Combining Sender and Receiver Payments in the Internet

David D. Clark

M.I.T. Laboratory for Computer science

Version 3.1.2

Presented at Telecommunications Research Policy Conference, October 1996

Abstract

There have been a number of proposals for usage-sensitive payment schemes for the Internet. This paper considers one aspect of payment for service in the Internet, which is the problem of allocating payment between the sender and the receivers of traffic. It proposed a scheme, called *zone-based cost sharing*, in which the sender and the receiver can pay some proportion of the cost for use of the Internet. The paper proposes that any practical scheme must permit cost sharing among the participants, and also proposes that this sort of pricing should be applied to enhanced services, such as real time traffic or predictable capacity under congestion.

1. Introduction -- the current Internet pricing model

The current payment model for the Internet might be called a *regional payment* scheme. Users attached to the Internet make payments to their access provider, who in turn makes payments to the regional or wide area provider who serves them, and this pattern has the effect that each user contributes to the payments for the part of the Internet that is "nearby". In the "center" of the network, the various wide area service providers interconnect at *revenue neutral* points, at which no money flows in either direction. The assumption is that these points represent a form of equilibrium in which the providers on each side have been adequately (and, by some approximation) equivalently compensated by the flow of money in from the edges, and thus there is no residual imbalance at these points in the "middle" of the Internet.

This scheme seems to work today, to the extent that it has allowed a healthy competitive market for Internet service providers to come into existence, and seems to provide a adequate means to allocate the payments from the users to the appropriate providers.

This paper explores more complex approaches to allocation of cost among users and providers. To motivate this endeavor, it is useful to start by discussing the limits of the current regional payment scheme, since if there is nothing wrong with the scheme, there is no reason to modify it.

1.1. What is wrong with today's scheme?

The basic issue with the regional payment scheme with negotiated static revenue neutral points is that it provides no explicit way to measure or determine from user behavior what value each customer actually associates with the various uses of the network, and what their willingness to pay for service is. It assumes that the overall value structure of users can be deduced by allowing the users to select among different service offerings in a competitive world, and then connecting these offering together.

There are three specific problems with the static regional payment scheme in use today.

First, there is no way of demonstrating that the revenue neutral boundaries are actually in the correct place in the Internet. The assumption that this boundary is at the points of connection among the major wide area service providers is just a simplifying approximation made by the various providers in order to avoid disputes.

Second, there is no effective way to rationalize the payment for expensive components in the "middle" of the network. A specific example is a trans-oceanic link, for example between the U.S. and Europe. This link occurs between major Internet service providers, and would, if this were a terrestrial connection, be assumed to be a revenue neutral point. But the trans-oceanic link is a very expensive facility, and this triggers extensive debate as to how the cost should be shared.

In the absence of any explicit way to determine the value of this link to either side, the providers are again forced to make a negotiated assumption. Today, there are several models that dictate how trans-oceanic links are paid for. If a small country connects to a large country, it is often assumed that most of the value is to the smaller country, (and that the small country has little bargaining power) so the small country pays for the whole link. In the case of major Atlantic and Pacific links, it is often assumed that since most users want to get to Web pages, and most Web sites are in the US, that most of the value of the link is to the users in the foreign countries, which thus pay the whole cost. (This cost appears to be a noticeable component of the higher prices seen for overseas Internet service, along with higher local access costs.) In other cases, the two ends of a trans-ocean link share the cost in some proportion. However, in all these cases, the balance of payment is not derived from actual indication of value to the different users.

Third, it forces all users to fit into the same assumed balance of values. There is no way to identify those users who would pay more to have better service and serve them appropriately. The example of trans-oceanic access to Web pages illustrates this issue. The pragmatic assumption today is that users who access a Web page extract most of the value from the transaction, and thus pay for it. This assumption, for some Web pages, may be exactly backwards. A highly commercial Web page, which has lots of advertising revenues, might derive value whenever it is accessed. The provider of that Web site might be willing to pay a larger fee to increase the number of users from greater distances that get to its site. The static regional scheme today, which imposes a uniform assumption of payment flow, precludes serving this customer according to his needs. There might be substantial additional revenues to be extracted from customers, if there were a mechanism to serve and bill them according to their individual service objectives.

There are some benefits to the current static scheme, as well as problems. First, of course, it is simple to implement, requiring no data gathering, traffic monitoring or control inside the network. Second, it forces each region to be as cost-effective as possible, since all the costs of each region must be recovered from that region. However, there may be circumstances in which there is a benefit if funds could flow across some of the presumed revenue neutral boundaries. For example, if there is a particular region of the Internet that has higher intrinsic costs to serve, that region is forced to carry all the costs of provisioning that region. There is no way for funds to flow in from outside, even if the collective users of the network would see value in having this happen.

2. Different users -- different value

Consider the example above of a commercial Web site, who is very interested in attracting users, and is willing to pay substantial fees to the network provider if this successfully subsidizes customers who thus are willing to explore the site. In contrast, another class of Web server, offering content on the network essentially as a public service, may not be able to afford to pay any substantial fees to support access to this data by users. The assumption in this case that if a user assigns sufficient value to the information to pay for its retrieval, then the transfer is justified.

