The Political Economy of Congestion Charges and Settlements in Packet Networks

William H. Lehr
Graduate School of Business
Columbia University
New York NY 10027
wlehr@research.gsb.columbia.edu

Martin B.H. Weiss
Telecommunications Program
Department of Information Science
University of Pittsburgh
Pittsburgh PA 15260
mbw@tele.pitt.edu

September 8, 1995

Abstract

This paper examines the case for usage-based pricing in the Internet. We extend earlier work by Mackie-Mason and Varian [21] on congestion-based pricing in a single network to the case of multiple, competing carriers, and examine the problem of settlements between service providers that arises. With optimal congestion prices, it will probably be necessary to allocate revenues to the multiple, competing carriers who comprise the Internet and PSTN. The settlements mechanism that is applied may have important implications for service providers incentives to interconnect and invest in network capacity. The tumultuous history of the settlements process for local access facilities and to fund universal service in the PSTN [5] suggests that this problem will not be trivial and provides a useful background for identifying the new challenges posed by settlements in packet-based networks.

1 Introduction

The dramatic growth of Internet traffic and the expectation that ATM services will play an increasingly important role in the future of the Public Switched Telecommunications Network (PSTN) are attracting new interest in the economics of pricing for packet-based services[1]. Since the costs of these networks are largely fixed, optimal usage prices will differ from zero only to the extent that there are congestion costs. Varian and Mackie-Mason [18, 19, 20], Bohn *et.al.* [3], Cocchi *et.al.* [7], Parris *et.al.* [24] and others have proposed several approaches for implementing congestion sensitive pricing in computer networks.

In this paper, we extend the earlier work to the case in which end-to-end network service must be supplied by multiple, independent carriers who may neither have the information nor the incentive to cooperate in setting prices or preparing investment strategies that are optimal for the overall network-of-networks. We show how it may yet be possible to set optimal congestion prices using only local information on costs and traffic.

[1] This is a draft document. Please do not quote it or cite it without permissions. The authors welcome all comments.

Our analysis begins by examining the case for congestion pricing in a variety of network contexts. We will consider some of the economically relevant features of packet switched communications networks and how these differ from the environments faced by traditional cable and telephony networks. We extend the modelling framework presented by Mackie-Mason and Varian [21] based on a single network domain to the case of N networks and show how the optimal pricing solution may be decentralized. In addition, we examine the settlements problem that arises with multiple networks and consider what its implications might be for the effective implementation of congestion prices.

1.1 Congestion Pricing

Congestion pricing is necessary in many communications networks because even when there are no additional costs imposed on the carrier due to congestion, costs may be imposed on other users when the network becomes congested. In the absence of usage prices, consumers fail to take into account the full social costs of their traffic. These include the reduction in service quality that may be experienced by all subscribers as the network becomes more congested.

Historically, congestion or capacity pricing schemes have been necessary for communications networks, including telephone networks, cable television networks, and the Internet because the marginal cost of service is near zero. Uniform marginal cost pricing will not allow service providers to recover their costs and is inefficient. This has lead to wide use of non-linear pricing strategies in communications networks. These usually take the form of multipart tariffs that include separate charges for access and usage. Flat access fees can be used to recover network fixed costs, while usage fees can serve to recover variable service costs.

In the absence of capacity constraints, one user's traffic is unlikely to affect another user and so optimal usage fees should be near zero when the variable network costs of usage are near zero as is the case in most electronic communication networks. In the face of capacity constraints, however, bandwidth becomes a scarce resource and increased congestion may reduce the quality-of-service offered to customers whose traffic is blocked, delayed or experiences higher error rates. User traffic is no longer independent and while the direct variable costs to the network provider may remain unchanged, there may be large social costs associated with increased usage.

There are a number of solutions available for allocating scarce bandwidth among competing users. One of the most obvious is "first come, first served". In traditional connection-oriented telephone and cable TV networks, each customer receives a fixed allocation of bandwidth until capacity is exhausted. Later calls are blocked. While simple to implement, this strategy does not discriminate among traffic that may differ widely in its value to customers. This can lead to an inefficient allocation of bandwidth and can encourage wasteful investments by customers who must compete for the scarce bandwidth. High value uses may be driven to invest in private networks in order to guarantee access which could result in higher costs for those who are left on the public network.

A centralized call-admission or traffic-control policy could control this directly, but this would require too much information regarding the exact nature of consumer demands. One obvious alternative is to offer priority pricing: higher prices for higher quality-of-service and preferential access to bandwidth. This induces consumers to self-sort their traffic in order of value which can result in significant benefits to both classes of subscribers. Another alternative is peak load

or congestion pricing where users are charged prices that vary with time and the availability of resources. When capacity is scarce, prices should be higher to reflect the increased social costs of congestion. Telephone networks implement a version of this in the form of off-peak discounts for evening and weekend calling[2].

Specifying the appropriate congestion price makes it possible to decentralize decision-making by forcing subscribers to internalize the full social costs (i.e., excess congestion) imposed on all subscribers to the network. Below, we show that with appropriate assumptions, it may be possible to compute these prices using only knowledge about local demand and capacity cost conditions. While the rationale for positive congestion prices is derived from the negative impact congestion may have on all users of the network-of-networks, it is not usually necessary to know individual responses to increased congestion in order to set prices. This is important since the individual responses to congestion are not directly observable.

While these results are explained in more detail by MacKie-Mason and Varian [21], several additional points should be made, however:

1. Even at the optimal capacity, there will be congestion and congestion revenues will be collected. If the market is perfectly contestable (i.e., firms do not have market power), then the congestion revenues will help offset the network capacity costs, thereby enabling lower access fees than would be required in the face of pure capacity pricing.

2. The problem becomes significantly more complicated when one considers an environment with multiple interconnected semi-autonomous networks. This forces one to think much more carefully about what congestion means in a "network-of-networks" and how the type of traffic (on-net or internet) may influence the degree of congestion.

3. There is the additional problem of settlements with respect to how usage, and potentially, access revenues should be distributed to the collection of sub- networks. In a dynamically-stable long-run equilibrium, each of the sub-networks must recover sufficient revenues to recover its costs. In general, this will require transferring revenues among the carriers and the basis chosen for mediating these transfers will certainly be contentious. While it is logically possible to separate the mechanism chosen to establish congestion prices (which will require the sub-networks to share information regarding their congestion status) and the settlements process for distributing the pool of revenues, in practice these two are likely to be closely linked. Moreover, they are likely to be influenced by the strategy used to settle access revenue.

1.2 Why are Packet Networks Different?

Before examining the theory of congestion pricing in a network-of-networks more carefully, it is worth considering how packet networks in general (and the Internet in particular) differ from other types of electronic communication networks and what this means for pricing. This is accomplished most easily by comparing the Internet to traditional telephone and cable networks. Furthermore, it is easiest to understand how these differ by focusing first on the nature of the customer need

[2]When traffic patterns are relatively predictable, peak load prices, such as those used in telephony, are possible. When the congestion is unpredictable, dynamic prices are necessary.

they seek to satisfy and then considering how this is reflected by the choice of technology, industry structure and pricing mechanism.

