Effects of pricing in multiple-service networks with resource allocation

Luiz A. DaSilva[a], David W. Petr[b] and Nail Akar[c]

[a]Virginia Polytechnic Institute and State University, Falls Church, VA 22043

[b]Information and Telecommunication Technology Center, University of Kansas, Lawrence, KS 66045

[c]Sprint Corporation, Overland Park, KS 66212

ABSTRACT

The recent deployment of broadband networks that accommodate applications with diverse quality of service requirements presents new challenges to the pricing of network services. Pricing can and should be used to influence customers to choose services that fit their application needs, maximizing the statistical multiplexing capability of the network. This paper presents a framework for studying the issues involved in pricing for multiple-service networks. Our results illustrate how we can affect customers' choices by charging according to the amount of resources that are allocated to a connection.

Keywords: network pricing, quality of service (QoS), traffic management, asynchronous transfer mode (ATM), networking games

1. INTRODUCTION

The promise of a Broadband Integrated Services Digital Network (BISDN) is gradually becoming a reality, propelled by advances in semiconductor and optical technology.[1] A BISDN must meet diverse application requirements while making efficient use of network resources; this can be accomplished by offering users a number of service choices, with various quality of service (QoS) guarantees. For instance, in Asynchronous Transfer Mode (ATM) networks, service classes range from best-effort service to services with QoS guarantees including maximum cell transfer delay and cell loss ratio.[2] The question then becomes one of how to influence users to choose services that are appropriate for the needs of their applications while at the same time utilizing the network efficiently: too little resource allocation might not meet QoS requirements; too much would be wasteful of resources. The natural solution arises through the design of an appropriate pricing policy.

The pricing of services is an important factor in the deployment of commercial multiple-service networks. In addition to its more traditional roles in the recovery of costs and in competitive strategy, pricing has a direct effect on engineering issues such as the dimensioning of the network (through its influence on offered traffic) and its statistical multiplexing capabilities (through the level of resource allocation requested by users).

The QoS obtained by a user depends not only on her traffic and choice of service but on those of other users as well, increasing the complexity of the pricing problem. The authors have previously studied the impact of pricing on priority-based networks.[3] In this paper, we describe a framework for studying the impact of pricing on user behavior and on the performance of integrated networks with resource allocation. Furthermore, our results illustrate how a pricing policy can encourage users to exhibit behavior that is beneficial to the network as a whole.

Many previous works on pricing for multiple-service networks have focused on *dynamic* policies, where prices change dynamically according to the current status of the network. Dynamic pricing provides some elegant solutions to the pricing problem; however, it may encounter resistance from customers due to the difficulty in budgeting for an expense that is not known *a priori*. Most pricing schemes in place today for commercial networks are of a *static* nature; in this paper, we focus on such policies.

Other author information: (Send correspondence to D.W.P.)
L.A.D. performed this work while affiliated with the Information and Telecommunications Technology Center at the University of Kansas
D.W.P. E-mail: dwp@ittc.ukans.edu – N.A. E-mail: akar@sprintcorp.com

This paper is organized as follows. We start by presenting in Sect. 2 a brief overview of the recent literature on network pricing. Section 3 describes a model of customer behavior using utility functions, as well as the game-theoretic framework employed to predict users' actions based on the existence of a Nash equilibrium. Section 4 discusses networks with allocation of resources to particular services and the simulations used to model such networks; the results of these simulations are presented in Sect. 5. In Sect. 6 we offer an analytical interpretation of these results. Finally, Sect. 7 summarizes our findings and discusses the open questions presently being addressed in our research.

2. PRICING FOR INTEGRATED SERVICES NETWORKS

Due to the effect of pricing on traffic management issues, the subject is addressed increasingly as part of the study of telecommunications network engineering and not purely from an economic perspective. Network engineering issues that may be affected by pricing include:

- **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.[4–10] 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.

- **Connection Admission Control (CAC)** – Some authors have proposed that CAC decisions take into account the elasticity of the demand for bandwidth, so that some calls would be postponed to times when the network is less heavily loaded.[7] It is also possible to charge users according to the accuracy of their traffic descriptor, rewarding customers who provide better information on the characteristics of their offered traffic. In static policies such as described in this paper, pricing will influence how much bandwidth is requested by each user, thereby affecting call admission.

- **Resource Management** – Static time-of-day and dynamic pricing policies will certainly influence the volume of traffic offered by users and the distribution of traffic over the day. The ability to affect the expected load can be used by 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.[5]

- **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 the engineering of the network.

A number of articles have recently been published in engineering and economic journals and conferences addressing the subject of pricing for multi-service networks, most notably with applications to ATM. Most of them concentrate on dynamic policies; we next summarize some of these.

Low and Varaiya[11] study a network that offers for rent its bandwidth and buffers; prices are periodically adjusted based on monitored user requests for resources with the objective of maximizing social welfare. Murphy and Murphy[12] suggest a pricing algorithm where 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 utilize. In Ref. 13, the authors consider the problem of setting up VP capacities, arguing that pricing and user self-regulation can be used as a means of allocating bandwidth in ATM networks. In Ref. 8, Murphy *et al* extend the concept of smart market pricing, first advanced by Mackie-Mason and Varian[7] in the context of single-service networks, to multi-service networks. Smart market pricing addresses the important problem of congestion externalities by having users inform the network of their willingness to pay for service; users are admitted if their bids exceed the market clearing price, which will be higher at times when the network is congested. Wang *et al*[14] propose an optimal pricing scheme for ATM networks using a time-varying price schedule rather than a price for each service. The optimal price is then determined based on demand elasticity and opportunity cost of providing the service. Courcoubetis *et al*[15] describe a policy in which prices per unit bandwidth are iteratively adjusted at every link at the beginning of each charging interval according to current demand. Jiang and Jordan[16] calculate the price per unit of effective bandwidth that maximizes total user benefit. Kelly[17] describes a system in which users reveal how much they are prepared to pay per unit time; the network then determines allocated transmission rates so that the transmission rates per unit price are proportionally fair.

