ELSEVIER

Computer Networks and ISDN Systems 30 (1998) 951–962

COMPUTER NETWORKS and ISDN SYSTEMS

Resource allocation in multi-service networks via pricing: statistical multiplexing

G. de Veciana *, R. Baldick [1]

Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712, USA

Accepted 2 December 1997

Abstract

In this paper we investigate the multiplexing of heterogeneous applications and the pricing of transmission services on an ATM network. Pricing is of interest for network management to promote efficient utilization of resources. A framework is presented in which users select from a menu of services, are peak-rate policed, and are charged according to their usage and connection time. We illustrate the considerable improvement in network utilization possible compared to a traditional reservation-based approach. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Pricing; Network management; Resource sharing; Quality of service; ATM networks

1. Introduction

The rapid growth of high-speed networks has led to many technological as well as social developments. The growth is likely to result in a large infrastructure to provide broadband services and global connectivity. As the networks supporting these services shift from the experimental phase to commercial operation, the pricing of limited resources will become an important problem. For example, software has recently appeared to allow ''telephone'' conversations to take place over the Internet [6]. If the marginal price to users of sending such traffic remains at zero – as it was under government subsidization of the Internet and as it continues for many users – then there is a grave potential to overwhelm the Internet.

Our particular interest lies in the area of Broadband Integrated Services Digital Networks (B-ISDN) based on the Asynchronous Transfer Mode (ATM). Such networks are similar to classical circuit-switched telephone networks in that they are organized using virtual circuits through which information flows in an orderly fashion. However, in contrast to circuit-switched networks, the traffic streams consist of small packets (cells), arriving at

* Corresponding author. E-mail: gustavo@mocha.ece.utexas.edu.
[1] E-mail: baldick@mocha.ece.utexas.edu.

PII: S0169-7552(97)00128-1

possibly time-varying rates. These cells are *statistically multiplexed* through links and switches sharing multiple ''congestible resources'', such as buffered links.

One traditional approach to supplying bandwidth on a network is to reserve capacity for users. If there are no usage based charges, and ''busy signals'' are to be avoided, then the network must be sized to meet near to the peak demand, which will typically occur during a relatively small proportion of time. For most of the rest of the time, however, the network will be operated at far below its peak capacity, so that the utilization of the capital investment in the network will be low.

Sizing a network to cope with peak demand is a traditional approach in several ''public utility'' industries, including telecommunications [1,5,7]. However, the resulting low level of capacity utilization represents a waste of resources during periods when demand is low. Deregulation of the telecommunications industry has put particular pressure on telecommunications providers to lower their costs. We believe that improving capital utilization is *the* most important step that can be taken to improve competitiveness and it is this goal we pursue in this paper. We propose to do it through a pricing scheme for resources that allows better utilization of capital.

Low and Varaiya [17] propose a reservation-based approach to pricing both bandwidth and buffer in an ATM network. A very attractive aspect of a pricing approach is that by adapting prices to variations in demand, scarce capacity can be rationed primarily by price rather than by ''busy signal''. This means that instead of building the network to meet the peak in the demand or using a busy signal to ration supply, demand can instead be spread out over longer periods. Reducing the peak demand allows more customers to be served for a given amount of bandwidth and capital expenditure. This is, of course, analogous to time-of-use pricing as currently used in telephone long-distance charges. By making the price high at times of peak usage, the demand is spread, less capacity is needed to meet the resulting peak demand, and capital utilization is higher [5]. Low and Varaiya's innovation is to formulate this framework for an ATM network and consider prices for bandwidth and buffer that are regularly updated to adapt to demand changes.

In summary, Low and Varaiya's approach can help to improve efficiency of capital utilization by shaving peak demands and so allowing the resulting peak demand to be met with a smaller installed capacity. The net gain in capital utilization possible from peak-shaving depends on customer flexibility, but a reasonable estimate for the gains would be of the order of 10 to 20 per cent [10]. In a competitive environment, this translates to lower average costs to customers and ultimately more market share for the efficient provider. Against this gain must be set the costs of measuring usage and accounting; however, we believe that even when these costs are included, significant net reductions in average cost are possible compared to a network that must meet peak demand at a fixed or zero demand charge.

However, there is a significant unexploited efficiency in Low and Varaiya's approach to provision of ATM services: reservation neglects the efficiency gains that are possible through sharing of resources. We present a simple example calculation in Section 2. In the example we show that by taking advantage of multiplexing while accepting a modest bit loss rate, the effective capacity of the network can be more than doubled. In other words, taking advantage of multiplexing can more than *double* the capital efficiency compared to a reservation-based approach. This increase far surpasses the improvements possible through peak-shaving that Low and Varaiya present. (Of course, it is possible to take advantage of *both* statistical multiplexing and peak-shaving.)

