In IFIP TC6 International Conference on Broadband Communications (BC'98), Stuttgart, Germany, April 1-3, 1998

A study of simple usage-based charging schemes for broadband networks*

C. Courcoubetis[1], *F. P. Kelly*[2], *V. A. Siris*[1], *and R. Weber*[2]

[1]*ICS-FORTH and Dept. of Computer Science, University of Crete, P.O. Box 1385 GR 711 10 Heraklion, Greece.*
email: {courcou,vsiris}@ics.forth.gr

[2]*University of Cambridge, Statistical Laboratory, 16 Mill Lane, Cambridge CB2 1SB, UK.*
email: {fpk,rrw1}@statslab.cam.ac.uk

Abstract

Operators of high-speed networks are interested in implementing simple charging schemes with which they can fairly recover costs from their customers and effectively allocate network resources. This paper describes an approach for computing such charges from simple measurements (the duration and transferred volume of a connection), and relating these to bounds of the effective bandwidth. A requirement for usage-based charging schemes is that they capture the relative amount of resources used by connections. Based on this criteria, we evaluate our approach for Internet Wide Area Network traffic. Furthermore, its incentive compatibility is displayed with an example involving deterministic multiplexing, and the effect of pricing on a network's equilibrium is investigated for deterministic and statistical multiplexing.

Keywords

Usage-based charging, effective bandwidths, incentive compatibility, ATM, Internet

1 INTRODUCTION

A method for charging and pricing is an essential requirement in operating a high-speed network. Pricing is not only needed for recovering costs. There are compelling reasons that pricing is needed as a method of control. The congestion that has plagued the Internet, where pricing is based largely on *flat rate* pricing, highlights the fact that without usage-based pricing it is difficult to control congestion or divide network resources amongst

*This work was supported in part by the EC under ACTS Project CASHMAN (AC-039).

customers in a workable and stable way (Mackie-Mason and Varian 1995, Mackie-Mason and Varian 1995, Gupta *et al.* 1994). Furthermore, in a competitive environment, besides offering sophisticated service disciplines, providers will need to price services in a manner which takes some account of network resource usage (Parris *et al.* 1992, Cocchi *et al.* 1993).

There are many considerations that influence the price of network services, such as marketing and regulation. However, these considerations are not particular to the operation of a communications network which is closely related to technological constraints (e.g., the quantities of services that it can support with a given network installation). A special consideration arises from the fact that a broadband communications network is intended to simultaneously carry a wide variety of traffic types and to provide certain performance guarantees. For example, in ATM networks a traffic contract is agreed among the customer and the operator. The customer agrees that his traffic will conform to certain parameters (e.g., which bound his peak rate and the size of his bursts), while the operator guarantees to carry this traffic with a particular quality of service (expressed, e.g., in terms of delay and cell loss ratio). The traffic contract gives the operator information by which he can bound the network resources that will be required to carry the call.

This paper is concerned with just one important part of the charging activity: that part which aims to assess a connection's resource usage. To avoid repeatedly having to qualify our remarks with a reminder that this is the focus, we shall henceforth simply refer to this component as *"charging"* and of computing a *"charge"*.

Some desired properties of tariffs

The role of tariffs is not only to generate income for the provider, but to introduce feedback and control. This happens via the mechanism that is automatically in effect as each individual customer reacts to tariffs and seeks to minimize his charges. For example, tariffs may be set which make it economical for some customers to shape their traffic, and by their doing so the overall network performance may be enhanced. This is the key idea of *incentive compatibility.* Tariffs should guide the population of cost-minimizing customers to select contracts and use the network in ways that are good for overall network performance (e.g., to maximize social welfare (Low and Varaiya 1993)). Tariffs which are not incentive compatible give the wrong signals and lead customers to use the network in very inefficient ways.

Well-designed tariffs should also have what we call the *fairness* property.† By this we mean that charges should reflect a customer's *relative* network usage. This raises the interesting question of when one charging scheme is more accurate than another, where accuracy is measured not in terms of the absolute value of the charges, but in terms of their correspondence to true network usage.

The above remarks naturally lead one to ask whether it possible to design tariffs that are sound, both in terms of incentive compatibility and fairness, but which are also not too complex, and whose implementation does not require the network operator to make overly sophisticated or unrealistic measurements. Incentive compatibility will be hard to achieve if tariffs are too complex, since customers will find it difficult to determine what

†In the case of differential pricing and/or time-of-day pricing, the fairness property is considered for customers of the same "class" which use network services at the same time period.

effect the decisions under their control, such as whether or not to shape their traffic, might have on the charges they incur.

Contribution of the paper

In this paper we provide the framework for constructing incentive compatible charges that reflect effective usage. Our approach is based on the notion of effective bandwidth as a proxy for resource usage. In this sense our work differs from (Low and Varaiya 1993, Sairamesh *et al.* 1995) which investigate optimal pricing strategies assuming that network resources (buffer and capacity) are charged separately, and (Wang *et al.* 1996) which also deals with optimal pricing, but does not address the issue of measuring the amount of resources used by connections.