Of particular relevance to this paper is the work of Lazar *et al*,[18] who analyze a non-cooperative game in which users reserve capacity for their VPs with the objective of minimizing some cost function. As in Ref. 18, we utilize a game-theoretic framework for the problem. However, important distinctions exist between the two models. The pricing policy studied in Ref. 18 is dynamic. Their model accounts for the reservation of resources, with no statistical multiplexing. Furthermore, the approach to users' objective functions is significantly different, allowing different assumptions to be made and often resulting in quite different conclusions.

These dynamic solutions to the pricing problem are flexible enough to react to changes in offered traffic and can track the optimal price to be charged by the network at any given time. However, these policies are usually costly to implement and often require modifications to users' applications. The desirability of static pricing policies stems from the fact that they are considerably easier for the network provider to implement and for users to understand and accept. Perhaps this explains why virtually all pricing policies now in place for commercial networks are of a static nature.

The published works about *static* pricing for multi-service networks distinguish themselves not so much for the novelty of the proposed policies but rather for the analysis of how these policies affect users' service choices, network utilization, performance, user satisfaction and provider revenue.

In Refs. 19,4, 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. This study motivated the extension of the model and the more analytical approach taken by the present authors in Ref. 3. Parris *et al* [9] compare several pricing policies through simulation and conclude that peak load pricing is a useful tool for congestion avoidance and that having a set-up charge in addition to per-packet charges seems advantageous from a performance standpoint.

In this paper, we assume a static pricing policy for an allocation-based network. We claim that, by adopting an appropriate pricing policy, the service provider can induce users towards service choices that are socially desirable. Since analytical results for queueing systems that allow the allocation of resources are scarce, the results reported here are accomplished initially through simulations. Analysis supporting the simulation results is then presented.

3. USER MODEL AND NETWORKING GAMES

In any study of pricing, it is essential that we characterize the tradeoffs customers are willing to make between their satisfaction with a service received and its price. A common approach is to recognize that individual users associate a value to each level of service; this value, hereby referred to as the user's *utility*, can be interpreted as the amount the user is willing to pay for a given QoS.

Utility functions have been employed in Refs. 4,17,12,10; an essentially equivalent approach is taken in Ref. 20, expressing performance objectives in terms of cost functions. Exact values for customers' utilities are difficult, if not impossible, to obtain.* We take the usual approach of postulating utility functions that reflect the known QoS demands of each application.

Consider the case of real-time video. For such applications, the QoS measures of interest are the number of packets lost in transit (due, for instance, to buffer overflow at the switches), and the number of packets delivered with excessive delay, which cannot be used for playback in real time.

The RACE (Research on Advanced Communication in Europe) consortium deems delay in excess of 1000 ms as unacceptable for broadcast video.[1] Taking into account a propagation delay of up to 25-30 ms (typical of coast-to-coast transmission in the United States), and the number of ATM switches between transmitter and receiver in the range of 4 (corresponding to a fully-meshed, core/edge topology) to 8, we postulate the following utility as a function of delay at a particular switch:

$$U(T1, T2) = \max(0, U_0 - \alpha_1 T1 - \alpha_2 T2) \tag{1}$$

where U_0 is the maximum amount the user is prepared to pay for the service, T1 is the proportion of packets whose delay exceeds 160 ms (including un-delivered packets, with "infinite" delay) and T2 is the proportion of packets delivered with delay between 100 and 160 ms. The constants α_1 and α_2 represent the customer's sensitivity to changes in QoS; clearly, $\alpha_1 > \alpha_2$. This function is illustrated in Fig. 1, with $U_0 = 100$, $\alpha_1 = 10^4$ and $\alpha_2 = 500$.

*Some limited information can occasionally be obtained. For instance, Arnott and Small[21] have studied the economics of highway traffic and through drivers' actions have been able to determine their willingness to pay for saved travel time.

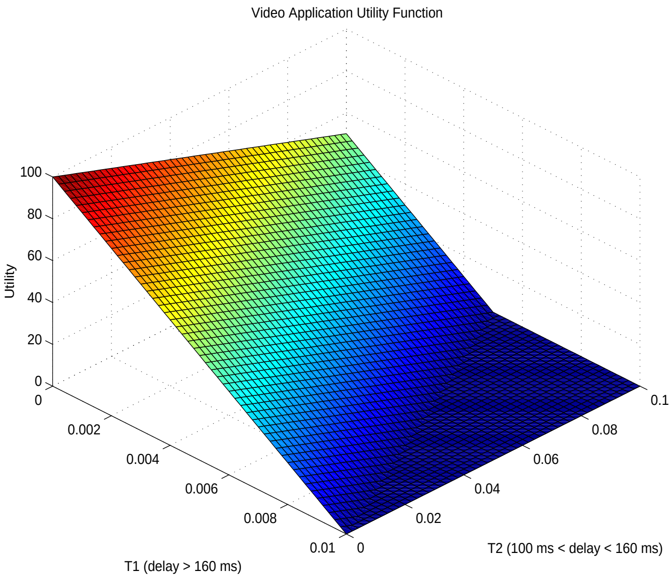

Figure 1. Proposed utility function for video applications.

The heuristic arguments behind the choice of this particular utility function are as follows. Real-time video, like voice, is stream traffic, so utility essentially depends upon the percentage of traffic with delay exceeding some bound that corresponds to the constant built-out delay of the connection (for a single switch, we assume this delay bound to be 160 ms). Since the simulation was conducted using short video segments, cells arriving with delay between 100 and 160 ms were used as an indication of an increased probability of future cells being delivered after their playback time.

We define *user surplus* as the difference between the utility derived from a given service choice and the price paid for the service; as in Refs. 17,11,12, surplus maximization is considered to be users' primary objective. A negative surplus indicates that the consumer is no longer willing to use the service. It is clear that a user's surplus will depend not only on her service choice but also on those of other network users; the interdependency among all users' decisions makes this problem particularly suited for a game-theoretic approach.