One approach to taking advantage of multiplexing is to grant access to the network on a packet-by-packet level as suggested by Murphy et al. [20]. However, the overheads of such an approach are much greater than required in Low and Varaiya's scheme and would probably outweigh the gains in utilization of the network. In general, the advantages of multiplexing come at the cost of implementing sufficiently sophisticated control schemes to:

1. avoid congestion and busy signals, and,
2. provide an array of qualities of service (QoS) to integrate the needs of various sources.

If such control schemes are expensive to implement, then the overheads due to control mechanisms could outweigh the advantages of multiplexing. In selecting an appropriate mechanism one must compromise between

economic efficiency, practicality of implementation, and the cost of the overheads. We believe that overheads make a per-packet approach impractical.

To avoid such large overheads, we adopt pricing of average quantities, such as average bandwidth, that are cheap to measure. We follow Low and Varaiya's approach, but differ from it in two ways:

1. we price only bandwidth, leaving buffer to be allocated by the network, and,
2. we avoid reservation so as to take advantage of statistical multiplexing.

We anticipate that in typical networks both bandwidth and buffer capacity will eventually become scarce. We also believe that, generally speaking, scarcity is better rationed through prices rather than a busy signal because price-based rationing is consistent with a competitive marketplace [5,7].

However, we also take the position that bandwidth is a natural commodity to price, while pricing buffer is less satisfactory from a user's point of view. Furthermore, performance is usually more sensitive to bandwidth than to buffer so that achieving optimal allocation of bandwidth, rather than buffer, is the most important goal. An economic interpretation of this observation is that the costs of overhead to explicitly price buffer are probably greater than the likely gains from its optimal allocation. Moreover, in providing real-time services with low latency, we need to minimize buffering. In this paper we will adopt Low and Varaiya's analytical framework [17] to show that it is possible to incorporate multiplexing gains if we restrict ourselves to explicitly pricing bandwidth.

In Section 2 we adapt Low and Varaiya's framework, illustrate the efficiency gains possible with statistical multiplexing, and define real-time and best-effort services. In Section 3, we describe pricing structures to achieve efficient utilization of the network capacity. We conclude in Section 4 with some suggested extensions.

2. Towards a framework

We begin in Section 2.1 by describing our model of the service provider and ''real-time'' and ''best-effort'' services. We then discuss the notion of a ''user'' who derives benefit from, and is prepared to pay accordingly for, the use of a network. In Sections 2.2 and 2.3, we discuss real-time and best-effort traffic in detail. Finally, in Section 2.4, we qualitatively describe how pricing is used to reconcile user demand for services with finite network capacity; that is, as a mechanism for network management aiming to improve ''efficiency''.

2.1. Service provider and services

In our framework, the service provider presents users with a menu of services. They are broadly divided into two types: *real-time* and *best-effort*. All real-time connections are guaranteed the same low-delay and relatively small probability of loss. Best-effort users are given no guarantees for delay or loss but may be able to estimate their average delays. We point out that other kinds of services can and should be envisaged, but the two polar extremes we describe are useful for illustrating the salient issues. For further discussion of service classes, see for example [12,22].

Real-time services are further divided into several types, J, where each type, $j \in J$, has an expected mean and peak number of packets per time slot, m_j and p_j, respectively. The user indicates his desires and traffic characteristics upon setting up a connection, through the so-called ''service contract'' between the user and service provider, by choosing one of the types $j \in J$. Given the customer's selection the network will in turn police real-time traffic to ensure s/he does not exceed the agreed upon peak-rate. Enforcing of a peak-rate bound can be easily done at the user network interface (UNI) and is an effective way to protect the network from hostile users while allowing some flexibility in the manner in which traffic is sent [4,16].

It is unlikely that users with policed connections will underestimate their peak-rate, as otherwise the traffic will be delayed or dropped at the network edge, resulting in a degradation of end-to-end performance. Moreover, a user will typically not want to overestimate the peak as this is likely to increase the price charged to

the user. The mean rate will not be policed, but incentives for declaring it truthfully can be provided through an appropriate implementation of tariffs, such as described in [16]. The menu of real-time services offered by the service provider should be sufficiently rich to support broad classes of users such as high and low quality video as well as audio.

2.2. Congestion control and efficiency for real-time traffic

Because resources are shared and limited, part of the role of the network manager is to avoid congestion. The primary mechanism we propose for this is pricing; however, parts of the network will still occasionally become congested and therefore a busy signal is still occasionally necessary. Thus, in order to avoid degradation in performance, the network will reject connection requests when resources are over-subscribed. Following Hsu et al. [12], we specify that real-time traffic is:

1. given service priority in accessing an outgoing link with capacity C packets/slot and
2. subject to relatively little buffering.

