The Role of Responsive Pricing in the Internet

Jeffrey K. MacKie-Mason, Liam Murphy, and John Murphy

Abstract

The Internet continues to evolve as it reaches out to a wider user population. The recent introduction of user-friendly navigation and retrieval tools for the World Wide Web has triggered an unprecedented level of interest in the Internet among the media and the general public, as well as in the technical community. It seems inevitable that some changes or additions are needed in the control mechanisms used to allocate usage of Internet resources. In this paper, we argue that a feedback signal in the form of a variable price for network service is a workable tool to aid network operators in controlling Internet traffic. We suggest that these prices should vary dynamically based on the current utilization of network resources. We show how this responsive pricing puts control of network service back where it belongs: with the users.

1. Introduction

A communications network is as good, or as bad, as its users perceive it to be. Network performance should therefore be measured in terms of overall user satisfaction with the service they receive. However network performance is usually expressed in terms of network engineering measures such as average packet delay or loss rate. These engineering measures are an imperfect reflection of overall user satisfaction because user requirements vary widely, in every service dimension and over time.

Example 1: some real-time interactive applications are able to tolerate relatively frequent packet loss without significant quality degradation, whereas some command and control functions require essentially lossless transmission.

Example 2: interactive communications usually have an upper limit on total delay and delay variation corresponding to the limits of human perception (e.g. 400-500 millisec maximum delay, with maximum variation an order of magnitude lower), whereas some offline data transfers are essentially insensitive to delays.

Example 3: packet delay or loss may be valued differently by different users even if they are running the same application. Similarly, a user's valuation of their quality of service (QOS) may vary depending on destination or time of day.

Example 4: some users want deterministic worst-case performance guarantees, whereas others are satisfied with average-case statistical guarantees. Some users may be content with "best-effort" service, for which the network offers no guarantees on loss or delay, especially if there are periods in which network utilization is low enough that even best-

effort traffic is reliably transferred.

Accounting for individual QOS requirements makes network operation and control considerably more complicated. In practice, user objectives are averaged across all users and over time. These averaged objectives are reduced to engineering measures and then are used to drive the network control process (see Figure 1): the users are not in the loop when making operational decisions.

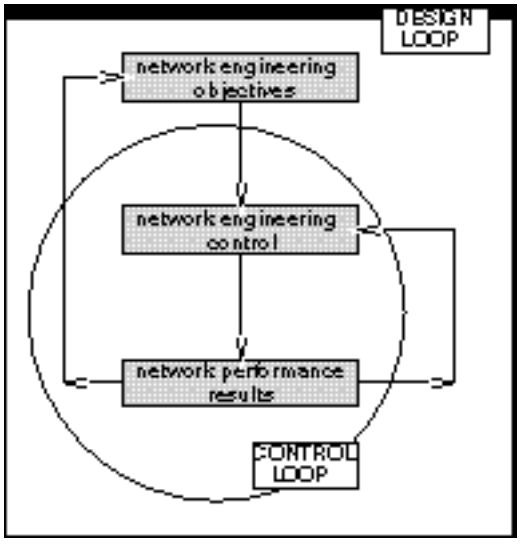

Figure 1: Network design and control loops (a)

In an effort to reflect variations in QOS requirements, many researchers divide usage into classes according to application requirements and traffic characteristics; for example, real-time video, real-time audio, one-way video playback, or off-line file transfer. Each class is then regarded as having a single representative user for analytical and control purposes. However, this approach ignores substantial heterogeneity within application classes and across users.

Heterogeneity across time: a user's valuation of a given application will be different at different times, and thus the user's requirements for network performance for the same application will vary over time.

Heterogeneity across users: different users will differently value a given application and its QOS. A common – but we believe erroneous – assumption is that video applications should receive network priority because the performance degrades more drastically with delay than for, say, a World Wide Web session. In fact, some users may place sufficiently high value on low latency Web usage that total user valuation of the network

would be increased by giving higher priority to their Web sessions than to some video sessions.

Therefore we think it is increasingly important for network operators to develop flexible tools that can support heterogeneity in usage types and user valuations.

One such tool is the ever-increasing intelligence located within the users' end-stations. It is already feasible for many traffic sources and sinks to do some basic processing on the transmitted data; for example, TCP congestion control schemes [Jacobson 1988] respond to network feedback by adjusting the traffic inputs. And as technology advances, these capabilities will continue to expand. This intelligence represents a network resource that, if properly managed, could enable a "tighter" network control loop than before.[1]

As a natural extension of existing network feedback control mechanisms, we propose bringing users back into the loop and thereby ensuring that performance measures are user-oriented (see Figure 2). We propose a form of feedback which we call responsive pricing and argue that it represents a particularly useful mechanism for maximizing network value. Users would gain by obtaining service more closely matched to their needs; network operators would gain through improved network utilization and increased user satisfaction with the service they receive. In particular, responsive pricing helps network operators by its "value discovery'' function: it reveals how the valuations for QOS differ across time, users and applications. Summary network engineering measures will continue to be important, but we believe that user preferences should be the primary consideration driving resource allocation and congestion control schemes.

[1] Our emphasis on increasingly controlling network congestion at the periphery echoes the views expressed by Dave Clark in his contribution to this volume [Clark 1996].