Game theory has been used for years as a tool of economic analysis, to understand and predict what will happen in economic contexts[22]; more recently, its concepts were applied to networking problems as diverse as flow control,[23] congestion control,[10] routing,[24] provisioning of resources[25] and pricing.[4,3,18]

The basic idea behind the model we use was delineated in the pioneering work of Cocchi *et al.*.[4] The problem is treated as a non-cooperative game, where users independently choose the appropriate strategy (*i.e.*, service class) to maximize their surplus. A *Nash equilibrium* is then defined as a joint strategy where no user can unilaterally increase her surplus by changing her strategy.[22] If a unique Nash equilibrium for the game exists, we consider it to be a good prediction of the outcome. Besides uniqueness, another desirable property of an equilibrium is that it be efficient. We use the concept of *Pareto optimality* in order to determine the efficiency of an equilibrium;[22] also, if multiple equilibria exist but only one is Pareto optimal then we consider it to be the most likely outcome. A strategy is Pareto optimal if there is no other joint strategy which one or more users would prefer and to which all others would be indifferent.

The pricing policy will determine the location of the Nash equilibria. Conversely, if there is an equilibrium that maximizes some measure of total welfare, we contend that, in most cases, a pricing scheme can be designed so that this equilibrium is achieved.

4. NETWORKS WITH RESOURCE ALLOCATION

In order to ensure a certain minimum QoS to users, networks must allocate resources, typically bandwidth and switch buffers, to each connection. For example, in ATM networks, services for which cell losses are guaranteed to be below

Table 1. Source characteristics. Peak rate is measured over a sliding window of 10 frames. Equivalent bandwidth is calculated as by Guerin,[28] for a maximum overflow probability of 10^{-8}.

	Description	Average Rate [Mb/s]	Peak Rate [Mb/s]	Mean Burst Length [cells]	Equivalent Bandwidth [Mb/s]
TV1	1996 NCAA Basketball	3.46	4.1	218.1	3.9
TV2	ABC Evening Shows	3.04	4.11	173.6	3.48
TV3	Headline News	3.44	5.18	202.3	3.87

a certain threshold include Constant Bit Rate (CBR) and Variable Bit Rate (VBR) services; users may also choose Unspecified Bit Rate (UBR) services, for which no such guarantees are made.[2]

It stands to reason[†] that in networks that allow the allocation of resources, the quantity allocated (as well as call duration) will be a major factor in determining the price of the service. Conversely, the price of allocating a given resource will influence users' choices of service, and pricing can be used in order to guide users towards choices that are mutually beneficial to users and providers.

In this fashion, it is in everyone's best interest that each user allocate neither more nor less than is needed to achieve the QoS required by her application. Over-allocation increases the likelihood that future, possibly more profitable, calls will be blocked by CAC, to the service provider's disadvantage; from a user's point of view, it also decreases consumer surplus. Similarly, under-allocation may result in poor QoS for users (decreasing utility).

We illustrate these principles through the simulation described next.

4.1. Simulation Model

We simulate the case of three independent sources competing for the same output link on a single network node. Each user (source) has the choice of three levels of bandwidth allocation:

- Best effort service (no allocation of resources);
- Allocation of some measure of "equivalent capacity" (a value between the average and the peak source rates);
- Allocation of enough bandwidth to accommodate the peak source rate.

We denote these as service classes 0, 1 and 2, respectively. In the context of ATM networks, these service classes can be interpreted as rough approximations to UBR, VBR and CBR service, respectively. Recently, there has been some discussion about adding a new service class definition to ATM.[26] This service class, sometimes called UBR+, would provide a minimum service rate guarantee, possibly determined by the Minimum Cell Rate (MCR) parameter. An alternate way of interpreting the service definitions above in the context of ATM would be as UBR+ services with varying MCR values.

The sources, denoted as TV1, TV2 and TV3, are video traces generated from television shows[27] using JPEG coding.[‡] A statistical profile of the sources is listed in Tab. 1. The frame generation rate is depicted in Fig. 2. The values of peak transmission rate in Tab. 1 are obtained over a sliding window of ten frames, smoothing some of the peaks in Fig. 2.

The equivalent capacity for service type 1 was calculated using the method developed in Ref. 28, as an approximation for the bandwidth requirements of each source. Variations of this method are currently used by several commercially available ATM switches.

Traffic scheduling is accomplished in our simulation through the use of a weighted round robin (WRR) scheduler. This scheduler implements a service discipline that allows the allocation of a portion of the output link bandwidth to each virtual circuit (VC); whenever a VC is not using its allocated bandwidth, it becomes available for use by other

[†] We provide formal analytical justification in Sect. 6.

[‡] JPEG is the image compression standard developed by the Joint Photographic Experts Group.

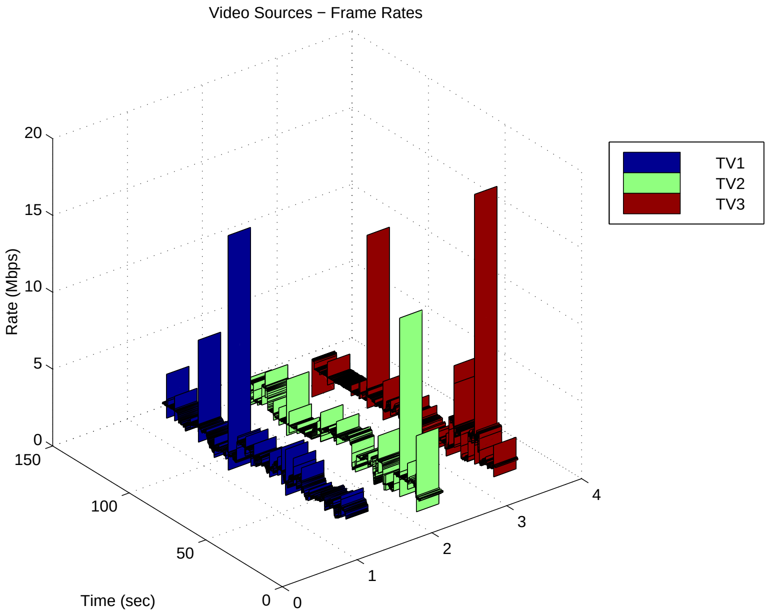

Figure 2. Frame rates for the three video sources.