In order to ensure that packets are rarely lost on a link, we need to control the probability, A, that the aggregate arrivals per slot for current real-time connections exceeds, C, the capacity of the shared link. Suppose we parameterize the acceptable link overflow probability by e^{-L}. Then the following operational constraint must be satisfied:

$$P(A > C) \leq e^{-L}. \tag{1}$$

In many applications, we might have a relatively stringent requirement, such as $e^{-L} = 10^{-12}$, meaning that losses due to overflows should occur rarely, or at least be on par with the typical packet error rates achieved by the links. However, for particular applications, such as video, the loss constraint will depend on whether it is high or low quality video, and the amount of effort at the decoder to perform loss concealment. The range of acceptable cell loss rates quoted in the research community are typically from 10^{-3} for very low quality service to 10^{-9}. We shall see that for our purposes almost any relaxation from the no loss case will be useful.

The rationale for providing little or no buffering to real-time services is as follows. For bursty traffic streams, large buffers are typically required in order to achieve good performance. However, large buffers must be avoided for traffic with a low-latency requirement. As a first cut at this issue, we therefore eliminate buffering altogether for such users [12,14].

Let n_j denote the number of connections of type $j \in J$. In [13,15] it is shown that the overflow constraint (Eq. (1)) translates to an approximately linear constraint on the admissible numbers of connections that may concurrently use the link:

$$\sum_{j \in J} n_j b_j(\delta) \leq \gamma, \tag{2}$$

where:

- $b_j(\delta) = (1/\delta) \log \mathbb{E} \exp[\delta A_j]$,
- A_j is a random variable denoting the packets/slot for a class j stream,
- $\gamma = C - L/\delta < C$, and
- $\delta > 0$ is parameter resulting from the linear approximation to the feasible region.

The work of [2,12] shows that, given the mean and peak of the source, it is conservative to consider on/off traffic behavior with instantaneous rate either zero or p_j, in which case the functions $b_j(\delta)$ are given by:

$$b_j(\delta) = b(\delta, m_j, p_j) = \frac{1}{\delta} \log\left[1 + \frac{m_j}{p_j}(e^{\delta p_j} - 1)\right].$$

Now, considering a link supporting a large number of users, in which the aggregate *average bandwidth* corresponding to each class of real-time traffic $j \in J$ is denoted by x_{rj}, we can state the above operational constraint as:

$$\sum_{j \in J} x_{rj} \frac{b_j(\delta)}{m_j} \leq \gamma. \tag{3}$$

This constraint is conservative in several ways:

1. The constraint (Eq. (3)) corresponds to a linear approximation to a non-convex admission region. [2]
2. The region corresponds to worst case statistics, i.e., on/off for specified mean and peak. We justify this choice by arguing that parsimony and simplicity are required in the traffic specification of real-time traffic.
3. The bound on the probability of overflow is conservative. For better approximations in more general settings, see [3,12,18], but these can again be approximated by appropriate linear regions.

Despite the conservatism of our approximation, using Eq. (3) is far less conservative and therefore more capital efficient than is possible with a reservation-based approach. For example, consider a $C = 155$ Mbps link, carrying MPEG-2 coded video streams. Frames are generated every 40 msec, and have a peak-rate $p = 5.66$ Mbps and mean $m = 0.86$ Mbps, based on empirical statistics [11]. Network buffers are assumed to be small, say on the order of 40 msec or less, whence we focus on frame statistics, and ignore frame correlations. With peak-rate allocation there is sufficient capacity for 27 sources, with no cell losses.

Alternatively, by allowing a very modest probability of overflow over a frame time of 10^{-12} and using the constraint (Eq. (2)) we find that roughly 66 streams can be admitted. The ratio of 66 to 27 implies a more than doubling of the effective capacity of the system compared to a reservation-based approach. If the overflow constraints could be relaxed to 10^{-9},10^{-6}, or 10^{-3} then the the admissible number of connections would be 84, 90, and 139 respectively, showing that for a wide range of overflow constraints one can expect gains in effective capacity in the range of two to four compared to peak-rate allocation. These statistical multiplexing gains increase with the ratio of the peak to mean rates of the traffic. In our video example, the ratio of peak to mean rate is about 5. For larger ratios, the gains are even more dramatic.

Simply put, an existing network could more than double the rate of calls it handles if, instead of using peak-rate allocation, it uses statistical multiplexing. The role of pricing in the rest of this paper is to facilitate statistical multiplexing to achieve these gains.