At the other extreme, there may be a class of potential users who assign so little value to the overall use of the Internet that under the static regional payment scheme, they refrain from subscribing. However, having these customers on the network may be of sufficient marginal value to the commercial content providers that they would be willing to see their payments flow towards these users to subsidize their basic access.

Today, all of these customers are required to purchase service within the context of the same regional payment scheme. There is no real way (except by attaching to different providers with different degrees of service capacity) for the various classes of users to purchase different sorts of services, and certainly no current way for one service provider to serve all of them with the same Internet infrastructure. One of the goals of the mechanisms described in this paper is to permit this flexibility in service offering.

There are parallels to this range of assumptions in other media. Television broadcast, at least in the U.S., is free to consumers, under the assumption that the advertisers derive sufficient value to pay for the transmission. Cable and satellite companies challenged this assumption, and proved that they can extract substantial payments from consumers for access to television. Public broadcasting attempts to fit into this structure without advertising revenues, and thus has to step outside the normal framework of revenue flows, and ask the consumers to contribute directly to the payment for the production and transmission in a rather *ad hoc* way.

The telephone system reflects this variation in willingness to pay, with callers paying for certain of their long distance calls, while merchants offer 800 service, in which they pay the per-call costs, in order to attract customers to call. The 800 service has added substantially to the total revenues derived from phone usage, and by analogy, a service with similar objectives might add to the total revenues derived from the Internet.

3. Assigning value to flows

While a service that captures the same objectives as the telephone 800 service might be useful in the Internet, one cannot carry the details of the 800 service directly into the Internet. The Internet has no concept of a call, packet flows in each direction are not linked, and today there is no indication in a packet as to how value should be associated with it. A scheme with different structure will have to be devised to match the nature of the Internet.

To begin, let us introduce the term *flow*, which is often used to describe a sequence of packets going from a specific source to a specific destination (which could be a multicast address describing multiple physical destinations) that share a common set of service requirements and value assumptions. A flow could be associated with the retrieval of a Web page, a file transfer, a video stream, one direction of a telephone call, and so on. While there are often packets flowing in both directions as a part of a network interaction, we model this as two flows, one in each direction. The goal of a pricing scheme must be to deduce how much value is associated with a flow by each *participant* (sender and set of recipients), and charge the various parties accordingly.

One approach to offering and procuring service is that the end nodes will perform a separate real-time determination of their willingness to pay for each flow of packets that they deal with (like dealing separately with each collect call). This degree of dynamic interaction is probably not what the customers want. Instead, they probably want to enter into some long term agreement that they will pay in some proportion for flows of various sorts (like an 800 service), perhaps in the form of a fixed fee for a certain level of service. This requires that there be some way for the customer to describe the sorts of flows they wish to pay for.

Assume that the users of the network -- the sources and destinations of flows -- had the capability to declare explicitly what value they associated with their flows. What information would such an expression of value contain?

In general, it would capture two aspects of the value equation, first, the performance parameters of the service that the flow should receive, and second the price that the user will pay. It might be that the price is not expressed directly, but is derived as a function of the requested service. However, **both predictable price and predictable service must be part of the service offering**. Since there are at least two participants to a flow -- the sender and one receiver, and in the case of multicast there can be many receivers, **it should be possible for each of these participants to share in paying the cost of the flow**. These points are discussed below.

3.1. It must be possible to specify the desired service for a flow.

In contrast to the telephone system, where all calls require the same network capacity and the same quality of service, flows in the Internet can differ widely in their need for capacity, control of delay, or other features. Especially in the context of associating value with enhanced services, it must be possible for the user to describe the service he desires.

There have been a number of proposals for how services should be characterized. Some proposals, such as the committed information rate (CIR) of Frame Relay, allow the subscriber to specify a service in terms of a minimum bandwidth [Stallings 1992]. Some, such as the discard eligible (DE) bit of Frame Relay or the cell loss preference (CLP) bit of ATM [ATM 1996], allow the user to specify a loss preference for his various packets in time of overload. The expected capacity scheme described in [Clark 1995] allows a user a wide variation in the burstiness and bandwidth of the service that he requests. The guaranteed service [Shenker 1996] of the Internet, intended for *real time* flows, allows the user to specify the average rate and maximum burst size of his traffic, and then find out the end-to-end worst case delay for the packets in the flow. In contrast, the controlled load service of the Internet [Wroclawski 1996], also intended for real time traffic, allows the user to specify the average rate and maximum burst size, but, while trying to reduce the delay of the packets, does not provide to the user any explicit indication of the maximum delay to anticipate.

This paper is not concerned with which sorts of services are available, or which approach to specification is used, so long as the definition is consistent across the Internet, and conforms to certain requirements, as discussed later. The point is that any proposal for usage-sensitive payment schemes must be expressed in the context of payment for *something*. The range and nature of the services that might be invented in the future is difficult to predict. A general pricing scheme must thus be able to control the use of enhanced services, without knowing in detail what these services are. What is needed is to understand enough about the general characteristics of these services to build suitable cost allocation mechanisms.

3.2. Benefit of shared payment schemes

Consider again the current payment practice in the Internet, the static regional payment plan. The consequence of this pattern is that the origin and destination of a flow *share* the cost of moving the data.

