A Practical Review of Pricing and Cost Recovery for Internet Services

Peter Reichl[1], Simon Leinen[2], Burkhard Stiller[1]
[1]Computer Engineering and Networks Laboratory TIK, ETH Zürich, Gloriastrasse 35,
CH – 8092 Zürich, Switzerland, E-Mail: [reichl | stiller]@tik.ee.ethz.ch
[2]SWITCH, Zürich, Limmatquai 138, CH – 8001 Zürich, Switzerland, E-Mail: leinen@switch.ch

Abstract

Suitable pricing models for Internet services represent one of the main prerequisites for a successfully running implementation of a charging and accounting tool. This paper introduces general aspects influencing the choice of a pricing model and presents a survey of relevant approaches to be found in the scientific literature. Based on cost model investigations some detailed insight into price and cost issues from an Internet Service Provider's (ISP) point of view is given. Moreover, current challenges as well as problems are discussed in a practical context as investigated in the Swiss National Science Foundation project CATI – Charging and Accounting Technology for the Internet.

1 Introduction

It is expected that many Electronic Commerce applications will make use of the Internet as a transport infrastructure. In contrast to more traditional communication networks such as the Public Switched Telephone Network (PSTN), the current public Internet lacks well-defined pricing models and cost recovery schemes, both between Internet Service Providers (ISPs) and their clients and between interconnecting ISPs.

As Electronic Commerce becomes more widespread, businesses will increasingly recognize the impact of the quality of the Internet infrastructure for revenue. This will spur a competition for improved Internet service between connected businesses, which will in turn lead to differentiated Internet service offerings from ISPs. As the Internet will move away from a pure "best-effort" service to such a more differentiated service network, new schemes for pricing and Service Level Agreements (SLAs) will necessarily evolve. A major goal of this work is to anticipate and investigate such schemes.

As the variety of Internet services, such as native IP (Internet Protocol) service, Mbone service, WWW (World Wide Web) services, or FTP (File Transfer Protocol) service, show extremely different traffic characteristics, sensible pricing models for integrated and differentiated services and for multicast traffic in the Internet are essential. Therefore, the development of appropriate SLAs, pricing models, and suggested settlement schemes for inter-provider connections (peerings) become a key issue. For example, the costs of operation of the Internet for Swiss universities and research has to be recovered almost entirely from its users. This requires clearly defined methods for charging and accounting of Internet services as well as sound cost-sharing models and appropriate pricing schemes. Improvements on the existing model of cost sharing are required to achieve the level of detail for all involved cost factors of a given IP network. This includes various refinements to optimize resource usage in the Internet. In addition, various connected sites, organizations, and enterprises on the Internet have expressed a need for better cost attribution within their local administrative domain with respect to utilized IP services.

This paper is organized as follows. While Section 3 highlights pricing models in practice, Section 2 deals with a classification scheme for Internet pricing models in terms of scientific research. An overview of significant related scientific work can be found in Section 4. Based on this background, a concrete approach for the cost recovery of an ISP is discussed in Section 5. Finally, Section 6 summarizes the work and focuses on requirements on economically viable and technically efficient approaches for pricing models and cost recovery schemes as investigated in the framework of the project CATI.

2 Pricing Models in Practice

Only a few pricing schemes seem to be in wider use on the Internet today – especially form the ISP's point of view. Dial-up access to the Internet is often sold for a fixed monthly fee, including either unlimited use or a limited duration of the connection. Additional use of the connection is then billed at a per-hour fee. Volume-dependent charges have been used by some providers in the past, but seem to have lost popularity in recent years.

The price of a fixed Internet connection is usually a monthly charge depending on the bandwidth of the connection. In many cases the customer has to provide the circuit to the Internet service provider's nearest point of presence. In most cases, this circuit will be a leased line, which adds another bandwidth-dependent recurring fee with an additional distance component.

For fixed connections, volume-based charging is more common, especially outside North America. There is usually still a fixed recurring component based on access

IEW'99: 2nd Internet Economics Workshop, Berlin, Germany, May 28-29, 1999.

capacity, but an additional cost per Megabyte (mostly with volume discounts) of data transferred over the connection. Such offers vary widely in the distribution between fixed and volume-dependent components, price/volume curves, and other parameters such as whether both directions of traffic are considered for the volume fee. In most cases, the bandwidth-based fixed part of the price includes some positive amount of volume that can be transmitted without charge.