2.3. Best-effort traffic

The discussion in Section 2.2 illustrates the gains possible from multiplexing real-time traffic. In this section, we consider the added gains due to best-effort traffic. From the network management point of view, best-effort traffic is used to exploit idle bandwidth resulting from fluctuations in real-time traffic arrivals. Best-effort traffic needs to be buffered so that it is ready to ''fill in the gaps'' in the aggregated real-time traffic, resulting in further improvements in the utilization of network capacity.

We let x_b denote the average throughput, in packets per slot, achieved by best-effort traffic. A further operational constraint is that:

$$x_b + \sum_{j \in J} x_{rj} \leq C, \tag{4}$$

that is, the aggregate average throughput cannot exceed the link capacity.

In order to aid best-effort users, the network might offer several buffering/storage alternatives, see [23], allowing users to optimize their transmission strategies to achieve their desired throughput at minimal cost. For

[2] We believe that subject to sufficiently strong conditions, a pricing mechanism can determine the best linear approximation.

example a user that could tolerate delay might postpone transmitting his traffic until an off-peak period to reduce the cost. [3] The benefits of such mechanisms to the network performance can be significant as the discussion and simulation studies of [10,21] show. For simplicity here, however, we suppose that the network allocates a fixed buffer of size B_b to best-effort traffic. Note that by Little's law a pessimistic bound for the average delay incurred by best-effort traffic would be B_b/x_b.

2.4. User model

Each user (or software application) derives benefit from communicating across the network via one or several channels. The question of *how much* benefit a communication entity derives is likely to depend on the throughput and QoS maintained throughout the call. Indeed it may depend on the temporal behavior of the connection, such as call blocking, burst blocking, cell loss, or the average, max, or tail delay characteristics as well as the jitter experienced by the traffic stream. The benefit is reflected in the user's ''willingness to pay'' for the offered network services; that is, the value, in money units, of the network services to the user. The willingness to pay, in turn, defines a user's demand function, which measures the amount of service demanded versus its price.

Individual customers choose whether or not to send traffic on the basis of their individual preferences. The higher the price, the fewer the number of customers that are prepared to pay the price, and the lower the amount of traffic. We *aggregate* the demand functions of users involving the same geographical access points and requesting the same service class from the provider's menu. The demand of a given population of users is represented by the aggregate *average bandwidth* of a given service class requested by the latter.

Following standard practice in economics, we assume that the aggregate demand is a smooth non-increasing function of the price [24]. Note that a particular user's demand function may not be smooth. For example, above a certain price s/he may abandon transmitting altogether. This is, however, not inconsistent with the assumption that aggregating such users results in a demand profile that is essentially smooth.

The aggregate demand for each type of bandwidth depends on the prices per packet posted by the network for real-time and best-effort traffic, w_{rj}, $j \in J$ and w_b, respectively. The aggregate demand for best-effort bandwidth is denoted by $D_b(w_b)$ (in cells/slot) and similarly for real-time traffic $D_{rj}(w_{rj})$, $j \in J$. In particular, we assume that the demand functions have the following functional forms:

$$D_b(w_b) = \nu_b \exp[-e_b w_b], \qquad D_r(w_{rj}) = \nu_{rj} \exp[-e_{rj} w_{rj}], \quad j \in J, \tag{5}$$

with parameters $\nu_b > 0$, $e_b > 0$, $\nu_{rj} > 0$ and $e_{rj} > 0$ for $j \in J$, that are known to the network. The parameters e_b and e_{rj}, $j \in J$ are called the price elasticities. These functional forms are commonly used in economic theory to approximate aggregate behavior and should be interpreted as ''small-signal models'' about an operating point. Herein we assume the price elasticities are known; however, by monitoring connection requests and usage at the user network interface, the network can infer such information, see [8].

3. Prices for services

Prices are set by the network to adjust demand, permitting the network to induce an optimally efficient operating regime. The goal in setting the prices is to maximize the ''efficiency'' (social welfare) of the allocation of resources in the network. We therefore do not address cost recovery. [4]

[3] An appropriate model for such mechanisms has yet to be explored. In particular, this requires a demand function that represents the performance (delay) and price tradeoffs underlying the decision to postpone transmission.

[4] A standard approach to cost recovery is to include fixed charges in the tariff in addition to the prices we describe. We refer the interested reader to the public utility pricing literature for a discussion of cost recovery and associated issues [1,5,7].

Much of our notation and method of proof is based on Low and Varaiya [17]. We relegate the proofs to the appendix. In Section 3.1, we present the analysis. Section 3.2 describes the algorithm and various issues are discussed in Section 3.3.

