Resource usage and charging in a multi-service multi-QoS packet network

Krishnan Kumaran Michel Mandjes
Debasis Mitra Iraj Saniee
Bell Labs, Lucent Technologies

700 Mountain Avenue, Murray Hill, NJ 07974
{kumaran|mmandjes|mitra|iis}@lucent.com

November 11, 1999

Abstract

We present a general economic equilibrium model of users and service providers in a multi-class multi-QoS network. The model describes utility maximization by users and revenue optimization by service providers. The key parameters in the model are the T_{spec} and Q_{spec}, which characterize traffic and Quality of Service (QoS) respectively. The model yields equilibrium prices, which, in the presence of competing providers, reduce to the marginal costs. Costs are shown to be related explicitly to resource allocation. Clearly, this is valid only if the admission control mechanism is consistent with the resource allocation.

To obtain resource allocation and subsequently prices for individual streams, online measurements may be used. However, measurements do not uniquely characterize T_{spec}. We demonstrate this additional freedom by examining trade-offs among dual leaky bucket parameters and choice of measurement time-scales corresponding to a traffic trace. We further provide guidelines to choose these parameters optimally for resource allocation and pricing.

Key words: Packet networks – pricing – usage-sensitive pricing – resource allocation – online measurements

1 Introduction

One of the most interesting debates in Internet economics is the issue of 'flat-rate pricing' versus 'usage-sensitive pricing'. Anania and Solomon [2] argue that flat rates, e.g., a usage-insensitive access price, is preferable since users typically favor simplicity and predictability. Proponents of usage-sensitive pricing [3, 13, 14], however, point out that charging users based on their effective resource utilization results in greater fairness and more manageable traffic. Additionally, usage-sensitive pricing provides the mechanism to give users the 'right' incentives. For example, the smoother the traffic offered to the network, the easier it is for the network to handle it. Therefore, prices should be increasing in the 'burstiness' of the traffic flow. Brownlee [3] demonstrates feasibility of usage-sensitive pricing in a moderately sized experiment. MacKie-Mason and Varian [13, 14] argue that usage-sensitive pricing could facilitate congestion control as it leads network resources to their most valuable uses. Kelly [11] finds ways to account for burstiness, relying on the concept of 'effective bandwidths', which can be computed from the traffic characteristics of a variable bit rate stream. The effective bandwidth thus computed lies between the mean and peak rates – the higher the burstiness of traffic stream, the stronger the tendency towards the peak.

The above debate began at a time when the Internet carried all types of traffic the same way – 'best effort'– and effectively only one level of quality of service (QoS) was provided. In particular, the Internet was not suitable for real-time traffic which, for example, requires stringent bounds on delay. Hence increasing attention is being paid within the IETF to 'service differentiation'. As advocated by Shenker [17], the philosophy behind offering a 'premium quality' mode and a 'best effort' mode is the maximization of 'social welfare'. Compared to the situation in which the classes were treated equally, the premium customers are charged more, but their utility increases, whereas the best effort customers are charged less. However, the *net* utilities (utility minus charge) of both service types increase: the delay-sensitive service users get the needed enhanced performance, whereas the customers that do not care much about the increased delay get more or less the familiar quality of service. Users that do not need premium quality are therefore encouraged to use the best effort class. Price regulates the traffic and acts as a traffic management agent.

However, we observe that purely usage based 'resource allocation' arguments break down once a social welfare criterion is in effect. This is because resource allocation is a function of both QoS guarantees and usage. *Within* classes standard resource usage arguments can be used to assign relative prices to the individual flows, but *between* classes the user's appreciation of the offered QoS should be taken into account. An overview on charging in an environment with multiple QoS levels is given by Walker *et al.* [21].

The above reasoning advocates pricing based solely on costs, derived from maximizing total welfare. While we understand that tariffing structures are often dominated by business and marketing arguments rather than costs, we believe it worthwhile to understand the cost structure as a guide for pricing purposes. This is particularly important in a multi-service network, since flows with identical bandwidths could have very different resource requirements based on their QoS needs.

In this paper we present a general microeconomic model consisting of users and service providers. Then we focus on the situation in which the number of suppliers grows large. With a simple model we show that the market tends to the situation of a perfectly competitive equilibrium, viz. the aggregate profit reduces with growing number of providers and eventually prices equal the marginal cost. This conclusion also holds with multiple QoS classes.
We also compute the costs to individual users. The cost of operating the network must also be recovered, and we are looking for a 'fair' way to assign these costs. This is not a straightforward task; as explained by Kelly [11] the usage of a network resource does not simply coincide with the total number of bits carried. The higher the burstiness of a flow, the more resources should be reserved to handle it. Thus, the allocation of operating costs to users need to reflect the traffic characteristics of the flows carried.
In addition, it should be emphasized that recovery of network costs through 'fair' pricing must be related to the way streams are accepted or rejected by the network: A poor admission policy, for example, could deprive a network of income from high revenue services admitted infrequently. This acceptance/rejection decision is made by an *admission control* algorithm. Clearly, this algorithm implicitly imputes the cost of admitting a stream, in terms of the shadow prices of the used resources. We recognize that there is an entire spectrum of possible admission control algorithms, and corresponding charging schemes. The most extreme are listed below.

(A) Suppose that the users are provided with strict QoS guarantees, which could be either deterministic or probabilistic. To achieve this, the user specifies a traffic descriptor (a peak rate, a maximum average rate, and the amount of traffic that can be sent uninterruptedly at peak rate). This contract is enforced by a dual leaky bucket regulator ('policer'). The maximum number of users that can be serviced is calculated for a framework that supports statistical multiplexing of 'worst-case traffic'. The admission control algorithm checks whether there are enough resources available to accommodate this traffic. An example of such an algorithm is Elwalid, Mitra, and Wentworth [7]. The allocated resources are completely determined by the traffic contract; hence the price per unit time can be known in advance.

(B) The other extreme is what could be called 'estimated' QoS. Connections are admitted as long as the measured load is less than, say, 90% of the capacity, cf. the measurement-based admission control (MBAC) algorithms of Jamin *et al.* [9]. Notice that the user does not need to provide the network with a traffic descriptor, but in return the guarantees will be quite 'soft'. Now the price per flow must be calculated completely on the basis of measurements of that flow. The resulting price cannot be known in advance.

Between these two extreme schemes, a large number of hybrid schemes can be distinguished.

The paper is organized as follows. Section 2 is devoted to the microeconomic model consisting of users and providers and the allocation of the operating costs. Section 3 focuses on the tight relationship between admission control and charging and a number of admission control variants are illustrated. In Section 4 we illustrate how traffic parameters on which prices are based can be estimated from traces for a leaky token bucket model. Section 5 concludes with suggestions for further research.

