to \infty$ we can reduce to networks that cater exclusively to type B and type A users, respectively. Intermediate choices are more interesting, though. We consider a few cases.

Example 1. $a = b = 3$, $c = 1$. We have $x_A = 0.9524456\ldots$, $x_{AB} = 2.784204\ldots$, while x_B does not exist. The network for type A users only produces traffic of $1.215175\ldots$, and revenues of $1.157389\ldots$ (in the arbitrary units we are using). A single network for type B and type A users would produce revenue of $0.3706693\ldots$ from traffic of $0.133132\ldots$, and so clearly would not be built, since both type A users and service providers would be much better off with a network just for type A users. On the other hand, consider a single physical network that has separate channels for the two types of users. Setting prices $y = 0.9$ and $z = 1.33865\ldots$ leads to total traffic of $1.942837\ldots$ (about 1.355 of type A and 0.587 of type B) and total revenues of $2.00630\ldots$, $1.2197\ldots$ from type A traffic and and $0.78659\ldots$ from type B traffic. Note that the gain to type A users from a network that accomodates type B users is relatively slight. The price they pay is reduced only by 5.5%. (The prices $y = 0.9$ and $z = 1.33865\ldots$ were selected to be close to those that maximize total revenue. Lowering the price y substantially below 0.9 quickly leads to declining revenues and soon after that there is no choice for z that will satisfy Eq. (A15).) The main benefit goes to type B users, who are offered a service they are want at a price they are willing to pay, and to network providers, whose revenue (and presumably profit) grows by 73%.

Example 2. $a = 20$, $b = 10$, $c = 1$. Then the optimal prices are $x_A = 0.424384\ldots$, $x_B = 1.56303\ldots$, and $x_{AB} = 0.85627\ldots$ for networks designed for type A traffic only, type B traffic only, and both types on the same network, respectively. We next consider a single physical network with logically separate networks for the two types of traffic. Total revenue is maximized with prices close to $y = 0.42$ and $z = 0.606846\ldots$. The traffic and revenue results of this choice for prices is shown in Table 1.

Table 1: Traffic on various networks in Example 2

network	traffic	revenue
A only	30.8293	13.0834
B only	1.3403	2.0950
$A + B$ on single network	14.8809	12.7422
$A + B$ on logically separate networks	40.2699	18.5916

As in Example 1, type A users experience a slight gain, while type B users find their price drops by a factor of 2.5 (compared to relying on a totally separate network just for their own traffic). Networks operators have a revenue gain of 22% (compared to running separate networks for the two types of users).

Example 3. $a = 10$, $b = 20$, $c = 1$. Then the optimal prices are $x_A = 0.55928\ldots$, $x_B = 1.04321\ldots$, and $x_{AB} = 0.85627\ldots$ for networks designed for type A traffic only, type B traffic only, and both types on the same network, respectively. A single physical network with logically separate networks for the two types of traffic and prices $y = 0.53$ and $z = 0.69381\ldots$ results in higher traffic and revenues, as is shown in Table 2. A revenue-maximizing network provider would be almost indifferent between having physically separate networks for the two types of users and a single one that gives all traffic the quality of service demanded by type B users. (Type B users would benefit from having a single network, type A users would lose from it.) However, a single physical network with logically separate channels would increase revenues by 24%.

Table 2: Traffic on various networks in Example 3

network	traffic	revenue
A only	10.2206	5.7162
B only	6.7545	7.0463
$A+B$ on single network	14.8809	12.7422
$A+B$ on logically separate networks	25.5093	15.8794

In all these examples, gains to type A users are small. This may help to explain why there has not been more pressure from users of the current Internet (whose applications almost by definition have to work reasonably well even in the presence of congestion) for higher quality of service.

In the three examples above, a and b are comparable, which means that the potential traffic from users of types A and B is assumed comparable. This might seem unrealistic, given that the bulk of current Internet traffic appears to be insensitive to congestion. However, the current distribution of traffic is unlikely to be typical of what would be seen if choices were offered. Much of Web surfing would surely move to higher-priced channels if those provided better quality of service. Furthermore, while the Internet is large and growing rapidly, it is still dwarfed by the private line and frame relay networks. Large fractions of the traffic from those networks could be diverted to the Internet if the latter could be improved.