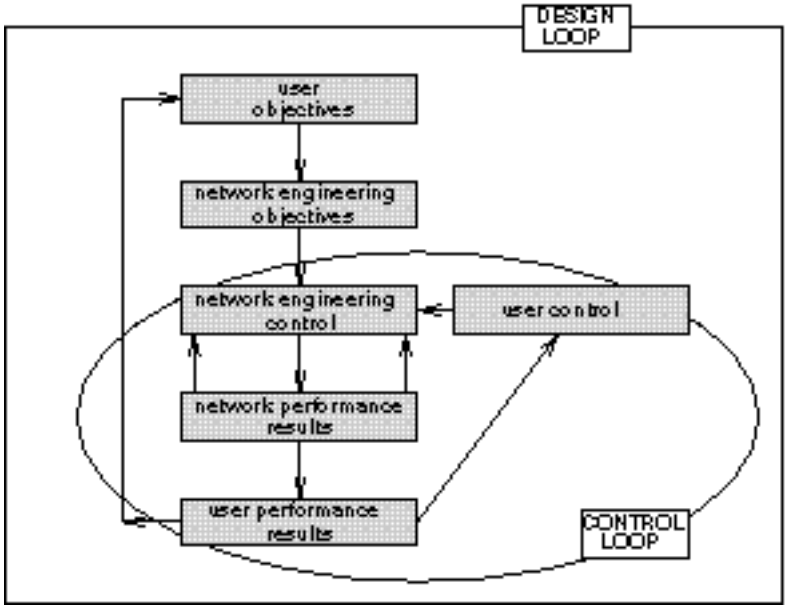

Figure 2: Network design and control loops (b)

2. Two definitions of efficiency

In focusing on user preferences, we need to distinguish two very different notions of efficiency:

Network efficiency refers to the utilization of network resources, such as bandwidth and buffer space.

Economic efficiency refers to the relative valuations users attach to their network service.

If a network can maintain a target level of service while minimizing the resources needed to provide this service, we say that its operation is network efficient. For example, by statistically multiplexing bursty transmissions, the bandwidth required can often be reduced from that of a pure circuit-switched approach while still meeting the delay requirements of the applications.

If no user currently receiving a particular QOS values it less than another user who is being denied that QOS, we say that operation is economically efficient. For example, if one user is willing to pay x per second for undelayed access to a 1 Mbps link, and a second user is willing to pay only x/2, and if only one of them can be accommodated, then in an economically efficient network the bandwidth will be allocated to the first user (whether or not they actually pay anything).

An obvious question is, why will either type of efficiency continue to be important? Some observers have suggested that the widespread deployment of fiber optic lines, and continuing

exponential decreases in processor and memory costs, will result in these network resources becoming essentially "free" so that efficiency in their use will not be important in the future, and all users can always be accommodated. We do not believe these arguments apply in the short or medium terms, if indeed they will ever apply. User demands are increasing exponentially, so that it is not clear when–if ever–network resources will be "free''. We share the dream of ubiquitous, two-way broadband connectivity at low or zero cost, but believe we must wait at least a few decades to achieve it, despite astonishing technological advances. Consider the cost of providing gigabit bandwidth not just to every home in the industrialized world, but for the other three-quarters of the planet's population as well. Add to that the cost of gigabit mobile communication that will follow each person around town, country and world. And experience suggests that application developers will have little difficulty in designing new services that use up all the available resources.

For the foreseeable future, we will continue to live in a world characterized by network resource scarcity. We will move most quickly towards "free'' service if we use our scarce network resources – whether public or private – efficiently in economic terms. The greater the value that users receive from scarce network resources, the more they will have to invest in building better and faster networks. Meanwhile, if commercial providers are not responsive to user valuations, they will not succeed in a competitive market. The same considerations apply even to private-access networks: the ultimate goal is to maximize some human measure of the value of using the network, such as profits, sales, shareholder value, and so on.

2.1. Feedback and adaptive users

Feedback is a well-established method of improving network efficiency. Users of current data networks respond to multiple forms of feedback, on various timescales. On the longest timescale users decide whether or not to use a particular network, perhaps based on the network's charging structure, or their previous experience with it. At the connection level, if a user observes that the network is usually heavily loaded at certain times of the day and lightly loaded at others, she may schedule her network usage accordingly. On these longer timescales, user responses are usually determined on economic grounds (although not always explicitly). For example, deciding whether or not to set up a connection involves weighing the expected value of using the network against the cost (in money, time, and/or degraded quality) of doing so. Most people are familiar with the decision that it is not "worthwhile" to use a network during busy periods, but instead to do something else and defer their network usage, without necessarily recognizing this process as economic decision-making.

During a connection, adaptive users can adjust their traffic inputs or QOS demands to respond to feedback signals from the network about the current state of network resources. TCP applications use various congestion control algorithms such as slow-start [Jacobson 1988] to adjust their input rates to the currently available bandwidth. The ATM Forum is developing an Available Bit Rate (ABR) service in which users who respond "appropriately" to dynamic feedback get loss guarantees from the ATM network [Newman 1994]. Since it is already accepted that user responses can be automated using pre-programmed network interfaces, fairly

sophisticated user behavior can be envisaged, and feedback strategies need not be limited by human user response times. The issue becomes one of choosing a feedback signal to modify user behavior in some desired manner.