2 Economic modeling

We now formulate a microeconomic model of multiple users and a single service provider in a multi-service multi-QoS network. Competition among multiple providers is addressed in the next section. The macroeconomic model consists of two optimization steps: utility maximization by the users and profit maximization by the provider.

2.1 Customer model

Consider a finite population of (potential) customers I indexed by $i \in I$, each of whom may use any application $j \in J$.

- The application determines the set of possible traffic specifications T_{spec}. For an application j we say that $T_{\text{spec}} \in T_j$. These traffic descriptors could be of many different kinds. A simple one could be the specification of a peak rate; another is the dual leaky bucket descriptor, with a peak rate, a sustainable rate, and a maximum burst tolerance, as mentioned in the introduction. In some cases, such as the dual leaky bucket descriptor, the customers still have some residual freedom in choosing the parameters, which must be accounted for carefully. This aspect is further elaborated in a later section devoted to estimating traffic parameters from traces.

- The Q_{spec} is the quality of service mutually agreed upon between user and service provider, and consists, say, of loss and delay. The service provider offers, say, a set Q of QoS levels; for convenience it is assumed to be a finite set.

Notice that the 'strict' and 'soft' guarantees mentioned in the introduction fit into this framework.

The utility the user assigns to the service on the one hand depends on the resources he gets, so the $T_{\rm spec}$. Think for instance of a data application: the more throughput is guaranteed, the larger the utility. The utility is, of course, also determined by the quality of service. In other words:

$$U_{ij} = U_{ij}(T_{\rm spec}, Q_{\rm spec}).$$

The price the customer has to pay also similarly depends on both $Q_{\rm spec}$ and $T_{\rm spec}$:

$$P_{ij} = P(T_{\rm spec}, Q_{\rm spec}).$$

Clearly, the better the offered $Q_{\rm spec}$, the higher the price; the less restrictive the $T_{\rm spec}$, the higher the price.

In this setting it is possible to compute — given the prices $P(T_{\rm spec}, Q_{\rm spec})$ that are exogenously determined (by the service provider) — the $T_{\rm spec}$ and $Q_{\rm spec}$ that customer i will choose when using application j. In fact the customer solves the problem

$$\max[U_{ij}(T_{\rm spec}, Q_{\rm spec}) - P(T_{\rm spec}, Q_{\rm spec})],$$

over all $Q_{\rm spec}$ that the service provider offers, Q, and all $T_{\rm spec}$ in T_j. In words: he maximizes his *compensated utility* or *surplus.*

In fact, the above optimization problem consists of an inner loop and an outer loop. (i) Fixing the $Q_{\rm spec}$, the problem the customer is faced with is to choose the $T_{\rm spec} \in T_j$ such that $U_{ij}(T_{\rm spec}, Q_{\rm spec}) - P(T_{\rm spec}, Q_{\rm spec})$ is maximized. For example, this may involve finding leaky bucket parameters that minimize the price for each customer separately. So every $Q_{\rm spec}$ has a fixed $T_{\rm spec}$. (ii) Then the outer loop performs the maximization over $Q_{\rm spec}$.

This model does not take into account that the higher the prices, the less a customer is inclined to make use of the provider's service. In fact, the above framework assumes that the customer has already decided to buy the service. To make this more realistic, we may think of the following solution.
Consider the model with a fixed set of customers as described above. The customer will not use any service that has a negative surplus. In fact, every customer i will have a threshold level u_{ij} such that the customer will make use of the provider's service only if the surplus is not less than this level. (Without loss of generality we can take $u_{ij} = 0$ by incorporating u_{ij} in U_{ij}.) This gives for a customer of type (i, j) the indicator variable D_{ij}, which is 1 if the user in fact accepts the service and 0 otherwise.
So the resulting variables for a customer i using application j are:

- D_{ij}: does the user use the service at all?
- $T_{\text{spec}}(i,j)$: given that he does, what traffic contract will he choose?
- $Q_{\text{spec}}(i,j)$: given that he does, what QoS level will he choose?

2.2 Service provider model

The above user model assumed given prices, as functions of T_{spec} and Q_{spec}. This subsection focuses on how these prices should be chosen. Here the provider is restricted by his (finite) set of resources R. Resources are typically buffer and bandwidth.
Then prices $P(T_{\text{spec}}, Q_{\text{spec}})$ determine the demands D_{ij}, and also T_{spec} and Q_{spec} of those customers. These demands determine the resources required by the population, that we denote by $r(T_{\text{spec}}(i,j), Q_{\text{spec}}(i,j), i \in I, j \in J)$.

Then the optimization problem for the supplier is

$$\max \left[\sum_{i,j} D_{ij} P(T_{\text{spec}}(i,j), Q_{\text{spec}}(i,j)) \right],$$

over prices $P(T_{\text{spec}}, Q_{\text{spec}})$, subject to

$$r(T_{\text{spec}}(i,j), Q_{\text{spec}}(i,j), i \in I, j \in J) \ \in \ R.$$

So the user reacts to the prices of the provider, and the provider reacts to the demand, traffic specifications T_{spec}, and QoS specifications Q_{spec} of the customers. Eventually this will tend to an equilibrium.

In fact, there are admission control policies that enable decomposition of R into bandwidth usage and buffer usage. Then we are left with two inequalities: the buffer usage should not exceed the buffer space B of the router, and the bandwidth consumption should not exceed its link capacity C:

$$b(T_{\text{spec}}(i,j), Q_{\text{spec}}(i,j), i \in I, j \in J) \leq B \text{ and}$$

$$c(T_{\text{spec}}(i,j), Q_{\text{spec}}(i,j), i \in I, j \in J) \leq C.$$

Then we can calculate the shadow prices of both resources (at a given demand), and that enables the computation of optimal prices.
It should be noted that R is in fact also an instrument of the provider: he can optimize his resources such to maximize his profits. However, he does so on a longer time scale; the choice

of, say, B and C could be better considered as part of the planning cycle. In the long run the situation of abundant resources cannot be optimal, but in the short run it may. Apart from that buffer and bandwidth are usually only available in a restricted number of values (bandwidth for instance: 45 Mbit/s, 150 Mbit/s, 620 Mbit/s, etc). This is also a reason why the optimization over B and C is not taken into account here.

2.3 Situation of many suppliers

It may be expected that if there are many suppliers prices will become cost-based. In this subsection we will use simple models to verify this statement.

If there is only one common QoS (i.e., Q consists of just one element) the prices will coincide with *effective bandwidths* (up to a multiplicative constant). An effective bandwidth E is a function of the traffic contract T_{spec}, which reflects the bandwidth usage (and implicitly also the buffer usage!) of the corresponding flow, given that the Q_{spec} must be met [7]. The effective bandwidth of a superposition of flows is just the sum of the effective bandwidths; in other words: the admissible region R is bounded by just one linear constraint. Due to this additivity property, effective bandwidths also reflect costs. In the case of perfect competition prices will be dictated by costs, and therefore we think that then prices agree with effective bandwidths: $P(T_{\text{spec}}) \sim E(T_{\text{spec}})$.