Network Type	**Technical Characteristics**	**Congestion Management**	**Economic Characteristics**
Telephone	Two Way	Call Blocking	Market Power
	Dedicated Bandwidth		Common Carrier
	Narrow Bandwidth		
	Dominant Application		
	Connection-Oriented		
Cable Television	One Way	N/A	Market Power
	Broadband		Integrated Content
	Broadcast		
Internet	Two Way	Delay	Government Support Legacy
	Mixed Bandwidth		Integrated Content
	Mixed Application		
	Connectionless and Connection-Oriented		

Table 1: Characteristics of Different Network Technologies

Table 1 provides an overview of some of the key characteristics of the different types of public networks. While this table captures some of the relevant features, it does not capture all. Because of the two-way nature of the voice network and the Internet, for example, there have been strong positive externalities associated with ubiquitous interconnection[3]. While competition has been emerging in many of the telephone markets, the legacy of market power remains. Furthermore, the historic separation of carrier and content is weakening in the present legal and legislative environment, making some of the distinctions less clear cut.

The threat of denying universal termination by a dominant carrier has provided strong justification for regulation of interconnection and access pricing, which in turn, creates the settlements issue. MacKie-Mason and Varian [21] only consider market power within the context of a single network. The debates over settlements and access pricing in telephony suggest that this becomes much more complex in multi-carrier environment. Since high congestion prices may be as effective in denying interconnection as a flat out refusal to interconnect, the opportunities for strategic manipulation of pricing when sub-networks have market power become much more complex.

The implication of packet switching *vs.* circuit switching means traffic effects are more global. The response of the packet networks (such as the Internet) has been to distribute congestion costs to all users in form of increased average delay rather than by blocking calls[4]. Once a network

[3]The emphasis on universal service and interconnection have been prominent in the recent debate in the U.S. Congress over telecommunications reform.

[4]In circuit switched networks, the quality of calls in progress is unaffected by congestion because, once estab-

is segmented into multiple networks, then we cannot be sure that the network that collects congestion revenue is where congestion revenue should be optimally allocated[5].

In the Internet, there has also been a mindset against usage-based pricing [17]. The well-documented growth of Internet users and traffic will make Internet face all of the challenges of traditional telephony (i.e., maintaining ubiquitous interconnection among semi-autonomous, profit-seeking firms), cable (i.e., integrated ownership of capacity and content with resulting confounding of pricing incentives) in a much more technically complex environment (mixed broadband and narrowband, connection-oriented and connection-less traffic). Confronting these issues now is therefore essential if pricing is to be properly incorporated into future network (and industry) structures.

2 Relevant Technical Issues

The economic analysis depends heavily on some of the key technical details of the reference network. We will describe the essential features of this network to clarify the economic arguments. In particular, we consider packet switched networks as opposed to circuit switched networks. Packet switched networks separate the information generated by a user into small blocks, or packets, that traverse the network independently of each other. They may do so in a *connection-oriented* way, or a *connectionless* way. In connection-oriented networks, all packets associated with a particular source and destination pair follow the same path through the network, over a *virtual circuit*, which must be established before information can be transmitted. In a connectionless network, no connection exists, and each packet can take a different path through the network.

2.1 Network Services

Telecommunications networks are designed to provide services to end users. A service is a collection of features that provide a communications capability to an end user. A service normally has a quality of service associated with it. From a technical perspective, quality of service includes throughput, packet loss rate, bit error rate, total transit delay, and *delay jitter* for a particular service. Delay jitter is the variation in transit delays between subsequent packets that arrive at the destination. An end user is a human being or a software program that uses a communications service.

2.2 Technological Overview of Packet Networks

Table 2 illustrates that different services require different levels of service quality. Generally, the service quality of a network is determined by the capacity of the network. A network consists of packet switches, links between packet switches, and network management/administration systems. While it is difficult to characterize the "capacity" of a network, one important set of controls

lished, a circuit switched network provides dedicated bandwidth to the user. When such a network is congested, only *new* calls are blocked.

[5] One model is that the Internet Service Provider's network bills the sender for connections and then (presumably) distributes revenues appropriately to the other service providers that the sender used for the connection. That is not the only model.

Service	Quality of Service Parameter		
	Delay	**Delay Jitter**	**Error Rate**
Voice	Low	Low	Moderate
Video	Moderate	Low	Moderate
`telnet`	Moderate	High	Low
`ftp`	High	Moderate	Low

Table 2: Quality of Service Requirements for Selected Services

over the quality of service in networks that engineers (and administrators) have is by managing the memory and processor speed of the packet switches, and speed of the links interconnecting the network[6]. In general, more memory reduces packet loss rate (hence the overall error rate) because fewer packets are dropped due to full memory buffers. Higher speed communications links and processors reduce the delay and delay jitter because packets spend less time in the memory buffers of the processor than they would using slower links and slower processors. In general, we can assume that the switch treats the buffer as a first-in, first-out queuing system[7].

The packet switches and the network management systems together perform the routing function. Routing is the process of directing packets or connections through the network in a way that is most efficient according to a range of possible criteria. Routing can be performed using static or dynamic techniques. In a static routing, routes between nodes are pre-computed using one of many available routing algorithms; these routes are stored in the form of routing tables in the packet switches. In a dynamic technique, the routing decision at each switch is made based on information about the current state of the network.

An individual system operator (or service provider) can control service quality by call admission control and packet flow control[8]. In call admission control, a service provider determines whether the network has the transmission and switch resources to handle a call with the quality of service requirements of the user/application. If the resources exist, the call is admitted; if not, the call is rejected. Further congestion control is often still required at the packet level because (1) traffic flows are stochastic and users sometimes exceed their announced bit rate and (2) non-real time and "best effort" connectionless traffic is sometimes carried by networks in addition to the real time traffic as a lower priority traffic class.

[6]The selection and "tuning" of the network's protocols also plays an important role in determining service quality.

[7]If the switch supports multiple service classes, then the buffer may be represented by several separate buffers, each devoted to a service class, and each operating on a first-in, first-out basis. Bertsekas and Gallagher review the formal analysis of these systems [2, Chapter 3]

[8]Call admission control implies a connection-oriented network service, because the concept of a "call" is meaningless in a connectionless network. Most telecommunications researchers agree that a connection-oriented network will be necessary to support real time, constant bit rate traffic such as voice or video.

2.3 Implications for Economics

From the point of view of economics, there are several general features to observe. In general, these pertain to the concerns about the economic implementation of some details.

An operator of a packet-switched network generally maintains several switches that are interconnected internally, as well as several that are connected to either users or other networks. In order to reduce costs, the operator seeks to optimize the investment in processors, memory for input and output buffers, and communications lines. One way to characterize this optimization is to minimize costs subject to being able to meet quality of service requirements. These costs are reduced by using as little memory as needed, processors that are just fast enough, and communications lines that are just fast enough to handle the anticipated traffic. While the operator can optimize the network's operation internally, it is generally not possible to optimize the delay on an end-to-end basis if several independent network operators are involved in the provision of end-to-end service to the user.

This is potentially a problem for real time networks, where performance guarantees are necessary. Ferrari and his collaborators [14], Field [16] and many others have been developing this theory. They have been implicitly assuming that the "nodes" cooperate; in a competitive network structure, it is less clear how an end-to-end performance requirement will be efficiently allocated among several competing subnetworks when each subnetwork is optimizing locally (see for instance, [11, 32]).

Different network service technologies result in different information flows among the various network service providers. For example, it is generally not meaningful to consider call admission control procedures for connectionless networks because there are no "calls" or "circuits" that are established. The specific implications relate to the nature of the information available to individual subnetworks. In competitive market structures, this information and any information asymmetries might be exploited by competing networks. The information flows described here also speak to the pricing and settlements problem, as we will discuss in Sections 4 and 5.