Adaptive users can help to increase network efficiency if they are given appropriate feedback signals. When the network load is high, the feedback should discourage adaptive users from inputting traffic; when the load is low, the feedback should encourage these users to send any traffic they have ready to transmit. In this way many of the congestion problems that occur if the offered load is regarded as fixed can be avoided. One possible feedback signal is a price based on the level of network load: when the load is high, the price is high, and vice-versa. Similarly, by associating a cost measure with network loading, all users can be signaled with the prices necessary to recover the cost of the current network load. Price-sensitive users–those willing and able to respond to dynamic prices–increase economic efficiency by choosing whether or not to input traffic according to their individual willingness to pay the current price. Users who value network service more will choose to transmit, while those who value it less will wait for a lower price. When the network is lightly loaded then the price will be close to zero, and all users can input traffic.

Price signals thus have the potential to increase both network and economic efficiency, though whether a particular pricing scheme increases either notion of efficiency depends on the implementation (see Section 5). In a public network, where the users cannot be assumed to be cooperative, the more traditional feedback schemes as currently used in TCP/IP networks are not robust to user manipulation: it is relatively easy to program a host to ignore the feedback signals. Of course it would be just as easy to ignore price signals; but since users would be liable for charges they incurred, there is some incentive to respond.

2.2. Price as a form of feedback

Congestion control and feedback control are difficult problems in network operations. Perhaps because of the technical challenges involved, most researchers have ignored two key issues (or relegated them to "policy issues") which are nevertheless crucial:

- how should congestion be defined and measured? This is a difficult question because individual user requirements vary considerably, so that one user may think the network is congested while another does not; and because in internetworks the responsibility for detecting congestion may be distributed among several network operators, each of which applies a different test at their bottleneck points. This problem will become more difficult in a multiple-QOS network, in which several different performance characteristics are relevant to different applications (e.g. maximum delay, delay jitter, packet loss rate, etc.).

- how should limited resources be allocated under congestion? Currently, randomization with First-In/First-Out queuing is used, but some proposals call for users to indicate the relative priority of their traffic – leading to the problem of providing incentives so that all

users will not choose the highest priority.[1]

Charges to an Internet user could have several components, such as a connection fee, a charge per unit time or per unit of bandwidth, premium charges for certain services, and so on. In this paper, we focus attention on only one type of pricing: a responsive component which varies with the state of network congestion. By responsive pricing we do not mean a charge which counts the number of bytes or packets regardless of the network conditions.[2] On the contrary, we propose charging only when network congestion indicates that some users may be experiencing QOS degradation, with the size of the charges related to the degree of congestion. If the network is lightly loaded and all users are getting acceptable QOS, the responsive prices would be zero.

Proposing a scheme to allocate network resources and service priorities is not a radical departure: allocation occurs today in the Internet. However, the current allocation is implemented on a first-come, first-served basis, without any consideration for whether some users value immediate access more highly than others. Given the heterogeneity of user requirements for network performance, responsive pricing would improve economic efficiency by inducing users with low priority traffic to delay it until a burst of congestion eases. Such time-smoothing would not upset users who can tolerate high latency, while it would improve the network's value to users who get the greatest benefits from immediate access.

Let us clarify one point: when most people think of prices, they think in monetary terms, e.g. dollars and cents. However, there is nothing inherently monetary in applying pricing principles to communication networks. As long as the appropriate cost and valuation functions can be defined, a pricing mechanism can be applied even if money is not directly involved. For example, in a private network where one organization controls all the users, the "prices" would be control signals which summarize the state of network resources. In this case the users (or their applications) are cooperative and can be programmed to obtain a desirable traffic mix.[3]

We recognize that many people are concerned about the use of pricing in network operations. Concerns range from questions about the feasibility and overhead of usage-sensitive pricing, to more philosophical issues such as profit opportunities and fairness. While some of these concerns may or may not be borne out by future developments, others are based on misconceptions of what is being proposed or on other non-technical grounds. We do not expect that decisions on pricing will be made solely on technical grounds, but we do believe that a clear understanding of the nature of what is being proposed is necessary on all sides. Therefore we first describe a framework for responsive pricing, and some of our work on analyzing and simulating various implementations. We then address some of the objections often raised in

---

1 See, for example, [Bohn et al. 1993], [MacKie-Mason and Varian 1995a], [Gupta et al. 1996], [Cocchi et al. 1992].

2 We use "packet" in a generic sense to mean a unit of data transmitted by a user.

3 Control theorists will recognize "prices'' as the costate variables or shadow values which provide the correct signals to optimize the allocation of resources in the network: in this case, that optimize a function of user valuations of network performance.

discussions of dynamic network pricing.

3. Modeling User Adaptation to Feedback

Our focus is on the interaction between user behavior and the efficiency of the network. Therefore, we want to model traffic types that a user can adapt to the state of the network. In this section we discuss the nature of such adaptive applications.

There are several ways in which a user can adapt her traffic load to the network state. Not all of these are equally obvious to the network. For example, a user may have a constant bit-rate (CBR) application that cannot tolerate either delay or loss. If the user does not like the service offered by the network, she may adapt by connecting to a different network, so the original network never sees the load at all.

Alternatively, the user may decide to make a connection, and use it for a CBR and delay-intolerant application. However, depending on the state of the network and the service guarantees it offers, the user may adapt the traffic at source. For example, the user could reduce the number of packets transmitted, by accepting lower fidelity or precision. Or the user could delay making the connection to a different time. In these cases, the network merely sees a CBR source; it does not directly observe that the user has adapted.