If there are multiple QoS levels possible, the region R is bounded by multiple linear constraints as in Figure 1 [4, 5, 6, 22], that depend on the queuing mechanism in the switch or router. In this situation there is no notion of effective bandwidth anymore; the shadow prices given by the bottleneck constraint will determine the prices [15]. But again, in the situation of many suppliers, the prices will be determined by costs. We illustrate this below in three steps. First we introduce a simple model, which still captures the crucial features of the user model and the provider model, as described in the previous subsections. In the first model there is just one supplier. In the second step we choose, for reasons of analytical tractability, special demand functions: one in which demand depends *linearly* on prices, and one in which there is *constant elasticity*. In the third step, we let the number of suppliers grow large.

This procedure resembles the standard approach presented in many microeconomics textbooks, see for instance Russell and Wilkinson [16, pp. 280 ff.]. It has its roots in the classical duopoly model of Cournot. This model has some intrinsic flaws. The lack of dynamics is particularly unrealistic: in reality the suppliers react to each other's actions, and a game-theoretic model is more suitable. However, in our opinion the procedure is effective in making the point that the market structure approaches one of perfect competition, when the number of providers is large.

Simplified equilibrium model with one provider. The following scheme optimizes the provider's profit. We only consider the case of two QoS classes, say Class 1 and Class 2. The same line of reasoning applies to the general case of multiple guaranteed QoS levels.

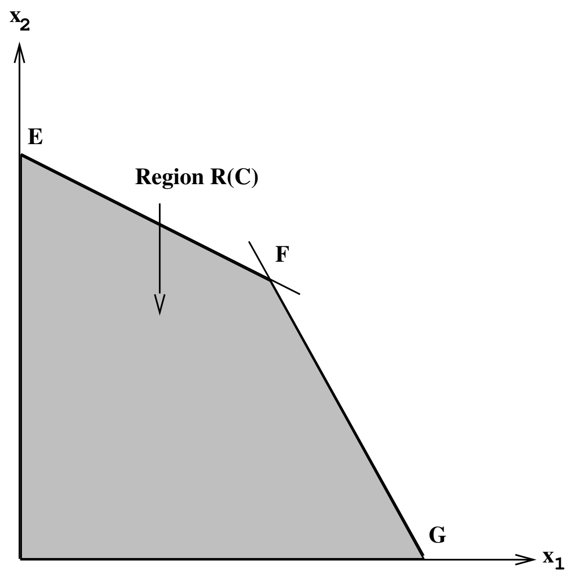
All Class 1 connections have identical traffic contract parameters. The same holds for Class 2 connections. Let x_1 (x_2, respectively) be the average number of Class 1 (Class 2) connections *simultaneously* present in the system. Notice that these numbers depend on the respective prices (per unit time) of these services, say p_1 and p_2. As indicated in the previous subsection, the optimization undertaken by the network provider involves choosing the prices (p_1, p_2), such that the profit is maximized.

Assume that the number of connections demanded by the customers is a known function of the price, say $x_1(p_1)$ and $x_2(p_2)$. Clearly $x_1(p_1)$ ($x_2(p_2)$, respectively) will be decreasing in p_1 (p_2). Denoting by $f(C)$ the cost of having a capacity C during one unit of time, the provider has to perform the following optimization:

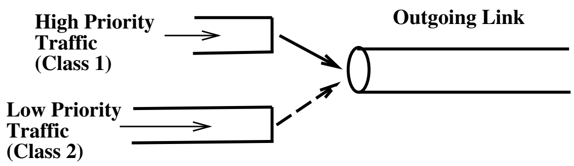
$$\max_{p_1, p_2} \left[\{p_1 \cdot x_1(p_1) + p_2 \cdot x_2(p_2)\} - f(C) \right], \tag{1}$$

for (x_1, x_2) in the admissible region $R(C)$.

We now consider admissible regions $R(C)$ which are bounded by linear segments, as shown in Figure 1(a).



(a)



(b)

Figure 1: (a) Example of a simplex as the admissible region $R(C)$ for two classes. (b) A two-priority, two-class multiplexing system.

The admissible region $R(C)$ consists of all (x_1, x_2) such that

$$\begin{cases} e_1^{(1)} x_1 + e_2^{(1)} x_2 & \leq & C \\ e_1^{(2)} x_1 + e_2^{(2)} x_2 & \leq & C \end{cases} \tag{2}$$

where $e_1^{(1)}, e_2^{(1)}, e_1^{(2)}$ and $e_2^{(2)}$ are non-negative constants satisfying $e_1^{(2)} < e_1^{(1)}$ and $e_2^{(1)} < e_2^{(2)}$. In the figure the segments FG and EF correspond to binding first and second constraints, respectively. Note that $e_i^{(j)}$ may be interpreted to be the effective bandwidth for Class i calls when the jth constraint is binding.

Approximations to the admissible regions in which the boundaries have linear segments arise in various contexts, see for example [4, 5, 6, 7, 22]. The special case wherein $e_2^{(1)} = 0$ is of particular interest, since it arises for the priority system shown in Figure 1(b), see [4, 5]. In this system Class 1 traffic has strict priority over Class 2, i.e. the contents of the buffer that is used exclusively for Class 1 traffic has priority access to the outgoing link. The above characterization of the admissible region for this system is given in [5] and the interpretation is as follows: in order to satisfy the QoS requirements of Class 1, at most $C/e_1^{(1)}$ connections of Class 1 are admissible, even if no Class 2 connections are present; however, up to $(C/e_2^{(2)})(e_1^{(1)} - e_1^{(2)})/e_1^{(1)}$ connections of Class 2 may also be admitted while satisfying the requirements of both classes. For less than $C/e_1^{(1)}$ connections of Class 1, the boundary of the admissible region, while provably nonlinear, is well approximated by a linear segment.

For other scheduling/priority systems (e.g. generalized processor sharing), the computations of the admissible regions are more complicated [6, 22], but their characterizations are not too dissimilar.

Linear and Constant Elasticity Demand Functions We illustrate the above optimization problem by two examples.