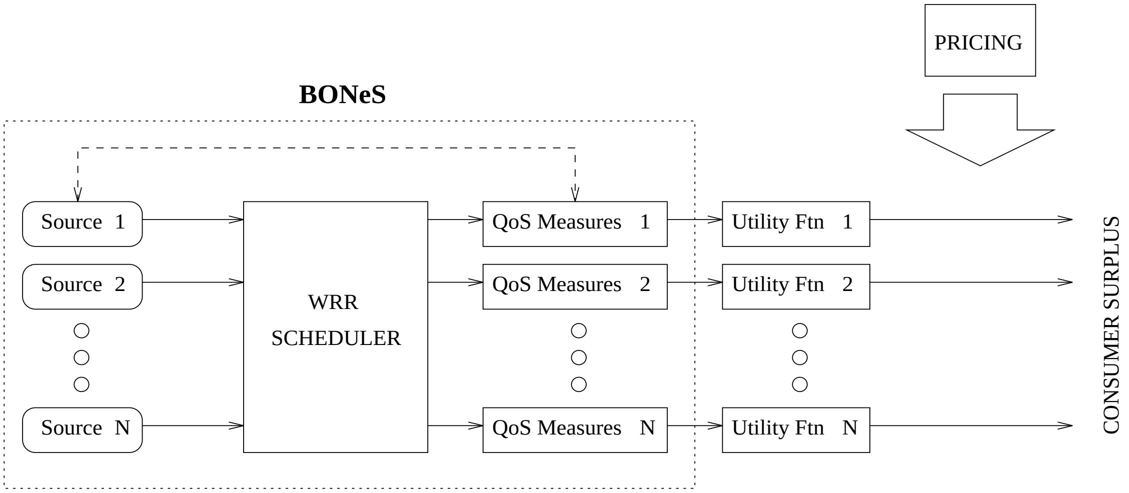

Figure 3. Block diagram of simulation.

VCs. A separate queue is maintained for each VC. At present, several ATM switches implement WRR algorithms, making WRR scheduling a popular method for resource allocation in currently deployed commercial ATM networks.

A block diagram of our simulation is shown in Fig. 3. The simulation was carried out using BoneS (Block Oriented Network Simulation), a software package for modeling and simulation of communication networks.[29] Each frame is segmented into cells, which in turn get time-stamped; cell headers also carry VC identifiers. The WRR scheduler has a high number of cell buffers (30,000); at its output, several QoS indicators are measured, including average cell delay, queue occupancy and a histogram of cell delay. QoS statistics allow us to calculate each user's utility according to equation 1; in conjunction with the pricing scheme, this will determine the user surplus.

The end result of a set of simulations is a matrix containing the user surplus for each service combination. If we have N users and S service classes, there are S^N possible joint strategies $(s_1, s_2, \ldots, s_N)$, $s_i \in \{0, 1, \ldots, S-1\}$. From this matrix, one can determine which service combinations constitute Nash equilibria.

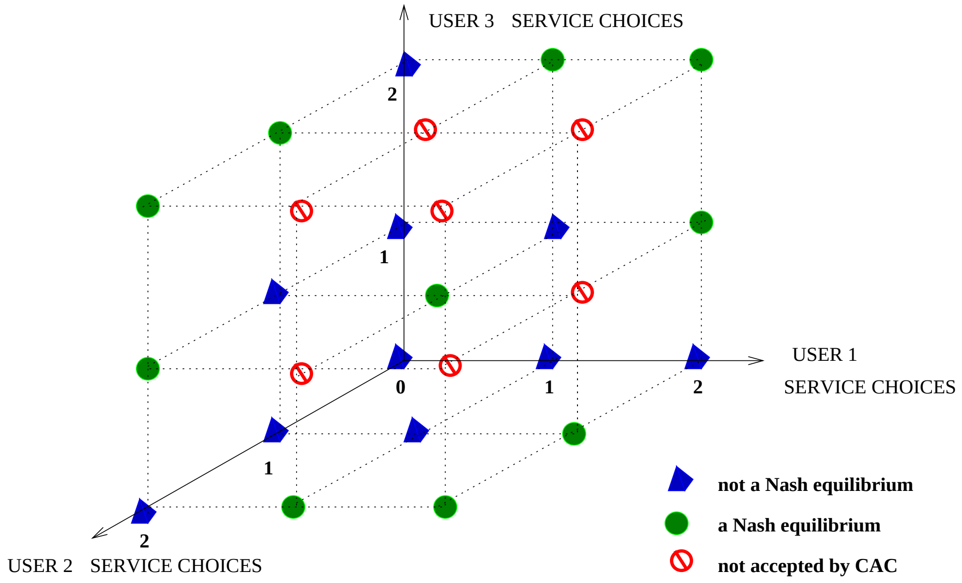

Figure 4. Location of Nash equilibria.

5. RESULTS

The results of the simulations described in the previous section depend upon the average link utilization ρ, calculated as the ratio of the average aggregate arrival rate at the node to the output link rate. Clearly, for low enough utilization, QoS is the same regardless of the service class chosen by each user. In particular, if the output link rate is greater or equal to the sum of the peak rates for all users, queue occupancy is always kept to a minimum and no bandwidth allocation is needed to ensure adequate QoS. In practice, since the sources are not synchronized, peak rates do not occur simultaneously (see Fig. 2) and even for values of link rate lower than the aggregate of the peak rates the previous observation still applies.

We concentrate therefore on the case of a heavily loaded network ($\rho = 0.88$). As a basis for comparison, let us first consider a service-insensitive pricing scheme: users pay a fixed access charge regardless of the service chosen. When this is the case, one would expect all users to choose service class 2; however, this is not possible since the bandwidth reserved would exceed the available capacity, and such a strategy combination would be rejected by CAC. The result is then illustrated in Fig. 4. The service triplet (2,2,2), the obvious candidate for a Nash equilibrium, is not accepted by CAC; neither are triplets such as (2,1,1) and (1,1,2). Indeed, one can observe in Fig. 4 that the equilibrium service combinations are those adjacent to the combinations rejected by CAC (and only those combinations are equilibria). The intuitive explanation is that, in the absence of pricing that is sensitive to the service choice, users will attempt to maximize the amount of resources allocated to their calls; if that service request is rejected, users will settle for the service category that allocates as much as possible of a resource to their calls.