In another case, a user may offer a load to the network, but accept a best efforts service quality. In this case, the user is abdicating the adaptation to the network. If the network accepts all best efforts traffic, then the offered load is not adapted to the network state. Rather, the burden the traffic imposes is adapted through varying the delay and packet loss. This form of adaptation can be quite costly if a higher layer protocol is resending lost packets: congestion breeds congestion.

Even without pricing, users obviously adapt their network usage in several ways. We propose taking advantage of this natural adaptability to improve efficiency over a short time horizon. We have elsewhere [MacKie-Mason et al. 1995], [Murphy and Murphy 1995] described several different types of adaptive users. So far we have modeled inelastic and elastic user types:

- Inelastic. An inelastic application requires a delay guarantee, but can tolerate loss and is adaptive. For example, this might be the second level of a two-level codec for video. The first level is likely to contain the minimum necessary information, and would be transmitted as a non-adaptive application. The second level consists of enhancement information. It is not essential that all of the information be delivered, and it is possible for the user to vary the amount of information transmitted in response to feedback. However, a delay guarantee is required: if the information does not arrive before the playback point, it is considered useless.

- Elastic. This type of user waits until feedback from the network indicates that they can input traffic, then transmits and requires that their cells are not lost in the network. Each elastic user decides individually what their transmission criteria are, e.g. the maximum

price per cell they are willing to pay. A possible example of an elastic user type would be a non-real-time data transfer with no ARQ capability, so that already-transmitted cells are not buffered at the sender.

With these two types we have heterogeneity across applications, and thus are modeling an integrated services network. Further, we are able to model within-type heterogeneity by specifying users of a given type who value their QOS differently. For our preliminary results see [Murphy et al. 1996].

4. Responsive Pricing Schemes

In our simulations we have been comparing three different schemes for allocating a simple network's resources. The first is a conventional approach that makes no use of feedback and user adaptation to the network state. The second is a closed-loop form of feedback and adaptation; the third is a closed-loop variation we call "tight loop'' because it shortens the delay in the control loop.

4.1. No feedback

Our proposal to improve network efficiency through involving the users in session control is somewhat novel, and certainly controversial. Most in the network engineering community seem to assume that a network will (and should?) be tuned for efficiency given a set of admitted user connections. The only room for interaction with the users in this setting is through the connection setup negotiation. Therefore, as a baseline, we simulate a network that does not provide feedback: users do not adapt to the network state.

A fixed number of inelastic sources are always admitted and active. In addition, a number of elastic applications are active at any given time. To simplify, the number of new elastic connections each period is held constant, but the number of packets to be delivered by each connection is random, so the number of active connections in any given period after the first will be random (as varying numbers of connections are completed). The distribution of elastic message sizes is chosen so that the average load being added to the network in each period is within the tolerance for a reasonable call admission algorithm. However, sometimes the amount of active elastic traffic will be large, and the network will suffer some performance difficulties (packet delays and losses).

4.2. Closed-loop feedback

Our first feedback network uses a simple scheme [Murphy and Murphy 1994], [Murphy et al. 1994]. The network state is measured by buffer occupancy at the gateway. This occupancy is converted into a price per packet, which is then transmitted back to each active adaptive application. The applications then decide on how many packets to send during the next interval, as described above.

In this network, users send some packets in period t, and network performance is affected by the aggregate number of packets received during an interval. At the end of period t the network

sends a signal back to users based on the network utilization in period t. Users then decide how many packets to send in period t+1, based on their observed period t performance and their application requirements. This is a closed-loop feedback system with at least a one-period lag between the state of the network and the effect on the user inputs.

4.3. Smart market pricing

A "smart market'' approach to adaptation has been proposed in [MacKie-Mason and Varian 1995a]. A user sends packets to the network interface which include in each header a token indicating how much the user is willing to pay to get that packet onto the network in the current interval. Then, during the pricing interval, the network gateway sorts the "bids'' on the incoming packets, and admits to the network only as many as it can accommodate without degrading performance below some bound.[1] The gateway admits packets in descending order of their bid. Users are charged not the amount that they declared they were willing to pay, but the value of the minimum bid on a packet that is admitted to the network. Thus, users pay just the congestion cost (the amount that the highest-value denied packet would have paid for immediate transport) but they get to keep all of the excess value that they attribute to delivery above the cutoff bid. This form of pricing by auction has several nice properties, described in [MacKie-Mason and Varian 1995a].

We call this mechanism a tight-loop because the user sends willingness to pay along with the packet, and the network admission and pricing is determined based on those reports and the current state of network congestion, without creating a feedback delay. In practice, there might be a one-period delay to allow the gateway to determine the approximate cutoff bid from packets presented in the prior interval.

4.4. Preliminary simulation results

To give a sense of the gains that are possible with responsive pricing, we offer some preliminary results that compare no feedback to the closed-loop pricing scheme.

---

[1] In practice, the gateway would probably estimate the cutoff bid that would admit the number of packets that can be accomodated, using recent information on the bid distribution and perhaps a sample of newly arriving packets. Then the gateway would merely route incoming packets into two (or more) queues – one for immediate handling, and one for buffering and re-entry in the next bid period. When the number of packets being held back exceeded buffer capacity, some would be dropped.

