IEEE COMMUNICATIONS SURVEYS
The Electronic Magazine of Original Peer-Reviewed Survey Articles

PRICING FOR QOS-ENABLED NETWORKS: A SURVEY

LUIZ A. DASILVA, VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

ABSTRACT

A complete solution to the problem of providing adequate quality of service (QoS) to heterogeneous users must take into account the issue of pricing. By adopting an appropriate pricing policy and by setting prices carefully, a service provider will be able to offer the necessary incentives for each user to choose the service that best matches his/her needs, thereby discouraging over-allocation of resources and maximizing revenue and/or social welfare. In this article, we survey some of the recent research in the literature dealing with service pricing in multi-service networks. All of the work surveyed here addresses the relationship between prices and traffic management functions such as congestion control, resource provisioning, and call admission control. We summarize proposed pricing policies for the Internet and for ATM networks, as well as some studies of pricing for general QoS-enabled networks.

For years, it has been clear that the integration of multiple services into a single network infrastructure has the potential to generate efficiencies in design, infrastructure and management. It also brings about the question of how to provide adequate quality of service (QoS) to heterogeneous applications.

Data networks have traditionally operated under a best-effort assumption, avoiding the reservation of resources and reaping the full benefits of statistical multiplexing. However, in a truly integrated network, some form of service differentiation (e.g., priority, drop precedence, bandwidth allocation) must be used in order to ensure appropriate QoS to each user. Service differentiation brings about a clear need for incentives to be offered to users to encourage them to choose the service that is most appropriate for their needs, thereby discouraging over-allocation of resources. In commercial networks, this can be most effectively achieved through pricing.

In this article we discuss the possible interactions between pricing and traffic management functions in QoS-enabled networks. In the study of network pricing, one must consider both the choice of a pricing policy (i.e., according to which factors to price a given traffic flow or call) and the setting of prices (i.e., how to find the optimal prices that will result in the desired behavior by users or in the fair allocation of resources).

Users may be charged for network services according to several factors, including service type (e.g., through the use of different price curves for different grades of service), utilization, resource allocation (or some measure such as effective bandwidth), call duration, access bandwidth, call start time, distance, and number of calls. More often than not, price is a combination of several of these factors. Furthermore, pricing policies may be *dynamic*, in which prices fluctuate as a result of some network condition, or *static*, in which prices are independent of network load. In this article we survey some of the important trends in the design of pricing policies for QoS-enabled networks and discuss how these policies are expected to affect user behavior and interact with traffic management functions.

Although cost is an important consideration in the final determination of prices, the subject of cost recovery in providing network services is beyond the scope of this article. For some discussion of the economic viability of QoS-enabled networks and the relationship between capacity costs and statistical multiplexing, the reader can refer to [1–2].

It is important to emphasize that the setting of prices for network services has always been (and will continue to be in the foreseeable future) primarily a marketing and strategic decision rather than an engineering concern. The author believes, however, that the engineering aspects of the problem should play an important role. At the very least, a thorough understanding of how the engineering issues relate to pricing decisions can inform the heuristic process of determining appropriate pricing policies. Proponents of dynamic pricing believe that one should go further and actually set prices according to the current state of the network; we will examine some of these proposals in the sequel.

This article is organized as follows. We will discuss the inter-dependency between pricing and traffic management functions. Widely used models for user behavior and optimization objectives to be achieved in setting prices are presented. The next three sections survey proposed pricing policies for multi-service networks. We will summarize schemes for Internet pricing and discuss Asynchronous Transfer Mode (ATM) pricing. We will list other pricing schemes that can be applied to the general family of networks offering

multiple service classes as we offer some closing remarks and point to some open issues in network service pricing.

PRICING AND TRAFFIC MANAGEMENT

Due to the effect of pricing on traffic management issues, the subject of network pricing has recently been embraced by researchers in the computer communications field, and not simply from an economic perspective. While the proposed implementations vary, the basic idea is that the appropriate pricing policy will provide incentives for users to behave in ways that improve overall utilization and performance. Some of the main network engineering issues that may be affected by pricing are:

Congestion Control — It has often been suggested that pricing can be employed to avoid over-utilization of network resources and as a mechanism for congestion control [3–9]. One way to achieve this effect is to dynamically set prices that reflect the current state of the network. As the network gets congested, prices increase, discouraging its use; when the load returns to manageable levels, prices decrease, encouraging users to offer additional traffic.

Call Admission Control — Some have proposed that the call admission decision take into account the elasticity of the demand for bandwidth and encourage some non-real-time calls to be postponed to times when the network is less heavily loaded [5, 10]. Discussions of the role of pricing in the connection establishment process can also be found in [11–14]. In networks where a traffic contract or service-level agreement (SLA) is established in advance, it is also possible to charge users according to the accuracy of their traffic descriptor, rewarding customers who provide better information on the statistical behavior of their offered traffic. Even with static pricing policies, prices will influence service choices and requested QoS guarantees, thereby indirectly affecting call admission.

