Paris Metro Pricing for the Internet

Andrew Odlyzko
AT&T Labs - Research
Florham Park, NJ 07974
(973) 360-8410
amo@research.att.com

ABSTRACT

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

PMP is the simplest differentiated services solution. It is designed to accommodate user preferences at the cost of sacrificing some of the utilization efficiency of the network.

1 INTRODUCTION

The Internet currently provides only best-effort service that treats all packets equally. However, there is wide dissatisfaction with the perceived performance, and there appears to be a wide consensus that new applications, especially real time ones such as packet telephony, will require changing how the Internet operates. Various QoS (quality of service) techniques are being developed. (For a general survey and references, see [12].) They will provide differentiated service levels. Many of these 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. 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 channels. In the basic design, each would have a fixed fraction of the capacity of the entire network. (Many variations on this proposal are possible and are discussed in Section 2.) All channels would route packets using protocols similar to the current TCP and UDP, with each packet treated equally. The only difference between the channels would be that they would charge different prices. Customers would choose the channel 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 channels with higher prices would be less congested than those with lower prices, and thus provide better service.

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 entire 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 that 1st class tickets cost twice as much as 2nd class ones. (The Paris regional RER lines continued to operate on this basis until September 1, 1999, when 1st class cars were eliminated.) 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.

Pricing is a crude tool. Different applications vary in requirements for bandwidth, latency, and jitter, for example. PMP would not provide any specific 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. This has been convincingly demonstrated on experimental networks such as vBNS, as well as on many corporate networks.

PMP inverts the usual order in which networks are designed. Usually an attempt is made to determine the QoS required by various applications, then the network is designed to provide that QoS, and finally 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 users, what QoS they will receive. The expectation is that the different logical channels would usually have predictable performance and

would provide sufficient QoS variety to satisfy most needs.

The pricing mechanism of PMP is as simple as that of any usage sensitive pricing scheme that has been proposed for the Internet. 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 infrastructure being required to handle the traffic management tasks.

The goal in designing PMP was to come up with a differentiated services scheme that catered to customer preferences, even at the expense of efficiency in the operations of the network. Section 5 discusses what users like, and the reasons PMP appears a good compromise between their desires and the need for differentiated services and usage sensitive pricing. The arguments for PMP are thus drawn from marketing as well as conventional economic concerns.

At a high level, PMP is similar to diff-serv, perhaps the most popular of the QoS techniques being developed. The difference is that diff-serv does not by itself say anything about assignment of priorities and pricing. It treats only the technical aspect of how the network should deal with packets with different markings. PMP integrates pricing with traffic management.

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). The general consensus seems to be that this is not feasible, and that differentiated services are required to overcome the problem of "the tragedy of the commons," with rapid growth in traffic demand leading to endemic congestion. When I first proposed PMP [23], I shared this view, but based on knowledge of how many networks are operated, felt that one should strive for maximal simplicity even at the expense of maximal efficiency in use of transport capacity. A recent series of studies [9, 13, 25, 26, 27] have led me to question the basic assumptions that underlie the work on differentiated services. Most of the Internet is very lightly utilized, most of the problems are not caused by link or switch congestion (which is what most QoS measures address), and "the tragedy of the commons" is much less of a problem than is commonly believed. The main demand is for low transaction latency, not for transmission of many bits. It appears that in the backbones of the Internet, providing uniformly high quality of service to all transmissions might be not just feasible, but optimal, given the full cost that most QoS measures, even PMP, would impose. However, it is impossible to be certain this will be the case, since it is not clear how rapidly advances in transmission technology will translate into lower prices. If prices do not decline (and they have been rising in recent years), differentiated services might be required even in the backbones. In that case, though, the studies mentioned above argue that nothing more complicated than PMP should be implemented. The reason is that networking is already too complicated. The behavior that has been observed (such as many network managers knowing practically nothing about the traffic on their networks, traffic staying on established private line networks instead of much less expensive Frame Relay services, and so on) shows that network staff already have too much to do, and it is unrealistic for them to assign proper priorities to different transmissions, say. Thus I feel that the arguments for PMP among all the QoS techniques are much stronger than before, but that hopefully even PMP will not be necessary.

PMP may be useful on the edges of the Internet. The arguments outlined above apply only to the backbones, where fiber optic technology does offer the hope of rapidly increasing capacity at rapidly decreasing prices. There will always be situations (such as wireless links) where resource constraints are stringent enough that it will be necessary to impose stronger constraints on users. In such settings, the arguments for simplicity mentioned above would argue for use of PMP or variants of it.

The general argument for PMP is that pricing and the end user interaction with the network should be as simple as possible. However, that does not mean that no QoS measures should be used. Techniques such as RED or WFQ, which are invisible to the end users, can be used to improve the operations of the separate channels in PMP. Since the core of the network will probably grow in total cost even as unit prices decline (as has happened with other high-tech products), there will be a strong incentive to run that core as efficiently as possible, and this will justify careful design and operation. The main point is that this quest for efficiency should not burden the end users.