Service triplets such as (2,2,0) are undesirable equilibria; they are wasteful of resources, since two of the users are allocating bandwidth in excess of what they need, while the other is not obtaining adequate QoS. Ideally, the network provider would like for triplet (1,1,1) to be the unique equilibrium, reducing the probability of blocking incoming calls while still providing adequate QoS to all users. By imposing a price P_{bw} per unit of reserved bandwidth, this unique equilibrium can be achieved. Figure 5 shows the relationship between the bandwidth cost and the number of equilibria in the system. We must note that, for any bandwidth cost, the service triplet (1,1,1) is the only combination that can be achieved with uniqueness. At the two extremes, the results are intuitive: when $P_{bw} = 0$, the equilibria are as shown in Fig. 4; when customer surplus goes to zero (high values of P_{bw}), every reachable service combination becomes an equilibrium. More importantly, there is a range for unit cost of bandwidth ($0 < P_{bw} \leq 3.31$) that yields the desired unique equilibrium. The exact range of P_{bw} is related to users' utility functions; however, the same general results were observed for several variations on the utility function in equation 1.

A simpler pricing policy would consist of assigning a fixed cost P_i per service class i. Figure 6 shows the number of Nash equilibria for several values of (P_1, P_2), while P_0 is fixed at 0. Again, appropriate choices of (P_1, P_2) yield the

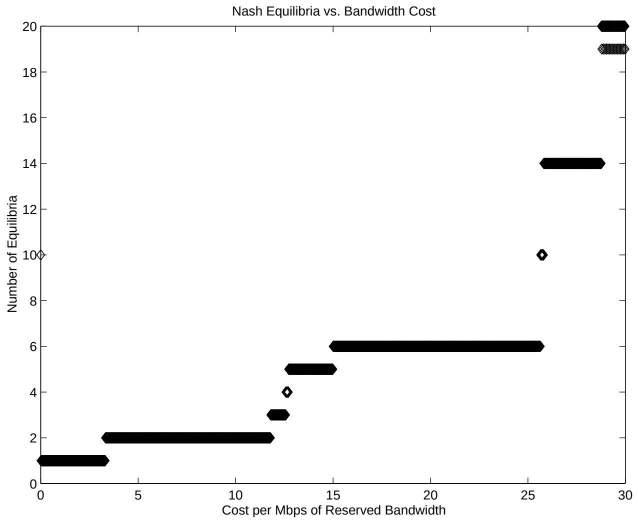

Figure 5. Nash equilibria as a function of unit bandwidth cost.

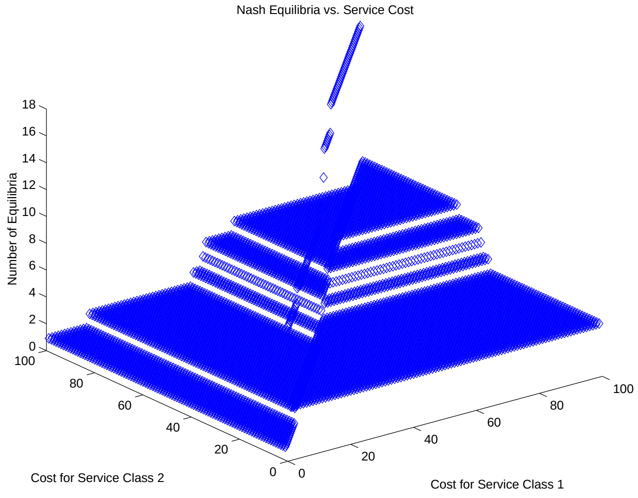

Figure 6. Nash equilibria as a function of service costs.

desired result. This pricing policy, however, is bound to be unfair when a more heterogeneous traffic mix is presented to the network.

Notice that, if the three sources were statistically identical, the first pricing policy would simply be a special case of the second. The results in Fig. 5 are essentially a cross section of those in Fig. 6 when the plot is bisected by a plane perpendicular to the horizontal plane and at an angle determined by the ratio of the bandwidth allocated for service class 2 to that allocated for service class 1.

Finally, a note on equivalent capacity. There is significant interest in the subject of determining, based on source statistics, how much bandwidth is needed to guarantee a certain maximum packet loss ratio or average delay. Our preliminary results seem to indicate that, in a quasi-stationary environment, a pricing policy that takes into account the amount of bandwidth reserved to a source may enable the user to make that determination through a learning process based on her consumer surplus.

6. ANALYTICAL INTERPRETATION OF RESULTS

It is possible to obtain a more rigorous justification for the simulation results discussed in the previous section. Consider a set of users $\mathcal{N} = \{1, 2, \cdots, N\}$ competing for a single link of bandwidth L. Users are allowed to request an allocation of bandwidth s_i for the duration of their calls. Notice that in this model, service "classes" are represented by a continuum of bandwidth, rather than (as in the simulation) discrete levels of allocation. A user's strategy space is then $\mathcal{S}_i = [0, L]$, and calls are admitted to the network as long as $\sum_{i \in \mathcal{N}} s_i \leq L$ (*i.e.*, overbooking is not allowed). We denote the joint strategy as $\mathbf{s} = \times_{i \in \mathcal{N}} s_i$. We also adopt the following notation: $\mathbf{a}_{-i}$ represents all components of vector $\mathbf{a}$ *except* its i^{th} component, and $(k, \mathbf{a}_{-i})$ represents vector $\mathbf{a}$ with its i^{th} component substituted by k.

The amount of bandwidth available to user i $(b_i(\mathbf{s}))$ is equal to the amount allocated (s_i) plus a portion of the total unallocated bandwidth; for simplicity, we assume unallocated bandwidth is uniformly distributed among all users. The traffic from each user is characterized by the transmission rate $\hat{b}_i(\mathbf{s})$, which depends on $b_i(\mathbf{s})$ as well as on the characteristics of the application itself; traffic policing may be employed to ensure that $b_i(\mathbf{s})$ is an upper bound on $\hat{b}_i(\mathbf{s})$. Sources are assumed to be greedy.