The long distance component of telephone service has a different pattern: either the sender pays all the costs, or (for collect or 800 calls) the receiver pays all. This "all or nothing" pattern has a drawback, which only occasionally becomes apparent. 800 numbers can only be called from within some scope, which is the region of the telephone system within which the receiver agreed to make full payment. If the caller is further away than this, the call cannot be completed. It is necessary, in that case, to find an alternative number for the receiver, a number associated with the "sender pays all" pattern, and call that number instead. This feature has caused callers in Europe a degree of frustration, since many U.S. companies only put their 800 number in their ads.

A more general payment structure would be for the sender and the receiver to *share* the service-specific costs, just as with today's static regional payment scheme, they share the overall network costs. This requirement is probably even more important in the Internet than in the phone system, because, as noted above, there are a wide range of services that the Internet offers, not just a phone service. This suggests that there will probably be a much wider variation in willingness to pay, depending on the service. Therefore, this paper proposes that any payment scheme proposed for the Internet should be a shared payment scheme, in which all the participants in a flow can potentially contribute to payment, each according to their willingness to pay.

3.3. How to define scope of payment?

There are a number of ways that the sender and receiver of a flow could express their partial willingness to pay. The scheme proposed in this paper is that each end of the flow (source and destinations) declare the region of the network in which they will pay the service-specific costs. For example, a Web site might agree to pay all costs for flows within the US. Another might only pay costs local to a city, while another might not pay any costs beyond their local site.

This pattern for service-specific payments is similar to the static regional payment scheme, and has the following features.

- At each point in the network, either the sender or some set of receivers are paying the cost. It is thus clear which way the compensation should flow at any point in the net, in particular, at an inter-provider connection points. (The scheme must be implemented in a way that makes this information available; see below.)

- The cost of agreeing to a certain service can be determined in advance. The cost of a service within a certain region is determined only by the prices charged within that region.
- There is an obvious way to deal with the case that the total willingness to pay from the sender and receivers is "not enough" to cover the total cost. In a regional payment scheme, the manifestation of insufficient willingness to pay is that a region of the network "in the middle" is uncompensated for providing service. A reasonable way to deal with this would be to give the flow only the lowest class of service across that region. For example, the traffic could be the first to be discarded in case of congestion. This is discussed below.

To put these benefits into perspective, consider an alternative (and inferior) scheme, in which each end of the flow expresses the per-cent of the cost that they would pay, for different ranges of addresses. Thus, a US Web site might say that they would pay 100% of the cost to U.S. clients, 60% of the cost to European sites, 40% of cost to other foreign sites, and so on.

- This approach is rather complex to implement, because to determine if the collective willingness to pay is enough to cover the total cost, it is necessary to implement a mechanism within the network to gather the information from all the participants in a flow (remember that for multicast, there is one sender but many receivers) and sum these values to see if they are sufficient.
- If the sum of the willingness is less that 100%, it is not clear what to do. Provide no service at all? Provide service for some arbitrary part of the path?
- The cost to each participant is not easily predicted. If, for example, some distant part of the path has a very high cost, a cost perhaps unknown and unexpected to the local site, some percent of that cost is nonetheless allocated to that site.
- There is no direct and easy way to look at the packets crossing an inter-provider boundary and determine which way compensation should flow in order to balance the payments. Only by determining the total costs of the flow and computing the percent each end is willing to pay, can it be determined which way compensation should flow.

3.4. Preserving the basic service of the Internet

As noted above: **if there is insufficient total willingness to pay, the consequence should be reasonable and predictable.** One feature of the current regional payment scheme is that it provides complete connectivity. In almost all cases, any two parties on the Internet are assumed to be able to exchange packets. Anything that disrupts this feature is considered a serious flaw. Per-user regional boundaries for payment runs the risk of having this consequence. If individual senders and receivers need to fully cover the value of a flow before it can be accomplished, then either a complex negotiation must occur before packets can flow, which is completely at odds with the current design of the Internet, which has no "setup" phase before starting to send, or there is a risk that, if a particular sender and receiver do not happen to associate enough value to a flow, it is simply not carried across the Internet.

One way to deal with this issue is to assume that there will be a range of network services -- a basic service and enhanced services. If we assume that the basic service is similar to what the Internet offers today, with universal connectivity, a simple subscription pricing scheme and no assurances of service quality, then users can assume that the Internet of tomorrow will at least be similar to the Internet of today. Then, more advanced pricing can be associated with more advanced services, which might, for example, give assured access during times of congestion and so on.

The proposal here is that the willingness of the user to pay should control the willingness of the network to offer *enhanced services* to the flow of packets.

4. Goals of a new payment scheme

The previous section proposed a specific scheme for allocating payment for service-specific costs in the Internet, by allowing each participant to indicate, for flows with certain performance characteristics (e.g. bandwidth) the region of the network within which that participant is willing to cover the costs. As a part of proposing this scheme, it also identified a number of design goals for any such scheme, as follows:

- **The scheme should minimize the need for dynamic, distributed protocols** to compute total end-to-end costs, or cost allocation formulas. Local costs should be locally determinable.
- **Each participant in the flow should be able to predict his component of the costs.**
- **It should be possible, at inter-provider connection points, to determine easily the direction in which compensation for the flow should go.**
- **An advanced pricing scheme must not break the current operation of the Internet.**

4.1. Other design goals for dynamic payment schemes

There are a number of additional goals that should be kept in mind in order to achieve a successful design, goals that derive more from the fundamentals of the Internet rather than from the specifics of a pricing scheme.