- Linear Demand: Assume that $x_1(p_1) = a_1 - b_1 p_1$ and $x_2(p_2) = a_2 - b_2 p_2$. The parameter a_1 (a_2) can be roughly interpreted as the number of customers when Class 1 (Class 2) services are free. The parameter b_1 (b_2) denotes the number of customers lost per unit increase of price; we assume $b_1 < b_2$. We associate Lagrange multiplier λ_1 (λ_2, respectively) with the first (second) constraint of (2) to find the optimum of (1), i.e. maximize:

  $$\begin{aligned} L(p_1, p_2, \lambda_1, \lambda_2) = \quad & p_1 \cdot x_1(p_1) + p_2 \cdot x_2(p_2) - \\ & \lambda_1 (e_1^{(1)} x_1 + e_2^{(1)} x_2 - C) - \lambda_2 (e_1^{(2)} x_1 + e_2^{(2)} x_2 - C) - f(C). \end{aligned}$$

  Then calculations yield as optimal prices:

  $$p_1 = \frac{a_1 + \lambda_1 e_1^{(1)} b_1 + \lambda_2 e_1^{(2)} b_1}{2 b_1} \quad \text{and} \quad p_2 = \frac{a_2 + \lambda_1 e_2^{(1)} b_2 + \lambda_2 e_2^{(2)} b_2}{2 b_2}.$$

Now there are four possibilities. In the first case, the optimal quantities lie in the interior of $R(C)$. Resources are abundant: $\lambda_1 = \lambda_2 = 0$; prices are simply $p_1 = a_1/2b_1$ and $p_2 = a_2/2b_2$. The second possibility is that the optimum is such that the first constraint is satisfied with equality and the second with strict inequality, corresponding to $\lambda_1 > 0$ and $\lambda_2 = 0$. In [4] it is more or less suggested for the priority system (i.e. $e_2^{(1)} = 0$), that the Class 1 user should get bandwidth for free here, but of course he has to pay $a_2/2b_2$. The third case is such that the second constraint is binding on the optimal quantities, whereas the first is not binding; then $\lambda_1 = 0$ and $\lambda_2 > 0$. In the last case, both resources are scarce, $\lambda_1 > 0$ and $\lambda_2 > 0$.

- Constant Elasticity Demand: Here $x_1(p_1) = A_1/p_1^{\epsilon_1}$ and $x_2(p_2) = A_2/p_2^{\epsilon_2}$ where ϵ_1 and ϵ_2 are constant elasticities for Class 1 and Class 2, respectively. Lanning at al [12] have provided evidence to justify such constant elasticity demand functions for voice and data services in which the elasticity exceeds unity, with the exceedance depending on the service. We too assume that the elasticities exceed unity.

  Here it will be more convenient to make (x_1, x_2) the decision variables. Note that $p_i = (A_i/x_i)^{1/\epsilon_i}$. Hence the objective is to maximize

$$\begin{aligned} L(x_1, x_2, \lambda_1, \lambda_2) = \quad & A_1^{1/\epsilon_1} x_1^{(\epsilon_1 - 1)/\epsilon_1} + A_2^{1/\epsilon_2} x_2^{(\epsilon_2 - 1)/\epsilon_2} - \\ & \lambda_1(e_1^{(1)} x_1 + e_2^{(1)} x_2 - C) - \lambda_2(e_1^{(2)} x_1 + e_2^{(2)} x_2 - C) - f(C), \end{aligned}$$

  where, as before, λ_1 and λ_2 are Lagrange multipliers associated with the constraints in (2). Setting the derivatives of L with respect to x_1 and x_2 to zero give

$$\left\{ \begin{array}{lcl} \frac{\epsilon_1 - 1}{\epsilon_1} \left(\frac{A_1}{x_1} \right)^{1/\epsilon_1} & = & \lambda_1 e_1^{(1)} + \lambda_2 e_1^{(2)} \\ \frac{\epsilon_2 - 1}{\epsilon_2} \left(\frac{A_2}{x_2} \right)^{1/\epsilon_2} & = & \lambda_1 e_2^{(1)} + \lambda_2 e_2^{(2)}. \end{array} \right. \tag{3}$$

  Hence, the optimum prices are given simply by the elasticities and the Lagrange multipliers:

$$p_1 = \frac{\epsilon_1}{\epsilon_1 - 1}(\lambda_1 e_1^{(1)} + \lambda_2 e_1^{(2)}) \quad \text{and} \quad p_2 = \frac{\epsilon_2}{\epsilon_2 - 1}(\lambda_1 e_2^{(1)} + \lambda_2 e_2^{(2)}).$$

  Moreover, simple criteria may be used to determine where the optimum is. First, we may rule out the possibility of the optimum occurring in the interior of the admissible region, i.e., where $\lambda_1 = \lambda_2 = 0$ in (3). The reason of course is that, with elasticity exceeding unity, revenue monotonically increases with demand.

Next consider the possibility of the optimum occurring in the interior of the segment EF, where only the second constraint is binding. For this to happen, (3) must hold with $\lambda_1 = 0$ and $\lambda_2 > 0$, i.e.,

$$\frac{1}{e_1^{(2)}} \frac{\epsilon_1 - 1}{\epsilon_1} \left(\frac{A_1}{x_1}\right)^{1/\epsilon_1} = \frac{1}{e_2^{(2)}} \frac{\epsilon_2 - 1}{\epsilon_2} \left(\frac{A_2}{x_2}\right)^{1/\epsilon_2} = \lambda_2.$$

Thus, this point, if it exists, is uniquely determined by the condition that the rates of revenue increase with demand (in bandwidth units) for both classes are equal to the common cost of unit bandwidth, λ_2. Tacitly $e_i^{(2)}$ is the bandwidth consumed per call of Class i.

Exploiting monotonicity, we may give the following equivalent condition for the optimum occurring in the interior of the segment EF. Let (x_1', x_2') denote the coordinates of the point F in Figure 1(a), where both constraints in (2) are satisfied. If

$$\frac{1}{e_1^{(2)}} \frac{\epsilon_1 - 1}{\epsilon_1} \left(\frac{A_1}{x_1'}\right)^{1/\epsilon_1} < \frac{1}{e_2^{(2)}} \frac{\epsilon_2 - 1}{\epsilon_2} \left(\frac{A_2}{x_2'}\right)^{1/\epsilon_2} \tag{4}$$

then the optimum is in the interior of EF.

Similarly, if

$$\frac{1}{e_2^{(1)}} \frac{\epsilon_2 - 1}{\epsilon_2} \left(\frac{A_2}{x_2'}\right)^{1/\epsilon_2} < \frac{1}{e_1^{(1)}} \frac{\epsilon_1 - 1}{\epsilon_1} \left(\frac{A_1}{x_1'}\right)^{1/\epsilon_1} \tag{5}$$

then the optimum is in the interior of the segment FG.Finally, if the inequalities in (4) and (5) are reversed, then the optimum occurs at the point F.

In the special case of $e_2^{(1)} = 0$, which corresponds to the priority system, the optimum does not occur in the interior of the first constraint.