2.3.1 Connectionless Networks

As we observed above, each packet that makes up a message can take a different path through a network in connectionless networks. Individual packets may therefore exhibit a significant variation in network service quality depending on the route taken by a packet and the state of congestion of the intermediate nodes. An end user only has the opportunity to observe this delay variation; she does not even know the overall level of delay without some kind of synchronizing message from her peer at the source.

In many connectionless packet services, each individual node could observe the source addresses and the destination addresses, and could calculate the rate at which packets from specific sources and to specific addresses are being carried. The service provider could also calculate the packet rate with each of its peer networks. In any case, no service provider is able to observe the delay added by any other service provider[9]. In the essentially cooperative TCP/IP networks, congestion can be observed via the Internet Control Message Protocol (ICMP) messages [8, Chapter

[9]It is possible to estimate this information using tools such as `traceroute` (see also [4]), but these require explicit actions on the part of users or network service providers, and may be detectable and blockable by other network service providers.

9]. While TCP/IP continues to be used in the privatized Internet, it is distinctly possible that the cooperative nature of the "old" Internet service providers may be tempered by the economic motivations[10].

As a result, the information on service quality that an individual network service provider has is generally local. Furthermore, it is generally not possible for the user to identify sources of delay without explicitly asking for that information from the service providers. This gives service providers an opportunity although perhaps not a motive to misrepresent their state of congestion. These information asymmetries should be considered explicitly in the design of network services, similar to the work Shenker did for the user/network interface [26].

2.3.2 Connection Oriented Networks

The relevant economic features of connection oriented networks are different than for connectionless networks. In connection oriented networks, virtual or physical connections must be established prior to any information transfer. For real time networks, where minimum performance standards exist, the connection establishment process often requires some performance guarantees[11]. Congestion management in such networks is frequently performed using some form of call admission control.

In addition to being able to observe traffic flows, under call admission control network service providers can also get explicit clues about congestion in each others networks[12]. When a user wishes to establish a connection, a path must be found through the network that can meet the user's performance requirements. In networks under a single administrative domain[13], this might be done by the network's control system. Under multiple administrative domains, the connection must be established through a communications process between the domains' network control systems.

For non-real-time networks in a single administrative domain, many techniques have been proposed (see [2, 28] for summaries). Some of the existing research can be readily extended to fit multiple administrative domains for non-real-time networks. However, techniques for real-time networks under a single administrative domain are still being proposed and discussed, and consensus has not yet emerged (see [16] for a summary).

For real time networks, the multiple administrative domain problem becomes more difficult because of the need to manage end-to-end network performance. As summarized by Field [16] numerous approaches have been proposed to achieve this. Since there is no generally accepted technical approach for real time environments, it is difficult to precisely formulate the economic problems associated with these environments.

[10] Shenker has performed this sort of analysis for the user-network relationship [26].

[11] This concept is articulated by Ferrari [15, 14]. Many implementations have been proposed; see, for example, Field [16]. Most researchers to date have considered a network under a single administrative domain and do not account for multiple competitive providers.

[12] This is particularly relevant for real time networks. In non-real-time connection-oriented network types, such as X.25 and Frame Relay, the explicit congestion signals may be less clear since no quality of service information needs to be exchanged between the networks.

[13] For a single administrative domain to exist, a single service provider is a virtual necessity. A single service provider may even choose to have multiple internal administrative domains.

The economic features of connection oriented networks depend on the the way in which this communication takes place. Some possibilities for this are:

1. All networks constantly provide current congestion price information to all other networks. An originating network then selects a least cost path (based on congestion prices and quality of service requirements) through the networks. If the congestion prices change during the duration of the connection, the network may reroute the call (at an overhead cost) or stay with the possibly more costly route.

2. The originating network may have a preferentially-ordered set of networks for each destination, and may solicit a connection sequentially as defined by the ordered set over a private channel. For example, each node may have a set of routing tables that define paths between endpoints. These routing tables may be pre-computed.

The response of a network to a request for a channel provides information about the state of congestion of the network. If a network refuses a connection, or announces a high price, it can be inferred that the network is congested.

The knowledge of congestion information may be symmetric or asymmetric. In Case 1, there is a possibility of symmetric congestion information. In all other cases, knowledge of congestion information is asymmetric because the originating network collects specific congestion information about other networks without having to reveal its own congestion level.

3 Congestion Pricing for Interconnected Networks

In considering the problem of settlements in a network of networks, it is necessary first to consider the problem of congestion pricing in such a network structure. Only when we understand congestion prices in this more general problem can we consider the question of settlements among network service providers. In fact, these two issues are not as separable as this suggests because the settlements mechanism affects the profit, and hence social welfare computation. We begin our discussion by considering congestion pricing with zero settlements.

An analysis of congestion pricing in a single network can be found in MacKie-Mason and Varian [21]. They assume that all traffic is homogeneous and that only the originator of a connection benefits, but that increased network traffic leads to increased congestion and a lower quality of service for all subscribers[14] Therefore, they can characterize the utility of each subscriber as increasing in the quantity of traffic originated and decreasing in the level of network congestion that is increasing in the level of network utilization. Network utilization is measured as $Y = X/K$, where X is aggregate traffic and K is network capacity. In their framework, it is relatively straightforward to demonstrate that the efficient uniform congestion price which forces subscribers

[14] We follow Mackie-Mason and Varian in neglecting the benefit that may accrue to the called party. While this may be appropriate in some contexts (e.g., a remote `telnet` session), it is clearly not appropriate in others (e.g., e-mail or voice telephony). The benefits that accrue to the recipient of a connection are a positive externality of connection originations which, ideally, should be reflected in the social welfare accounting and will at least partially offset the negative congestion externality. By ignoring these benefits we are able to determine unambiguously that the externality from increased calling is negative. Srinagesh notes that one of the rationales for zero settlements among Internet service providers was the benefits that accrue to both call originators and recipients [29].

to internalize the congestion costs they impose on other subscribers is a function of aggregate demand, X, total capacity costs, $C(K)$, and network capacity, K. From a practical perspective, there are three main points that we wish to highlight here:

- Congestion is difficult to define in real networks since not all traffic is homogeneous, therefore we do not interpret Y as utilization but as a measure of congestion such that $Y(i, \mathbf{X}, \mathbf{K})$ is a function that may vary by subscriber, i, and increases when traffic increases (vector $\mathbf{X}$) and decreases when capacity expands (vector $\mathbf{K}$). This is straightforward extension to [21];
- Once it is clear that different types of traffic have potentially different implications for congestion, optimal congestion charges will differ so that more congesting traffic is charged a higher price (i.e., priority congestion pricing is needed); and,
- users whose utility drives the need for high congestion charges may not be attached to the part of the network that is congested. This is not a problem in a single network.

.