Resource Management — Time-of-day pricing and dynamic pricing policies will influence the volume of traffic offered by users and the distribution of traffic over the day. The ability to affect the expected load may be useful to providers when dimensioning the network as well as for managing existing resources. Of particular interest is the problem of how to engineer a more efficient network through pricing [3].

Billing — Billing requires the collection, maintenance, and consolidation of network usage information. The nature of the processing that must be done at network access points as well as the additional traffic produced for the consolidation of billing information are important issues in network engineering. Important requirements of a billing system include that it should impose minimal changes to existing protocols and applications [15].

Other traffic management functions such as routing and policing are also tied in to the issues above and can be directly or indirectly impacted by the choice of pricing policy and the setting of prices.

When designing a pricing policy, one must model user objectives (which will influence service choice and offered traffic) and provider objectives (which will influence prices and policy). We discuss some alternatives for these models next.

UTILITY AND OPTIMIZATION OBJECTIVES

Network users' preferences may be modeled through **utility** functions, which describe how sensitive users are to changes in QoS. It is sometimes useful to think of utility as the amount of money a user is willing to pay for certain QoS guarantees.

Ideally, utility should be expressed as a function of actual QoS parameters, such as delay or packet losses; such is the approach in [16–18]. In most real networks, however, such quality measures are virtually impossible to predict in advance and are closely dependent upon factors such as traffic models, scheduling disciplines and network topology. Therefore, utility is often expressed as a function of the amount of a resource made available to a user by the network, as in [19–24]. In this fashion, the utility still indicates, albeit indirectly, a user's sensitivity to changes in QoS.

Applications exhibit varying degrees of sensitivity to QoS parameters. For instance, real-time voice and video are very sensitive to delay and jitter, while traditional data applications are more sensitive to losses. In order to characterize user behavior under a given pricing policy, it is often necessary to capture the performance sensitivity of various applications into utility functions.

In [16], Cocchi *et al.* propose utility functions for four application types: e-mail, file transfer, remote login, and real-time voice. For e-mail applications, they assume utility to be a decreasing function of both average delay and the percentage of messages not delivered within a delay bound of five minutes; for remote login, user satisfaction decreases with average packet round-trip time; for real-time voice, the important parameters are the average one-way delay and the percentage of voice packets not obeying a tight delay bound of 100 ms. Jiang and Jordan [12] express user benefit for real-time applications as a function of pre-transmission loss (quality degradation that may result from user decisions prior to transmission), while benefit for non-real time service is a function of completion time. Honig and Steiglitz [4] suggest a continuous, monotonically decreasing function of delay to represent each user's utility. Peha and Tobagi [25] quantify users' performance objectives through a cost function, a function of queuing delay.

Instead of characterizing utility as a function of actual predicted QoS, another possibility is to use a function of resources allocated to the flow or the call, which serve as an indication of expected performance. In this case, it is useful to consider both elastic and inelastic applications [24]. Real-time voice and video applications that employ constant bit rate coding require a fixed amount of bandwidth for adequate QoS. These applications are *inelastic* in their demand for bandwidth and therefore it is reasonable to model their utility as a step function, as in Fig. 1a. On the other hand, traditional data applications such as e-mail are *elastic*: they tend to be tolerant of variations in delay and can take advantage of even minimal amounts of bandwidth. The utility function in this case may be represented as in Fig. 1b. There are numerous ways in which real-time applications can be made tolerant of changes in available bandwidth through adaptive coding; however, some minimum bandwidth is nevertheless required. For these *partially elastic* applications, utility functions may take a shape such as in Fig. 1c.

We should note that the assumption that users' utility functions are known *a priori*, while not uncommon, is somewhat controversial. The precise modeling of customer behavior can be, after all, the most critical part of the pricing problem, and by assuming a given utility function we run the risk of oversimplifying the complexities of customer choices and sensitivity to price.

An important example of a project that attempts to experimentally characterize customer behavior in relation to network services is the Internet Demand Experiment (INDEX) [26], conducted at the University of California at Berkeley. In

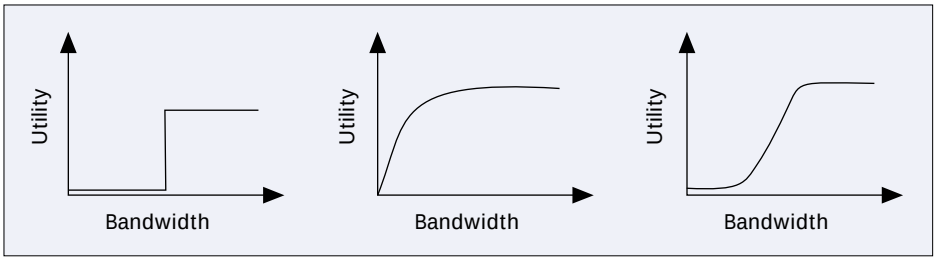

FIGURE 1. *Utility functions [24]: a) inelastic applications; b) perfectly elastic application; and c) partially elastic application.*

this trial, residential users are given differentiated services access to the Internet; users can then choose among several quality/price combinations. The results published to date point to inefficiencies in the current flat-price model adopted by most Internet service providers (ISPs).