Oligopoly and perfect competition. Although the analysis presented above assumes one provider – a monopoly – it can be adapted easily to a situation with a (small) number of competitors, an oligopoly, and the situation of perfect competition.
Let there be n providers, each with C/n capacity. Let $p_1(x_1)$ be the linear price-demand relation introduced above; $p_2(x_2)$ is defined analogously. Then the optimization problem of the ith supplier becomes

$$\max_{x_i^{(1)}, x_i^{(2)}} \left\{ p_1 \left(\sum_{i=1}^{n} x_i^{(1)} \right) x_i^{(1)} + p_2 \left(\sum_{i=1}^{n} x_i^{(2)} \right) x_i^{(2)} \right\} - f\left(\frac{C}{n}\right),$$

over $x_i^{(1)}, x_i^{(2)} \in R(C/n)$. Calculations yield that the prices become

$$p_1 = \frac{a_1}{(n+1)b_1} + \frac{n}{n+1}(\lambda_1 e_1^{(1)} + \lambda_2 e_1^{(2)}) \quad \text{and} \quad p_2 = \frac{a_2}{(n+1)b_2} + \frac{n}{n+1}(\lambda_1 e_2^{(1)} + \lambda_2 e_2^{(2)}).$$

We see that the profits of all providers leak away as n grows large: prices become $(\lambda_1 e_1^{(1)} + \lambda_2 e_1^{(2)})$ and $(\lambda_1 e_2^{(1)} + \lambda_2 e_2^{(2)})$, respectively, exactly equal to the costs. The aggregate demand converges to $\{a_1 - (\lambda_1 e_1^{(1)} + \lambda_2 e_1^{(2)})b_1\}$ for Class 1 and $\{a_2 - (\lambda_1 e_2^{(1)} + \lambda_2 e_2^{(2)})b_2\}$ for Class 2, $n \to \infty$. Similar calculations can be done for the constant elasticity demand function, but the computations are more involved.

3 Resource usage, admission control, and charging of individual streams

The previous section described conditions under which prices for services are cost-based. Given this objective, still a number of practical matters have to be addressed. In the first place, the resulting prices critically rely on the shape of the admissible region R, i.e., α, β, and γ in the previous section. In other words: *the chosen admission policy has a direct impact on the prices.* This will be the major subject of this section. For reasons of simplicity we consider the situation in which all customers get the same QoS.
We focus on two extreme levels of QoS: strict QoS guarantees, and what may be called 'predictive' QoS guarantees.

A. Strict QoS guarantees. As long as users want to have strict guarantees about the QoS level they get, a crucial role is played by the policing function. If all flows are subject to these regulators, there is an upper bound on the traffic they transmit. The (dual) leaky bucket regulator is commonly used, which comprises three parameters. In the first place, a *peak rate* P is defined. Further, the *sustained rate* r is an upper bound to the long-term average rate. Finally, the *maximum burst size* B_T is the amount of traffic that can be uninterruptedly sent at peak rate.
In particular, a *worst case* profile can be determined: this type of adversarial traffic characteristic is — within the limits imposed by the parameters of the regulator — what maximizes the loss probability for fixed maximum delay. It is generally believed that, when traffic is offered to a link, loss probabilities are maximized by streams of on-off type with deterministic on and off times: i.e., a stream transmits at rate P during on-times of length $B_T/(P-r)$ and at rate 0 during off-times of fixed length such that the average rate is r.
For a queue with given buffer B and link rate C, an *effective bandwidth* can be calculated [7]:

This yields effective bandwidth

$$c_i = \max\left\{r_i, \frac{B_{T,i}}{(B/C)(1 - r_i/P_i) + B_{T,i}/P_i}\right\}, \tag{6}$$

where $(r_i, P_i, B_{T,i})$ are the parameters of source i. As long as the sum of the c_i of the connected flows is smaller than the link speed C, there is no loss. This procedure can be easily enhanced to allow a small loss fraction, for instance 10^{-5}.

Crucial is that the guarantees given by such a procedure are strict, and in fact this implies that also the cost of a connection is predictible. User i should be an amount that is proportional to his c_i per unit time.

B. 'Predictive' QoS guarantees. The traffic contract is usually only a rough 'envelope' of the traffic actually sent by the user. This is mainly due to the fact that the traffic descriptor that is used in the traffic contract typically consists only of three parameters, being simply too few to give an accurate description of the traffic pattern. In the second place, in general it is hard for a user to *a priori* describe his traffic pattern, as it very much depends on the specific actions the user undertakes during his session. Hence, a reliable AC based on the traffic contract – usually called 'static' AC – will be very pessimistic, in that it leads to a low utilization of network resources.

Based on the above line of reasoning, clearly static AC should only be used in the case that very strict QoS guarantees are required. Fortunately, a significant set of services can cope with temporary QoS degradations (e.g., data services using TCP/IP) and can be subjected to a less conservative admission control approach. In this way the gap between the traffic contract and the real traffic profile can be exploited: one can allow more flows to the network than justified by AC based on worst-case profiles.

Particularly for applications that tolerate such a 'predictive' QoS, recent literature has focused on *measurement-based admission control* (MBAC): admission based on real-time load measurements, see e.g. Jamin *et al.* [9]. As long as the load is below a certain thresholds new connections are admitted.

Clearly, these schemes offer only very 'soft' QoS indications. As the admission decision is done on the basis of load measurements, it seems reasonable to charge customers on the basis of the volume of traffic they sent. However, one could also think of punishing the streams that essentially caused the temporary service degradations that now and then occur. In this respect, we mention the work by Gibbens and Kelly [8]. They propose to mark all packets when the resources approach overload, where the user has to pay for every mark. In this way the users are given the right incentive to use the network efficiently. Here the charge is not predictable; the amount a particular user has to pay also depends on the load imposed on the resources by other users.

Between these two extreme schemes, a large number of hybrid schemes can be distinguished. For instance, the charging algorithm proposed by Kelly [11, 19] is somewhere in the middle of this spectrum. His charging formula is $aT + bV$, where T is measured connection time, V is measured volume, and a and b are constants that depend on the traffic contract. Extensive feasiblity tests of this approach were given in [18].

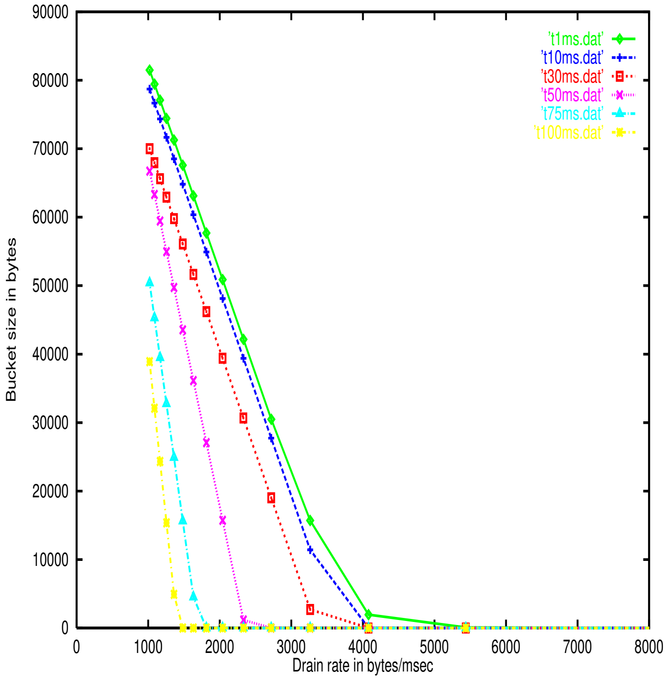
Another hybrid variant could be the following: do MBAC as described above, but accept as long as the measured load increased by the peak rate of the new flow is below, say, 90%. This variant is close to scheme (B), but somehow the peak rate of the new flow must be incorporated in the charge.