We extend MacKie-Mason and Varian's model to the case of M network domains. Once we have two or more networks, it is no longer clear how one should measure the quality-of-service or congestion experienced by a subscriber. In principle, we might expect it to vary depending on the type of calls made by a subscriber (i.e., on-net or internet), the route followed by the call and the capacities of the various sub-networks. We generalize the Mackie-Mason and Varian framework as follows:

- Let there be M networks, each of which has N_j subscribers. A type "ij" call originates on network i and terminates on network j. Assume each consumer makes a unique type of call and that the N_{ij} ($N_{ij} \in \{0, 1, 2, \ldots\}$) consumers who make type "ij" calls have identical preferences. This will simplify the notation considerably[15]. Note that $N_j = \sum_{i=1}^{M} N_{ji}$. An "on-net" call is type "ii"; all other calls are internet calls.
- Let x_{ij} be the volume of type "ij" calls. Assume that there is a unique route for each type of traffic through the network and let $R(ij)$ be the subset of networks that are included in the route of call "ij"[16].
- Let $U^{ij} = U^{ij}[x_{ij}, Q^{ij}]$ be the utility of a consumer who makes type "ij" calls. Assume that $\frac{\partial U^{ij}}{\partial x_{ij}} \geq 0$ and $\frac{\partial U^{ij}}{\partial Q^{ij}} \leq 0$, i.e., consumers value increased calling but dislike the lower quality that comes with increased congestion[17];

[15] This is not a very restrictive assumption for several reasons. First, if the fixed access fees are zero, then a "real world" user who uses the network for many destinations can simply be represented by multiple "model" users without loss of generality. Furthermore, it is likely the case that the demand from the "real world" user for traffic to various destinations is each subject to different demand. If access fees are non-trivial with respect to congestion fees, this representation becomes more problematic.

[16] This applies most directly to connection-oriented traffic, although the theory could be extended to include connectionless traffic if $R(ij)$ applies only for a short time interval.

[17] We will assume that the individual consumer neglects the effect his traffic has on overall congestion since N_{ij} is large.

- Let Q^{ij} provide a measure of the quality-of-service experienced by "ij" callers. This could be measured in a wide variety of ways such as the level of average delay, the maximum potential delay, the bit error rate, delay jitter, blocking, or some weighted average of all of these. We will assume here that it refers to the level of network congestion. In general, we might expect it to be a weakly increasing function of each type of traffic and a weakly decreasing function of each network's capacity.

  We further specialize the analysis by assuming that congestion is measured in terms of the average end-to-end delay and that this is simply the sum of the average delay expected at each switching node along the route from origination to termination, or,

$$Q^{ij} = \sum_{k \in R(ij)} D[Y_k] \tag{1}$$

  $D[Y_j]$ is the average delay on the j^{th} network, which is a function of network j's utilization, $Y_j = \frac{X_j}{K_j}$. K_j is the capacity of the j^{th} network. We assume that this is a monotonically increasing function of network utilization.

- Let $X_j = X_j^{or} + X_j^{te} + X_j^{tr}$, be the total traffic that is carried by network j, where the following definitions hold:
  - $X_j^{or} = \sum_{i \in M} N_{ji} x_{ji}$ is the traffic that *or*iginates on j;
  - $X_j^{te} = \sum_{i \in \{M \neq j\}} N_{ji} x_{ij}$ is the traffic that *te*rminates on j (but originates elsewhere);
  - $X_j^{tr} = \sum_{kl \in L(j)} N_{kl} x_{kl}$ is the pure *tr*ansit traffic that passes through network j; $L(j)$ is the subset of traffic on network j that neither originates nor terminates on j.

  Note further that $X_j^{on} = N_{jj} x_{jj}$ is the on-net traffic that originates on network j and $X_j^{off} = X_j^{or} - X_j^{on}$ is the internet traffic that originates on j. $X = \sum_{j=1}^{M} X_j^{or}$ is the aggregate traffic originated on all networks[18].

- Assume two-part tariffs and voluntary participation and that the "sender-pays", so that the consumer surplus realized by consumer ij is $U^{ij} - p_{ij} x_{ij} - T_j \geq 0$ in equilibrium, where p_{ij} is the total congestion charge for a type "ij" call and T_j is the fixed access fee for network j.

- Let $C^j(K_j)$ be the fixed cost of capacity for network j; and $\sum_{i=1}^{M} C^i(K_i)$ are total network costs;

- Assuming zero settlements, the profit of the j^{th} network service provider can be computed as:

$$\Pi^j = N_j \, T + \sum_{i=1}^{M} (N_{ji} \, x_{ji} \, p_{ji}) - C^j(K_j) \tag{2}$$

[18] Note that this is *not* all traffic carried by all networks, since some traffic is internet traffic, and may transit multiple networks.

- Total social welfare can be computed as:

$$W = \sum_{i=1}^{M}\sum_{j=1}^{M} N_{ij}(U^{ij} - p_{ij}x_{ij} - T_j) + \sum_{j=1}^{M} \Pi^j \tag{3}$$

As in Mackie-Mason and Varian, one finds the optimal congestion prices from inspection of the first order condition for maximizing social welfare with respect to each type of traffic, x_{ij}. This yields a series of equations of the form:

$$\frac{\partial W}{\partial x_{ij}} = 0 = \frac{\partial U^{ij}}{\partial x_{ij}} N_{ij} + \sum_{lk \in \{lk \neq ij\}} \frac{\partial U^{lk}}{\partial Q^{lk}} \frac{\partial Q^{lk}}{\partial x_{ij}} N_{lk} \tag{4}$$

The second term is the negative externality imposed on other network subscribers from increased congestion when type "ij" consumers increase their calling. In order to induce a type "ij" subscriber to internalize the effects of her calling, congestion prices should be set so that:

$$p^*_{ij} = -\sum_{lk \in \{lk \neq ij\}} \frac{\partial U^{lk}}{\partial Q^{lk}} \frac{\partial Q^{lk}}{\partial x_{ij}} N_{lk} - (N_{ij} - 1)\frac{\partial U^{ij}}{\partial Q^{ij}} \frac{\partial Q^{ij}}{\partial x_{ij}} \tag{5}$$

The first term on the right side of Equation 5 represents the congestion effect of type "ij" traffic on all others in the network, while the second term is the effect of type "ij" traffic on other type "ij" traffic. Substituting further for Q^{ij} in (5) and re-arranging yields

$$p^*_{ij} = -\sum_{j=1}^{M} \frac{D^j_Y}{K_j} \sum_{j \in R(lk)} N_{lk} U^{lk}_Q \tag{6}$$

where $D^j_Y = \frac{\partial D(Y_j)}{\partial Y}$ and $U^{lk}_Q = \frac{\partial U^{lk}}{\partial Q^{lk}}$. Note that, since network utilization may vary, we cannot assume that the marginal increase in delay is constant for all networks. Therefore, we retain the j superscript to remind ourselves that D_Y ought to be computed for each network. If we further assume that network service providers earn zero profits (i.e., that the markets are contestable [1]), then we can compute the optimal access charge incorporating the optimal values for X, p and K into the service providers' profit functions.[19]

With a single network as in Mackie-Mason and Varian, this reduces to,

$$p^* = -\frac{N-1}{K}\frac{\partial U}{\partial Q}\frac{\partial D}{\partial Y} \tag{7}$$

In the case where M=2, there are only four types of calls: "11" and "22" on-net traffic; and "12" and "21" internet traffic. We can use the above formulas to compute the optimal congestion prices for the three types of traffic as follows:

$$p^*_{11} = -\frac{D^1_Y}{K_1}\left(N_{11}U^{11}_Q + N_{12}U^{12}_Q + N_{21}U^{21}_Q\right) \tag{8}$$

[19] One must check that at with this access charge and congestion price that each consumer's surplus is weakly positive such that participation is not an issue. We assume that this is the case.

$$p_{22}^* = -\frac{D_Y^2}{K_2}\left(N_{22}U_Q^{22} + N_{12}U_Q^{12} + N_{21}U_Q^{21}\right) \tag{9}$$

$$\begin{aligned} p_{12}^* &= -\frac{D_Y^1}{K_1}\left(N_{11}U_Q^{11} + N_{12}U_Q^{12} + N_{21}U_Q^{21}\right) \\ &\quad -\frac{D_Y^2}{K_2}\left(N_{22}U_Q^{22} + N_{12}U_Q^{12} + N_{21}U_Q^{21}\right) \\ &= p_{21}^* \end{aligned} \tag{10}$$

From inspection of these prices it should be obvious that the optimal congestion price for internet calls should be exactly equal to the sum of the congestion prices for on-net calls. This is intuitively obvious because an internet call congests both networks, whereas an on-net call congests only the network that carries it. This result generalizes to the case of N networks: to find the optimal congestion price for a call "ij", one should add the optimal on-net congestion prices for each node along the route (i.e., for the subset of networks in $R(ij)$).