3.1. Analysis

As developed in the previous section, network operation is subject to the following constraints:

$$\sum_{j\in J} x_{rj} + x_b \le C, \tag{6}$$

$$\sum_{j\in J} x_{rj}\frac{b_j(\delta)}{m_j} \le \gamma. \tag{7}$$

Naturally, the allocated bandwidth is no greater than the demand:

$$x_b \le D_b(w_b),\ \ x_{rj} \le D_{rj}(w_{rj}),\ \ j\in J. \tag{8}$$

Allocated bandwidth can be less than demanded bandwidth if calls are rejected.

Following standard practice in economics, we define social welfare, $W(x_b, x_{rj}, w_b, w_{rj}, j\in J)$, to be the sum of the user surplus and the payment to the network [24]. Our goal is to maximize social welfare subject to the capacity and demand constraints:

$$\max_{\substack{x_b\ge 0,\\ x_{rj}\ge 0,\\ w_b\ge 0,\\ w_{rj}\ge 0,\\ j\in J}} \left\{ \begin{array}{l} \displaystyle\int_{w_b}^{\infty} \min\left[x_b, D_b(u)\right] du + \sum_{j\in J}\int_{w_{rj}}^{\infty} \min\left[x_{rj}, D_{rj}(u)\right] du + x_b w_b + \sum_{j\in J} x_{rj} w_{rj} : \\ \displaystyle x_b + \sum_{j\in J} x_{rj} \le C,\ \sum_{j\in J} x_{rj}\frac{b_j(\delta)}{m_j} \le \gamma,\ x_b \le D_b(w_b),\ x_{rj} \le D_{rj}(w_{rj}) \end{array} \right\}. \tag{9}$$

Following Low and Varaiya [17]:

Definition 3.1. A set of non-negative services produced and non-negative prices charged,

$$\left(x_b^*, x_{rj}^*, w_b^*, w_{rj}^*, j\in J\right),$$

is called an equilibrium if:

1. x_b^* and x_{rj}^* satisfy the capacity constraints (Eqs. (6) and (7)) and also $x_b^* = D_b(w_b^*)$ and $x_{rj}^* = D_{rj}(w_{rj}^*)$ at the stated prices; that is, supply meets demand, and,
2. for any other non-negative $x_b, x_{rj}, j\in J$ meeting the capacity constraints (Eqs. (6) and (7)), we have $x_b^* w_b^* + \sum_{j\in J} x_{rj}^* w_{rj}^* \ge x_b w_b^* + \sum_{j\in J} x_{rj} w_{rj}^*$.

Proposition 3.1. A set of non-negative services produced and non-negative prices charged,

$$\left(x_b^*, x_{rj}^*, w_b^*, w_{rj}^*, j\in J\right),$$

is an equilibrium if and only if there exist $\lambda_b^* \cong, \lambda_r^* \ge 0$ such that:

$$x_b^* = D_b(w_b^*), \tag{10}$$

$$x_{rj}^* = D_{rj}(w_{rj}^*),\quad j\in J, \tag{11}$$

$$x_b^* + \sum_{j\in J} x_{rj}^* \le C, \tag{12}$$

$$\sum_{j \in J} x_{rj} \frac{b_j(\delta)}{m_j} \leq \gamma, \tag{13}$$

$$w_b^* = \lambda_b^*, \tag{14}$$

$$w_{rj}^* = \lambda_b^* + \lambda_{rj}^* \frac{b_j(\delta)}{m_j}. \tag{15}$$

The total revenue at the equilibrium is given by $x_b^* w_b^* + \sum_{j \in J} x_{rj}^* w_{rj}^* = C\lambda_b^* + \gamma\lambda_r^*$.

Proof. See the appendix.

We interpret the prices $w_b^* = \lambda_b^*$ and $w_{rj}^* = \lambda_b^* + \lambda_r^*(b_j(\delta)/m_j)$ as follows. For all classes these represent cost per transmitted cell during a given slot. In particular:

- a best effort user would be charged λ_b^* per transmitted packet, while,
- a real-time user of type j would, in addition, pay a premium, $\lambda_r^* b_j(\delta)/m_j$, per packet, depending on his/her service class.

In practice we expect operational constraints to be binding in which case the prices depend on the class of service $j \in J$ that is delivered.

We now observe that there exists an optimizer of the welfare function that is an equilibrium point:

Proposition 3.2. There exists a welfare optimizing set of non-negative services produced and non-negative prices charged, $(x_b^*, x_{rj}^*, w_b^*, w_{rj}^*, j \in J)$, that is also a set of equilibrium services and prices.

Proof. See the appendix.

Proposition 3.2 means that suitable prices can induce welfare optimal behavior. For any production $(x_b, x_{rj}, j \in J)$ the network has some leeway in the selection of the prices that will induce welfare optimal behavior. We assume, however, that it always chooses equilibrium prices such that supply equals demand; that is, $x_b = D_b(w_b)$, $x_{rj} = D_r(w_{rj})$.