Customer surplus is defined as the difference between the maximum amount a customer is willing to pay for a given grade of service and the amount he or she is charged for it, representing what a consumer gains from the trade. In other words, it is the difference between utility and price; a common model [12, 16, 17, 20–22, 27] is that of users whose objective is to maximize their surplus.

Given some expected behavior model for users, setting prices is often treated as an optimization problem, with various possible objective functions. The network may be viewed as a public good, in which case a benevolent service provider would set prices that maximize some measure of social welfare [12, 27] such as aggregate utility or surplus. Alternatively, the network may be viewed as a private good; in this case, one would expect the service provider to be revenue maximizing, as in [11, 18, 21, 28, 29].

Since much of the current interest in the field of network pricing has been sparked by the discussion of how to charge for Internet access, we start by discussing some recent progress on pricing structures for the Internet.

THE FUTURE OF INTERNET PRICING

The study of the technology, economics, and policy surrounding the Internet has intensified over the past few years, due to a combination of several factors. When the National Science Foundation (NSF) decommissioned the NSFnet backbone that served as the core of the Internet, it forced a transition from a government-funded to a commercial Internet. All the while, traffic has been growing at a faster rate than the network infrastructure, and it is widely believed that some sort of congestion control must be exercised in order to ensure adequate use of available resources and avoid a collapse of service. Finally, the Internet evolved under a best-effort assumption; even though the current version of the Internet Protocol (IPv4) provides some functionality for service differentiation, this functionality is by and large not implemented. If the Internet is now to support a variety of heterogeneous applications with acceptable QoS, it must evolve toward a multi-service architecture. Several QoS architectures for the Internet are under study by the Internet Engineering Task Force (IETF), in particular in the Integrated Services (IntServ) [30] and Differentiated Services (DiffServ) [31] working groups.

Each of these factors by itself opens up new questions regarding pricing policies to be adopted in the near future. The transition to a commercially operated Internet brings about the need to recover fixed infrastructure costs through access charges as well as, possibly, usage-based charges. An appropriate pricing policy may also be able to fulfill at least partially the role of congestion control mechanism, encouraging efficient use of a precious resource: bandwidth. Moreover, a multi-service Internet will have to provide appropriate incentives for an efficient alignment of users' needs to offered services. In this article, we focus on this last factor.

One of the debates in Internet pricing has to do with whether flat rates or usage-sensitive charges should be leveled on users. By *usage-sensitive pricing* we mean that prices are a function of the amount of traffic that actually flows through a connection, while *flat rate* refers to a tariff that is independent of the amount of traffic produced as well as of grade of service. A brief summary of the main advantages and disadvantages of each of these two approaches is presented in Table 1.

A flat rate is the method by which Internet access is currently charged in the United States. Low-load users (e.g., e-mail, occasional Web browsing) may therefore be penalized with respect to high-load users (e.g., multimedia applications, frequent downloading of large files). Although fixed costs can be recovered, congestion costs cannot. Also, the unbridled consumption of capacity is likely to lead to overuse of resources, in what has been termed *server overgrazing* [28].

Usage-sensitive pricing addresses the problems of congestion cost recovery and avoids a "tragedy of the commons." However, there is anecdotal evidence that Internet users do not react favorably to usage-sensitive pricing schemes.[1] Certainly, such policies tend to make it more difficult for customers to budget for a network expense that is uncertain. Furthermore, additional management and billing costs to the network provider may be substantial. Finally, usage-sensitive pricing tends to discourage the use of the Internet, a notion that many in the research and academic communities find objectionable. Most of these objections need to be addressed before usage-based pricing schemes become widely used.

A well known recent proposal for usage-based pricing for the Internet is due to MacKie-Mason and Varian [5]. They introduce the concept of *smart market pricing*, where, prior to transmission, users inform the network of how much they are willing to pay for the transmission of a packet; packets are admitted if their bids exceed the current cutoff amount, determined by the marginal congestion costs imposed by an additional packet. Users do not pay the price they actually bid, but rather the market-clearing price, always lower than the bids of all admitted packets. This is an elegant solution for the pricing of a (single-service) network, as it addresses the important problem of *congestion externalities*.[2] The implementation of this type of scheme requires major structural changes not only to network management but also to users' applications, which must be able to submit bids to the network; on the other

[1] In [32] the authors recount that when the Department of Defense decided to implement a usage-based pricing scheme for its inter-agency Internet, the different branches of the military developed their own IP networks, for which they paid the traditional flat-fee model. Similar examples of adverse reaction to usage-based pricing exist in the private sector, where an Internet service provider in Italy had to revert to a flat rate in order to avoid the loss of customers to a competitor [33].