When the above pricing results are combined with the first order conditions used to compute the welfare maximizing levels of capacity for each of the M networks, we obtain the following relationship:

$$\begin{aligned} \frac{\partial W}{\partial K_j} = 0 &= \sum_{j \in R(lk)} N_{lk} U_Q^{lk} \frac{\partial Q^{lk}}{\partial K_j} - \frac{\partial C^j(K_j)}{\partial K_j} \\ &= -\frac{D_Y^j}{(K_j)^2} \sum_{j \in R(lk)} N_{lk} U_Q^{lk} - \frac{\partial C^j(K_j)}{\partial K_j} \end{aligned} \tag{11}$$

or,

$$p_{ii}^* = \frac{\partial C^j(K_j)}{\partial K_j}\frac{K_j}{X_i} \tag{12}$$

This is analogous to the result in Mackie-Mason and Varian, and shows that it is possible to compute the optimal on-net congestion charge based on local information (i.e., without direct knowledge of the utility functions for all N subscribers). As long as each sub-network charges each packet it carries p_{ii}^*, the total congestion revenues collected by network i will provide the proper signal for when to expand capacity (i.e., when congestion revenues exceed the value of the sub-network's capacity valued at the marginal cost of additional capacity).

Two points are worth noting about this solution. First, the optimal solution requires that internet traffic should face higher end-to-end congestion charges because it results in more congestion per minute than does on-net traffic. In general, each type of traffic that has a different impact on overall congestion should face a different end-to-end congestion price. This is a form of "congestion priority pricing" which is analogous to other priority pricing schemes in its intent but is motivated by a slightly different need. In the traditional approach, priority pricing is used to effect an efficient allocation of scarce capacity: those who are less congestion sensitive accept a lower quality of service in return for a lower price. In the example cited above, it would be optimal to charge different rates for internet and on-net traffic even if all consumers were identical.

The second important point to note about the above solution is that the sub-networks will need to account for all of the traffic that passes across their network in order to set efficient local congestion prices, and subscribers will have to be billed for the sum of these prices along the least cost route. One solution is to have a "pay-as-you-go" billing scheme, where each network charges each packet handled its on-net congestion price and bills the consumer directly. Alternatively, the customer could be billed by the originating network, but then the originating network would need to know what the sum of the congestion prices are along the rest of least cost route (i.e., $p^*_{ij} - p^*_{ii}$) in order to set the appropriate price for a type "ij" call.

If there are at most two networks involved in every internet call (i.e., there are no transit networks), networks could bill each other for terminating calls.[20] This would provide each sub-network with the information about the appropriate termination charge for a call and the total congestion revenue collected would provide an accurate signal of whether it was advisable to expand capacity.

Another solution is to have the networks continuously update each other regarding their congestion charges which would allow the originating network to compute p^*_{ij} directly. This may be the case in a least cost routing environment. If routing is hop-by-hop, then the appropriate congestion charge could be passed back up the chain if each node billed traffic the sum of its on-net cost plus the cost charged to terminate the call at the next link in the chain. For example, in a call that will be routed from 1 to 2 to 3, network 2 should charge network 1 $p_{22}{}^* + p_{33}{}^*$ which will allow network 1 to compute the appropriate end-to-end charge without direct knowledge of network 3's congestion status.

In all of these solutions, it is possible for the networks to exchange the required information in the form of traffic accounting data without actually making what might amount to sizable revenue transfers in both directions. However, it is important for the networks to account for the congestion charges associated with terminating or transmitting traffic that originates on other networks. A naive application of the local congestion pricing formula defined above may result either in on-net congestion prices that are too high or the failure to invest in adequate network capacity when such investment is appropriate.

In the telephony context, Brock [6] argued why termination charge settlements are unlikely to be very important. He first argued that the marginal network cost of termination is small, and even when this is not the case, it is only necessary to settle the net balance of traffic between two networks (which will be zero if the networks terminate reciprocal volumes of traffic). Neither of these arguments applies in the present case since Brock was concerned with settlements motivated by the need to recover network termination costs, which we are assuming are zero. The justification for settlements here is to provide information to the sub-networks to enable them to make efficient pricing and capacity expansion decisions.

Since there does not need to be an actual transfer of revenues in order to implement efficient decentralized congestion pricing, it may seem as if there is no settlements problem. Such a conclusion is wrong, however, as will be discussed in the subsequent section[21].

[20] With three networks, the pure transit network could bill the customer and then pay the originating and terminating congestion charges. A version of this occurs in long distance telephone when the long distance company pays the originating and terminating local exchange carriers a per minute access charge.

[21] Before proceeding, we should note that these results depend heavily on the specialized assumptions that underlie both the earlier paper by Mackie-Mason and Varian and the extension to N networks discussed here. In

4 Optimal Congestion Prices and Settlements

To understand why a settlements problem arises in a network-of-networks, it is sufficient to consider a very simple example with just two networks. Assuming no settlements, optimal congestion prices, and origination-network billing, each network will earn profits of:

$$\Pi^1 = N_1 T_1 + X_1{}^{on} p_{11}{}^* + X_1{}^{off} p_{11}{}^* + p_{22}{}^* - C^1(K_1) \quad (13)$$

$$\Pi^2 = N_2 T_2 + X_2{}^{on} p_{22}{}^* + X_2{}^{off} p_{11}{}^* + p_{22}{}^* - C^2(K_2) \quad (14)$$

If the network-of-networks is to recover its costs without external subsidies then the sum of the profits of the constituent networks must be weakly positive. This would be the case either if the markets for network services is contestable (free-entry) or if there is rate of return regulation subject to the constraint that industry profits are non-negative. In the absence of settlements, the profits of each network must be weakly positive. This imposes a stronger constraint on the optimization problem and may require distorting the optimal solution in order to be satisfied. Setting $\Pi^1 = 0$, substituting for the efficient congestion prices and re-arranging yields the following result (which is analogous to the result in Mackie-Mason and Varian [21]):

$$\frac{T_1 N_1}{C^1(K_1)} = 1 - \frac{\partial C^1(K_1)}{\partial K_1} \frac{K_1}{C^1(K_1)} + \frac{p_{11}^* X_2^{off} - p_{22}^* X_1^{off}}{C^1(K_1)} \quad (15)$$

The left hand side gives the share of network costs which must be recovered via the flat access fees in order for the network to recover its costs. The second term on the right drops out if there is only one network, or if traffic flows are balanced and the optimal congestion prices are identical. In either case, the share of network costs which are recovered via the flat access fee increases towards one as the ratio of marginal to average capacity costs goes to zero. In the multiple network case, it is unlikely that traffic flows would be identically balanced or that the optimal congestion prices will be equal.