Another variant that seems to have become popular recently is a "bursty" rate, where the Internet service provider periodically, *e.g.*, every hour, measures the volume of data transferred over the connection. For each charging interval, *e.g.*, every month), all samples are sorted by volume, a fixed percentage, *e.g.*, 5%, of the highest samples are discarded to eliminate unusual peaks, and the highest remaining sample is used to define the bandwidth at which the connection is charged.

It is quite rare to see further differentiation of usage-based charging, such as short and long distance, or peak and off-peak rates. Notable examples of those are volume-based charges on specific bottleneck links such as the trans-oceanic connections of New Zealand's and Britain's research networks [1], [12].

3 Pricing Model Classification

The classification of pricing models must encounter at least four basic dimensions, to allow for (1) a clear service dependency in terms of the technical network type, (2) a suitable distinction of price components that need to be considered for pricing models, (3) a selection of parameters applicable to charging, and (4) identification of pricing intentions.

3.1 Dimensions

In order to classify, characterize, and distinguish the different proposals for Internet pricing and charging mechanisms, four dimensions have been identified. In addition, a general research type is introduced. They can be thought of as being almost independent of each other, thus creating the domain within which any discussion on pricing and charging issues has to be placed.

3.1.1 Service Categories

Among the proposals there is an apparent distinction between connection-oriented and connectionless approaches. In the history of Internet pricing, *packet-based models* were the first to distinguish between several qualities of service, focussing mainly on single packets as being the entity to be charged. Quality-of-Service (QoS) levels are only designed in the form of relative best effort service classes. Adding guarantees to the QoS classes has lead to the *DiffServ* approach. Here, the focus is on service classes characterized by guarantee parameters (abandoning the per packet view). DiffServ is commonly based on IP technology, i.e. connectionless and hence at least questionable w.r.t. real-time applications. In contrast, the *IntServ* approach has changed to connection-oriented mechanisms, thus allowing high level QoS as now users may be rejected during times of heavy traffic. But IntServ can no longer be run by IP as a reservation-based one like RSVP is necessary.

Note that another terminology distinguishes between *soft QoS* papers (proposing different service categories for connectionless traffic) and *hard QoS* papers (focussing only on real-time connection-oriented approaches).

3.1.2 Price Components

Common telecommunication pricing consists of three basic elements, i.e. access fees, setup fees and usage fees. Combining these three leads to a classification of pricing mechanisms as flat fee, usage-based, reservation-based, volume-based, service class-based, bandwidth-based etc. In the same way, there is a couple of basic approaches for Internet pricing mechanisms to be distinguished, like flat rate, usage sensitive or volume-based pricing, packet or flow auctions, service or user profile-based mechanisms, edge pricing versus multilateral contracts.

3.1.3 Charging Parameters

This dimension concerns the question of which parameters are or should be available for use in charging and pricing mechanisms, starting from priority flags and packet tagging over peak, nominal bit or average flow rates as well as effective bandwidths to parameters like expected path and congestion cost or dynamic bid-prices per packet or resource unit, to name but a few.

3.1.4 Pricing Classification

Here, the basic function of pricing has to be investigated. Pricing may be used for reasons of network efficiency, i.e. maximizing the utilization of resources (bandwidth, buffer space), or of economical efficiency, i.e. value to the user. Hence pricing maximizes either provider revenue (by efficient resource sharing and access control) or user satisfaction. Incentive compatibility is a further aspect to be considered here.

3.1.5 Type of Research

A last dimension may describe whether the research is conducted from a more theoretical or more practical and application-oriented point of view. Both types of research are able to deliver important input especially for the development of an architecture general enough for implementing different pricing models.

3.2 Further Aspects and Requirements

Beyond such general dimensions, there are much more aspects and requirements to be considered while looking for viable pricing models, reaching from different types of applications (*e.g.*, burstiness issues) over technological and economical issues (*e.g.*, sender or receiver based payment, marginal cost, congestion/responsive pricing) to more practical ones (*e.g.*, like transparence, predictability, practicability, fairness, user acceptance and user friendliness) and will later be investigated in detail.