4 Estimation of Leaky Token Bucket Parameters from Traffic Traces

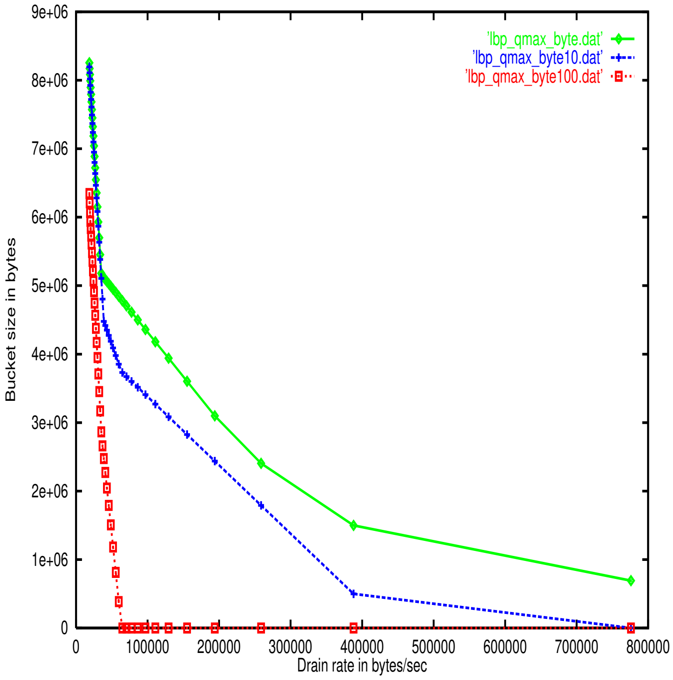
In the previous three sections, relationships between pricing and admission control were described and models were formulated for revenue and utility optimization. These models assume the existence and universality of traffic descriptors (T_{spec}) that characterize services. A traffic descriptor not only provides a satisfactory description of each service and its resource needs through a compact set of parameters, but also gives an adequate characterization and quantification of the service so that network providers can manage and utilize their network efficiently.

In this section we propose use of leaky bucket parameters – together with loss and delay associated with a service – as the T_{spec} used in the pricing model described in Section 2 and admission mechanism (based on effective bandwidth) described in Section 3. Leveraging extensive work on leaky bucket parametrization (LBP, see [1, 20]) of variable bit-rate ATM services [7], we argue LBP is a compact and natural basis for pricing as it reflects the amount of resources to be allocated in simplest terms. At least as a first exercise in the application of the general theory developed so far, the LBPs reveal the critical issues that need to be addressed in pricing of variable rate services and pave the way for more specialized descriptors proposed in the traffic engineering literature.

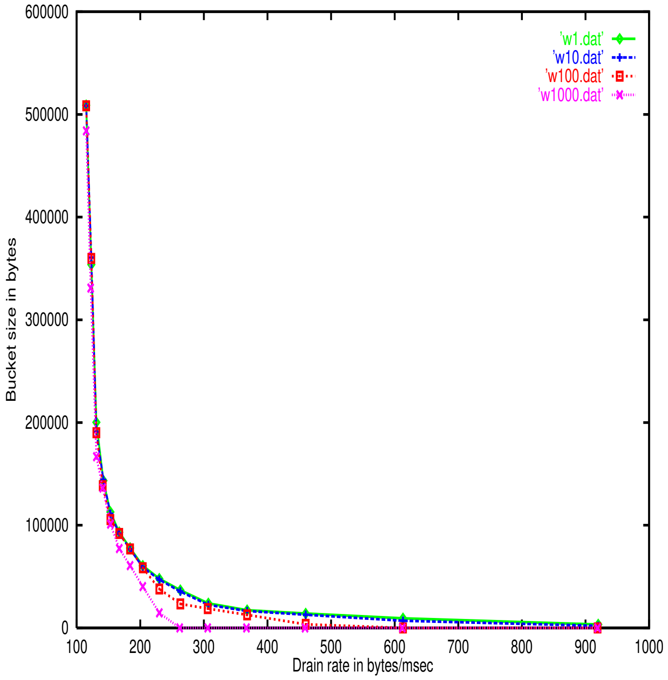
Our proposal differs from [10, 11] in that the authors use *a priori* traffic descriptors together with *a posteriori*(measured) connection duration and data volume to determine charges. In our proposal, an *a priori* descriptor is not required as the LBPs are meant to be estimated on-line from the data collected by policing mechanisms at the edge of network. Indeed, in the rest of this section we demonstrate how this is done and list the difficulties that need to be resolved.