Source Type		% Loss	User Value	% Decr. Value	% Incr. Value
Unpriced	Inelastic	0	240		
	Elastic	30.4	146		
	Combined	**19.1**	**386**		
Priced	Inelastic	4.4	239		
	Elastic	0.1	204		
	Combined	**1.7**	**443**	**91.0**	**14.8**

Figure 3: Performance and Economic Gains from User Feedback (preliminary)

In the simulations we have generated 20 video sources with random frame sizes to represent the inelastic traffic, and between 1 and 39 elastic data sources with random frame sizes, with the number of active flows at a given moment chosen randomly. Under these conditions our network capacity has experienced an average of 80% utilization. When closed-loop pricing is implemented packet loss drops from 19% to under 2%, while the net benefits perceived by the users increase by nearly 15%.[1]

5. Objections to responsive pricing

We explore some common arguments against responsive pricing in network operations in this Section, and provide some counterpoints. Some previous work along these lines is contained in [MacKie-Mason and Varian 1995b] and [Murphy and Murphy 1995a].

5.1 Myths

- Why do we need to introduce prices? The Internet is free now – let's keep it that way.
- *Counterpoint*: The Internet is not free now, though it seems that way to many users whose universities or organizations pay the access fees. The issue is not whether Internet usage should be priced: it already is. The issue is how the Internet should be priced so that its value to the users is maximized.
- Network resources will soon be essentially free, therefore Internet congestion and the accompanying QOS degradation will not be a problem in the future.
- *Counterpoint*: This represents an optimistic view of the future, but we do not believe that

[1] See [Gupta 1996] for related simulation results. They show a greater improvement due to the type of pricing they simulate. Of course, our results and theirs are illustrative, as they depend on the reasonableness of the many simulation approximations and parametrizations.

this will come true in the short or medium terms, if ever. See our arguments on why efficiency will continue to be important in Section 2.

- With any form of responsive pricing, it's the small users who will suffer the most. Rich users could behave as they want since they have the resources, and could effectively limit the network access of smaller users. The role of the Internet as a medium for information exchange between all-comers will be lost!

- *Counterpoint*: Absolutely not! Quite the opposite. Suppose that the network is supported only by connection fees. The connection fee will then be set based on the average usage for a connection of a given size. Then the small users will be paying more than their share to support the heavy users. A corollary to this myth is that the user cost of the Internet will increase if responsive pricing is introduced. The whole purpose of responsive pricing is to make the network more efficient, and to raise the value for users. Thus, if implemented intelligently, we will get more value out of a network of given cost. For a network of fixed size, we can lower the connection fees by an amount equal to the congestion fees and still recover costs, so total outlays are the same but the network has higher value. Or we could use the congestion revenues to invest in a bigger, more valuable network facility. It is also worth remembering that with responsive pricing, you pay for your actual usage in terms of the cost it imposes on other users. If all you want to use the Internet for is email and netnews, your charges under responsive pricing would be zero because these are flexible applications and do not require real-time performance or guaranteed bandwidth. As for rich users being able to afford to ignore dynamic prices, this is true under any pricing scheme and is a larger issue concerning the distribution of wealth in a society.[1]

- Responsive pricing is just another way for network operators to make more money. Users will lose out as network operators maximize their profits.

- *Counterpoint*: It's true that there is the potential for profiteering whenever prices are charged, especially when the conditions under which prices are set are not immediately accessible to ordinary users. But in a competitive environment, the market disciplines network operators whose revenues exceed actual cost by more than the minimal amount necessary to stay in business. Of course, market discipline is limited in the case of a monopoly provider or a cartel of price-fixing providers. But then the outcome depends on policy and regulatory decisions rather than on the specific pricing scheme.

[1] We do not mean to dismiss income distribution problems as unimportant, but rather to say that network pricing (or non-pricing) is not the right venue for solving them.

5.2 Objectives for Responsive Pricing

- If congestion is caused by bandwidth-intensive users such as real-time video, why don't we just keep these users off the Internet, or limit their number so that they don't cause congestion problems?

- *Counterpoint*: Keeping these users off the Internet means keeping the Internet low-tech and continuing the best-effort no-guarantees paradigm. That is, can't we do better than a network that cannot support reasonable quality for real-time video and audio? This runs contrary to the trend towards integrated-services networks, and may cause the Internet to miss out on innovative information transfer and retrieval mechanisms. Apart from the administrative issues, why should "low-tech" users be allowed to veto "high-tech" users? What will the general public want when they come online? Which administrative bodies do we want to empower with rationing authority?[1]

- Why won't some non-pricing scheme be enough? Administrative controls can be used to impose some appropriate notion of fairness, for example; or users can choose a traffic priority level which matches their requirements.

- *Counterpoint*: Who decides what is fair? The network operator can; but according to a user-oriented objective, fairness should be determined collectively by the users. We might all agree that telesurgery is more important than email, but what about interactive video games versus email? Also, every time a new application is developed it has to be slotted into the priority order, an increasingly complex process. Further, the value of an application to a given user will vary over time. Priorities for different applications will sometimes – perhaps usually – incorrectly order valuations. Suppose the network simply supports priority levels and allows each user to choose their own level. Why wouldn't they all choose the highest priority? To guard against such abuses, there would have to be some penalty for "inappropriate" declarations, implying the need to define "appropriate" priority levels or to assign increasing charges to higher priorities (e.g., [Bohn et al. 1993]). A user's choice of priority level would then be based on economic considerations: balancing the benefits of higher priority against the costs and/or the penalties for inflating their application's perceived priority level. This is the essence of a pricing scheme.