4 Related Work on Pricing Models

Although the issue of pricing Internet services has gained its actuality only recently, there is already a couple of related work to be found in the scientific literature. First ideas during the early 90‘s usually preferred some simple sort of service priority models, until in 1994, [10] introduced the idea of using auction mechanisms. In 1995, [13] formulated an important model based on the intserv approach. Both those papers provided major stimuli to the research of the following years (cf. especially the collection of relevant papers in [11]), but it was not until 1997/98 that first steps towards designing “real world” charging mechanisms took place, mainly based on the RSVP protocol. The following section contains an overview of important Internet pricing models as investigated over the last five years. It presents the idea of edge pricing, concentrates on different types of auction mechanisms, deals with user and service profiles, and concludes with an outlook on practically oriented approaches.

4.1 Edge Pricing

The fundamental idea of edge pricing [3], [14] is to charge the user only by the first network provider along a data path that might use also services from other providers. The charge to be paid includes expenses for all different providers handling the respective data. Thus, multilateral contracts are reduced towards a sequence of bilateral ones, the complexity is reduced enormously and user transparence is provided. In the basic approach, the user defines the maximal total price she is willing to pay as a sender or a receiver of data, respectively, as well as an upper bound for the maximal number of hops. The charging information can be transmitted as part of a signalling protocol, *e.g.*, the RSVP header [4], [5].

4.2 Auction Mechanisms

The seminal work of [10] deals with the question of how an efficient pricing structure allows to manage congestion, encourage network growth and guide resources to their most valuable use. As the marginal cost for transporting packets over the network is essentially zero as long as the network is not congested, usage-sensitive pricing schemes appear to be a good candidate for congestion control mechanisms, as they approach the allocation of scarce Internet resources in an economic context. Note that the objective is not to raise profits, but to find a pricing mechanism yielding most efficient usage of existing resources. Current pricing schemes usually offer no incentives to flatten peaks or mechanisms for bandwidth allocation during congestion, whereas ideal prices should reflect the resource costs that the user generates so that she can make informed decisions about resource utilization. Such costs include fixed costs for network infrastructure, costs of connecting to the network as well as sending extra packets, and finally the social costs of delaying other users’ packets during congestion periods. A congestion pricing scheme (where packets are charged if and only if the network is congested) could be implemented by using a “smart market” where the price for sending a packet varies on a very short time-scale, thus reflecting the current degree of network congestion. Each packet header contains a bid field, and the packet is admitted if the bid exceeds the current marginal cost of transportation. Note that the user does not pay the actual bid but only the (lower) market-clearing price. In this kind of “second-bid” auction (Vickrey auction) the optimal strategy for the user is to bid her true evaluations - fooling the market results only in disadvantages. Alas, the mechanism guarantees only relative priority, no absolute QoS. Other critical issues include the question of how accounting should be done without yielding too much an overhead, how bursts will be handled and how the user will react to maybe rapidly fluctuating bandwidth prices.

The model [8] allows guaranteeing multiple QoS (especially for inelastic traffic) by scheduling resources in advance. In order to maximize the sum of user utilities, a routing problem has to be solved by standard multi-commodity flow techniques. The notion of “effective bandwidth” allows to aggregate a broad range of source types in form of a one-dimensional bandwidth reservation. Decoupling routing and usage optimization and solving the resulting linear problem and its dual, the latter allows interpreting the optimal solution in terms of spot prices for inserting (or extracting, respectively) traffic at a certain node. This allows to express the marginal system cost for traffic from node A to node B in terms of only two numbers: the nodal spot price for the source and for the sink of the flow; the user no longer needs to know about the optimal route. Note, however, that finding the optimal spot prices still requires solving the full central planning problem. Decentralization requires that the users truthfully reveal their preferences so that an Pareto efficient allocation can be calculated. The mechanism proposed is again a “smart market”, i.e. generalized Vickrey auctions (GVA), as presented in [10].