Control of complexity

Today's network components -- routers, access points, and so on, are subject to very high demands for throughput, cost reduction, and generality. Any pricing scheme that adds a significant complexity to these devices will cause problems in acceptance of the scheme. Any scheme that requires the implementation of large amounts of new software for the end notes will be very difficult to deploy.

Incremental deployment

There is no way to add a new service feature to all parts of the Internet at the same time. Only schemes that can be deployed in partial incremental steps can succeed.

Minimal perturbation of the current network

Proposals that require changes to standards are difficult and slow to deploy, because of the need to get agreement on the scheme. Schemes that can exploit existing mechanisms in new ways are much easier to put into operation.

Control of abuse

The operation of the Internet today is dependent on the proper implementation and operation of software in the computers attached at the end-points. Any scheme designed to collect revenues for enhanced service will attract attention as a target for fraud and attempts to avoid payment. It is not worth deploying a scheme that is not resistant to these obvious abuses.

5. Implementing this proposed scheme

The previous section proposed a specific approach to assigning value and collecting costs for packet flows in the Internet. This section discussed how this might actually be implemented, given the existing details of the Internet design.

5.1. A framework for implementation

Each router inside the net, when forwarding the packet, must be able to determine two things easily: what enhanced service is requested for the packet, and whether the sender or some set of receivers is willing to pay for that service. Because routers are high performance devices, this determination must be easy to make, and must not require any computational complexity.

As discussed above, there are a number of proposals for describing the service desired from the network. This paper is not concerned with how this information is expressed in the packet. We return below to one relevant and very important aspect of this issue, which is how receivers can express their service-specific willingness to pay. But at this point, consider the other half of the problem, which is how to express the scope of the willingness to pay.

Each router, on receiving a packet, must be able to determine whether the router is within the sender's region of willingness to pay for that packet, the receiver's region, or neither. This information has to come from somewhere. Mechanically, there are only two ways that a router can have access to information about a flow. Either it is stored in the router, or in the packets of the flow. There is no other option.

Information about a flow stored in a router, called *per-flow state*, is in general not a preferred design approach for the Internet. Either it is put in place in the routers as it is needed, which requires a dynamic protocol for setting up the state, or it is stored as long-term tables, which might get very large, since there are tens of millions of end points today, and the net is growing rapidly. The only viable approach to the storage of long term state information in the routers of the Internet is to group users with similar willingness to pay into large aggregates, rather than describing each end-node separately.

The alternative to per-flow state is to put the information in the packet. This has the advantage that the flow is *self-describing*, but has the disadvantage that the amount of information that can be carried in a packet is very limited, since as a practical matter there are at most a very few bits in the header that could be assigned to this purpose, This implies that any practical scheme must involve some creative approximation, which allows for a reasonable flexibility in expressing willingness to pay, while still reducing the information that must be in both the routers and the packets. However, if a reasonable approximation can be devised, the avoidance of per-flow state is very desirable. So a starting point for a design should be to maximize the use of packet-borne information.

A basic (over-simplified) scheme is easy to describe. Consider a field in each packet which can have one of three values:

- Sender will pay
- Receiver will pay
- Neither will pay

The latter indicates that the router should not provide the requested enhanced service at this point in the path. If one of the first two values holds, the enhanced service is provided, and (depending on which value) the direction in which the compensation should flow is noted.

This scheme is fine, until we ask the question as to how the setting of the field is changed as the packet crosses region boundaries. As the packet crosses region boundaries, this field must be modified, and this will require additional information at that point in the network about the payment regions of the sender and receiver. Some more complex notation will be required to specify the region across which the sender and receiver are willing to pay.

One approximation for expressing a region of willingness to pay is a *zone* system. The Internet could be divided into zones within which service is provided at a uniform, distance insensitive way. A possible set of zones might be:

- Within site
- Within city
- Within region
- Across wide area in country
- End-to-end within country
- Trans-ocean
- End-to-end

Many systems use zone approximations to provide cost tiers, and it seems a reasonable approach here. In fact, the current regional payment scheme could be viewed as a very coarse zone system. The zones, however defined, would have to have meaning that is globally agreed, but in different parts of the world there could be different bindings for different zones. This paper will use this proposal, which we will call *zone based cost sharing*, as an example of a scheme that takes very few bits to express, but potentially has enough flexibility to express a reasonable range of user options.

5.2. The crux of the implementation problem

Proposing a zone system does not solve the implementation problem. There is an essential asymmetry in the network: packets flow from sender to receivers. This very obvious point has important implications. When a packet leaves the sender, the sender can put information in the packet that directly expresses that sender's willingness to pay. For example, it could contain an indication of the payment zone of the sender. But it is much harder for the packet to carry information about the receiver, for several reasons. The most obvious reason is that the sender may not know about the zone within which the receiver is willing to pay.