[2] When making decisions regarding their traffic and QoS demands, users will typically take into account prices and their own performance, ignoring the congestion they may impose on others. This phenomenon is known as a congestion externality [5, 34].

hand, it is quick to adapt to changes in network load.

A single-service network would need to be extremely over-provisioned in order to support the performance needs of heterogeneous applications. It seems inevitable that multiple service levels will eventually be offered for the Internet, possibly with the addition of latency and jitter specifications, bandwidth allocation capabilities, and quantitative or qualitative packet delivery guarantees. Moving from a single-service to a multiple-service architecture adds new dimensions to the pricing problem. It is quite clear that a flat rate will not provide adequate incentives for users' choices of services; furthermore, the provider has no knowledge of the value of the information carried in each flow, making it difficult to prioritize traffic for graceful QoS degradation at times of congestion.

Kirkby [13] proposes a separate, circuit-switched, dynamically-priced cut-through layer that allows users to set up QoS-guaranteed circuits for sensitive traffic. The end-user can then choose between traditional flat-rate best-effort service and this premium circuit by pressing an "Internet Turbo" button on her screen. Congestion-based prices are quoted in real-time to the user. It is not difficult to imagine an automated decision process that would use artificial intelligence to make the service choice decision based on current prices and application needs without direct user involvement.

Courcoubetis and Siris [28] present an approach for pricing differentiated services based on an upper bound for effective bandwidth. The network sets prices to reflect the demand for effective bandwidth and publishes a family of pricing curves. With knowledge of these curves, each user selects a traffic contract that includes information of peak bit rate and token bucket parameters. Prices are adjusted within long time scales and can be assumed fixed for the duration of the service level agreement (SLA). This approach can be extended to other network architectures that offer multiple service classes.

Interactions between pricing and emerging standards for IP QoS, such as DiffServ, are still unclear. In fact, current work in the IETF DiffServ working group is aimed at standardizing queuing and scheduling mechanisms for differentiated treatment of datagrams by routers, the so-called *per-hop behaviors* (PHB). It is up to network operators to define end-to-end services built on these PHBs. One possibility is that pricing mechanisms may relate directly to these services, and indirectly to the PHBs themselves. Pricing for IP QoS is at present very much an open problem that must be addressed prior to wide-scale deployment of QoS in the Internet.

We should also note that in some instances IP-level services may be built on top of ATM services, with ATM providing the lower-layer technology. QoS guarantees must then be mapped between the two layers [35]; in effect, pricing will also get mapped. It is not unusual for an Internet service provider to make use of an ATM infrastructure; in this case, the ATM service provider will establish a pricing structure that in turn will affect the prices the ISP will charge users for premium services.

While Internet QoS is still in its infancy, ATM provides a reasonably mature architecture where QoS guarantees are tied into users' choices of service class [36]. Next, we explore some pricing alternatives for ATM.

	Pro	Con
Flat rate	• Easy to implement • Little overhead for billing	• Unfair to light users • No recovery of congestion costs • Server overgrazing • Not appropriate for differentiated QoS
Usage-sensitive pricing	• Can play a role in congestion control • Increased fairness	• Adverse response from customers • Difficult to budget for • Increased billing complexity • May discourage usage

Table 1. *Arguments for and against flat rates and usage-sensitive pricing.*

ATM Pricing

All the large inter-exchange carriers, as well as several local exchange carriers and competitive access local carriers, currently offer ATM service in the U.S. At present, there is no standard pricing policy for the ATM world. Most providers charge a monthly recurring charge (MRC), usually dependent on the access rate (T-1, DS-3, etc.). This MRC can also vary with the distance to the next ATM cross-connect switch or the number of permanent or switched virtual circuits (VCs) used. Most carriers add a component to the price according to how much bandwidth must be allocated to the VC, and some charge according to utilization.

Before we proceed to a discussion of some of the pricing schemes that have been proposed for ATM in the recent literature, we should discuss the concepts of static and dynamic pricing.

Static versus Dynamic Pricing

We can classify pricing policies into *dynamic* policies, in which prices fluctuate as a result of some network condition, and *static* policies, in which the pricing function is independent of current network utilization. Notice that this definition of static policies is broad enough to encompass time-of-day pricing. Time-of-day policies attempt to take advantage of demand elasticity by utilizing historical information about expected peak load periods. However, since prices do not depend upon *current* network load, these schemes still fit our definition of static pricing.

Dynamic policies are more complex and in general require more sophisticated accounting and processing; consequently, they tend to be more costly to implement. Their biggest advantage is that they offer flexibility to react to changes in offered traffic and are better equipped to track the optimal prices to be charged by the network at any given time.

A common criticism of dynamic pricing mechanisms is that they often require a level of computational complexity that may be impractical. Even more important, there are substantial obstacles to user acceptance of dynamic pricing: users may find such policies difficult to understand and to budget for. This explains why virtually all pricing policies now in place for multi-service networks are of a static nature. On the other hand, the main shortcoming of static pricing is that it cannot guarantee optimality in revenues or social welfare.