A related approach of using auctions as a proper method for decentralizing the decision-making in packet-switched multiservice integrated networks is presented in [7]. Here, Vickrey auctions are generalized yielding a mechanism that is designed to be stable, simple, efficient and fair. While the original “smart market” proposal [10] uses one-dimensional bids (price per packet) and thus requires the central setting of the market clearing price based on explicitly assumed utility functions for the users, this approach requires the possibility of per-flow resource reservations. The resulting two-dimensional bids (price and quantity) allow to determine the clearing price directly from the bids only. Instead of dividing the resource into many small units and handling each of them as an indivisible object subject to a Vickrey auction (which yields a considerable loss of flexibility and scalability), here allocations are assumed for arbitrary shares of the total available resource quantity. Player preferences are given in the utility function describing the individual “value” of quantity/price vectors for the individual user. The Progressive Second Price (PSP) rule generalizes the idea of Vickrey auctions: you pay a price per unit which is calculated from all other players’ bids, where each of them is weighted by how much the allocation of that player is decreased by the existence of your bid. Hence, for each infinitesimal share of the resource, the player who is getting it pays the maximum amount which the player who is denied it would have been willing to pay for it). This rule can be shown to have a number of nice properties, ultimately leading to the existence of a fair and efficient Nash equilibrium.

Summarizing the approaches described, second price auctions appear to be a useful concept for determining actual market prices in case of network congestion which is one of the main prerequisites for a truly dynamic pricing scheme. However, it must be noted that these papers are kept rather theoretically and may need major adaptations before using their ideas for practical purposes.

4.3 Profiles and Classes

[2] deals with the question of how to provide different QoS with high predictability while still running usual best effort. Instead of allocating capacity to users by explicit reservations, the “Expected Capacity” framework defines service profiles for each user and separates demands into those within the profiles and those outside. Treating these two types of packets differently (i.e. favoring traffic that obeys the respective profile) allows the network to offer different levels of service with high predictability. In such a scheme, packets from a user that behaves correctly are tagged “in”, whereas packets from a user exceeding her profile are tagged “out”. During congestion a suitable dropping scheme is used to preferentially drop “out” packets. Note that this approach prevents traffic being separated at the routers into different flows or queues.

In contrast, [6] does not classify users but services. Inside each service class every customer receives equal service, but higher service classes offer significantly better service than any lower service classes and are therefore charged higher prices. The nominal bit rate (NBR) provides the underlying parameter for a monthly fee. Congestion is recognized by monitoring the load level of output buffers in the nodes; the system reacts by discarding some packets, preferably from flows with actual bit rate to NBR ratio being high. Each packet carries drop preference and delay indication bits, based upon which the system decides about the discarding of packets.

4.4 Charging Framework for RSVP

[4] deals with the question of how to enhance the Resource Reservation Protocol RSVP with information for payment and prices. Here, one major aspect is that, depending on the underlying economic model, this information may be interpreted in different ways. The idea is to use the PATH messages (path messages are directed from senders to receivers and contain a sender-offered set of QoS specifications for a flow) and RESV messages (reservation messages are directed from receivers to senders and contain the requested QoS specification for one flow) of RSVP in order to transmit price information. Extended PATH messages carry a field with price information that is initially set to zero. At each hop with an outgoing link, the current market price for the requested QoS is added to the price field. Hence, if the PATH message has arrived at the receiver, an approximate picture of the market situation is delivered, albeit the final price may still vary slightly, due to small changes on the market. Finally, the RESV messages are sent back and show the actual price to the sender. Those messages may also be used for payment information, i.e. the PATH messages may contain sender provided payment, whereas the RESV messages may carry receiver provided payments.

5 Cost Recovery within SWITCH

SWITCH is an Internet service provider for research and higher education being organized as a non-profit foundation with little centralized funding. In yearly budget discussions, representatives of foundation members define a charging scheme for the next year. These charges are set to recover all operating costs for the network.

In recent years, the tariff for leased line connections consisted of an access charge based on the capacity of the connection, and a charge per Giga-Byte of data transferred from SWITCH towards the connected organization. The fixed charge and volume-based charge were set to account for about one third and two thirds of the total.

The reasons for only counting traffic towards customers are as follows: The underlying transmission links of the network are always sold symmetrically, but more traffic flows towards the universities than from them. Thus, there is usually no congestion in the outbound direction. Secondly, the pricing is set to encourage sharing of information.