In the case of unicast, it is possible in principle. A control packet could be specified that carried information about the receiver back to the sender, and this information could then be put into the packets headed towards the receiver. However, for multicast, this approach does not work. There are an unknown number of receivers, potentially in the thousands. The packet does not have room in the header to carry information about each of them, and the control information necessary to carry the information about all of these receivers back to the sender could be overwhelming. In fact, it is a design goal of Internet multicast that the sender does not need to know the identity of the receivers. So, at least in the case of multicast, information about the receiver's willingness to pay cannot be carried in the packet from the source. Only in the very restricted case that every receiver is required to have exactly same willingness to pay for each service could this work, and this restriction is at odds with giving different users the ability to select different services. Instead, for multicast, the information about the willingness of the receiver to pay must be stored as state within the routers along the multicast tree, and the control messages will have to flow to these points to install this information.

At worst, this would imply that there would be a control message sent for every enhanced service flow in the network, and state stored for every node in every multicast tree in the network. However, there is a technique that can be used to reduce the need for control information. Once the packet leaves the region in which the sender is willing to pay, it initially carries no information about the payment zone of the receiver. But it still carries the indication that it requires enhanced service. As it passes through the net towards the receiver, it encounters a sequence of routers, each of which attempts to provide that service. Routers keep track of how loaded they are, and at any instant are in one of two modes, *idle* and *loaded*. An idle router is sufficiently underloaded that there is no difference in the service that different packets receive. If there is no service queue in the router, for example, there is no benefit to an algorithm that reorders the queue to favor some packets. A loaded router, in contrast, would have to shift some service degradation to some other packet to offer the enhanced service to this one.

If the router is idle, it forwards the packet, and no accounting is made of the fact that the packet asked for enhanced service. This decision corresponds to the economic principle that when a network is underloaded, the short-run marginal cost of forwarding the packet should be zero. But if the router is loaded, as the packet is forwarded according to the request for enhanced service, the router marks the packet with the zone in which this first happened. The packet then makes its way to the final receiver. At the attachment point for the receiver, there is a new component in the network, called a *receiver traffic meter*, which has been set up to know the receiver's willingness to pay. It examines the incoming packets, and sees if, for the zone and the enhanced service indicated in the packet, there is willingness to pay. If there is, the packets in the flow are allowed to continue unmodified. If, on the other hand, the receiver is not willing to pay for this enhanced service in the indicated zone, a *zone exception* control message is sent back, either to the source or to the suitable point on the multicast tree, reporting the zone within which the receiver is willing to pay for this enhanced service. This fact can then be remembered in routers as needed, or (in the case of unicast) written into future packets at the source. Thus, a control message need be sent, and a state entry recorded in a router, only in the case where there the zones of the sender and receivers do not properly overlap, or where there is a mismatch in the enhanced service that the various participants are willing to pay for.

This zone-based scheme makes use of both per-flow state in the router, as well as information contained in the packets. How poorly does this scale? In the case of unicast, the sender (once it has been informed by the zone exception message) can note the receiver's zone directly in future packets. So only the sender (or the traffic meter at the source) needs to remember the receiver zone, and the complexity there scales as the number of ongoing transfers from the sender within some time period. This is not an excessive amount of information to remember. In the cast of multicast, information must be remembered at each point in the multicast tree where the tree branches, and recipients down each branch have different zones of willingness to pay. So the zone exception control message must make its way backwards up the multicast tree until it reaches a branch point, and trigger the installation of a piece of information. While this is perhaps more complex than would be desired, it turns out that the multicast protocols already have a message with exactly this behavior, the *prune* message. And in fact the zone exception message is exactly a form of prune; it does not prune the receiver from the tree, but it prunes it from the enhanced service along (part of) that tree. So adding this new control message and router state information does not add a new form of overhead to multicast.

6. Inter-provider payments

As described, the scheme above provides a crucial piece of information that is needed at inter-provider points of connection. If a packet requests an enhanced service, then it carries, in the zone fields, an explicit indication of which end is willing to pay the service-specific costs for this flow at this point in the net. The providers can look at these indications to determine whether there is an imbalance in the expressed value flowing in each direction. The providers will still have to negotiate the price each charges the other to carry various sorts of enhanced services to/from specific end points. But the raw information as to the direction of the value flow is now present in the packets.

The cost of carrying a packet will depend on how far it is going, and thus a simple count of packets in each direction is not adequate. However, the information needed for accounting is very simple. For a packet in the region of the sender, where the compensation is flowing in the direction of the packet, the total cost of the packet depends on how much further it is going, which is determined solely by the sender zone. In the region of the receiver willingness to pay, where the compensation is flowing counter to the packet, the total cost depends on where the packet came from, which again can be determined looking only at the receiver's zone. Thus, what the two providers need to keep track of is how many packets (or bytes) marked with which zones have crossed the interface. More detailed information, such as the source and destination addresses, are not required.

7. Market Forces

If we had a scheme like this in place, how might the market actually evolve? Consider a particular example to illustrate the issues that might emerge.

7.1. Determining the costs

A claim made above is that each participant must be able to determine what the cost of a service will be. The region scheme, in principle, isolates the costs to the region of willingness to pay. But it does not necessarily make it easy to determine what the total cost will be. Consider the zone example above of "end-to-end within country". From the sender's perspective, this would cover the Internet access provider (IAP) of the sender, some number of wide area Internet service providers (ISPs), and the IAP serving the receiver of the flow at the "other end" of the network. Using the U.S. as an example, there are hundreds of IAPs serving customers, and each might set a different price to deliver a flow to an attached user. One of these distant IAPs could set a very high fee for incoming service, and the sender would then discover that he has just paid this very high fee for a transfer to that region. How can an arrangement be established, such that a sender of information across the Internet could agree to pay for service-related charges across a distant service provider, and still have some assurance of what the cost would be.