We consider each user's utility to be a function of the bandwidth available to the user: $U_i = U_i(b_i)$. This preserves, albeit indirectly, the concept of utility as a function of achieved QoS. We say a utility function is *bandwidth-limited* if there exists a least upper bound (l.u.b.) on the intervals for which it is a strictly increasing function of b_i; we denote this l.u.b. as $\tilde{b}_i$.

Finally, we consider price to be a combination of three components: a fixed connection charge; an allocation-based component; and a usage-based component. This is represented in equation 2, with f and g being monotonic increasing functions:

$$P_i = c + f(s_i) + g(\hat{b}_i) \tag{2}$$

Several of the results in this section depend on the following lemma.

LEMMA 6.1. *Take* $\mathbf{s}^* \in \mathcal{S}$ *and* $i \in \mathcal{N}$. *Then*

1. *For* $0 < \epsilon \leq s_i^*$, $b_i(s_i^* - \epsilon, \mathbf{s}_{-i}^*) < b_i(\mathbf{s}^*)$;
2. *For any* $\epsilon > 0$ *such that* $(s_i^* + \epsilon, \mathbf{s}_{-i}^*) \in \mathcal{S}$, $b_i(s_i^* + \epsilon, \mathbf{s}_{-i}^*) > b_i(\mathbf{s}^*)$.

If any individual user decreases her allocation while the others' remain fixed, then the amount of bandwidth available to her is also decreased. The converse is true if the user increases her allocation. We omit the trivial proof.

We are ready to formally justify the assertion that pricing is more effective in influencing user behavior when aggregate demand for bandwidth is high.

PROPOSITION 6.1. *Let* $b_+ = \max_{i \in \mathcal{N}} \tilde{b}_i$. *If* $L \geq N b_+$, *then* $\mathbf{s}^* = \mathbf{0}$ *is a Nash equilibrium.*

If, additionally, f *is strictly increasing, then* $\mathbf{s}^* = \mathbf{0}$ *is Pareto optimal.*

If demand is low enough as compared to the total available bandwidth, not allocating any bandwidth is an equilibrium regardless of the pricing policy. Now suppose the allocation-based component of price is strictly increasing. Whenever bandwidth is plentiful, not allocating bandwidth is not only a Nash equilibrium but is also efficient (Pareto optimal).

We make an analogy with traffic management: if utilization is very low, traffic management becomes superfluous, since the likelihood of congestion is low and users are bound to obtain acceptable service quality. It is at times of high utilization that precise traffic management is needed; fortunately, at such times pricing can play an important part in inducing users to reveal their true bandwidth requirements and/or traffic characteristics, very valuable information for the management of the network.

PROOF 6.1. *Notice* $b_i(\mathbf{s}^*) = L/N \geq b_+ \geq \tilde{b}_i \ \forall \ i$, *so* $\hat{b}_i(\mathbf{s}^*) = \tilde{b}_i$.

Take $0 < a \leq L$, *and notice that* $b_i(a, \mathbf{s}_{-i}^*) \geq \tilde{b}_i$ *by Lemma 6.1. Therefore, since* $f(\cdot)$ *is non-decreasing,*

$$C_i(a, \mathbf{s}_{-i}^*) = U_i(\tilde{b}_i) - c - f(a) - g(\tilde{b}_i) \leq U_i(\tilde{b}_i) - c - f(0) - g(\tilde{b}_i) = C_i(\mathbf{s}^*) \tag{3}$$

The statement above shows $\mathbf{s}^*$ *is a Nash equilibrium. Now suppose it is* not *Pareto optimal. Then* $\exists \ \hat{\mathbf{s}} \in \mathcal{S}$, $\hat{\mathbf{s}} \neq \mathbf{0}$, *such that* $C_i(\hat{\mathbf{s}}) \geq C_i(\mathbf{s}^*) \ \forall \ i \in \mathcal{N}$ *and* $C_k(\hat{\mathbf{s}}) > C_k(\mathbf{s}^*)$ *for some* $k \in \mathcal{N}$. *Assuming* f *to be strictly increasing, if* $\hat{s}_i \neq 0$ *and* $b_i(\hat{\mathbf{s}}) \geq \tilde{b}_i$, *then* $C_i(\hat{\mathbf{s}}) < C_i(\mathbf{s}^*)$. *On the other hand, if for some* i $\hat{s}_i \neq 0$ *and* $b_i(\hat{\mathbf{s}}) < \tilde{b}_i$, *then* $\exists \ j \in \mathcal{N}$

such that $\hat{s}_j \neq 0$ and $b_j(\hat{\mathbf{s}}) > L/N \geq b_+ \geq \tilde{b}_j$. In either case, for at least one user the surplus is lower *with joint strategy $\hat{\mathbf{s}}$. We have reached a contradiction, showing that $\mathbf{s}^*$ is Pareto optimal.*

The simulation results also indicated that, in the absence of an allocation-based pricing component, users can be expected to allocate as much bandwidth as allowed, regardless of actual need. We offer the following propositions:

PROPOSITION 6.2. *Under flat rate pricing, if for each $i \in \mathcal{N} \nexists \hat{s}_i > s_i^*$ such that $(\hat{s}_i, \mathbf{s}_{-i}^*) \in \mathcal{S}$, then $\mathbf{s}^*$ is a Nash equilibrium.*

This proposition says that, under a flat rate, it is always a Nash equilibrium for the aggregate bandwidth allocated by all users to equal the total bandwidth offered in the system, even if the bandwidth allocated by a particular user exceeds her needs. Clearly, in this case the induced Nash equilibrium is not necessarily Pareto optimal.