The large part of usage-based charges has caused a high level of awareness of the costs of network usage within the organizations connected to SWITCH. Several creative methods have been devised by universities to reduce the amount of traffic and control the cost associated with their network connection. This includes reducing WWW bandwidth consumption by use of caching proxy servers (in some cases made mandatory through blocking of direct connections), or converting USENET News feeds from “push” to “pull” mode.

Some of these measures to reduce traffic volume have actually been counterproductive in the sense that they reduced traffic only on those lines that were not congested anyway, so they did n0t actually reduce cost of operation. In some cases, efforts to save volume even made the service more expensive for everyone because or increased load on servers.

If the concept of volume-based charging is maintained in the future, it is likely that the prices will have to be further differentiated to encourage bandwidth savings in a way that is more focused to the expensive and congested or highly sollicited parts of the networks – currently the international links and in particular the connection to the U.S.

SWITCH participants have often expressed their intention to charge individual users (or organizational units such as departments or institutes) for the volume of network traffic generated. So far, the technical and administrative complexity involved with this has prevented them from doing so. A few universities provide tools through which individual users can inform themselves about their amount of network usage. This has been found very useful and could obviate the need of performing actually accounting and billing within such organizations.

5.1 Cost Structure of SWITCH’s Services

In many respects, SWITCH is not exactly representative of Internet service providers: the essentially closed user group, organizational structure as a non-profit foundation, and emphasis on the communication needs of the academic community make it different from most ISPs which operate on a commercial basis and offer services to a wide range of individual or corporate users in a very competitive market. However, the cost structure of operating SWITCH's network service is quite similar to those of other large providers with mainly leased-line customers.

SWITCH operates a national backbone to transport data between connected sites. This national backbone is built from routers operated by SWITCH, connected with leased lines or ATM connections. For both types of connections, SWITCH pays monthly volume-independent fees to carriers.

For traffic between SWITCH-connected sites and other networks, there are two different cases: Some providers' networks are reachable at low cost because a “peering agreement” permits mutual exchange of traffic at no charge. In those cases, the router ports and lines connecting the two networks are the only costs involved. Where peering agreements cannot be established, SWITCH has to pay other networks for “transit” to other parts of the Internet. Currently, SWITCH uses transit service from a commercial ISP in the USA and leases its own line (actually an ATM virtual circuit) to New York. This transit subscription and leased line represents a very significant fraction of the cost of operating SWITCH.

Because international circuits are much more expensive than domestic ones, the costs for international connectivity exceed the costs for the national backbone. Since international transit connections are used to ensure connectivity to those parts of the Internet that cannot be reached by peerings, there is a strong incentive to peer with other networks to which there is significant traffic.

5.2 Peering Agreements

This explains why Internet service providers agree to exchange mutual traffic without settlements. However, such agreements are hindered by the fact that the savings are usually not perfectly balanced, i.e. one of the potential peers would benefit from the peering more than the other. An ISP who believes that the potential peer would reap larger benefits may see the peering as counter-productive because it would make the other ISP more competitive at the first one's expense.

In cases where a difference in benefits is perceived, the peering parties usually negotiate some kind of compensation, such as one party paying for the entire line between the two networks, or the provision of free transit to a third network by one of the potential peers. Where the benefits are seen as widely differing, the provider with only small perceived benefits will likely not go through the effort of setting up the peering.

Assessing the mutual benefits of existing peerings is often done when upgrades have to be negotiated. The metric that is used most often is the amount of traffic flowing in both directions. Traffic sent from ISP A to ISP B is generally thought of as beneficial to ISP B, because presumably it represents data that has been requested by ISP B's customers. This crude measure still corresponds to the prevalent ways the Internet is used today. However, it may be challenged by the emergence of unsolicited information such as bulk e-mail or paid advertisements.

While this metric is useful enough to serve in arguments on establishing or upgrading settlement-free peerings, it would probably not be sufficient as a basis for settlements between peering ISPs. In fact, peering with settlements doesn't seem to be popular at all. The only scheme used in practice is, that one (smaller) ISP pays another (larger) one to connect to their network, but without the right to transit to other networks. The authors are not aware of cases where settlements between peers are set in both directions and based on a common metric.