Our charging schemes are simple and can be cast in the same formats that are used today, namely the charge depends on *static* contract parameters (access line speed, policing parameters, anticipated average rate) and on *dynamic* parameters of the connection (actual average rate). Our approach is quite general, and can also be used to design that part of a tariff which prices the network usage of large customers connected to an Internet service provider. Furthermore, it can be complemented with other pricing mechanisms such as time-of-day pricing (Shenker *et al.* 1996).

The novelty of the approach lies in the following two points. First, we provide an interpretation of effective bandwidths that is right for our purposes. In (Courcoubetis, Kelly and Weber 1997) we provide the mathematical foundation of our charging framework where we show that the effective bandwidth of a connection depends on the actual state (composition of the traffic mix) of the links in a network, hence can not be defined in isolation. Furthermore, this dependence is only through a pair of parameters (the s, t parameters discussed in Section 2.1). The same connection will potentially exhibit different effective bandwidths at different times of the day. An important consequence of the approach is that it treats deterministic and statistical multiplexing in a unifying way.

The second contribution is in the way we transform simple tariffs of the form $a_0T + a_1V$, where T is the duration and V is the volume of a connection, into sound approximations of the effective bandwidth of the connection, by casting all the information from the static contract parameters and the operating point of the network into the coefficients a_0, a_1.[‡] Based on experimentation, we believe that our simple tariffs can serve their purpose well and can provide the right incentives for efficient and stable network operation.

The rest of the paper is organized as follows. In Section 2 we briefly explain our charging methodology by reviewing some key notions and results for the simpler case of a network consisting of a single shared link. In Section 3 we discuss issues related to the fairness of charging schemes, based on which we evaluate our approach for Internet Wide Area Network traffic. In Section 4 we discuss the incentive compatibility of the approach and work through a complete example in the simpler, but illuminating, case of deterministic multiplexing. Our conclusions and some open issues are discussed in Section 5.

[‡]The theory developed in (Courcoubetis, Kelly and Weber 1997) allows for the construction of more elaborate charging schemes where the network measurements can be arbitrarily complex.

2 A THEORY FOR USAGE-BASED CHARGING

2.1 Effective bandwidths as a measure of resource usage

Suppose the arrival process at a broadband link is the superposition of independent sources of J types: let n_j be the number of connections of type j, and let $n = (n_1, \ldots, n_J)$. We suppose that after taking into account all economic factors (such as demand and competition) the proportions of traffic of each of the J types remains close to that given by the vector n, and we seek to understand the relative usage of network resources that should be attributed to each traffic type.

Consider a discrete time model and let $X_j[0,t]$ be the total load produced by a source of type j in epochs $0, \ldots, t$. We assume that the increments of $\{X_j[0,t],\, t \geq 0\}$ are stationary. Then, the *effective bandwidth* of a source of type j is defined as

$$\alpha_j(s,t) = \frac{1}{st} \log E\left[e^{sX_j[0,t]}\right] , \tag{1}$$

where s, t are *system defined* parameters which depend on the characteristics of the multiplexed traffic and the link resources (capacity and buffer). Specifically, the *time* parameter t (measured in, e.g., msec) corresponds to the most probable duration of the buffer busy period prior to overflow. The *space* parameter s (measured in, e.g., kb^{-1}) corresponds to the degree of multiplexing and depends, among others, on the size of the peak rate of the multiplexed sources relative to the link capacity. In particular, for links with capacity much larger than the peak rate of the multiplexed sources, s tends to zero and $\alpha_j(s,t)$ approaches the mean rate of the source, while for links with capacity not much larger than the peak rate of the sources, s is large and $\alpha_j(s,t)$ approaches the maximum value of $X_j[0,t]/t$.

Let $L(C,B,n)$ be the proportion of workload lost, through overflow of a buffer of size $B > 0$, when the server has rate C and $n = (n_1, n_2, \ldots, n_J)$. Assume that the constraint on the proportion of workload lost is $e^{-\gamma}$ (we will assume that the Quality of Service -QoS- is expressed solely through this quantity). The *acceptance region* $A(\gamma, C, B)$ is the subset of $\mathbf{Z}_+^J$ such that $n \in A(\gamma, C, B)$ implies $\log L(C,B,n) \leq -\gamma$, i.e., the QoS constraint is satisfied.

If n is on the boundary of the region $A(\gamma, C, B)$, and the boundary is differentiable at that point, then the tangent plane determines a half-space which is well approximated, when C, B, and n are large, by (Kelly 1996)

$$\sum_j n_j \alpha_j(s,t) \leq C + \frac{1}{t}\left(B - \frac{\gamma}{s}\right) , \tag{2}$$

where (s,t) is an extremizing pair in the equation (called the *many sources asymptotic*; see (Courcoubetis and Weber 1996))

$$\lim_{N\to\infty} \frac{1}{N} \log L(NC, NB, nN) = \sup_t \inf_s \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - s(Ct+B) \right] . \tag{3}$$

The asymptotics behind this approximation assumes only stationarity of sources, and illustrative examples discussed in (Kelly 1996) include periodic streams, fractional Brownian input, policed and shaped sources, and deterministic multiplexing. Note that the QoS guarantees are encoded in the effective bandwidth definition through the value of γ which influences the form of the acceptance region.