3.2. Algorithm

The algorithm to achieve the welfare optimal allocation is as follows:

Users:	Request connections (bandwidth) in response to the network's posted prices w_b^* and w_{rj}^*, i.e., $D_b(w_b^*)$ and $D_{rj}(w_{rj}^*)$.
Network Þit > (Call admission)Þ/it > :	Allocate bandwidth in response to requests at posted prices.
Network Þit > (Billing)Þ/it > :	A best-effort user is charged a total of $w_b^* \times M$ if he sends M packets. A real-time user in the j^{th} service class is charged a total of $w_{rj}^* \times M$ if he sends M packets. Alternatively, since real-time users are assumed to have truthfully declared their mean rates, m_j, a real-time user is charged $w_{rj}^* m_j$ per unit time for the duration of his call.
Network Þit > (Management/Pricing):Þ/it >	Solve welfare optimization problem (Eq. (9)) based on recent estimates of the demand functions and post new prices for the next pricing interval. Select prices so that supply meets demand.

3.3. Discussion

The welfare optimization problem (Eq. (9)) is a nonlinear optimization problem that can be solved using nonlinear programming software or sequential linear or quadratic programming software [19], given values for the parameters in the demand models and other design parameters. It is important to realize that an approximate solution to Eq. (9) based on imperfect models of demand behavior will usually be adequate for setting prices. This is because, as evidenced by the example in Section 2.2, the improvements in capital utilization by using statistical multiplexing can be huge. In other words, the prices do not have to be exactly optimal, nor does the allocation of resources have to be exactly optimal, for there to be large improvements in capital utilization compared to peak-rate allocation. Improvements by a factor of two in capital utilization are possible with even rough approximations to the optimal prices.

The prices that are set for each pricing interval are based on demand response for the previous pricing interval. That is, they will be calculated based on slightly out-of-date information. However, as we have indicated, inaccuracies in the prices are easily tolerated, so that this does not pose a major problem so long as prices are updated occasionally.

4. Extensions and outlook

Several issues remain, some of which are addressed in an extended version of this paper [8]. First, demand functions are typically not known explicitly, hence mechanisms for estimating these characteristics based on user participation need to be developed. This means that requests for service will very occasionally be met with a busy signal because the anticipated average demand was smaller than the actual demand at a particular time. Second, the results need to be extended to the network setting. If we make the conservative assumption of nodal decomposition – roughly speaking that the resources are decoupled [9] – then we believe that a similar cost structure can be applied to multi-link networks [8]. Resolution of this issue and several implementation issues concerning measurement, accounting, and billing need to be addressed.

In our view there are several approaches to network control each suggesting different pricing mechanisms. The first option is to attempt to provide QoS through best-effort management schemes via preferential scheduling; in this case an auction in which priority is directly related to value makes sense; however, there are significant communication overheads in this approach. The second is to attempt to make reservations of bandwidth and buffer for particular users; here a pricing scheme renting bandwidth and buffer independently as suggested by Low and Varaiya might work; however, the potential gains from statistical multiplexing are eliminated.

A third option is to allow statistical multiplexing of resources while at the same time placing the burden of guaranteeing QoS on the network. In this case, we believe the pricing approach we have developed leads in the right direction. Our framework is based on a simple traffic descriptor: a policed peak-rate and a mean rate. In practice, users would select from a menu of services, which would require updating to match changing technologies such as compression. In our scheme, bandwidth and buffer are implicitly bundled together to deliver various levels of QoS. Based on user demands and usage, the network computes prices for the offered services that will induce efficient utilization of network resources. The increases in capital utilization are dramatic. The price structure developed is simple and intuitive: a price for average bandwidth plus a premium for the selected real-time services.

Acknowledgements

The authors would like to thank Dr. Steven Low for discussions during the course of this work. G. de Veciana is funded in part by the National Science Foundation under Grant NCR-9409722. R. Baldick is funded in part by the National Science Foundation under grant ECS-9457133.

Appendix A. Proofs of propositions

A.1. Proposition 3.1

First suppose that $(x_b^*, x_{rj}^*, w_b^*, w_{rj}^*, j \in J)$ is an equilibrium. Then, Eqs. (10)–(13) are satisfied. Furthermore, x_b^* and x_{rj}^*, $j \in J$ solve the linear program:

$$\max_{x_b \geq 0, x_{rj} \geq 0, j \in J} \left\{ x_b w_b^* + \sum_{j \in J} x_{rj} w_{rj}^* : x_b + \sum_{j \in J} x_{rj} \leq C, \sum_{j \in J} x_{rj} \frac{b_j(\delta)}{m_j} \leq \gamma \right\}. \tag{A.1}$$