- Bits/bytes/cells/packets are not the correct units to charge for – it's information that users care about. Any scheme which proposes to look inside every packet to determine how it relates to other packets is likely to be too complex to be justified. Also, lower-layer

[1] It is interesting to note that for a while in 1995 the EUNet backbone in Europe administratively forbade unrestricted use of the CU-SeeMe videoconferencing software, requiring that users apply in writing, in advance, for permission to establish a session.

mechanisms (such as Ethernet collisions) or packet losses requiring retransmissions make it difficult to predict how much "raw" data has to be transferred to transmit a given amount of information. Should users be charged for retransmissions that they have no control over, or packets that are dropped by the network?

- *Counterpoint*: Our proposal involves pricing for transport, not for content. The "importance" of a particular packet, and its relation to other packets, is a higher-layer issue determined by the application (or ultimately by the users). We are not proposing that the network be aware of these issues; on the contrary, with responsive pricing it's up to the users to decide how packets are used to transfer information. It's true that it is in general impossible to determine beforehand exactly how many packets are required to transmit a block of information, but again this is a higher-layer issue. Indeed, there will be some efficiency gains from providing a financial incentive to software developers to make more efficient use of network packet transport. The important question is whether the users or the network should bear the uncertainty arising from variations in congestion. If the network is expected to offer a "file transfer" service, the file transfer charge per megabyte could be computed by averaging over many such transfers. If the user is expected to pay for all transmitted packets, they could define a maximum number of packets they are willing to transmit per megabyte of information, and invoke an application-layer process if this threshold is exceeded.

- Once a network is installed, any load-dependent costs of transferring data are minimal – the fixed costs of network management and maintenance dominate. These fixed costs can be efficiently recovered through connection fees and capacity prices (proportional to the size of the access link). Why implement an elaborate pricing mechanism to recover the relatively small variable costs?

- *Counterpoint*: Our point is not about current cost recovery. Most network production costs can and should be recovered through connection fees and capacity prices. We are concerned about the *congestion* cost which one user's traffic imposes on other users sharing the resources. Bandwidth or buffer space occupied by one user's traffic is not available to other users. When this reduces other users' quality of service (through increased delays, loss rates, blocking probabilities, and so on), they suffer congestion costs which may translate into significant actual costs of service degradation. One mechanism to capture these costs is a price which is sensitive to some indicator of congestion, such as load.

- Why are we so concerned with modifying individual user behavior anyway? Surely one user can't do that much damage to the Internet?

- *Counterpoint*: One user, or a relatively tiny number of users, can now do a lot of damage

to the Internet. A single interactive video connection can take up as much bandwidth as thousands of traditional Internet applications. Without some incentives to take other users into account (and/or penalties when they do not), a small fraction of the user base could bring large regions of the Internet to a standstill. In any case, the collective behavior of lots of individuals, acting without concern about the effects of their traffic on others, can easily lead to congestion. We think it most natural and efficient to attack the problem at source, but it may be that feasible responsive pricing schemes are more practical if imposed at a higher level of aggregation.

- There is already a penalty for heavy network usage: my application runs slower. Why should I pay again, in real money?

- *Counterpoint*: Your application running slower represents a penalty to you, but what about other users' applications which are also running slower? In order to efficiently share resources, you have to be made aware of the costs your usage imposes on other users. If there is enough of the shared resources, these congestion costs can be insignificant. But we believe that the Internet cannot rely on these costs being essentially zero, at least not for the foreseeable future. Meanwhile, some forms of responsive pricing give the user a choice: either pay in delay, or pay in money to avoid delay. This choice is available on a gross scale today: we can use the Internet with uncontrolled delays, or use a low-delay private leased network. We think it is possible to offer this pay-or-delay choice to users within the Internet, making them better off by giving them a wider range of service choices.

5.3 How would it work?

- Most users will want to know their charges in advance, and will not want to deal with prices that change during the lifetime of a typical connection. Why won't flat-fee prices (per minute connected, or per kbps of the access link) be enough?

- *Counterpoint*: We are not advocating that all users must face responsive prices. Any user can choose not to face dynamic prices, even if their application is adaptive. They would then be charged according to some other pricing scheme. For example, a user might be allowed to pay zero responsive prices in exchange for getting only best efforts service with lower priority than other users who pay a positive price. In any realistic pricing scheme it would be possible for a user to set the maximum charge they are willing to pay, which is what is usually required for budgetary purposes. We should also point out that flat-fee pricing is really long-term usage pricing, so even "flat" fees include a usage-based component: it's just averaged over a period much longer than a connection

lifetime.[1]

- Even if we want to allocate according to congestion costs, how can the network determine what actual costs the current load is imposing on users who probably have widely varying service requirements? Getting users to reveal these costs is likely to be extremely complicated, if not impossible.