We must stress the network engineering implications of the above results. For any given traffic stream, the effective bandwidth definition (1) is nothing more than a template that must be filled with the link's operating point parameters s, t in order to provide the correct measure of effective usage. Furthermore, experimentation has revealed that the values of s, t are, to a large extent, insensitive to variations of the traffic mix (percentage of different traffic types) (Courcoubetis, Siris and Stamoulis 1997). Since during different times of the day the traffic mix at a given link is anticipated to remain relatively constant, we can assign particular pairs (s, t) to different periods of the day. These values can be computed off-line using (1) and (3), where the expectation in (1) is replaced by the empirical mean which is computed from traffic traces.

2.2 Charges based on effective bandwidths

We have argued above that effective bandwidths can provide a way to assess resource usage, and hence can be used for constructing the usage-based component of the charge. There are two extreme methods by which this can be done.

Consider sources of type j, where "type" is distinguished by parameters of the traffic contract and possibly some other static information. The network could form the empirical estimate $\alpha'_j(s, t)$ of the expectation appearing in formula (1), as determined by past connections of type j. A new connection of type j would be charged at an amount per unit time equal to $\alpha'_j(s, t)$. This is the charging method adopted in an all-you-can-eat restaurant. At such a restaurant each customer is charged not for his own food consumption, but rather for the average amount that similar customers have eaten in the past. Under such a charging scheme, each customer may as well use the maximum amount of network resources that his contract allows, which will result in $\alpha'_j(s, t)$ eventually becoming the largest effective bandwidth that is possible subject to the agreed policing parameters. Customers who have connections of type j, but whose traffic does not have the maximal effective bandwidth possible for this type, will not wish to pay as if they did, hence will seek network service providers using a different (more competitive) charging method.

At another extreme, one might charge a customer wholly on the basis of measurements that are made for his connection, i.e., charge the value of the effective bandwidth of the traffic actually sent. This has a conceptual flaw which can be illustrated as follows. Suppose a customer requests a connection policed by a high peak rate, but happens to transmit very little traffic over the connection. Then an *a posteriori* estimate of quantity (1), hence his charge, will be near zero, even though the *a priori* expectation may be much larger, as assessed by either the customer or the network. Since tariffing and connection acceptance control may be primarily concerned with expectations of *future* quality of service, the distinction matters. This is the case because such a charging scheme does not account for the resources reserved at call setup, which is unfair for the network operator.

Our approach lies part way between the two described above. We construct a charge that is based on the effective bandwidth, but which is a function of both *static* parameters (such as the peak rate and leaky bucket parameters) and *dynamic* parameters (these correspond to the actual traffic of the connection, the simplest ones being the duration and volume of the connection); we *police* the static parameters and *measure* the dynamic parameters; we bound the effective bandwidth by a linear function of the measured parameters, with coefficients that depend on the static parameters; and we use such linear functions as the basis for simple charging mechanisms. This leads to a charge with the right incentives for customers, which also compensates the network operator for the amount of resources reserved.

2.3 Charges linear in time and volume

Suppose that a connection lasts for epochs $1, \ldots, T$ and produces load $X_1, \ldots, X_T$ in these epochs. Imagine that we want to impose a *per unit time* charge for a connection of type j that can be expressed as a linear function of the form

$$f(X) = a_0 + a_1 g(X), \tag{4}$$

where $g(X)$ is the measurement taken from the observation $X = (X_1, \ldots, X_T)$ corresponding to $(1/T)\sum_{i=1}^{T} X_i$. In other words, the total charge is simply a function of the total number of cells carried, and, through a_0, the duration of the connection. This is practically the simplest measurement we could take and leads to charging schemes based on just time and volume.

We argued in Section 2.2 that the usage-based charge of a connection should be proportional to the effective bandwidth $\alpha(s,t)$ of the connection, for appropriate s,t. Next we describe how linear functions of the form (4) can be constructed so that the expected charge bounds the effective bandwidth.

Let $\bar{\alpha}(m, \mathbf{h})$ be an upper bound for the greatest effective bandwidth possible subject to constraints imposed by the traffic contract $\mathbf{h}$, while the mean rate is m. Consideration of $\bar{\alpha}(m, \mathbf{h})$ is partly motivated by the remark that this is what we would charge to a customer with mean rate m who makes maximal use of his traffic contract.

We define our tariffs in terms of the charging function f parameterized with $m, \mathbf{h}$. Mathematically, this corresponds to the tangent of $\bar{\alpha}(m, \mathbf{h})$ at m:

$$f(m, \mathbf{h}; X) := \bar{\alpha}(m, \mathbf{h}) + \lambda_m (g(X) - m), \tag{5}$$

which is of the form $a_0 + a_1 g(X)$, where $a_0[m, \mathbf{h}] = \bar{\alpha}(m, \mathbf{h}) - \lambda_m m$, $a_1[m, \mathbf{h}] = \lambda_m = \frac{\partial}{\partial m}\bar{\alpha}(m, \mathbf{h})$. These coefficients depend on the customer's choice of m. Because $\bar{\alpha}(m, \mathbf{h})$ is concave in m (Courcoubetis, Kelly and Weber 1997), one can show that the expected value of the charging rate for this connection is $Ef(m, \mathbf{h}; X) \geq \bar{\alpha}(Eg(X), \mathbf{h})$, with equality if $m = Eg(X)$ (the actual mean rate of the connection). Hence, the customer minimizes his expected charge if he chooses the tariff $f(Eg(X), \mathbf{h})$.