This linear program has dual:

$$\min_{\lambda_b \geq 0, \lambda_r \geq 0} \left\{ C\lambda_b + \gamma\lambda_r : \lambda_b \geq w_b^*, \lambda_b + \lambda_r \frac{b_j(\delta)}{m_j} \geq w_{rj}^*, j \in J \right\}. \tag{A.2}$$

Since the primal feasible region is bounded, it has a finite optimal solution, and so the dual attains the same value and the total revenue is $x_b^* w_b^* + \sum_{j \in J} x_{rj}^* w_{rj}^* = C\lambda_b^* + \gamma\lambda_r^*$, where λ_b^* and λ_r^* solve Eq. (A.2). Note that $w_{rj}^* \geq w_b^*, \forall j \in J$, for else the second part of Definition 3.1 of equilibrium is violated by reducing x_{rj}^* and increasing x_b^* by equal amounts. (Such a change keeps the left hand side of Eq. (6) constant, reduces the left hand side of Eq. (7), and increases the objective if $w_{rj}^* < w_b^*$.) Explicitly solving the dual program, noting that $\frac{b_j(\delta)}{m_j} \geq 1$ and $C > \gamma$, we find that $w_b^* = \lambda_b^*, w_{rj}^* = \lambda_b^* + \lambda_r^* \frac{b_j(\delta)}{m_j}$ so that Eqs. (14) and (15) are satisfied.

Conversely, suppose that Eqs. (10)–(15) are satisfied. By Eqs. (10) and (11), x_b^* and x_{rj}^*, $j \in J$ satisfy the first part of Definition 3.1. By Eqs. (12)–(15), they also satisfy the optimality conditions of Eqs. (A.1) and (A.2), which in turn means that the second part of Definition 3.1 is satisfied. □

A.2. Proposition 3.2

Consider the objective of the welfare optimization problem (Eq. (9)):

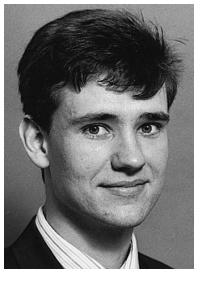
$$\begin{aligned}
&\int_{w_b}^{\infty} \min\left[x_b, D_b(u) \right] \mathrm{d}u + \sum_{j \in J} \int_{w_{rj}}^{\infty} \min\left[x_{rj}, D_{rj}(u_j) \right] \mathrm{d}u_j + x_b w_b + \sum_{j \in J} x_{rj} w_{rj} \\
&\leq \int_{w_b}^{\infty} D_b(u)\, \mathrm{d}u + \sum_{j \in J} \int_{w_{rj}}^{\infty} D_{rj}(u_j)\, \mathrm{d}u_j + D_b(w_b) w_b + \sum_{j \in J} D_{rj}(w_{rj}) w_{rj}, \\
&= \nu_b \exp\left[-e_b w_b \right] \left(\frac{1}{e_b} + w_b \right) + \sum_{j \in J} \nu_{rj} \exp\left[-e_{rj} w_{rj} \right] \left(\frac{1}{e_{rj}} + w_{rj} \right),
\end{aligned} \tag{A.3}$$

which is strictly decreasing in w_b and w_{rj}, $j \in J$. Moreover, there are feasible solutions to Eq. (9). Hence, Eq. (9) has a well defined and finite optimum, which can be found using nonlinear programming. So, let $(x_b^*, x_{rj}^*, \hat{w}_b, \hat{w}_{rj}, j \in J)$ be a welfare optimizing set of services and prices. By the definition of welfare, they must be non-negative services and prices. Suppose that $x_b^* < D_b(\hat{w}_b)$ or $x_{rj}^* < D_{rj}(\hat{w}_{rj})$ so that supply does not meet demand. We claim that we can construct a non-negative set of services and prices that satisfy Eq. (8) with equality and which produce the same welfare: increase the prices until the constraints are met with equality. Let the new prices be w_b^* and w_{rj}^*. The objective of Eq. (9) stays constant, and no other constraints are violated. Since the services and prices $(x_b^*, x_{rj}^*, \hat{w}_b, \hat{w}_{rj}, j \in J)$ maximize welfare, then these new services and prices $(x_b^*, x_{rj}^*, w_b^*, w_{rj}^*, j \in J)$ also maximize welfare.