In the fully symmetric case with equal numbers of on-net and internet callers and identical costs for each network, the optimal congestion prices, access fees, traffic and capacity for each network will be identical. There will not be a settlements problem. Consider what happens, however, if the subscribers are distributed asymmetrically such that a larger share of the internet callers are located on network 1. Under our assumptions, the network congestion caused by a call depends on the route followed but not the direction of the route (i.e., call "12" causes the same congestion as call "21"), so this change should not affect the optimal access and congestion charges faced by consumers.[22]). Under the original solution, however, network 2 will fail to recover its costs.

---

a more general model, we might not expect network costs to be separable as assumed here. Also, we might expect much more complex interactions among different types of traffic and capacity in the determination of call-specific congestion. Furthermore, computing the least cost route when alternate routing is feasible is likely to be quite difficult since it amounts to optimally routing traffic so as to minimize congestion costs.

[22]We are assuming here that the level of network capacity costs depends on traffic patterns and not on the number of subscribers. Although in general, we might expect network costs to depend both on the number of subscribers and the capacity, K_j, which itself may depend on the number of subscribers, this need not be the case for several reasons:

- K_j refers to the capacity that is relevant for determining the level of network congestion; this might be the size of the switch, which may depend largely on X and not on the number of subscribers that generate X;

In the absence of settlements, there are a number of approaches that may be used to resolve this problem.

1. If participation is not an issue, we could allow asymmetric access charges, with network 2 charging an access fee that is sufficient to recover its higher costs. [23] While this solution may be efficient, it may not be perceived as equitable. One could argue that it is unfair that consumers on network 2 face higher access charges since consumers on network 1 also benefit from the reduction in overall congestion when network 2's capacity expands.

2. If we constrain ourselves to uniform access pricing, it may still be possible to implement the efficient capacity and congestion pricing solution by charging higher access fees to all subscribers. In this case, we would need to prevent entry competition for network 1 since it will earn positive profits at p^* and the new, higher T^{**}.

3. If we constrain ourselves both to free-entry and to uniform pricing, then it will generally be optimal to modify both usage and access fees, and in general, we will not be able to achieve the same level of total surplus as in the unconstrained problem. This problem arises because in a zero-profit equilibrium, it is possible that sizable congestion revenues will be collected from subscribers in order to induce them to properly internalize the welfare implications of their originating-call behavior. These congestion revenues will permit firms to charge lower access fees than would be necessary in the absence of congestion charges, but the sum of these congestion charges and access fees may be insufficient to recover the costs of all of the networks in the optimal solution. In a "sender-keep-all", "no settlements" world it would be possible for an uncongested, upstream network that originates a disproportionate amount of traffic to collect most of the congestion revenue.

While this illustrates that a settlements problem may exist in multiple networks, it does not propose solutions. The design of these mechnisms is complex and beyond the scope of this paper. Section 5.2 addresses some of the issues that must be considered in such a design.

5 Implications of Decentralized Congestion Prices and Settlements

The analysis in Sections 3 and 4 suggest important implications for the practical implementation of congestion pricing. Broadly, these can be classified as technical and strategic. In the following two sections, we discuss some of these in qualitative terms. Our goal is more to identify what some of the problems might be and to suggest a research agenda for the future.

- capacity may have to be added in fixed increments and so equal capacity may be optimal for differing numbers of subscribers over a relatively large range; and
- all traffic may be internet traffic, in which case both networks need identical congestion capacity, because all calls transit both networks.

[23] If consumers could move freely, then we would end up with the fully symmetric case. However, this may not be possible for many subscribers.

5.1 Technical Implementation Considerations

The result that decentralized congestion pricing is optimal is important from a practical perspective. This means that optimal congestion prices can be computed if each network service provider simply computes local optimal congestion prices, without considering the congestion state of other service providers. In Section 2, we demonstrated that sufficient local information exists in today's packet networks to compute optimal congestion prices. Thus, from a practical perspective, the result of Section 3 is encouraging, as it is the simplest and most straight-forward solution to implement.

While this result is encouraging, it does not eliminate all practical problems. Estrin and Zhang [13] have considered some of the practical problems inherent in internetwork pricing. These include the interaction among application types, network architecture, and accounting; type of service considerations, and accounting overhead. These concerns (and others) have given rise to arguments that simple packet counting is not an adequate basis for settlements [24].

In addition to these issues, there are several additional considerations that emerge as a result of this work:

1. Congestion prices work if the user internalizes the congestion externality caused by his or her use of the network. If the total congestion price of a packet is the sum of the congestion prices of the networks it traverses, the user must be aware of the congestion price before the packet is sent. This requires (1) that all price information is continuously available to all users (or subnetworks to which users are attached) *and* (2) that the user (or subnetwork) know the route a packet will take in advance. Requirement (1) places an information flow requirement on all of the networks that may be substantial, depending on how the congestion pricing scheme is implemented. Requirement (2) is reasonable for connection-oriented network services but may not be for connectionless network services, depending on the routing scheme used and the rate of change of congestion prices.

   If a congestion price can be represented by a single broadcast packet, and we assume that such a packet is B bits (including the IP header), then aggregate price information would consume approximately BM/τ for M networks, where τ is the interval between congestion price changes. Thus, if τ is relatively large and the packet size is small[25], the price dispersion overhead is not large, and the latency between a price change and the receipt of the updated price information is not a significant issue.

   If the price information changes rapidly (e.g. $\tau \leq 0.1$ sec), however, these issues (and others) can become problematic and must be dealt with in a practical system. For example, should the traffic be reconfigured as the congestion prices change (in a connection oriented network, it means re-calculating routes and changing the routing, in the middle of a session, of at least a subset of all calls). Alternatively, if users have fixed price contracts, as suggested by Ferrari and his collaborators, it opens the possibility of arbitrage by reselling potentially cheap fixed cost bandwidth at "spot" congestion prices.

[24] See, for instance remarks attributed Vinton Cerf [9]. This report also raises the issue of different "business models" of the internet service providers, arguing that MCI's Internet network, as a predominant "transit" network is currently unprofitable, raising the pressure for some sort of settlements scheme.

[25] Without designing a specific scheme, it seems reasonable to assume that such a packet would probably be less than 500 bits, including IP overhead.

2. Even if congestion prices are implemented, and price information is dispersed appropriately, there is still the question of billing for network service. There have been a number of approaches that have been proposed for accounting and billing in networked information systems (see, for instance, [12, 22, 25, 27]). Before any of these approaches can be applied, however, an overall collection and billing strategy must be identified. For example:

   (a) Does the user receive a single price from the network to which he or she is attached?

   (b) Does the user see separate price information (and subsequently bills) from each network through which a packet will pass?

   (c) What, exactly, is billed? Are users billed for the packets required to set up a call? Are users billed for other network management packets (such as ICMP packets) related to their call? Are users billed for acknowledgment packets if they are included in the communications protocol they are using?

   In case (a), there is a simpler user interface, but there are likely to be transfer payments required among service providers.[26]. In case (b), there is no transfer payment problem, but a more complex (expensive) user interface is needed. Cases (a) and (b) do not consider the issues in (c), but these issues can affect the user's choice of protocol. In the above cases, there may still be a settlements problem even when there is no transfer payments problem, because the settlements problem is really about dividing the *total* congestion revenues among service providers in a network to optimize total social welfare.

3. Computing $\frac{\partial C^j(K_j)}{\partial K_j}$ is unlikely to be trivial in a complex subnetwork consisting of many components. While we use "capacity" fairly freely here, its precise definition is more elusive, since "capacity" can be affected by network management, congestion control techniques, etc. in addition to direct investments in communications lines and equipment.