As we intended, the coefficients $a_0[m, \mathbf{h}], a_1[m, \mathbf{h}]$ depend upon both static information, as well as the customer's expectation regarding his mean rate (which is measured by the network). The dependence of the charge on m provides the customers with the right incentives for avoiding the "all-you-can-eat restaurant" effect mentioned before.

Approximations for $\bar{\alpha}(m, \mathbf{h})$

Let m be the mean rate of a source, and $\bar{X}[0,t]$ be the maximum amount of traffic produced in a time interval of length t. Since the source is policed by parameters (ρ_k, β_k), $k \in K$, we have

$$\bar{X}[0,t] \leq H(t) := \min_{k \in K} \{\rho_k t + \beta_k\} \,. \tag{6}$$

The last constraint together with the convexity of the exponential function implies that

$$\bar{\alpha}(m, \mathbf{h}) \leq \frac{1}{st} \log \left[1 + \frac{tm}{H(t)} \left(e^{sH(t)} - 1 \right) \right] = \tilde{\alpha}_{\mathrm{sb}}(m, \mathbf{h}) \,. \tag{7}$$

We call the right hand side of the above equation the "simple" approximation. This equation is illuminating for the effects of leaky buckets on the amount of resource usage. Each leaky bucket (ρ_k, β_k) constraints the burstiness of the traffic in a particular time scale. The time scale of burstiness that contributes to buffer overflow is determined by the index k which achieves the minimum in (6).

If t=1, then the bound (7) reduces to

$$\tilde{\alpha}_{\mathrm{pm}}(m, \mathbf{h}) = \frac{1}{s} \log \left[1 + \frac{m}{h} (e^{sh} - 1) \right] \,, \tag{8}$$

which is appropriate when the buffers are small and the argument minimizing expression (6) corresponds to the peak rate h. We refer to this as the "peak/mean" bound. Charges based on this bound have been considered in (Kelly 1994).

In many cases (Courcoubetis, Kelly and Weber 1997), the worst case traffic (for given values of s, t) consists of blocks of an inverted T pattern repeating periodically or with random gaps. In this paper we consider the periodic pattern shown in Figure 1, which gives the following effective bandwidth approximation (referred to as the "inverted T" approximation):

$$\tilde{\alpha}_{\perp}(m, \mathbf{h}) = \frac{1}{st} \log E \left[e^{sX_{\perp}[0,t]} \right] \,, \tag{9}$$

where $X_{\perp}[0,t]$ denotes the amount of load produced by the inverted T pattern in a time interval of length t. The expected value in the right-hand side of (9) can be computed analytically.

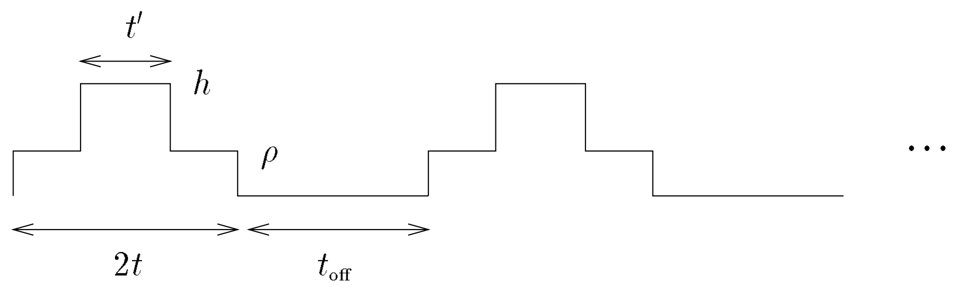


Figure 1 Periodic pattern for the inverted T approximation. $t' = \frac{\beta}{h-\rho}$, $t_{\mathrm{off}} = \frac{(2t-t')\rho + t'h}{m} - 2t$

3 EVALUATING THE CHARGING SCHEME

In Section 2.3 we introduced the class of tariffs $f(m, \mathbf{h}; X) = \bar{\alpha}(m, \mathbf{h}) + \lambda_m(g(X) - m)$, where $\mathbf{h}$ are the policing constraints in the traffic contract, $g(x)$ is the measured mean rate of the connection, and m is the anticipated value of this mean rate by the customer. For simplicity we assume that the customer knows his mean rate, hence his charge will be equal to $\bar{\alpha}(m, \mathbf{h})$, which can be approximated by (7), (8), or (9). In this section, we evaluate the performance of these approximations.