If we ever do see differentiated services, they may well evolve towards (or degenerate into, depending on one's point of view) PMP. This would be the result of users abandoning all the non-essential features in the interests of simplicity.

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 deals with the public's aversion to usage sensitive schemes, and the way in which PMP might overcome it. Finally, Section 6 briefly references 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. There appear not to be any serious quantitative models of the various QoS proposals that are being developed, including diff-serv, which is the current front-runner. The appendix presents some simple economic models of the gains that one could obtain from schemes such as PMP or diff-serv.

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 actions 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 channels in PMP should be small, possibly just two, but more likely three or four. Having few channels 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. [34]) that in many situations, most of the economic gains from subdivision into different classes of service can be gained with just a few classes.

The basic version of PMP mentioned in the Introduction assigns to each channel a fixed fraction of the capacity of the entire network. One can also use priorities. In the proposals [5, 19], 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 [12]. One could also use different approaches in different parts of the network. One can even mix these approaches on the same link.

In general, assignments of capacities and prices to the channels 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 channels. However, it would likely 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.

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 experience of both the Paris Metro and of pricing of interactive computer services [16] suggests that prices should jump by a substantial factor, around two, from one channel to the next.

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. 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. Experimental networks such as vBNS, which have low utilization levels, are able to handle all applications. 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. 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 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. 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 (even if it were to flourish, which seems exceedingly questionable) 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 channels 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 channels, the effect could be comparable to that of conventional networks.

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 already been considered in the literature for other pricing schemes, and the solutions proposed there could be adopted for PMP. Yet another question is to decide how frequently to vary the capacities and prices of different channels in PMP.

Would PMP survive in a competitive market? There is an analysis of a greatly simplified version of PMP by Gibbens, Mason, and Steinberg [18] which shows that in their model, PMP would be optimal for a monopolist, but a carrier offering PMP would lose to one offering undifferentiated service. However, this issue is not settled, since competition in information goods in general is hard to model, and most analyses predict destructive price wars (see [14], for example). (In general, there is also the tricky question of how any QoS measures are to be implemented in the Internet, which consists of many heterogeneous subnetworks.)

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 channels. 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 channels 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 channel to send it on, I expect that some would go by some general expectation of quality of service for different channels, 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 channels has to be predictable, at least on average. Unfortunately, the fractal nature of data traffic [21] 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 channels, 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, or to send data at 10 Kbps for 10 seconds. 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 channels. For example, the lowest-priced channel (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 [10], 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 [24, 33].) Methods of this type could be used to induce a more even load on the separate channels, 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. A class of "premium ISPs" is developing, which provide higher quality of service. Customers with connections to several ISPs would then have a choice similar to that in PMP. 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. (It is thus also possible to implement a form of PMP without usage sensitive charges, but having customers commit to using a fixed channel for extended periods of time, weeks or months. This option would still have the advantage of multiplexing of traffic and avoiding of per-packet charges. However, it would not promote the separation of traffic flows that can tolerate congestion from those which cannot.)

PMP would be easy to introduce. As with diff-serv, it would not be necessary to wait for the deployment of IPv6 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 [5, 2].) Since the number of channels 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) channel.

At least initially, the cost per packet on the lowest cost channel would undoubtedly be zero. That would make this channel look like the current Internet, and so make the transition easier. It might also be possible to have zero prices on this channel in the long run during slack periods.

Eventually applications, such as videoconferencing software, would be rewritten to give users the choice of channel (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 current diff-serv QoS efforts in the IETF provide all the technical tools for implementing PMP.

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.

As with most other pricing schemes, there are still areas requiring further research. For example, how should one charge for multicasting? (Cf. [20].) 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

If we are going to have differentiated services on the Internet, it appears we will need to have usage sensitive pricing. Such pricing has economic logic behind it. Unfortunately, it collides with users' unshakeable preference for flat-rate pricing. The problem is how to reconcile the two.

Usage as well as satisfaction with goods or services depend in large part on customers' subjective reactions to pricing schemes (cf. [6]). Consumer preference for flat-rate pricing has attracted considerable attention recently, especially when AOL 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. 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 (see the discussion and references in [14]). 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.)

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 [1], 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 [15].)

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 [3, 14, 24, 32]. 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. [3]). 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 [14]. 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, differentiated services require usage sensitive charging. The problem is how to reconcile these conflicting tendencies.

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. In Internet transmissions, there have been many instances of charging for the amount of transmitted data [7, 28]. In particular, the largest Australian ISP, Telstra, charges by the byte (but only for bytes received, since their traffic is very asymmetrical). It seems it might 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 [22].