We next list a representative, while by no means comprehensive, set of recent works on ATM pricing.

Proposed Schemes for ATM Pricing

Table 2 summarizes some of the studies and proposals advanced by some of the leading researchers in the field for the pricing of ATM services. The list illustrates the variety of different approaches as well as the popularity of dynamic pricing schemes as an active research topic.

Reference	Scenario	Summary of general approach and results
Low and Varaiya [27]	Network that offers for rent its bandwidth and buffers	Network periodically adjusts prices based on monitored requests for resources with the objective of maximizing social welfare. Users reserve resources based on individual traffic parameters and delay requirements so as to maximize utilities subject to budget constraints.
J. and L. Murphy [22]	Dynamic adaptive intertemporal priority scheme for ATM networks	At the start of each pricing interval the network announces the price per unit of bandwidth on each virtual path (VP); users then decide how much bandwidth to use. Based on buffer occupancies, prices are updated after each interval to match the marginal cost of each buffer with respect to its total traffic input rate.
J. and L. Murphy, Posner [37]	How to set up VP capacities?	Pricing and user self-regulation can be used as a means of allocating bandwidth in ATM networks.
Murphy *et al.* [7]	Smart market pricing for ATM	The concept of smart market pricing, discussed previously, is applied to ATM networks.
Wang, Peha, and Sirbu [29]	Time-varying price schedule, with the optimal price determined according to demand elasticity and the opportunity cost of providing the service.	For best-effort service, network constantly updates the cutoff price on a per-cell basis, which takes into account both current buffer occupancy and predicted willingness to pay for future calls. Calls are accepted for transmission if and only if the user is willing to pay the cutoff price. For guaranteed services, price reflects the opportunity cost of providing the service, taking into account service characteristics and shadow prices of reserving/using bandwidth. Optimal price is a function of service type, call starting time, and service duration.
Courcoubetis, Siris, Stamoulis [38]	Prices per unit of bandwidth are adjusted at every link at the beginning of each charging interval according to demand.	Based on the prices announced by the network, the user of each connection specifies a bandwidth request. This policy is designed for available bit rate (ABR) service, and the suggested implementation makes use of mechanisms originally provided by the ATM Forum for rate-based flow control.
Anerousis and Lazar [34]	Two services: VP service, equivalent to a leased line; and VC service, for more transitory connections.	Prices per unit of bandwidth would be lower for VP connections. An iterative procedure based on increases and decreases in revenue is used to arrive at optimum prices.
Ramesh *et al.* [39]		Study of the efficacy of using the cell loss priority (CLP) bit to carry streams with differential QoS requirements in an attempt to maximize revenue.
Songhurst and Kelly [14]		Users pay in proportion to the volume of traffic and duration of the call. A connection charge is imposed, and the choices of service and traffic contract parameters also affect the total service.

Table 2. *Summary of some significant proposals from the research community for pricing of ATM services. We list the scenario studied and briefly summarize the general approach.*

One can also envision a static pricing scheme for ATM services, for instance as discussed by this author in [20]. In this scenario, there is a call set-up charge for each call; in addition to this charge, CBR calls incur an allocation-dependent charge; VBR and ABR charges have both allocation and utilization-dependent components; and UBR is not charged according to utilization, in an effort to provide incentives to best- effort traffic, since it imposes low opportunity cost to the network. Such a scheme does not guarantee optimality in either a social or economic sense; on the other hand, it is reasonably easy to understand and implement.

A few recent European projects, such as Charging and Accounting Schemes in Multi-Service ATM Networks (CA$hMAN) [40] and Contract Negotiation and Charging in ATM Networks (CANCAN) [41], have also been looking in detail at the question of ATM pricing.

Other Pricing Schemes for QoS-enabled Networks

Several other works in the literature address the problem of pricing services in QoS-enabled architectures, without assuming a specific underlying set of services. The results and conclusions of these studies often find direct application in Internet differentiated services, ATM, or other architectures.

In [42], Lazar *et al.* analyze a non-cooperative game in which users reserve capacity for their VPs with the objective of minimizing some cost function. They discuss properties of a Nash equilibrium under a dynamic pricing scheme where the price charged per unit bandwidth depends upon the total amount of bandwidth currently reserved by other users.

Jiang and Jordan [12] propose that users be charged a price per unit of effective bandwidth. Assuming that the network knows its trunk capacities and virtual path routing, as well as every user's benefit function and traffic stream characterization, the optimal price that maximizes total user benefit can be calculated.

Ji *et al.* [18] develop a QoS-based pricing scheme that results in efficient utilization of network bandwidth and buffers. Essentially, each class of traffic is charged an amount equivalent to the QoS degradation caused to other users sharing network resources. Price is therefore a function of network utilization as well as individual utilities.