One important criterion for a pricing scheme, which is based on some approximation $\tilde{\alpha}$ of the bound $\bar{\alpha}$, is *fairness*. Ideally we would like the relative charges using $\tilde{\alpha}$ to be as close as possible to those using the actual effective bandwidth α. Hence, if (with a slight abuse of notation) we denote by $\tilde{\alpha}(x)$ and $\alpha(x)$ the corresponding charges for a connection x, then we would like to have $\tilde{\alpha}(y)/\tilde{\alpha}(x) \approx \alpha(y)/\alpha(x)$, for any two connections x, y. A reasonable measure of the *unfairness* of an approximation for a set of connections is the standard deviation of $\tilde{\alpha}(x)/(\mu\alpha(x))$, where μ is the average of $\tilde{\alpha}(x)/\alpha(x)$ as x ranges over the connection set. We will refer to this as the *unfairness index* $\mathcal{U}$. For example, an approximation that consistently overestimates the true effective bandwidth by some constant will have $\mathcal{U} = 0$, hence would be preferable than some other approximation which, on the average, is closer to the true effective bandwidth, but whose ratio $\tilde{\alpha}(x)/\alpha(x)$ varies (hence $\mathcal{U} > 0$).

We have done extensive experimentation involving the three approximations introduced in Section 2.3, with different types of traffic (e.g., MPEG video). In this paper we consider the case of Internet Wide Area Network (WAN) traffic using the Bellcore Ethernet trace BC-Oct89Ext[§] (Leland and Wilson 1991), which has a duration of 122797 seconds. We assume that a customer is policed by two leaky buckets $\mathbf{h} = \{(h, 0), (\rho, \beta)\}$, and initially assume that traffic is shaped in a 200 ms buffer. This reduces the peak rate to $h =$ 0.88 Mbps. The pairs (ρ, β) for which no traffic is discarded by the policer corresponds to the indifference curve G (Figure 2). Finally, we assume that all users are "rational", i.e., they select the pair (ρ, β) that minimizes their charge.

From the initial Bellcore trace we created a set of 15 non-overlapping trace segments, each with duration 8186 seconds (approximately 2.5 hours). For this set, we wish to compare the three different charging schemes based on approximations (7), (8), and (9) according to the unfairness index $\mathcal{U}$ defined above.

As discussed in Section 2.1, the parameters s, t characterize the link's operating point. We consider a link with capacity $C = 34$ Mbps and a target overflow probability equal to 10^{-6}, and use equations (1) and (3) to compute "typical" values of s, t, where the expectation in (1) is replaced by the empirical mean which is computed from the trace.

Figure 3 shows that the unfairness for the simple bound and inverted `T` approximations is close, and much smaller than that for the peak/mean bound. Furthermore, while the unfairness for the former two approximations decreases when the buffer size increases, this is not the case for the peak/mean bound. This is expected because the peak/mean bound becomes accurate for small values of t, which are realized for small buffer sizes.

Figure 4 shows the unfairness for the three approximations in a neighborhood of values for s, t when $B = 0.25 \times 10^6$ bytes. Observe that both the simple bound and the inverted `T` approximations are fairer and more robust (the surface is "flatter") compared to the

---

§Obtained from The Internet Traffic Archive, <http://www.acm.org/sigcomm/ITA/>.

peak/mean bound. Furthermore, increasing the link capacity and buffer size increases the fairness and robustness of the schemes.

4 INCENTIVE COMPATIBILITY

As we have already mentioned in the introduction, the operating point of the link and the posted tariffs are interrelated in a circular fashion. The network operator posts tariffs that have been computed for the current operating point of the link, expressed through the parameters s, t. These tariffs provide *incentives* to the customers to change their contracts in order to minimize their anticipated costs. Under these new contracts, the operating point of the system will move, since the network operator must guarantee the performance requirements of these new contracts. Hence, the network operator will calculate new tariffs for the new operating point. This interaction between the network and the customers will continue until an equilibrium is reached. We validate below, for a simple example, that if the network operator uses our charging approach, then an equilibrium does exist and that it is a point maximizing social welfare, as measured in this example by the number of customers admitted to the system.

For simplicity, we assume that all customers have identical profiles, are policed with a single leaky bucket (ρ, β), and have identical indifference curves $G = \beta(\rho)$. We assume that G is convex, tends to infinity when ρ goes to the mean rate m, and is zero for $\rho = h$. The network consists of a shared link with capacity C and buffer B, and uses deterministic multiplexing for loading the link.

In the case of deterministic multiplexing (zero cell loss), our effective bandwidth theory suggests that the value of the parameter s should be ∞ (this follows from (2) when $\gamma = \infty$), and that the effective bandwidth of a connection policed with (ρ, β) is $\alpha_j(\infty, t) = \bar{X}[0, t]/t = \rho_j + \frac{\beta_j}{t}$ for $t > 0$ and $\alpha_j(\infty, 0) = \beta_j$. Simple algebra shows that the acceptance region A (one-dimensional in our case) is defined by the constraints

$$\sum_j \rho_j \leq C \text{ and } \sum_j \beta_j \leq B\,, \tag{10}$$

on which the effective bandwidth is defined for $t = \infty$ and $t = 0$, respectively.