One possibility is that the wide area ISPs will find themselves in the role of rate setters, in order to create the market. An ISP will explore the price structure of the various of the IAPs, and then set some fixed amount it will be willing to pay to any distant IAP to carry flows with a given enhanced service. If a distant IAP will not carry the traffic for the rate that the ISP will pay, then the ISP will turn off the bits in the packet requesting the enhanced service. The distant IAP will then receive no payment, and the flow will not receive enhanced service across that portion of the network.

If the ISP sets too low a price that it will pay distant IAPs to carry sender-compensated traffic, then many distant IAPs will be excluded from the payment region because their costs are too high. Senders will be then be frustrated, because a large proportion of their potential receivers cannot be reached with the enhanced service they are willing to pay for. If, on the other hand, only a few high price outlier IAPs are excluded, this will be more acceptable. The ISP would have to balance these considerations to determine the service price that the ISP will pay the distant IAPs and factor into its own pricing to the senders within its own region.

This possible market structure raises a number of questions. Does this situation leads to a stable market with the right incentives. There are a number of issues. For example, how could a low cost IAP benefit from its cost structure? If the ISP offers a set price for payment to any IAP for sender-compensated service, a lower internal cost structure cannot translate into lower costs for the senders paying for flows into that IAP, except by providing a downward pressure on the set price that the ISP agrees to pay. However, a low price could lead to higher profits, and a lower attachment cost for users. This would have the benefit of attracting additional customers, which would benefit the provider. Alternatively, there might be several tiers of pricing, and a sender could agree to pay for traffic only into regions below a certain tier.

Does this give too much power to the ISPs? Could an ISP distort the market by setting a payment price that is too high or too low? Will competition among ISPs control the setting of this price in an effective way? It would seem that if one ISP set the price it would pay for remote service inappropriately, this would be corrected in the market. If the price is too high, payers will move to another ISP. If it is too low, IAPs will refuse to accept payment, and ISP will not be able to provide the offered service.

7.2. Just how far should the money flow?

A payment scheme of the sort described here provides a way for participant with high willingness to pay to cross-subsidize those who attach lower value to the service. In the extreme, an IAP could set itself up with a price structure that had very low attachment costs for users, hoping to recover most of its costs from the income derived across the network from distant participants willing to pay. This pattern would match somewhat the provision of "free" broadcast television, and might permit pricing that would attract potential users with very low willingness to pay. Is this either desirable or feasible in the Internet, using this proposal?

There are two ways that such an IAP could structure his prices. In one, the subscriber pays a monthly fee sufficient to cover the IAPs costs, and then receives a rebate as his use of distant sources brings in revenues. This structure seems unreasonable, as any scheme that pays a user to generate network load must have its incentives backwards and invite abuse.

The other approach is for the IAP to set fixed, usage-independent rates for subscribers, based on overall expected income from ISPs, and take the risk internally that there is not enough income from the attached providers to cover its costs. It seems unlikely that a IAP would take this risk in large proportion until the various providers have enough experience with usage patterns and expected levels of resulting revenues.

How could this payment situation come into existence, absent regulation? How could providers of content be persuaded to pay a high enough fee for the transfer of their content across the network that the resulting revenues subsidize the attachment costs of users? The argument would have to be that this subsidy attracts additional users to the network, who would otherwise not attach at all. It is difficult to make the case, without knowing something about the elasticity of demand from these users. However, with the current experiments in pricing for retail Internet service, some relevant experience may be accumulating.

Today, the assumption of most IAPs is that they must recover all their costs from the subscriber payments. They make no assumptions of any other income, either from regulation or from some free market structure. This assumption, if sustainable, will greatly help the market, since it removes the need for substantial fees to flow across the IAP interface. But the question of whether the market will "stall" is an interesting one. Additional sources of revenue into the IAP, from remote participants with high willingness to pay, or from some other source, such as advertising, as discussed below, might provide a means to lower attachment costs and bring a larger pool of users onto the network.

8. Other roles of pricing

Another role of pricing is to shape user behavior in socially desirable ways. Attaching some incremental cost to network usage tends to discourage low-value usage. This is viewed as a disadvantage when it discourages exploration of new uses for which the value is not clear. It is viewed as an advantage if it discouraged annoying usage such as junk e-mail. In this context, it is worth noting that some usage is sender-driven, and some receiver-driven. Junk e-mail is sender driven, so the pricing scheme would have to include a "sender-pays" mode to discourage mail sending. Casual Web browsing, which can clog servers, is receiver initiated, and could only be controlled by a "receiver-pays" mode. So if we anticipate the use of payments as a social control, both sender and receiver modes must be supported. This leave unanswered how the correct payment mode would be forced upon the user for the particular uses.

In general, it has been hypothesized that the incentives for effective and efficient use of the network are better aligned if the receiver is responsible for (at least much of) the willingness to pay at each point in the net. For example, if receivers must always pay for the cost of delivery, this naturally encourages practices such as multicast and caching. If a provider caches a Web page, he pays to receive it only once, but potentially receives payment more than once, if multiple receivers request it. This incentive is lost if the sender pays for delivery, in which case it benefits the network provider to have the page delivered as many times as possible.