Frank Kelly [21] describes a system in which users reveal how much they are prepared to pay per unit time; the network then determines allocated rates so that the rate per unit charges are proportionally fair. Kelly determines that the system optimum in this case is achieved when users' choices of charges and the network's choices of rates are in equilibrium.

Parris *et al.* [8] compare three pricing policies through simulation: per-packet pricing, where users are charged on the basis of the number of packets transmitted, regardless of service class; set-up pricing, where, in addition to packet charges, there is a set-up charge for establishing a connection; and peak-load pricing, with specified periods of time designated as peak and off-peak and different prices associated with each period. The authors compare the call blocking probability and peak utilization for each policy for a fixed generated revenue.

In [16], Cocchi *et al.* use simulations to study the problem of customer decisions in a two-priority network where a fixed per-byte price is associated with each priority class. They determined that, through the use of class-sensitive pricing, it is possible to set prices so that all users are more satisfied with the combined cost/benefit provided by the network. This study motivated the extension of the model and the more analytical approach in [17]. In [17, 20], DaSilva *et al.* argue for a policy with three price components (set-up, allocation, and usage) for multi-service networks that allocate resources according to a SLA.

Honig and Steiglitz [4] study a simple model in which each user decides whether to transmit based on announced price per packet and QoS. A price schedule containing two entries, a "day price" and a "night price," is considered; results show that traffic smoothing is achievable through proper choice of the price differential.

CONCLUDING REMARKS

The development of the means for price and QoS negotiation consists of a fundamental building block in the implementation of dynamic pricing in commercial networks. Recent work by Wang and Schulzrinne [43] proposing a pricing and resource negotiation protocol and by Kirkby [13] discussing a dynamically priced control protocol (DP-CP) start addressing this important issue.

It is possible to envision a future where QoS levels with arbitrarily fine granularity are available to all Internet users, where prices are set in real time according to current load and taking full advantage of demand elasticity to maximize efficiency and fairness, and where customers are billed for network use in real time using micro-payment systems developed for e-commerce. The research work surveyed here, along with the work of many others, lays the groundwork toward this future. For this futuristic scenario to become reality, however, several open issues must still be resolved, including:

Scalability — In a commercial network, maintaining per-call or per-flow information to be used for billing may not be feasible. The increased processing and storage resources needed to collect usage or allocation information for each switched virtual circuit or (in the case of connectionless networks) for each flow may well negate the potential benefits of usage and allocation-based pricing from a traffic management perspective. Coarser granularity can be used to mitigate the scalability problem.

Hierarchy — It is likely that there will continue to be a hierarchy in network service provision. For instance, an individual residential customer will enter into a contract with an Internet service provider (ISP), which in turn may purchase services from another provider. In this scenario, multiple different pricing schemes may be utilized simultaneously: the network provider may charge the ISP according to usage and allocation, while the ISP charges its customers according to a flat rate for a set number of hours of access plus hourly charges for additional access. Such a scheme would address the scalability problem, with usage and allocation pricing based on the aggregate of traffic produced by the ISP, although it does not offer an incentive for individual users to react to network conditions.

Impact of QoS architectures on end-to-end application performance — In order for QoS-sensitive pricing to be implemented effectively, the relationship between the metrics that are being charged for (bandwidth, bytes transmitted, etc.) and the performance obtained (as measured by response time, subjective quality, etc.) must be made clear to the end user. It may be up to the ISP to provide an abstraction of these lower-layer metrics and map them into metrics that are meaningful to the average user.

Maturity of the market for network services — It is not surprising that the pricing of network services is often done in a heuristic fashion. The market for these services is changing extremely rapidly; furthermore, end users are likely to resent any pricing model that deviates from the current "flat rate." The introduction of QoS differentiation in the Internet will force the pricing structure to evolve into a multi-tiered structure, since without pricing there is no incentive for users to choose the appropriate service level. At first, we speculate that this structure will still be static and fairly simple; however, this may break the barrier toward more complex pricing policies, possibly including dynamic pricing.

ACKNOWLEDGMENTS

The author would like to thank Drs. David Petr and Nail Akar for their support and input. The author is also grateful to the three anonymous reviewers, who contributed valuable, insightful suggestions.

REFERENCES

[1] M. E. Anagnostou *et al.*, "Economic Evaluation of a Mature ATM Network," *IEEE JSAC*, vol. 10, no. 9, Dec. 1992, pp. 1503–509.

[2] I. Saniee, A. Erramilli, and C. D. Pack, "The Economics of Statistical Multiplexing for Broadband Networks," *Teletraffic Contributions for the Information Age*, V. Ramaswami and P. Wirth, Eds., Elsevier, 1997.

[3] A. Gupta *et al.*, "Streamlining the Digital Economy: How to Avert a Tragedy of the Commons," *IEEE Internet Computing*, vol. 1, no. 6, Nov./Dec. 1997, pp. 38–46.