We assume that the system proceeds in lock-step and customers have identical requirements. Hence, at any point in time their choices will coincide. Due to this, the above constraints become $n\rho \leq C$ and $n\beta \leq B$, where n is the total number of customers.

Consider the point $Q \in G$ where the two constraints coincide and the number of customers n is maximized (welfare optimum). This point is also defined by the intersection of the line with slope B/C (that passes from the origin) with G. One can easily see that for any point M in G (i.e., initial choice of (ρ, β) by customers) which is below Q, the system will fill so that the active constraint will have a corresponding value $t = \infty$ for the calculation of the effective bandwidth, whereas if M lies above Q, then $t = 0$.

Assume now that our charging approach is used by the network. If the customers choose a point M below Q, then the first constraint will be active ($t = \infty$) and the charge will be proportional to ρ; this will guide customers to reduce ρ and move towards Q. If the customers choose a point M above Q, then the second constraint will be active and the charge will be proportional to β; this will guide customers to reduce β and move towards

	Deterministic mult.			Statistical mult.		
B (bytes)	ρ (Mbps)	β (bytes)	$n_{\max}$	ρ (Mbps)	β (bytes)	$n_{\max}$
0.5×10^6	0.615	10600	33	0.475	29100	1530
1×10^6	0.553	18300	54	0.399	52800	1650
5×10^6	0.373	62500	80	0.202	175500	2070
10×10^6	0.285	95500	105	0.162	341100	2170

Table 1 Equilibrium under deterministic and statistical multiplexing.

Q. Assuming that, in order to avoid oscillations, customers are allowed to make small changes to their traffic contracts, the point Q will be eventually reached. At Q, since both constraints are active, the charge will be proportional to a linear combination $\lambda_1\rho + \lambda_2\beta$ of the effective bandwidths corresponding to the active constraints at Q (i.e., both ρ and β), where λ_1, λ_2 are the shadow prices of the optimization problem which maximizes the number of users under constraints (10). One can check that the above charges correspond to the tangent of G at Q, hence Q is an equilibrium since the user minimizes his charge by remaining there.

In the case of statistical multiplexing, the above arguments can be extended to show a similar user-network behavior. We have calculated such equilibria for a range of buffer sizes and for a target overflow probability 10^{-6} (Table 1). As expected, the utilization in the case of statistical multiplexing is much higher than in the case of deterministic multiplexing.

Effects of traffic shaping

We are now in a position to make some interesting observations about the effect of customers delaying their traffic into the network. As we will argue, for the anticipated buffer sizes, shaping has a surprisingly small effect on the overall multiplexing capability of the network.

First, observe in Figure 2 that for large values of β, the indifference curve $G(d)$ is not greatly affected when the shaping delay is smaller than 500 msec . Second, observe in Table 1 that in the case of statistical multiplexing and for buffer sizes greater than 1×10^6 bytes, at the equilibrium we have $\beta > 50000$ bytes. Combining these two observations we see that for buffer sizes greater than 1×10^6 bytes, the equilibrium point will not be affected by traffic shaping, when the shaping delay is less than 500 msec .

Of course a customer can use shaping to make a contract with a lower peak rate. However, contrary to the intuition, this will not affect his effective bandwidth as seen by the network, since the time parameter t at the equilibrium is always large enough so that $ht > \rho t + \beta$. In this case, the effective bandwidth is determined largely in terms of the values (ρ, β) (e.g., if the customer sends traffic close to the maximum amount allowed by the simple bound (7)) which, as argued previously, remain practically unaffected by shaping.

The above discussion demonstrates how the theory described in Section 2 clarifies the effects of various time scales and the importance of the various traffic and network parameters on the amount of resources used by connections.

5 CONCLUSIONS AND OPEN QUESTIONS

This paper has dealt with one important part of the charging activity: the part which aims to access a connection's network resource usage. In this direction, we have provided a framework for constructing incentive compatible charges that reflect effective resource usage. Our charging schemes are based on bounds on the effective bandwidth and involve only measurements of the duration and volume of connections. The schemes are simple in the sense that they are easily understood by the customers. Furthermore, they can be cast in the same formats that are used today, namely, charges depend on static contract parameters (e.g., access line speed, leaky bucket policing parameters, anticipated average rate), and on dynamic parameters of a connection (e.g, actual average rate). We have displayed the incentive compatibility of the proposed schemes through an example involving deterministic multiplexing, and have presented numerical results, with real broadband traffic, that display the fairness of the schemes. It is important to note that our approach is quite general and can be used to charge for effective usage at many levels of network access, ranging from individual users to large organizations. It can be applied to any packet switching technology and can be used under both deterministic and statistical multiplexing.

The extension of our approach to networks consisting of more than one link raises several further issues which we hope to treat in the future. Important choices concern whether a user sees a single charge from its immediate service provider, or whether a user might see several charges arising from various intermediate networks. We simply note here that charges linear in time and volume remain so under aggregation.