- *Counterpoint*: It is true that providing users with the right incentives to reveal their actual costs of service degradation is complicated. It is not impossible however: truthful revelation is one of the properties of the smart market mechanism in [MacKie-Mason and Varian 1995a]. With any prices that increase with the degree of congestion in the network, users will be induced to prioritize their traffic. Only users who value their traffic at least as much as the current price will transmit. If congestion remains unacceptably high, then the associated price was too low; conversely if capacity is unacceptably underutilized, the price was too high. Thus, through a process of experimentation and dynamic adjustment, the network can shape the price schedule so that users approximately reveal their valuations for uncongested service through their responses to the price feedback.

- Suppose we institute some form of responsive pricing, and users (especially the high-bandwidth ones who will pay the most) leave the Internet and use other networks. Won't that reduce the value of being on the Internet, perhaps to the point where even small users leave and join the other networks?

- *Counterpoint*: We are discussing only charging users for the amount of congestion cost they impose on other users. Users get to decide whether they want to pay money to avoid congestion, or not pay money and bear congestion delays – in the current Internet, everyone is forced to accept the latter alternative. Costs are not just monetary: if the cost of congestion delay is severe, then we can expect that some users are already being driven away. Indeed, many network applications are restricted to private leased-line networks to avoid Internet congestion (e.g. most videoconferencing). By allowing transport priorities to be sorted based on who suffers the most from congestion delay, we will increase the value of the network, which should spawn additional growth and new uses.

5.4 Feasbility

- Dynamic pricing schemes are unworkable in practice due to the overheads involved in

[1] This assumes that the flat fees are set to recover some function of the usage cost, as they would be in a competitive market for service provision.

accounting and billing for usage on such a detailed level. In addition, a significant portion of the revenue raised is needed to defray the cost of doing dynamic pricing in the first place!

- *Counterpoint*: The costs of dynamic pricing may outweigh the benefits for a particular implementation but we do not believe this is necessarily true for all dynamic pricing schemes. In particular, online pricing mechanisms may reduce the actual cost to an acceptable level; there is no reason to think that current billing and accounting costs in other industries, such as telephone or electricity networks, will necessarily apply to dynamic pricing in the Internet. In particular, since data networks have vast distributed computing power in the form of smart end-user devices at the periphery, it may be possible to design distributed billing systems that have very low cost for large numbers of small transactions.

- Dynamic pricing is impractical because users cannot respond to prices which are updated many times per second. If the update interval is increased to the minimum period in which users can respond, congestion can arise and disperse in between price updates, so that prices no longer influence user behavior.

- *Counterpoint*: Our scheme assumes an intelligent network interface at price-sensitive user sites, so the processing necessary to respond to dynamic prices would be done automatically based on pre-programmed user preferences. For example, a user could have a default preference in her email program that instructed the software to hold outgoing email whenever the price exceeds 0.01 cents per packet. Such software would play a similar role to current TCP implementations, which respond to network feedback by adjusting their traffic inputs, except that the feedback in our case is the current price.

- Charging for transmission fails to capture cases where the benefit of a transfer is with the receiver. If senders are charged for receiver-initiated transfers, we could see a drastic reduction in the number of open-access servers with a corresponding decrease in the value of using the network.

- *Counterpoint*: The problem of allocating the benefits of a particular information transfer is a higher-layer issue. We do not believe that associating the charge for a transmission with the sender constrains the actual flow of money in any way. It is easy to imagine multiparty connection protocols which initially negotiate each party’s responsibility for the total charge, or "reverse-charges" servers which only transmit data once the receiver has indicated willingness to pay the resulting transmission costs. Just as in telephony, we can expect "1-800’’, "1-900’’ and other billing services to arise.

5.5 Cultural Effects

- By introducing responsive pricing, the traditional Internet culture (which emphasizes openness and sharing) will be destroyed, and it will become just another commercial service.

- *Counterpoint*: It's true that by changing the pricing scheme used in the Internet, the culture will also change. However the culture is changing anyway due to the strains imposed by the demands of the ever-increasing user population. The question is, how can this change be managed so that overall user satisfaction with the Internet is maximized? Insisting that the Internet remain a connection-fee only network consigns it to ever-lessening value as usage and congestion increases, and new, QOS-sensitive applications are developed that cannot be successfully implemented in a first-come-first-served network. Many users are already fleeing the increasingly noisy Internet; shifting some responsibility for congestion control out to the users, and treating them as smart rather than dumb devices, will help preserve those parts of the Internet that users most value preserving.

- We don't know what the future Internet will look like, so it would be a mistake to adopt a responsive pricing scheme which is so controversial – it could stifle innovation and cause the Internet to miss out on opportunities to enhance its value to society.

- *Counterpoint*: Equally, by not introducing some additional forms of congestion control, the Internet may miss out on future growth and improvements. For example, the Internet may be consigned to missing the widespread deployment of real-time interactive video if better mechanisms for controlling congestion are not developed. We propose one particular form of congestion control based on economic principles of pricing for resource allocation. Price is one possible feedback signal which has some attractive properties (compactness, quantifiable, etc.). Economists have developed a large body of theory of pricing mechanisms, and there is a lot of experience with the use of prices in real-world markets. However we do not rule out the possibility that there are other feedback mechanisms that, for one reason or another, may be preferable in communication networks.

6. Conclusions

Many proposals have been made to incorporate feedback into network control and resource allocation schemes, such as TCP congestion control and avoidance algorithms or ABR service in ATM networks. We suggest taking these proposals one step further by explicitly defining how that feedback is generated by the network, and what form it takes. In responsive pricing, the network announces a price which is based on the cost of using network resources, and price-sensitive users adjust their traffic inputs based on this price and their own specification of how

valuable network service is to them.