[4] M. Honig and K. Steiglitz, "Usage-based Pricing of Packet Data Generated by a Heterogeneous User Population," *Proc. IEEE INFOCOM*, vol. 2, Boston, MA, Apr. 1995, pp. 867–74.

[5] J. MacKie-Mason and H. Varian, "Pricing the Internet," Public Access to the Internet, B. Kahin and J. Keller, Eds., Prentice–Hall, 1994.

[6] J. MacKie-Mason and H. Varian, "Pricing Congestible Network Resources," *IEEE JSAC*, vol. 13, no. 7, Sept. 1995, pp. 1141–48.

[7] L. Murphy, J. Murphy, and J. K. MacKie-Mason, "Feedback and Efficiency in ATM Networks," *Proc. IEEE ICC*, Dallas, TX, June 1996, pp. 1045–49.

[8] C. Parris, S. Keshav, and D. Ferrari, "A Framework for the Study of Pricing in Integrated Networks," Technical Report, Int'l Comp. Science Institute, Berkeley, CA, 1992.

[9] J. M. Peha, "Dynamic Pricing as Congestion Control in ATM Networks," *Proc. IEEE GLOBECOM*, Phoenix, AZ, 1997, pp. 1367–72.

[10] D. Ferguson *et al.*, "An Economy for Flow Control in Computer Networks," *Proc. IEEE INFOCOM*, Ottawa, Canada, 1989, pp. 110–18.

[11] G. Fodor, E. Nordstrom, and S. Blaabjerg, "Revenue Optimization and Fairness Control of Priced Guaranteed and Best Effort Services on an ATM Transmission Link," *Proc. IEEE Int'l Conf. on Commun.*, vol. 3, 1998, pp. 1696–705.

[12] H. Jiang and S. Jordan, "A Pricing Model for High-Speed Networks with Guaranteed QoS," *Proc. IEEE INFOCOM*, 1996, pp. 888–95.