REFERENCES

Cocchi, R., Shenker, S., Estrin, D. and Zhang, L. (1993) Pricing in computer networks: Motivation, formulation, and examples. *IEEE/ACM Trans. on Networking*, **1**, 614–627.

Courcoubetis, C., Kelly, F. P. and Weber, R. (1997) Measurement-based charging in communications networks. Technical Report 1997-19, Statistical Laboratory, University of Cambridge.

Courcoubetis, C., Siris, V. A. and Stamoulis, G. D. (1997) Many sources asymptotic and effective bandwidths: Investigation with MPEG traffic. Presented at the *2nd IFIP workshop on traffic management and synthesis of ATM networks*, Montreal, Canada, September 1997. Extended version submitted for publication.

Courcoubetis, C. and Weber, R. (1996) Buffer overflow asymptotics for a switch handling many traffic sources. *Journal of Applied Probability*, **33**, 886-903.

Gupta, A., Stahl, D. O. and Whinston, A. B. (1994) Managing the Internet as an economical system. Technical report, University of Texas, Austin.

Kelly, F. P. (1994) On tariffs, policing and admission control for multiservice networks. *Operations Research Letters*, **15**, 1–9.

Kelly, F. P. (1996) Notes on effective bandwidths. In F. P. Kelly, S. Zachary and I. Zeidins, eds., *Stochastic Networks: Theory and Applications*, pp. 141–168. Oxford University Press.

Leland, W. E. and Wilson, D. V. (1991) High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection. In *Proc. of IEEE INFOCOM'91*,

pp. 1360–1366.
Low, S. H. and Varaiya, P. P. (1993) A new approach to service provisioning in ATM networks. *IEEE/ACM Trans. on Networking*, **1**, 547–553.
Mackie-Mason, J. K. and Varian, H. R. (1995) Pricing congestible network resources. *IEEE J. Select. Areas in Commun.*, **13**, 1141–1149.
Mackie-Mason, J. K. and Varian, H. R. (1995) Pricing the Internet. In B. Kahin and J. Keller, eds., *Public Access to the Internet.* Prentice Hall, Englewood Cliffs, NJ.
Parris, C., Keshav, S. and Ferrari, D. (1992) A framework for the study of pricing in integrated networks. Technical Report TR-92-016, International Computer Science Institute, Berkeley, CA.
Sairamesh, J., Ferguson, D. F. and Yemini, Y. (1995) An approach to pricing, optimal allocation and quality of service provisioning in high-speed packet networks. In *Proc. of IEEE INFOCOM'95.*
Shenker, S., Clark, D., Estrin, D. and Herzog, S. (1996) Pricing in computer networks: Reshaping the research agenda. *ACM Computer Communication Review*, **26(2)**, 19–43.
Wang, Q., Peha, J. M. and Sirbu, M. A. (1996) The design of an optimal pricing scheme for ATM integrated-services networks. In J. P. Bailey and L. Mcknight, eds., *Internet Economics*, Massachusetts, 1996. MIT Press.