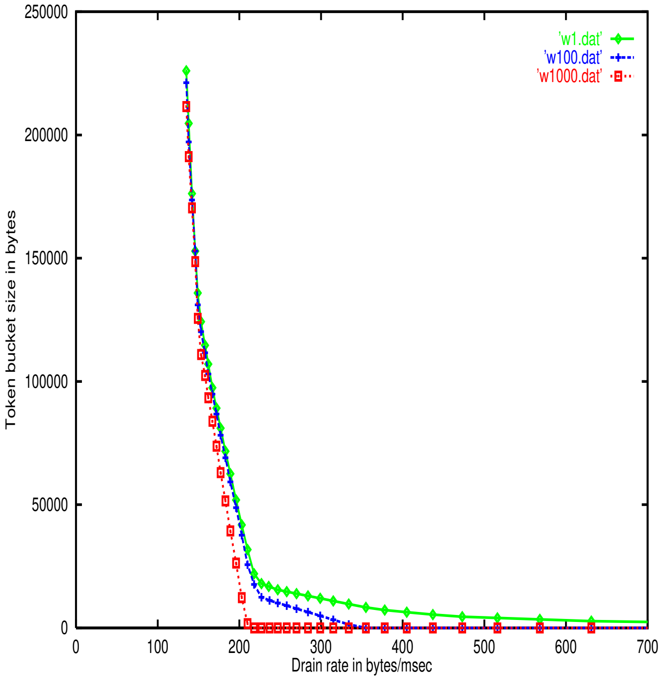
(a) MPEG2 Trace



(b) IP Trace



(c) TCP Trace



(d) UDP Trace

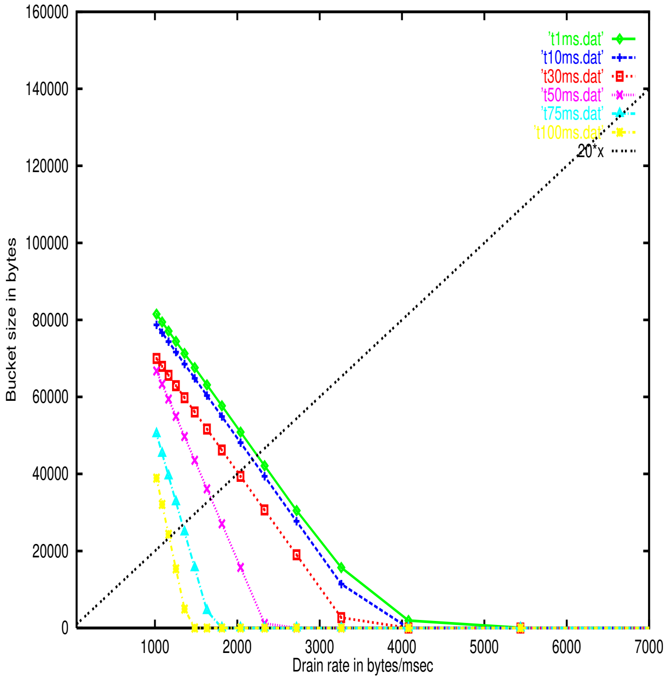
Figure 2: Maximum queue length versus output rate of a virtual queue designed to determine token bucket size as a function of token rate, for a variety of choices for the measurement interval, for 4 traffic traces of MPEG2 (5 min), IP (1 hour), TCP (4 min) and UDP (4 min)

Four traffic traces * were used. To estimate the peak rate, sustained rate and token bucket size, a discrete time *virtual fluid* queuing model was used. The size of the clock increment was set to the resolution of the data, the token rate was set equal to the drain rate of the queue and the maximum queue size, corresponding to lossless LBP, was recorded for each token rate. Thus a curve trading leaky bucket size against leaky bucket token rate results, see . 2(a)-(d). The experiment was repeated for higher levels of aggregation of data corresponding to different measurement intervals and higher – or more aggregate – clock increments. These different measurement intervals could correspond to the granularity of the policing mechanisms at the edge of the network that collect online traffic. Fig. 2(a)-(d) show the resulting bucket size versus token rate plots, each depicting a set of curves corresponding to different measurement intervals. A summary of the observations from these plots follow:

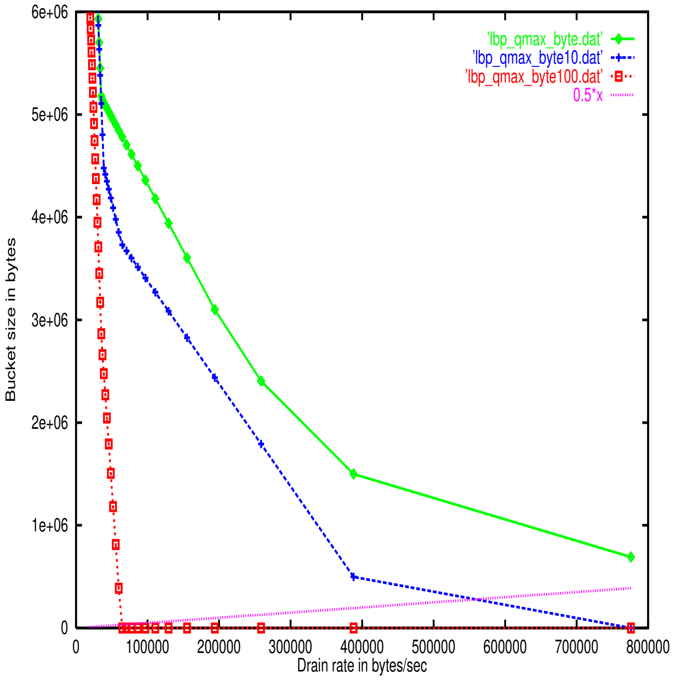
- Trade-off between bucket size and token rate – There's a natural trade-off between the leaky bucket size and the token rate, where any point on the "iso-service curve" as seen in each of Fig. 2(a)-(d) gives rise to a feasible T_{spec}. Each point is a valid LBP for the traffic flow/service. A lower bound on the token bucket rate is clearly the mean rate of the service, as measured from the trace and its maximum is the peak rate, at which point the token bucket size is zero. Given a delay bound, a better lower bound may be imposed on the token rate. Clearly a fair pricing mechanism needs to capture this tradeoff by selecting a single –lowest cost– point from the iso-service curve, given the loss and delay requirements of the application. We derive the effective bandwidth corresponding to the optimum price from the lossless multiplexing formula of [6] below.

- Measurement intervals – T Whereas LBPs are defined in continuous time, in practice measurements are performed at discrete times. Taking different measurement intervals, however, can result in widely differing LBPs for the same service: long intervals lead to mean-rate dominated characterization, whereas shorter units have a bias towards the peak rate, as seen in the MPEG2, IP and UDP iso-service curves in Fig. 2(a)-(d). Thus, there is a need for predetermination of measurement interval prior to pricing. It may be argued that the measurement interval itself belong to the extended LBP parametrization, e.g. together with loss and delay. We assume T_{spec} contains the measurement interval as given, thus resolving the natural tendency to do long term averages typically detrimental to service recipients.

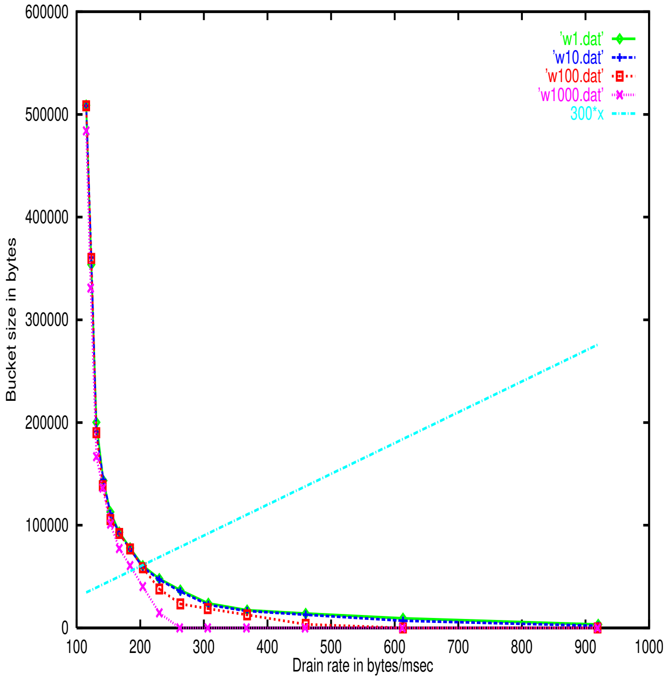
*Two publicly available traces for TCP and UDP were derived from data at the Internet Traffic Archives http://ita.ee.lbl.gov/html/contrib/DEC-PKT.html, and two private traces from Knoltex Corporation (courtesy of Amarnath Mukherjee) and Lucent Technologies, respectively.