[13] P. Kirkby, "Business Models and System Architectures for Future QoS Guaranteed Internet Services," *IEE Colloquium on Charging for ATM — The Reality Arrives*, 1997, pp. 11/1–11/8.
[14] D. Songhurst and F. Kelly, "Charging Schemes for Multiservice Networks," *Teletraffic Contributions for the Information Age*, V. Ramaswami and P. Wirth, Eds., Elsevier, 1997.
[15] R. J. Edell *et al.*, "Billing Users and Pricing for TCP," *IEEE JSAC*, vol. 13, no. 7, Sept. 1995, pp. 1162–75.
[16] R. Cocchi *et al.*, "Pricing in Computer Networks: Motivation, Formulation and Example," *IEEE/ACM Trans. Net.*, vol. 1, no. 6, Dec. 1993, pp. 614–27.
[17] L. A. DaSilva, D. W. Petr, and N. Akar, "Equilibrium Pricing in Multiservice Priority-Based Networks," *Proc. IEEE GLOBECOM*, Phoenix, AZ, Nov. 1997, pp. S38.6.1–5.
[18] H. Ji, J. Y. Hui, and E. Karasan, "QoS–Based Pricing and Resource Allocation for Multimedia Broadband Networks," *Proc. IEEE INFOCOM*, 1996, pp. 1020–027.
[19] Z. Cao and E. W. Zegura, "Utility Max-Min: An Application-Oriented Bandwidth Allocation Scheme," *Proc. IEEE INFOCOM*, vol. 2, 1999, pp. 793–801.
[20] L. A. DaSilva, D. W. Petr, and N. Akar, "Static Pricing and Quality of Service in Multiple Service Networks," *Proc. 5th Joint Conf. Information Sciences*, Feb. 27–Mar. 3, 2000, Atlantic City, NJ, vol. 1, pp. 355–58.
[21] F. P. Kelly, "Charging and Rate Control for Elastic Traffic," *European Transactions on Communications*, vol. 8, 1997, pp. 33–37.
[22] J. Murphy and L. Murphy, "Bandwidth Allocation by Pricing in ATM Networks." Technical Report, Dublin City University, Ireland, July 1994.
[23] J. Sairamesh, D. F. Ferguson, and Y. Yemini, "An Approach to Pricing, Optimal Allocation, and Quality of Service Provisioning in High-Speed Packet Networks," *Proc. IEEE INFOCOM*, Boston, MA, Apr. 1995, pp. 1111–19.
[24] S. J. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE JSAC*, vol. 13, no. 7, Sept. 1995, pp. 1176–88.
[25] J. M. Peha and F. A. Tobagi, "Cost-based Scheduling and Dropping Algorithms to Support Integrated Services," *IEEE Trans. Commun.*, vol. 44, no. 2, Feb. 1996, pp. 192–201.
[26] R. J. Edell and P. P. Varaiya, "Providing Internet Access: What we Learn from the INDEX Trial," INDEX Project Report 99-010W, April 1999.
[27] S. H. Low and P. P. Varaiya, "A New Approach to Service Provisioning in ATM Networks," *IEEE/ACM Trans. Net.*, vol. 1, no. 5, Oct. 1993, pp. 547–53.
[28] C. Courcoubetis and V. A. Siris, "Managing and Pricing Service-Level Agreements for Differentiated Services," *7th Int'l Wkshp. on QoS (IWQoS'99)*, 1999, pp. 165–73.
[29] Q. Wang, J. M. Peha, and M. A. Sirbu, "The Design of an Optimal Pricing Scheme for ATM Integrated Services Networks," *Journal of Elec. Pub.*, 1995, Special Issue on Internet Economics.
[30] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: An Overview," RFC 1633, 1994.
[31] S. Blake *et al.*, "An Architecture for Differentiated Services," IETF RFC 2475, Dec. 1998.
[32] L. W. McKnight and J. P. Bailey, "Internet Economics: When Constituencies Collide in Cyberspace," *IEEE Internet Computing*, vol. 1. no. 6, Nov./Dec. 1997, pp. 30–37.
[33] J. P. Bailey and L. W. McKnight, "Internet Economics: What Happens when Constituencies Collide," *INET'95*, Honolulu, HI, June 1995, pp. 659–66.
[34] N. Anerousis and A. A. Lazar, "A Framework for Pricing Virtual Circuit and Virtual Path Services in ATM Networks," V. Ramaswami and P. E. Wirth, Eds., *Teletraffic Contributions for the Information Age*, vol. 2b, Elsevier, 1997, pp. 791–802.
[35] L. A. DaSilva, "QoS Mapping Along the Protocol Stack: Discussion and Preliminary Results," *Proc. 2000 IEEE Int'l Conf. on Commun. (ICC 2000)*, June 18–22, 2000, New Orleans, LA, vol. 2, pp. 713–17.
[36] The ATM Forum Technical Committee, Traffic Management Specification Version 4.0, April 1996 (af-tm-0056.000).
[37] J. Murphy, L. Murphy, and E. Posner, "Distributed Pricing for Embedded ATM Networks," *Int'l. IFIP Conf. Broadband Communications (BB-94)*, Paris, France, Mar. 1994.
[38] C. Courcoubetis, V. A. Siris, and G. Stamoulis, "Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks," *Proc. IEEE GLOBECOM*, London, 1996, pp. 644–48.
[39] S. Ramesh, C. Rosenberg, and A. Kumar, "Revenue Maximization in ATM Networks Using the CLP Capability and Buffer Priority Management," *IEEE/ACM Trans. Net.*, vol. 4. no. 6, Dec. 1996, pp. 941–50.
[40] CA$hMAN: Charging and Accounting Schemes in Multi-Service ATM Networks, web site, http://www.uk.infowin.org/ACTS/RUS/PROJECTS/ac039.htm
[41] CANCAN: Contract Negotiation and Charging in ATM Networks, web site, http://www.teltec.dcu.ie/cancan/
[42] A. A. Lazar, A. Orda, and D. E. Pendarakis, "Virtual Path Bandwidth Allocation in Multi-User Networks," *Proc. IEEE INFOCOM*, Boston, MA, Apr. 1995, pp. 312–20.
[43] X. Wang and H. Schulzrinne, "RNAP: A Resource Negotiation and Pricing Protocol," *9th Int'l Wksp. Network and Operating Systems Support for Digital Audio and Video (NOSSDAV '99)*, Basking Ridge, NJ, June 1999.

ADDITIONAL READING

[1] K. Lindberger, "Cost-Based Charging Principles in ATM Networks," V. Ramaswami and P. E. Wirth, Eds., *Teletraffic Contributions for the Information Age*, vol. 2b, Elsevier 1997, pp. 771–80.
[2] T. Monk and K. Claffy, "Cooperation in Internet Data Acquisition and Analysis," *Coordination and Administration of the Internet*, Cambridge, MA, Sept. 1996.
[3] D. Morris and V. Pronk, "Charging for ATM Services," *IEEE Commun. Mag.*, vol. 37, no. 5, May 1999, pp. 133–39.

BIOGRAPHY

LUIZ A. DASILVA (http://www.ee.vt.edu/~ldasilva/) joined Virginia Polytechnic Institute and State University (Virginia Tech) in the fall of 1998 as an assistant professor in the Bradley Department of Electrical and Computer Engineering. He obtained his Ph.D. at the University of Kansas in 1998. Previously he worked for IBM for six years. He teaches computer engineering and communications courses and is active in asynchronous and distance learning. He also conducts research at Virginia Tech's Alexandria Research Institute. His research interests include QoS, network performance, and traffic management. He is currently involved in funded research projects dealing with performance in ADSL access, IP QoS architectures, and networking issues in the deployment of smart antennas. Current and recent research sponsors include NSF, the Office of Naval Research, ECI Telecom, Sprint, and LGE. He is a senior member of IEEE, a member of ASEE, and a past recipient of the Frontiers in Education New Faculty Fellow award and the Paul Huebner award for excellence in teaching.