Now fix the prices at w_b^* and w_{rj}^* $j \in J$ and consider varying the services. Note that revenue increases with increasing x_b and x_{rj} and that x_b^* and x_{rj}^* are the maximum feasible values of x_b and x_{rj} at the given prices. Hence, by definition, $(x_b^*, x_{rj}^*, w_b^*, w_{rj}^*, j \in J)$ are an equilibrium set of services and prices. □

References

[1] S.V. Berg, J. Tschirhart, Natural Monopoly Regulation: Principles and Practice, Cambridge Surveys of Economic Literature, Cambridge University Press, Cambridge, 1988.
[2] M. Bonatti, A. Gaivoronski, Worst case analysis of ATM sources with application to access engineering of broadband multiservice networks, in: Proc. Mascots, 1993.
[3] D.D. Botvich, N.G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers, Queueing Syst. 20 (1995) 293–320.
[4] J. Brassil, Peak rate regulation maintains service quality in an ATM LAN/WAN interconnection, in: Proc. IEEE INFOCOM, 1994, vol. 2.
[5] S.J. Brown, D.S. Sibley, The Theory of Public Utility Pricing, Cambridge University Press, Cambridge, 1986.
[6] J. Browning, I.T., phone home: cheap calls on the Internet shake everyone up, Sci. Am. 273 (2) (August 1995) 35–36.
[7] M.A. Crew, P.R. Kleindorfer, The Economics of Public Utility Regulation, MIT Press Series on the Regulation of Economic Activity, 1st ed., MIT Press, Cambridge, MA, 1986.
[8] G. de Veciana, R. Baldick, Resource allocation in multi-service networks, Technical Report SCC-94-06, U.T. Austin, ECE Department, 1994.
[9] G. de Veciana, G. Kesidis, J. Walrand, Resource management in ATM networks using effective bandwidths, IEEE J. Sel. Areas Commun. 13 (6) (August 1995) 1081–1090.
[10] R.J. Edell, N. McKeown, P.P. Varaiya, Billing users and pricing for TCP, IEEE J. Sel. Areas Commun. 13 (7) (September 1995) 1162–1176.
[11] D.P. Heyman, T.V. Lakshman, A. Tabatabai, Statistical analysis and simulation study of MPEG2-coded variable bit rate video traffic, preprint, 1995.
[12] I. Hsu, J. Walrand, Admission control for ATM networks, in: Proc. IMA Workshop on Stochastic Networks, 1994.
[13] J.Y. Hui, Resource allocation for broadband networks, IEEE J. Sel. Areas Commun. 6 (9) (1988) 1598–1608.
[14] Chia-Lin Hwang, S.-Q.Li, On the convergence of traffic measurement and queueing analysis: a statistical-match queueing (SMAQ) tool, in: Proc. IEEE INFOCOM, 1995.
[15] F.P. Kelly, Effective bandwidths of multi-class queues, Queueing Syst. 9 (1) (1991) 5–15.
[16] F.P. Kelly, On tariffs, policing and admission control for multiservice networks, Technical Report, Research Report No. 93-2, Statistical Laboratory, Cambridge, England, 1993.
[17] S. Low, P. Varaiya, A new approach to service provisioning in ATM networks, IEEE/ACM Trans. Netw. 1 (5) (1993) 547–553.
[18] M. Montgomery, G. de Veciana, On the relevance of time scales in performance oriented traffic modeling, in: Proc. IEEE INFOCOM, 1996, vol. 2, pp. 513–520.
[19] J.J. Moré, S.J. Wright, Optimization Software Guide, Society for Industrial and Applied Mathematics, Philadelphia, 1993.
[20] J. Murphy, L. Murphy, E. Posner, Distributed pricing for embedded ATM networks, in: Proc. ITC 94, 1994.
[21] C. Parris, S. Keshav, D. Ferrari, A framework for the study of pricing in integrated networks, Technical Report, International Computer Science Institute, Berkeley, CA, March 1992.
[22] S. Shenker, Service models and pricing policies for an integrated services Internet, in: Public Access to the Internet, Harvard, Cambridge, MA, 1993.
[23] K. Sriram, Methodologies for bandwidth allocation, transmission scheduling, and congestion avoidance in broadband ATM networks, Comput. Netw. ISDN Syst. 26 (1993) 43–59.
[24] Hal Varian, Microeconomic Analysis, 3rd ed., W.W. Norton & Company, New York, 1992.



Gustavo de Veciana (S'88–M'94) received his B.S., M.S, and Ph.D. in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993 respectively. Since 1993, he has been an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Texas at Austin. His research focuses on issues in the design and control of telecommunication networks. He is the recipient of a 1996 National Science Foundation CAREER Award.

Ross Baldick received his B.Sc. and B.E. from the University of Sydney, Australia and his M.S. and Ph.D. from the University of California, Berkeley. From 1991–1992 he was a post-doctoral fellow at the Lawrence Berkeley Laboratory. In 1992 and 1993 he was an Assistant Professor at Worcester Polytechnic Institute. Since 1994 he has been an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Texas at Austin. He is the recipient of a 1994 National Science Foundation Young Investigator Award.