PMP offers a simple 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. 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. [4]), but they are not entirely convincing. It seems likely that the cause lies more in the realm of consumers' seemingly irrational economic behavior, many instances of which have been documented by Kahneman and Tversky and others. 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 [24, 35]. 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 channels 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

Many proposals have been made for usage sensitive pricing. Extensive information can be found on the Web site [31] and in the collection of paper edited by McKnight and Bailey (of which the reference [1] below is one). Further references, short summaries, and criticisms can be found in [8, 29, 30]. There are also interesting new proposals, such as the elegant PFP (Proportional Fair Pricing) one of Gibbens and Kelly [17]. PMP differs from all those proposals in not maximizing any simply quantifiable objective function. Instead, it strives for maximal simplicity for the user, and is designed to accommodate strong user preferences that have so far proved hard to model in quantitative form.

7 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

[1] Anania, L. and Solomon, R. J. 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/⟩.

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

[3] Bakos, Y. and Brynjolfsson, E. 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⟩.

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

[5] Bohn, R., Braun, H.-W., Claffy, K. C. and Wolff, S. 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⟩.

[6] Brittan, D. 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⟩.

[7] Brownlee, N. 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/⟩.

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

[9] Coffman, K. G. and Odlyzko, A. M. The size and growth rate of the Internet, *First Monday,* 3(10) (October 1998), ⟨http://www.firstmonday.dk/⟩. Also available at ⟨http://www.research.att.com/∼amo⟩.

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

[11] Edell, R. J. and Varaiya, P. P. Providing Internet access: What we learn from INDEX. To be published. Available at ⟨http://www.path.berkeley.edu/∼varaiya/papers_ps.dir /networkpaper.pdf⟩.

[12] Ferguson, P. and Huston, G. *Quality of Service: Delivering QoS on the Internet and in Corporate Networks,* Wiley, 1998.

[13] Fishburn, P. C. and Odlyzko, A. M. Dynamic behavior of differential pricing and Quality of Service options for the Internet, pp. 128-139 in *Proc. First Intern. Conf. on Information and Computation Economies (ICE-98),* ACM Press, 1998. Available at ⟨http://www.research.att.com/∼amo⟩.

[14] Fishburn, P. C., Odlyzko, A. M. and Siders, R. C. 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⟩.

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

[16] Gale, W. A. and Koenker, R. Pricing interactive computer services, *Computer Journal,* vol. 27, no. 1 (1984), pp. 8-17.

[17] Gibbens, R. J. and Kelly, F. P. Resource pricing and the evolution of congestion control, available at ⟨http://www.statslab.cam.ac.uk/∼frank/rate.html⟩.

[18] Gibbens, R., Mason, R. and Steinberg, R. Multiproduct competition between congestible networks, available at ⟨http://www.soton.ac.uk/ ram2/papers.html⟩.

[19] Gupta, A., Stahl, D. O. and Whinston, A. B. 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/⟩.

[20] Herzog, S. Shenker, S. and Estrin, D. 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/⟩.

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

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

[23] Odlyzko, A. M. A modest proposal for preventing Internet congestion. Unpublished manuscript, available at ⟨http://www.research.att.com/∼amo⟩.

[24] Odlyzko, A. M. 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⟩.
[25] Odlyzko, A. M. Data networks are lightly utilized, and will stay that way. Available at ⟨http://www.research.att.com/~amo⟩.
[26] Odlyzko, A. M. The economics of the Internet: Utility, utilization, pricing, and Quality of Service. Available at ⟨http://www.research.att.com/~amo⟩.
[27] Odlyzko, A. M. The Internet and other networks: Utilization rates and their implications. Available at ⟨http://www.research.att.com/~amo⟩.
[28] OECD, Information infrastructure convergence and pricing: The Internet, report available at ⟨http://www.oecd.org/dsti/gd_docs/s96_xxe.html⟩.
[29] Shenker, S. 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.
[30] Shenker, S., Clark, D., Estrin, D. and Herzog, S. Pricing in computer networks: reshaping the research agenda, *Telecommunications Policy,* 20 (1996), 183-201.
[31] Varian, H. R. The economics of the Internet, information goods, intellectual property and related issues, reference Web pages with links, ⟨http://www.sims.berkeley.edu/resources/infoecon/⟩.
[32] Varian, H. R. Pricing information goods, available at ⟨http://www.sims.berkeley.edu/~hal/people/hal /papers.html⟩.
[33] Varian, H. R. Versioning information goods, available at ⟨http://www.sims.berkeley.edu/~hal/people/hal /papers.html⟩.
[34] Wilson, R. Efficient and competitive rationing, *Econometrica* 57 (1989), pp. 1-40.
[35] Zajac, E. E. *Political Economy of Fairness,* MIT Press, 1995.

APPENDIX: 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. For more detailed models, similar to the one presented here, but ones that take into account the temporal aspect of networks, with technological progress reducing costs and traffic volume growing, see [13].

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. 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. (See [13] for a detailed discussion.)

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

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

We let x_{AB} denote the minimal solution to (A12) (when one exists, which happens precisely for $27(a+b) \geq 8c^4 e^3$). 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 accommodates 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 were taken of comparable size, which means that the potential traffic from users of types A and B is assumed comparable. One can obtain other results by varying the assumptions.