The way in which these details are resolved matters. If the originating network supplies the end-to-end price to the user and performs the billing, settlements may be necessary. If each individual network announces price and bills separately, then additional user software is necessary (see, for instance [10]) to present a consolidated congestion price (and perhaps a bill) to the end user[27].

The congestion pricing we have analyzed here does not include multiple service classes, such as "real time" or "best effort". It is widely anticipated by computer science researchers that some form of performance guarantee will be needed to implement real time traffic [14, 16]. Parris and Ferrari [23] argued that different service classes require different prices. Stahl and Whinston [30] have considered client-server computing with priority classes. The structure of their analysis can inform the problem of multiple service classes in networks with congestion externalities as well.

[26] In the case where the networks are equally large, the traffic to and from each network is equal, and the congestion in each network is equal, transfer payments may not be necessary.

[27] This is, in effect, how the telephone network presently works. Users pay a fixed network access fee directly to the local telephone operating company, and receive a separate statement (often in a consolidated bill) from the interexchange carrier. This bill includes all settlements between the carriers.

5.2 Strategic Implementation Considerations

In all of the preceding analysis, we have assumed that network providers do not have market power and hence will not be able to bias their pricing, network capacity or interconnection decisions either to extract consumer surplus or to protect their market position. The historical record for the telephone and cable television industries suggests that this is assumption has not been valid in the past[5]. As a consequence, both industries have been subject to heavy regulatory oversight with respect to their investment, access and pricing policies. The economic basis for this regulation has relied heavily on assertions about the nature of network capacity costs and the difficulty of sustaining effective competition while maintaining the coordination required to support reliable, ubiquitous end-to-end services. Many analysts believe that advances in technology (e.g., open systems, modularization, technological convergence, etc.) have reduced entry barriers and much of the need for regulatory oversight. The trend has been towards relying increasingly on market forces to eliminate opportunities for excess profits, and therefore, incentives for wasteful investments in rent-seeking.

Market power may not be a significant issue in a privatized Internet. If it is, then there will be myriad ways in which service providers may seek to distort either congestion pricing or the settlements mechanism. For example, a transit network that controlled a bottleneck facility would have an incentive to distort its prices for access (interconnection) and usage fees in order to extract monopoly rents. It may charge lower or higher than optimal usage fees, depending on the relationship between inframarginal and marginal subscriber responses. The design of the settlements mechanism could thus be used to facilitate the collection of monopoly rents.

If the networks were able to form a cartel to collectively extract monopoly rents from subscribers then the settlements mechanism could provide a vehicle for distributing those rents. From the discussion above, it should be clear that monitoring individual behavior would be difficult and so individual cartel members may have an incentive to misrepresent their traffic/congestion status in order to capture a larger share of the settlements revenue. The breakdown of such a cartel may not be welfare improving if it causes the network of networks to fragment. Introducing settlements into network profit calculations will influence their behavior. Therefore, if the markets for network services are not contestable, then the design of the settlements mechanism must anticipate how individual behavior will be distorted. There is a principal agent problem which must be resolved.

While the difficulties posed by imperfect competition are worthy of significant research attention, they go beyond the scope of the present paper. However, even if we restrict ourselves to the (perhaps dubious) case of contestable markets, we cannot presume that all subscribers will be equally represented or influential in determining how future networks will evolve. For example, in our model, there is a fundamental tension between subscribers who make different types of calls. On-net and internet callers would like to see the other's traffic minimized, so each would like to see the other face higher prices. This may have implications for customer attitudes towards the efficient implementation of congestion pricing and toward the debate about emerging notions of "universal service" for the Internet.[28] As noted above, efficient prices should discriminate among

[28]There is a sizeable community of Internet users that oppose usage-based pricing. Many of these users are concerned about the effects of usage-based pricing on the modes of behavior (such as mailing lists) that they perceive to be valuable. See [17] for an example of this position.

on-net and internet traffic and non-zero settlements offer one mechanism for implementing these higher prices.

Let us suppose that the network community can be convinced of the advisability of congestion pricing, and that the debate has turned to the need to discriminate among different types of traffic[29]. Since efficient congestion pricing implies that internet traffic should face higher prices, these callers would have an incentive to argue against price discrimination while on-net subscribers would take the opposite position. Since it is likely that the settlements mechanism chosen is likely to affect the feasibility of implementing price discrimination, there may be a bias from "internet-type" callers in favor of zero-settlements mechanisms.[30] Consider what might happen in negotiations between the subscriber communities of a large and a small network. Assuming that connection between user pairs are uniform, the large network is more likely to originate on-net connections while the small network is more likely to originate by internet connections[31]. Thus, under congestion pricing, subscribers on the larger network should press for a complex settlements mechanism that facilitates charging for termination traffic, while subscribers on the smaller network may argue for zero settlements. The point of this discussion is to suggest how even in the absence of market power by service providers, the political debate over optimal pricing may be distorted by private economic interests.

The failure to adopt optimal congestion prices may influence the choice of where subscribers choose to originate their traffic, although not all subscribers are likely to face the same flexibility. For example, optimal congestion prices should be identical regardless of the direction in which a particular calling route is followed. If, however, $p^*_{ij} > p^*_{ji}$ then sophisticated callers will have an incentive to originate their calls from network j. It is not necessary for a caller to physically locate on another network since she could use an inexpensive call to set-up the return origination call.[32] Generally, rate arbitrage that results in similar end-to-end congestion charges for traffic with similar congestion (quality-of-service) characteristics would be welfare improving. However, such arbitrage may not occur on a sufficiently large scale and may leave unsophisticated subscribers at a disadvantage.

Content providers are another class of sophisticated subscribers who may seek to influence the setting of usage prices. Generically, we might presume that they would like to see relatively low network access and usage fees so that consumers have more surplus to spend on content. Ideally, they might like to see network services provided free (subsidized by general tax revenues which would include non-subscribers). Alternatively, if the typical content customer is an inframarginal consumer of network services, they may prefer higher than optimal access fees in return for lower than optimal usage fees. Although this scenario need not be the case, we suggest it to illustrate

[29] We ignore the accounting and implementation costs associated with usage pricing. These may be substantial and when included in the cost/benefit analysis may make it optimal to employ usage pricing. Assessing the magnitude of these costs is clearly an important area for further research.

[30] This bias may be partially (or, wholly) offset if the uniform on-net price or access fees rise in order for the network to recover its total costs.

[31] Clearly, this need not be the case. The logic for this assumption is that all callers are equally likely to call each other and since thus a higher proportion of the origination/termination pairs are on the larger network, most callers will be on-net callers on the larger network and internet callers on the small network. Both networks could be dominated by on-net calling if most calling is local and such internet calling as occurs is done by subscribers on the larger network.

[32] A number of entrepreneurs offered such services to international callers to arbitrage international telephone settlements that resulted in higher prices for calls that originated internationally.

why the establishment of usage pricing is likely to be contentious.

6 Summary and Conclusions

We believe usage pricing is both desirable and unavoidable for the Internet. We also believe that there is still much research that needs to be done to better understand both the theoretical and practical issues that arise in the context of a packet-based network, comprised of a collection of independent, and potentially competing, networks. This paper provides a preliminary analysis of the dual problem of congestion pricing and settlements in such an environment. We extend the congestion pricing analysis of MacKie-Mason and Varian [21] to the context of multiple networks. This analysis shows that the optimal congestion price between two users depends on the route followed, and, specifically, is the sum of the local congestion prices which should be set by each of the networks along the route. These prices may be computed using local cost and traffic information. This is important if network control is to be decentralized. In the absence of centralized coordination, the networks need to share congestion pricing information so that the originating networks can know what price to set for end-to-end service. A settlements process that requires networks to bill each other for terminating traffic offers one mechanism for conveying this information. This provides one rationale for the linkage between the two problems. A second rationale stems from the need for each network to recover its costs. If prices are set so as to induce optimal consumer behavior by forcing them to internalize the welfare implications of their behavior for the network-of-networks, then individual firms may fail to recover sufficient revenue in the absence of settlements.