What we propose is to give users incentives to consider the effects of their usage on other users. While users may or may not behave "considerately" in a privately-owned network, it appears that some incentives will always be necessary in commercial networks. We also address the issue of user valuation of the service, and allow for some sources to have more demanding traffic than others regardless of the type of applications involved. Simulations show that it is possible to gain both network efficiency and economic efficiency by using pricing. In other words, the network actually carries more traffic and carries more important traffic from the users' point of view.

References

(Bohn et al. 1993) R. Bohn, H.-W. Braun, K. Claffy, and S. Wolff. 1993. Mitigating the coming Internet crunch: Multiple service levels via precedence. Tech. rep., UCSD, San Diego Supercomputer Center, and NSF.

(Clark 1996) D. Clark. 1996. A Model of Cost Allocation and Pricing in the Internet. In Internet Economics, J. Bailey and L. McKnight, eds., MIT Press. Available from URL: http://www.press.umich.edu:80/jep/works/ClarkModel.html.

(Cocchi et al. 1992) Ron Cocchi, Deborah Estrin, Scott Shenker, and Lixia Zhang. 1992. Pricing in computer networks: Motivation, formulation, and example. Technical report, University of Southern California. Available from URL: ftp://ftp.parc.xerox.com/pub/net-research/pricing2.ps.Z.

(Gupta et al. 1996) Alok Gupta, Dale O. Stahl, and Andrew B. Whinston. 1996. A Priority Pricing Approach to Managing Multi-Service Class Networks in Real Time. In Internet Economics, J. Bailey and L. McKnight, eds., MIT Press. Available from URL: http://www.press.umich.edu:80/jep/works/GuptaPrior.html.

(Jacobson 1988) V. Jacobson. 1988. Congestion Avoidance and Control. Proc. ACM SIGCOMM '88 Symp.

(MacKie-Mason and Varian 1995a) J. MacKie-Mason and H. Varian. 1995. Pricing the Internet, in Public Access to the Internet, B. Kahin and J. Keller, eds., Prentice-Hall,

Englewood Cliffs, NJ. Available from URL http://www.spp.umich.edu/ipps/papers/info-nets/Pricing_Internet/Pricing_the_Internet.ps.Z

(MacKie-Mason and Varian 1995b) Jeffrey K. MacKie-Mason and Hal R. Varian. 1995. Some FAQs about usage-based pricing. Computer Networks and ISDN Systems 28. Available from URL: http://www.spp.umich.edu/ipps/papers/info-nets/useFAQs/useFAQs.html. Also in Proceedings of WWW '94, Chicago, Illinois, and in Proceedings of the) Association of Research Librarians 1994.

(MacKie-Mason et al. 1995) J. MacKie-Mason, J. Murphy and L. Murphy. 1995. ATM Efficiency Under Various Pricing Schemes. Proc. 3rd International Conference on Telecommunications Systems Modelling and Analysis, Nashville, TN, March.

(Murphy and Murphy 1994) J. Murphy and L. Murphy. 1994. Bandwidth Allocation By Pricing In ATM Networks. IFIP Transactions C: Communication Systems, No. C-24, p. 333-351. Available from URL http://www.eeng.dcu.ie/ murphyj/band-price/band-price.html.

(Murphy and Murphy 1995a) L. Murphy and J. Murphy. 1995. Feedback and Pricing in ATM Networks. Proc. IFIP TC6 Third Workshop on Performance Modelling and Evaluation of ATM Networks, Ilkley, England, July, p. 68/1-68/12. Available from URL http://www.eeng.dcu.ie/ murphyj/brad-price/brad-price.html.

(Murphy and Murphy 1995b) L. Murphy and J. Murphy. 1995. Pricing for ATM Network Efficiency. Proc. 3rd International Conference on Telecommunication Systems Modelling and Analysis, Nashville, TN, March, p. 349-356. Available from URL http://www.eeng.dcu.ie/murphyj/atm-price/atm-price.html

(Murphy et al. 1994) J. Murphy, L. Murphy and E.C. Posner. 1994. Distributed Pricing For Embedded ATM Networks. Proc. International Teletraffic Congress ITC-14, Antibes, France, June, p. 1053-1063. Available from URL http://www.eeng.dcu.ie/murphyj/dist-price/dist-price.html.

(Murphy et al. 1996) L. Murphy, J. Murphy and J. MacKie-Mason. 1996. Feedback and Efficiency in ATM Networks. Proc. Int'l Conf. Comm. (ICC '96).

(Newman 1994) P. Newman. 1994. Traffic Management for ATM Local Area Networks. IEEE Communications Magazine, pp 44-50, August.

Acknowledgements

We would like to thank the contributors to the com-priv Internet mailing list, and some of the attendees of the MIT Internet Economics Workshop, for their vigorous objections to usage-based pricing.

Author Information

Jeffrey MacKie-Mason, Dept. of Economics and School of Information, University of Michigan, Ann Arbor, MI 48109-1220, USA (also NBER, Cambridge, MA); jmm@umich.edu. Liam Murphy, Department of Computer Science and Engineering, Auburn University, AL 36849, USA; lmurphy@eng.auburn.edu. John Murphy, School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland; murphyj@eeng.dcu.ie.