PROOF 6.2. *Take $\mathbf{s}^* \in \mathcal{S}$ such that for each $i \in \mathcal{N} \nexists \hat{s}_i > s_i^*$ with $(\hat{s}_i, \mathbf{s}_{-i}^*) \in \mathcal{S}$. Notice from Lemma 6.1 that $b_i(\mathbf{s}^*) \geq b_i(s_i, \mathbf{s}_{-i}^*)$ $\forall$ s_i such that $(s_i, \mathbf{s}_{-i}^*) \in \mathcal{S}$. Since utility functions are monotonically increasing in b_i, it also holds under these conditions that $U_i(\mathbf{s}^*) \geq U_i(s_i, \mathbf{s}_{-i}^*)$. Surplus is the difference between utility and price; under flat rate pricing, price is independent of $\mathbf{s}$. Therefore, $C_i(\mathbf{s}^*) \geq C_i(s_i, \mathbf{s}_{-i}^*)$, and by definition $\mathbf{s}^*$ is a Nash equilibrium.*

Furthermore, instituting a usage-based charge is still not sufficient to prevent users from overallocating. This is made clear if we consider elastic users,[30] adopting a piecewise-linear approximation to their utility functions, with constant marginal utility A/λ_i for $b_i < \lambda_i$ and $\tilde{b}_i = \lambda_i$.

PROPOSITION 6.3. *Suppose pricing is simply a linear function of utilization with proportionality constant k_g. If for each $i \in \mathcal{N} \nexists s_i > s_i^*$ such that $(s_i, \mathbf{s}_{-i}^*) \in \mathcal{S}$, then $\mathbf{s}^*$ is a Nash equilibrium.*

As with flat-rate pricing, joint strategies that allocate all available bandwidth (regardless of need) are equilibria.

PROOF 6.3. *Take any s_i such that $(s_i, \mathbf{s}_{-i}^*) \in \mathcal{S}$. Then by the condition on the proposition, $s_i \leq s_i^*$, and by Lemma 6.1 $b_i(\mathbf{s}^*) \geq b_i(s_i, \mathbf{s}_{-i}^*)$. Let us now consider two separate cases:*

1. $b_i(\mathbf{s}^*) \geq \lambda_i$

   If $b_i(s_i, \mathbf{s}_{-i}^*) \geq \lambda_i$, then it follows trivially that $C_i(\mathbf{s}^*) = C_i(s_i, \mathbf{s}_{-i}^*)$. On the other hand, if $b_i(s_i, \mathbf{s}_{-i}^*) < \lambda_i$, then $\hat{b}_i(s_i, \mathbf{s}_{-i}^*) = b_i(s_i, \mathbf{s}_{-i}^*)$ and so

$$C_i(s_i, \mathbf{s}_{-i}^*) = (\frac{A}{\lambda_i} - k_g) b_i(s_i, \mathbf{s}_{-i}^*) < (\frac{A}{\lambda_i} - k_g)\lambda_i = C_i(\mathbf{s}^*) \tag{4}$$

2. $b_i(\mathbf{s}^*) < \lambda_i$

   In this case, $\hat{b}_i(\mathbf{s}^*) = b_i(\mathbf{s}^*)$ and $\hat{b}_i(s_i, \mathbf{s}_{-i}^*) = b_i(s_i, \mathbf{s}_{-i}^*)$, so

$$C_i(s_i, \mathbf{s}_{-i}^*) = (\frac{A}{\lambda_i} - k_g) b_i(s_i, \mathbf{s}_{-i}^*) \leq (\frac{A}{\lambda_i} - k_g) b_i(\mathbf{s}^*) = C_i(\mathbf{s}^*) \tag{5}$$

In both cases, $C_i(\mathbf{s}^*) \geq C_i(s_i, \mathbf{s}_{-i}^*)$ $\forall$ s_i such that $(s_i, \mathbf{s}_{-i}^*) \in \mathcal{S}$, and therefore $\mathbf{s}^*$ is a Nash equilibrium.

Finally, the problem is solved by the addition of a charge for allocated bandwidth. Keeping the assumption of elastic users, we can state:

PROPOSITION 6.4. *Suppose prices strictly increase with amount of bandwidth allocated. If $\mathbf{s}^* \in \mathcal{S}$ is such that $s_i^* > \lambda_i$ for some $i \in \mathcal{N}$, then $\mathbf{s}^*$ is not a Nash equilibrium.*

So, while fixed charges and usage-based charges allow the possibility that at an equilibrium users will allocate bandwidth in excess of what they can utilize, the provider can effectively prevent over-allocation by adopting allocation-based pricing.

PROOF 6.4. *Take $\mathbf{s}^*$ such that $s_i^* > \lambda_i$ for some i. Then it is possible to choose $\epsilon > 0$ such that $s_i^* - \epsilon > \lambda_i$ and $(s_i^* - \epsilon, \mathbf{s}_{-i}^*) \in \mathcal{S}$. The customer surplus $C_i(\mathbf{s}^*) = A_i - f(s_i^*) - g(\lambda_i) - c$, while $C_i(s_i^* - \epsilon, \mathbf{s}_{-i}^*) = A_i - f(s_i^* - \epsilon) - g(\lambda_i) - c$. Since $f(\cdot)$ is strictly monotonic, $C_i(s_i^* - \epsilon, \mathbf{s}_{-i}^*) > C_i(\mathbf{s}^*)$ and $\mathbf{s}^*$ is not a Nash equilibrium.*

7. CONCLUSIONS AND FURTHER WORK

Pricing is an important, and often neglected, factor influencing the performance of a BISDN. Whenever a network offers different grades of service (as integrated networks must, if they are to guarantee appropriate QoS to a diverse range of applications), the pricing policy must be carefully designed in order to make efficient use of resources.

In order to illustrate the claim that pricing can be effectively used to influence users towards a desired strategy, we have simulated the case of three video sources competing for a common link. The results of the simulations indicate that, at high network utilization, the pricing policy controls the location of the Nash equilibria, and the appropriate policy will yield the desired unique equilibrium. We expect that more sophisticated policies that take into account actual usage as well as bandwidth allocation will be needed to yield similar results when a more heterogeneous traffic mix is present.

Our goal is to generalize and extend the results presented here by using an analytical approach. To this end, in Section 6 we present a model for networks that allow the allocation of resources and are able to obtain initial analytical results that confirm the conclusions indicated by the simulation.

The next step is to apply such results to practical multiple-service networks, such as ATM, compiling a set of general recommendations to aid service providers in the design of an adequate pricing policy.