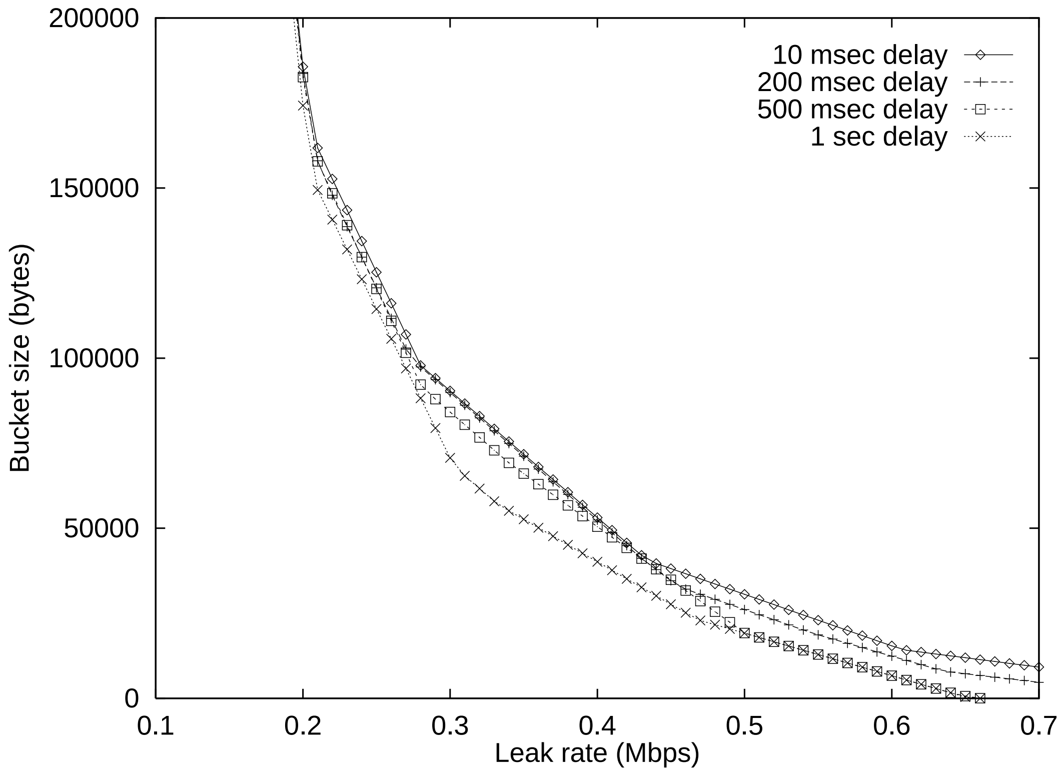
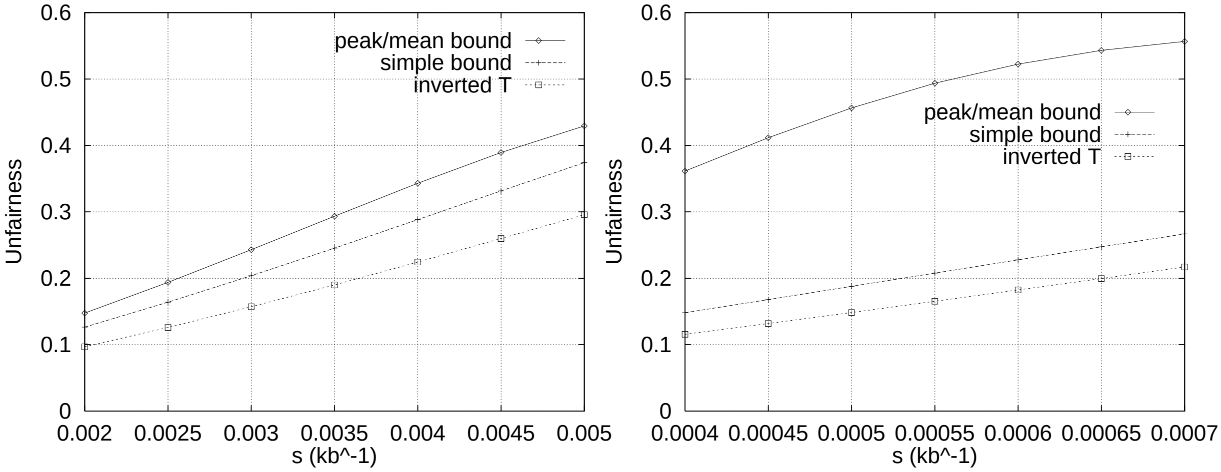
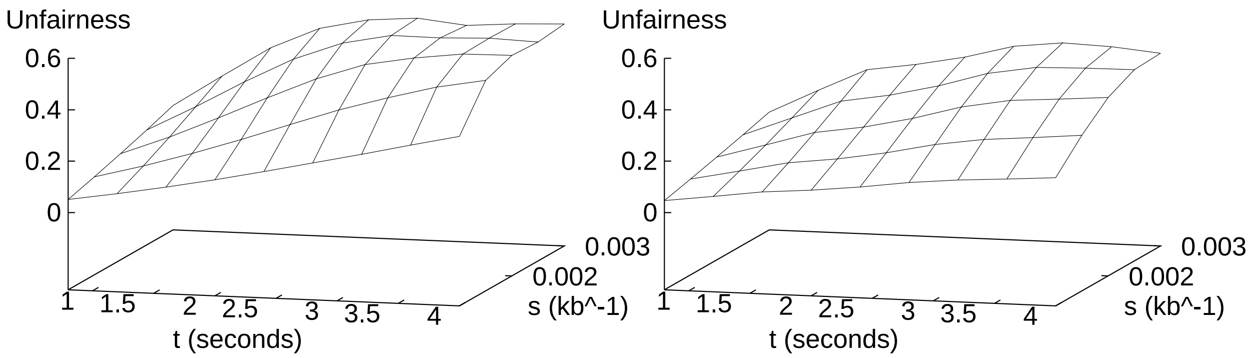


Figure 2 Indifference curve $G(d)$ for the Bellcore trace.

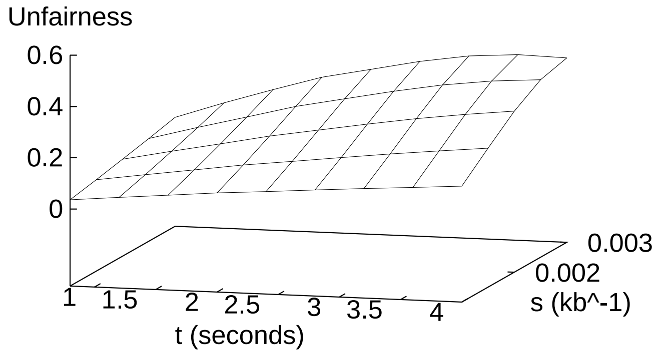


(a) $B = 0.25 \times 10^6$ bytes (b) $B = 1 \times 10^6$ bytes

Figure 3 Unfairness for $C = 34$ Mbps and two buffer sizes.



(a) peak/mean bound (b) simple bound



(c) inverted `T`

Figure 4 Unfairness for $C = 34$ Mbps, $B = 0.25 \times 10^6$ bytes.