Even in an world where firms do not have market power, revenue transfers among service providers (i.e., settlements) are likely to be necessary and since the amount of revenue transferred is likely to depend on both the volume of traffic and the price faced by consumers, congestion pricing and settlements issues are not readily separable. We demonstrate this using a simple case of two networks. Further, we argue that the nature of the settlements problem depends on the technology of the networks being used to deliver service as well as the design of the settlements mechanism.

In our analysis, we have assumed the following:

- None of the providers have market power. When providers have market power, this analysis becomes considerably more complex, since strategic interactions among the service providers must be considered.

- A network provides a single type of service, like the Internet. If multiple service classes exist, as may be necessary with the emerging ATM-based networks, or if a scheme such as the "smart market" [19] or "precedence" [3] is used to provide price-based priority, additional factors may need to be considered in the analysis of congestion pricing and settlements in multiple networks.

- Our analysis is static. As we suggested in Section 5, a dynamic analysis raises numerous technical problems that must be solved (not the least of which is the user interface). There are also a host of new economic issues that arise under a dynamic analysis, particularly if a settlements strategy is included explicitly in the analysis.

This analysis is a starting point in considering settlements in multiple networks. There is clearly much more work that needs to be done in the area of generalizing this analysis from an economic perspective and in applying it to specific network implementations, both statically and dynamically.

References

[1] Baumol, W., J. P., and Willig, B. *Contestable Markets and the Theory of Industry Structure*. Harcourt, Brace and Jovanovich, New York, 1982.

[2] Bertsekas, D., and Gallager, R. *Data Networks (2nd Edition)*. Prentice Hall, Englewood Cliffs, NJ, 1992. ISBN 0-13-200916-1.

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

[4] Braun, H.-W., Claffy, K., and Polyzos, G. C. Measurement considerations for assessing unidirectional latencies. Tech. rep., The San Diego Supercomputer Center and the University of California, 1992.

[5] Brock, G. W. *Telecommunication Policy for the Information Age*. Harvard University Press, Cambridge MA, 1994. ISBN 0-674-87277-0.

[6] Brock, G. W. The economics of interconnection. Tech. rep., Teleport Communications Group, Staten Island, NY 10311, April 1995.

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

[8] Comer, D. E. *Internetworking with TCP/IP, Volume I (Second Edition)*. Prentice-Hall, Englewood Cliffs, NJ, 1991. ISBN 0-13-46805-9.

[9] Cook, G. Summary of the september 1995 COOK report. Distributed on the `telecomreg` newsgroup, September 3 1995.

[10] Danielsen, K., and Weiss, M. B. User control modes and IP allocation. Tech. rep., University of Pittsburgh, Pittsburgh PA 15260, March 10 1995. Presented at the Internet Economics Workship.

[11] Danielsen, K., and Znati, T. A congestion management scheme for real time packet switched networks. In *6th ISMM Conference on Distributed System and Parallel Processing, Pittsburgh, PA*. (October 1992).

[12] Edell, R. J., McKeown, N., and Varaiya, P. P. Billing users and pricing for TCP. *IEEE Journal on Selected Areas in Communications* ((forthcoming)).

[13] Estrin, D., and Zhang, L. Design considerations for usage accounting and feedback in internetworks. *ACM Computer Communications Review 20*, 5 (October 1990), 56–66.

[14] Ferrari, D. Real-time communication in an internetwork. *Journal of High Speed Networks 1*, 1 (1992), 79–103.

[15] Ferrari, D., and Verma, D. Quality of service and admission control in ATM networks. Tech. rep., International Computer Science Institute, Berkeley, CA, 1989.

[16] Field, B. *A Network Channel Abstraction to Support Application Real-Time Performance Guarantees*. PhD thesis, University of Pittsburgh, Department of Computer Science, 1994.

[17] Love, J. Future internet pricing. gopher://essential.essential.org:70/0R0-12615–/pub/listserv/tap-info/950310, March 10 1995.

[18] MacKie-Mason, J. K., and Varian, H. R. Some economics of the Internet. Tech. rep., University of Michigan, MI, November 1992. Retrieved from [31].

[19] MacKie-Mason, J. K., and Varian, H. R. Pricing the Internet. Tech. rep., University of Michigan, MI, April 1993. Retrieved from [31].

[20] MacKie-Mason, J. K., and Varian, H. R. Economic FAQs about the Internet. *Economic Perspectives* (1994). Can be retrieved from [31].

[21] MacKie-Mason, J. K., and Varian, H. R. Pricing congestible network resources. Tech. rep., University of Michigan, MI, October 1994. Retrieved from [31].

[22] Mills, C., Hirsh, D., and Ruth, G. Internet accounting: Background. Tech. Rep. RFC 1272, Network Working Group, 1991.

[23] Parris, C., and Ferrari, D. A resource based pricing policy for real-time channels in a packet-switching network. Tech. rep., International Computer Science Institute, Berkeley, CA, 1992. Available from ftp: tenet.icsi.berkeley.edu/pub/tenet/Papers/ParFer92.ps.

[24] Parris, C., Keshav, S., and Ferrari, D. A framework for the study of pricing in integrated networks. Tech. Rep. Tech. Rept. TR-92-016, International Computer Science Institute, Berkeley, CA, 1992. Available from ftp: tenet.icsi.berkeley.edu/pub/tenet/Papers/PaKeFe92.ps.

[25] Ruth, G., and Mills, C. Usage-based cost recovery in internetworks. *Business Communications Review xx* (July 1992), 38–42.

[26] Shenker, S. Making greed work in networks: A game-theoretic analysis of switch service disciplines. In *Proceedings of SIGCOMM* (1994), ACM. Available from ftp://parcftp.xerox.com/pub/net-research/.

[27] Sirbu, M., and Tygar, J. Netbill: An internet commerce system optimized for network delivered services. In *Workshop in Internet Economics* (1995), MIT.

[28] Spragins, J. D., Hammond, J. L., and Pawlikowski, K. *Telecommunications: Protocols and Design*. Addison-Wesley, Reading MA, 1991. ISBN 0-201-09290-5.

[29] Srinagesh, P. Internet cost structures and interconnection arrangements. In *Toward a Competitive Telecommunications Industry: Selected Papers from the 1994 Telecommunications Policy Research Conference*, G. W. Brock, Ed. Lawrence Erlbaum Associates, Hillsdale NJ, 1995.

[30] Stahl, D. O., and Whinston, A. B. An economic approach to client-server computing with priority classes. Tech. rep., University of Texas at Austin, 1992.

[31] Varian, H. R. Economics of the Internet: Information service located on the Internet, accessible by `www` page `gopher.econ.lsa.umich.edu`.

[32] Yurcik, W., Tipper, D., and Banerjee, S. Local quality of service provisioning to meet end-to-end requirements in ATM networks. In *First International Symposium on Photonics Technologies and Systems for Voice, Video, and Data Communications* (October 1995), International Society for Optical Engineering (SPIE).