ACKNOWLEDGMENTS

This research has been partially supported by Sprint Corporation.

REFERENCES

1. M. de Prycker, *Asynchronous Transfer Mode – Solution for Broadband ISDN*, Prentice Hall, 3rd ed., 1995.
2. The ATM Forum Technical Committee, *Traffic Management Specification Version 4.0*, Apr. 1996. af-tm-0056.000.
3. L. A. DaSilva, D. W. Petr, and N. Akar, "Equilibrium Pricing in Multiservice Priority-Based Networks," in *Proc. of IEEE GLOBECOM*, pp. S38.6.1–5, (Phoenix, AZ), Nov. 1997.
4. R. Cocchi, S. J. Shenker, D. Estrin, and L. Zhang, "Pricing in Computer Networks: Motivation, Formulation, and Example," *IEEE/ACM Transactions on Networking* **1**, pp. 614–627, Dec. 1993.
5. A. Gupta, B. Jukic, M. Parameswaran, D. O. Stahl, and A. B. Whinston, "Streamlining the Digital Economy: How to Avert a Tragedy of the Commons," *IEEE Internet Computing* **1**, pp. 38–46, November/December 1997.
6. M. L. Honig and K. Steiglitz, "Usage-based Pricing of Packet Data Generated by a Heterogeneous User Population," in *Proc. of IEEE INFOCOM*, vol. 2, pp. 867–874, (Boston, MA), Apr. 1995.
7. J. K. MacKie-Mason and H. R. Varian, "Pricing the Internet," in *Public Access to the Internet*, B. Kahin and J. Keller, eds., Prentice-Hall, 1994.
8. L. Murphy, J. Murphy, and J. K. MacKie-Mason, "Feedback and Efficiency in ATM Networks," in *Proc. of the 1996 Intl Conf. on Communications (ICC'96)*, pp. 1045–1049, (Dallas, TX), June 1996.
9. C. Parris, S. Keshav, and D. Ferrari, "A Framework for the Study of Pricing in Integrated Networks," tech. rep., International Computer Science Institute, Berkeley, CA, 1992.
10. S. J. Shenker, "Making Greed Work in Networks: A Game-Theoretic Analysis of Switch Service Disciplines," *IEEE/ACM Transactions on Networking* **3**, pp. 819–831, Dec. 1995.
11. S. H. Low and P. P. Varaiya, "A New Approach to Service Provisioning in ATM Networks," *IEEE/ACM Transactions on Networking* **1**, pp. 547–553, Oct. 1993.
12. J. Murphy and L. Murphy, "Bandwidth Allocation by Pricing in ATM Networks," tech. rep., Dublin City University, Ireland, July 1994.
13. J. Murphy, L. Murphy, and E. Posner, "Distributed Pricing for Embedded ATM Networks," in *International IFIP Conference on Broadband Communications (BB-94)*, (Paris, France), Mar. 1994.
14. Q. Wang, J. M. Peha, and M. A. Sirbu, "The Design of an Optimal Pricing Scheme for ATM Integrated Services Networks," *Journal of Electronic Publishing* , 1995. Special Issue on Internet Economics.
15. C. Courcoubetis, V. A. Siris, and G. Stamoulis, "Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks," in *Proc. of IEEE GLOBECOM*, pp. 644–648, (London), 1996.

16. H. Jiang and S. Jordan, “A Pricing Model for High Speed Networks with Guaranteed Quality of Service,” in *Proc. of IEEE INFOCOM*, pp. 888–895, 1996.

17. F. P. Kelly, “Charging and Rate Control for Elastic Traffic,” *European Transactions on Communications* **8**, pp. 33–37, 1997.

18. A. A. Lazar, A. Orda, and D. E. Pendarakis, “Virtual Path Bandwidth Allocation in Multi-User Networks,” in *Proc. of IEEE INFOCOM*, pp. 312–320, (Boston, MA), Apr. 1995.

19. R. Cocchi, D. Estrin, S. Shenker, and L. Zhang, “A Study of Priority Pricing in Multiple Service Class Networks,” in *SIGCOMM Symposium on Communications Architectures and Protocols*, pp. 123–130, (Zurich, Switzerland), Sept. 1991.

20. J. M. Peha and F. A. Tobagi, “Cost-Based Scheduling and Dropping Algorithms to Support Integrated Services,” *IEEE Transactions on Communications* **44**, pp. 192–201, Feb. 1996.

21. R. Arnott and K. Small, “The Economics of Traffic Congestion,” *American Scientist* **82**, pp. 446–455, September/October 1994.

22. D. M. Kreps, *Game Theory and Economic Modelling*, Clarendon Press, 1990.

23. Y. A. Korilis and A. A. Lazar, “On the Existence of Equilibria in Noncooperative Optimal Flow Control,” *Journal of the ACM* **42**, May 1995.

24. A. Orda, R. Rom, and N. Shimkin, “Competitive Routing in Multiuser Communication Networks,” *IEEE/ACM Transactions on Networking* **1**, pp. 510–521, Oct. 1993.

25. Y. A. Korilis, A. A. Lazar, and A. Orda, “Architecting Noncooperative Networks,” *IEEE Journal on Selected Areas in Communications* **13**, pp. 1241–1251, Sept. 1995.

26. K.-Y. Siu, “A New Data Service in ATM?,” *IEEE Network* **11**, pp. 4–5, July/August 1997.

27. K. Liu, D. W. Petr, and C. Braun, “A Measurement-based CAC Strategy for ATM Networks,” in *Proc. of the 1997 Intl Conf. on Communications (ICC’97)*, pp. 1714–1718, (Montreal, Canada), June 1997.

28. R. Guerin, H. Ahmadi, and M. Naghshineh, “Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks,” *IEEE Journal on Selected Areas in Communications* **9**, pp. 968–981, Sept. 1991.

29. Comdisco Systems, Inc., *BONeS DESIGNER Modeling Reference Guide*, June 1993.

30. S. J. Shenker, “Fundamental Design Issues for the Future Internet,” *IEEE Journal on Selected Areas in Communications* **13**, pp. 1176–1188, Sept. 1995.