6 Summary and Remaining Issues

This paper provides a brief review of a large variety of possible pricing schemes that could be used in the context of Internet pricing and cost models. Moreover, major relevant issues concerning cost structures and cost recovery have also been enlightened from an Internet Service Provider's point of view. This is espcially important for judging the practical relevance of research work in a business environment.

As most of the work on pricing schemes so far is kept rather theoretically, the next step is to validate their practical applicability. Having future general Internet developments in mind, in our opinion an ideal implementation of a pricing model requires the essential integration of the following concepts:

- Sufficient support of temporal pricing aspects;
- Using some sort of second-price auction approach for dynamic pricing, but refining it in a way that the auctions and, therefore, the final price become more transparent to the user;
- Simplifying tariffs by offering a suitably defined traffic classification;
- Extending pricing models and the corresponding technical protocol-based signalling to the case of multiprovider networks;
- Keeping pricing and charging issues as far as possible at the border of the ISP networks.

Creating a platform that allows testing and using different pricing models in a practical multiprovider environment integrating both the IntServ and DiffServ worlds is one of the major goals of the SNF project CATI. The prototype currently developed focuses on a working charging and accounting framework to be used, *e.g.*, for tariffing IP telephony. The cooperation with the Swiss ISP SWITCH thereby guarantees that the project results will fit closely the features required by practical users of an Internet charging and accounting tool.

Acknowledgments

This work has been performed in the framework of the project Charging and Accounting Technology for the Internet – CATI (CAPIV 5003-054559/1 and MEDeB 5003-054560/1) which has been funded by the Swiss National Science Foundation, SNF, Berne, Switzerland. The authors like to acknowledge contributions of their colleagues G. Fankhauser and W. Huber as well as student's work of F. Floessl.

References

[1] Brownlee, N.: Internet Pricing in Practice. In: [11] pp. 77-90.

[2] Clark D.; Fang, W.: Explicit Allocation of Best-Effort Packet Delivery Service, IEEE/ACM Transactions on Networking, Vol. 6, No. 4, August 1998.

[3] Cocchi, R.; Estrin, D.; Shenker, S., Zhang, L.: Pricing in Computer Networks: Motivation, Formulation and Example. IEEE/ACM Transactions on Networking, Vol.1, No. 6, December 1993, pp. 614 - 627.

[4] Fankhauser, G;. Stiller, B.; Vögtli, C., Plattner, B.: Reservation-based Charging in an Integrated Services Network. 4th INFORMS Telecommunications Conference, Bocca Raton, Florida, U.S.A., March 1998.

[5] Karsten, M; Schmitt, J.; Wolf, L.; Steinmetz, R.: An Embedded Charging Approach for RSVP. International Workshop on QoS 98. Napa, California, U.S.A., May 1998.

[6] Kilkki, K. et al.: Internet Charging Reconsidered. 5th Annual Networld and Interop Engineers Conference. Las Vegas, Nevada, U.S.A., May 1998.

[7] Lazar, A.; Semret, N.: Auctions for Network Resource Sharing. CTR Tech. Report. Columbia University New York, February 1997.

[8] MacKie-Mason, J.: A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network. University of Michigan, September 1997.

[9] MacKie-Mason, J.; Murphy, J.; Murphy, L.: The Role of Responsive Pricing the Internet. Tech. Report, University of Michigan, June 95. In: [11].

[10] MacKie-Mason, J.; Varian, H.: Pricing the Internet. JFK School for Government, May 1993.

[11] McKnight, L; Bailey, J. (eds.): Internet Economics. MIT Press, Cambridge, Massachusetts, U.S.A.,1997.

[12] Rogerson, R.: Usage-related Charges for the JANET Network. JISC Circular 3/98, March 1998. http://www.jisc.ac.uk/pub98/c3_98.html].

[13] Shenker, S.: Some Fundamental Design Issues for the Future Internet. IEEE Journal on Selected Areas in Communications, Vol. 13, No. 7, September 1995 pp. 1176 -1188.

[14] Shenker, S.; Clark, D.; Estrin, D.; Herzog, S.: Pricing in Computer Networks: Reshaping the Research Agenda. ACM Computer Communications Review, Vol. 26, No. 2, April 1996 pp. 19 - 43.