There is thus a potential tension in the design space of pricing for the Internet. Receiver-based assertion of value may more naturally set incentives properly, but is harder to implement because packets, flowing from the sender, can easily carry information about the sender's willingness to pay, but not the receiver's. The zone-based cost sharing scheme described above has a higher overhead and complexity when dealing with the expression of receiver value.

9. Capturing the higher level money flows

The discussion above was in terms only of costs for network services, and did not tie this into any model of payments for higher level services and content. The normal model for payment for content has the eventual receiver of the content pay a fee to the provider, who then delivers the content to the recipient. In the terms of this paper, this would imply payment outside the scope of this mechanism from receiver to sender, and then a full willingness to pay for the transfer itself by the sender. For the Internet Service Provider, one of the potential advantages of putting a flexible pricing model in place is the speculation that by doing so, it might be able to also implement, and thus tap into, these higher level revenue flows, which may be substantial.

The zone-based cost sharing scheme is capable of expressing payment for higher level services, but in a rather restricted way. In the zone scheme, the packets carry no indication of actual payment required, but only the zone across which the participants are willing to pay. It is assumed that the cost associated with each zone can be computed in advance, and is a fixed value for each packet (or byte) sent. Not all fees for higher level services will fit this model. For certain services, one could imagine a well know fee that is fixed across a number of providers. But the zone scheme cannot easily express prices that depend on the service used, and that do not relate to the number of packets or bytes sent. Further work is required to characterize the limits of the pricing that this zone scheme can express.

10. What happens if nothing happens?

This zone based cost sharing scheme is rather complex, with new fields in packets and new protocols to allow the receivers to express their willingness to pay. This may not come to pass, and certainly will not come to pass unless there is enough anticipated benefit compared to the consequence of taking no action.

How might pricing evolve if no architectural extensions are put in place? The most obvious possibility is that service providers will offer separate networks (how the separation is achieved, and at what level of the protocol stack, is not relevant), and offer different levels of service on these different nets. Since there would be no specific flags in packets to select the enhanced service, all users of a given net would see the same service. The most obvious version of "enhancement", which is already beginning to show up, is a lightly loaded network, which simply makes "things work better". Since these networks will have different addresses, the addresses become an indirect way of expressing the scope and nature of the service and willingness to pay. All the subscribers have to accept the same scope of willingness to pay, since there is no field in the packet to convey individual preference.

For example, one provider might offer a network provisioned to offer uncongested access from the U.S. to points of attachment on the major continents. Local IAPs in other countries could then connect to this net (presumably at low cost) and the users on that net would discover that they had very responsive access to sites (such as commercial Web sites) attached to this high value network. But the network would not offer transit service back to other sites in the U.S., so another (perhaps much more congested) path would have to be used to reach sites that would not pay to attach to this high performance network.

The addresses on this "enhanced network" function somewhat similarly to 800 addresses in the phone system. Users connecting to these addresses would observe that connections "cost less", or "work better" than connections to other sites. Users might thus be tempted to use them preferentially. It might be possible to make this feature easily visible to users to convey this appeal[1]. These addresses have the additional advantage, relative to 800 numbers, in that they support a (static form of) co-payment, so that they can be used from any part of the Internet.

This sort of specialized ISP offering is bound to happen in the short run. So the interesting speculation is whether this approach will provide enough ability to allocate costs so that no further mechanism is demanded. With the simpler scheme, there is no way for a user to change the service it purchased from the network except by changing to a new network provider, which implies renumbering all the hosts of the user. There is no way to select among a range of enhanced services on a dynamic basis, so services such as real time telephony and bulk transfer cannot be differentiated. There is little likelihood that payment flow can be provided that reaches all the way from the sender to the receiver, to facilitate the class of "cross-subsidy" discussed above. But the scheme might still provide enough service and price differentiation to succeed in the marketplace.

11. Conclusions

11.1. Summary of mechanism

This paper proposes a set of mechanisms that together permit flexible pricing and cost allocation for services on the Internet. The key assumptions of the proposal are the following.

---

[1] In contrast to 800 numbers, where the meaning of the 800 prefix is well known, Internet addresses are not usually seen by users. Instead, users see higher level names, such as domain names and World Wide Web names. Those do not directly indicate the address to be used. So the sender's willingness to pay for the transfer needs to be made evident in some way other than the address prefix. Some service icon on the Web page pointing to the object is one approach.

- Both sender and receivers should be permitted to share in the payment for services. A simple approach for allowing the sender and receivers to share in the payment is to associate with each a region of the network within which they will pay. This matches the existing static payment scheme for basic network service. The approximation used in this paper is a *zone-based cost sharing* model.
- Usage sensitive payments should be associated with enhanced services, such as real time flows or flows that are assured some expected capacity in times of congestion, and not the basic service of today, which provided universal connectivity without much assurance of service level.
- Multicast, which implies that there may be more than two participants in a flow, is an important service. Any scheme should be compatible with multicast.
- Implementation of a "receiver-payment" mode is more complex than a "sender-payment" mode, because packets can more easily carry information about the sender than the receiver.