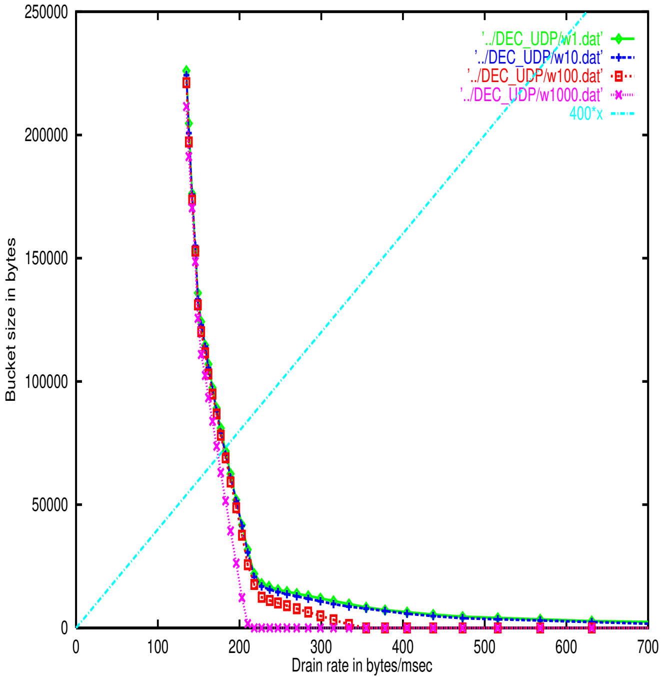
(a) MPEG2 Trace



(b) IP Trace



(c) TCP Trace



(d) UDP Trace

Figure 3: Effective bandwidth determined from the intersection of iso-delay lines with the iso-service curves for each of the 4 traffic traces (MPEG2, IP, TCP and UDP) with typical delay values of 20msecs for MPEG2, 0.5 sec for IP, 300msec for TCP and 400msec for UDP

4.1 Determination of the Effective Bandwidths

Using formula (6) we calculate the effective bandwidth for lossless multiplexing for each of the iso-service curves for each measurement interval, as seen in Fig. 3(a)-(d). Using approximate delay bounds for each service class (20msecs for MPEG2, 0.5 sec for IP, 300msec for TCP and 400msec for UDP) the effective bandwidth for each measurement interval is obtained from the intersection of the "iso-delay line" with the iso-service curves. The optimal effective bandwidths calculated above are used for pricing, resource allocation, and admission control as described before.

5 Conclusions

We presented a general economic equilibrium model of users and service providers in a multi-class multi-QoS network. The key parameters in the model were the T_{spec} and Q_{spec}, which characterize traffic and Quality of Service (QoS) respectively. Using trace-driven simulations, we derived the effective bandwidths used in pricing of four distinct services and showed the measurement intervals for collection of online traffic must be considered as part of the traffic descriptor. We showed that the imputed prices were valid only if the resource reservations were appropriately used by each service. This, we argued, could only be achieved via a consistent call admission control mechanism.

References

[1] ATM Forum UNI Specifications for CCITT Recommendation I371.

[2] L. Anania and R. Solomon. Flat—the minimalist price. In L. McKnight and J. Bailey, editors, *Internet Economics*. MIT Press, Cambridge, Massachusetts, 1997.

[3] N. Brownlee. Internet pricing in practice. In L. McKnight and J. Bailey, editors, *Internet Economics*. MIT Press, Cambridge, Massachusetts, 1997.

[4] C. Courcoubetis, F. Kelly, and R. Weber. Measurement-based charging in communications networks. Technical report, 1997-19, Statistical Laboratory, University of Cambridge. Available as http://www.statslab.cam.ac.uk/Reports/1997/1997-19.html, 1997.

[5] A. Elwalid and D. Mitra. Analysis, approximations and admission control of a multi-service multiplexing system with priorities. In *Proc. IEEE INFOCOM'95*, pages 1115–1128, 1995.

[6] A. Elwalid and D. Mitra. Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes. In *Proc. IEEE INFOCOM'99*, pages 1220–1230, 1999.

[7] A. Elwalid, D. Mitra, and R. H. Wentworth. A new approach for allocating buffers and bandwidth to heterogenous, regulated traffic in an ATM Node. *IEEE Journal on Selected Areas in Communications*, 13(6):1115–1127, Aug. 1995.

[8] R. Gibbens and F. Kelly. Resource pricing and the evolution of congestions control. Technical report, available as http://www.statslab.cam.ac.uk/ frank/evol.html, 1998.

[9] S. Jamin, P. Danzig, S. Shenker, and L. Zhang. A measurement-based Admission Control for integrated service packet networks. *IEEE/ACM Transactions on Networking*, 5:56–70, 1997.

[10] F. Kelly. Notes on Effective Bandwidths. In F. Kelly, S. Zachary, and I. Ziedins, editors, *Stochastic Networks: Theory and Applications*, 1996.

[11] F. Kelly. Charging and accounting for bursty connections. In L. McKnight and J. Bailey, editors, *Internet Economics*. MIT Press, Cambridge, Massachusetts, 1997.

[12] S. Lanning, D. Mitra, Q. Wang, and M. Wright. Optimal planning for optical transport networks. To appear in *Phil. Trans. Royal Soc., Part A*.

[13] J. MacKie-Mason and H. Varian. Pricing the Internet. In B. Kahin and J. Keller, editors, *Public access to the Internet*. Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[14] J. MacKie-Mason and H. Varian. Pricing congestible network resources. *IEEE Transactions on Selected Areas in Communications*, pages 1141–1149, 1995.

[15] M. Mandjes and N. van Foreest. Aspects of pricing in an integrated services network. In *Proceedings of ITC 16*, pages 1331–1340, 1999.

[16] R. Russell and M. Wilkinson. *Microeconomics. A synthesis of modern and neoclassical theory*. Wiley, New York, 1979.

[17] S. Shenker. Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communications*, pages 1176–1188, 1995.

[18] D. Songhurst. *Charging communication networks. From theory to practice.* Elsevier, Amsterdam.

[19] D. Songhurst and F. Kelly. Charging schemes for multi-service networks. In *Proceedings of ITC 15*, pages 781–790, 1997.

[20] J. S. Turner. Managing Bandwidth in ATM Networks with Bursty Traffic. *IEEE Network*, 1992.

[21] D. Walker, F. Kelly, and J. Solomon. Tariffing in the new IP/ATM environment. In *Telecommunications Policy*, pages 283–295, 1997.

[22] Z.-L. Zhang. Large Deviations and the Generalized Processor Sharing scheduling for a two-queue system. *Queueing Systems*, pages 229–254, 1997.