The prices being discussed in this paper are not necessarily dynamic per-flow prices. What is more likely is that providers will enter into long term contracts that set the rates that they will pay each other to offer various enhanced services. These inter-provider arrangements will specify such things as the peak loads or long term limits on the total amounts of enhanced service traffic, and other parameters, such as time of day that will permit the providers to make provisioning decisions for the service. These expectations about service levels and prices will then become part of the overall context in which individual participants in flows express a willingness to pay for service across different regions of the net. In most cases, those individual arrangements will also be long term agreements.

Other, simpler approaches to supporting users with different willingness to pay are possible. The most simple approach is the construction of different Internets with different degrees of over-capacity and different regions of coverage. These will attract different sets of customers. This approach can be understood as an approximation to the scheme being proposed here, with major limitations but the potential to be "good enough".

11.2 Support for explicit pricing schemes

The zone-based scheme assumed that the packet carried, in addition to the zone information, the indication of the enhanced service that the packet requested. There is another class of pricing scheme, in which the packet carries a bid for service, rather than describing the service parameters. These schemes include the smart market proposal of [MacKie 1993], and the priority scheme of [Gupta]. Some adjustment of these schemes may be required in order to fit them into a zone-based scheme.

The first issue is that, to be compatible with a cost-sharing model, any pricing scheme must come in two forms, which express respectively the willingness of the sender and the receivers to pay. The schemes mentioned above have been described in terms of packet fields that express the sender's willingness to pay. The second issue is that those schemes, by intention, compute the actual price to be paid for service in a very dynamic way at the points of congestion. This implies that there must be some means to reflect these charges back to the sender and forward to the receiver, and thus to relate the bid carried in the packet to the total cost, which depends on the price set at each point of congestion. In contrast, the zone scheme assumes that the cost for any specific enhanced service can be computed in advance for each zone, so that the network nodes can just count the packets or bytes going to and from each zone, and compensate each other based on the aggregated traffic flows. The sender and receiver can know in advance the cost of agreeing to a particular enhanced service across some particular zone. No per-flow accounting need be done inside the network.

11.3. Research agenda

This speculative design of a cost sharing model raises a number of questions that must be considered, in order to better understand the possible shape of the marketplace, and how to build mechanism to meet the needs.

The technical community would have a number of design tasks if this approach were undertaken. It would be necessary to articulate and agree on a standard set of *enhanced service building blocks* and define a standard way in the packet header to request these services. These could be used to create real time services, allocated capacity best effort services, and so on. An Internet *zone exception* message would have to be designed, to allow a receiver's traffic meter to request that a flow not receive some particular enhanced service, or receive in only within some zone. Another issue is the relationship between zones and the routing and topology of the Internet. It would be desirable to associate zones with addresses, based on routing information and known zone boundary points.

The economics and business community have a more uncertain set of questions to evaluate. The first step is to validate the need for a co-payment model in which the receiver as well as the sender can express his willingness to pay. Following that, the task is to explore the utility of a region scheme, and a zone scheme as an approximation to allocate shared payments, and consider what granularity of zones/regions will meet the need of the market.

Another set of questions is whether this market of providers compensating each other based on value of flow is similar to other markets that have useful or known properties? Is this a stable market with proper incentives? Will competition at the various levels cause the market to converge to a useful operating point?

Finally, a critical question is how much better this zone based scheme is than the "do-nothing" alternative. That model, with different ISPs building different networks with different service levels and different regions of coverage, is the likely outcome if no extensions to the Internet are undertaken.

12. Acknowledgments

This paper incorporates a number of very useful suggestions from Joe Bailey, at MIT. I appreciate his helpful comments. This research was supported by the Advanced Research Projects Agency of the Department of Defense under contract DABT63-94-C-0072, administered by Ft. Huachuca. This material does not reflect the position or policy of the U.S. government, and no official endorsement should be inferred.

References

ATM Forum Technical Committee. 1996. *Traffic Management Specification 4.0*.

Clark, D. 1995. "Adding Service Discrimination to the Internet", Proceedings of the 23rd Annual Telecommunications Policy Research Conference (TPRC), Solomons, MD, October 1995. Also appeared in *Telecommunications Policy*, Vol. 20, No. 3, April 1996.

Gupta, A., Stahl, D., and Whinston, A. To appear. Priority Pricing of Integrated Services Networks. In *Internet Economics*, ed. Lee McKnight and Joseph Bailey. Cambridge MA: MIT Press.

MacKie-Mason, J. and Varian H. 1993. Pricing the Internet. Presented at *Public Access to the Internet*. JFK School of Government. Available on line as ftp://ftp.econ.lsa.umich.edu/pub/Papers/Pricing_the_Internet.ps.Z

Shenker, S., Partridge, C. and Guerin, R. 1996. *Specification of Guaranteed Quality of Service*. Availiable on line as ftp://ftp.ietf.org/internet-drafts/draft-ietf-intserv-guaranteed-svc-06.txt.

Stallings, W. 1992. *ISDN and Broadband ISDN with Frame Relay and ATM*. Prentice Hall.

Wroclawski, J. 1996. *Specification of the Controlled-Load Network Element Service*. Available on line as ftp://ftp.ietf.org/internet-drafts/draft-ietf-intserv-ctrl-load-svc-03.txt.