*Telecommunication Systems,* 15(3-4):323-343, 2000

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

Costas Courcoubetis [a,b], Frank P. Kelly [c], Vasilios A. Siris [a] and Richard Weber [c]

[a] *Institute of Computer Science (ICS)*
*Foundation for Research and Technology - Hellas (FORTH)*
*P.O. Box 1385, GR 711 10 Heraklion, Crete, Greece*
E-mail: {courcou, vsiris}@ics.forth.gr
[b] *Athens University of Economics and Business, Athens, Greece*
[c] *University of Cambridge, Statistical Laboratory*
*16 Mill Lane, Cambridge CB2 1SB, U.K.*
E-mail: {f.p.kelly, r.r.weber}@statslab.cam.ac.uk

Operators of multi-service networks require simple charging schemes with which they can fairly recover costs from their users and effectively allocate network resources. This paper studies an approach for computing such charges from simple measurements (time and volume), and relating these to bounds of the effective bandwidth. To achieve economic efficiency, it is necessary that usage-based charging schemes capture the relative amount of resources used by connections. Based on this criteria, we evaluate our approach for real traffic consisting of Internet Wide Area Network traces and MPEG-1 compressed video. Its incentive compatibility is shown with an example involving deterministic multiplexing, and the effect of pricing on a network's equilibrium is investigated for deterministic and statistical multiplexing. Finally, we investigate the incentives for traffic shaping provided by the approach.

**Keywords:** usage-based charging, effective bandwidths, statistical multiplexing, incentives, broadband networks

## 1. Introduction

A method for charging and pricing is an essential requirement in operating a broadband network. Pricing is not only needed for recovering costs, but is also 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 and share network resources amongst users in a workable and stable way [13,12,7]. Furthermore,

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

in a competitive environment providers will need to price services in a manner that takes some account of network resource usage [14,2]. 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 for a given amount of network resources). 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 the user and the network operator negotiate a traffic contract. Under this contract, the user agrees that his traffic will conform to certain parameters (e.g., that bound his peak rate and the size of his bursts), while the operator guarantees a particular quality of service (expressed, for example, 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: the part that aims to assess the relative amount of resources used by a connection. We shall henceforth simply refer to this component as *charging* and of computing a *charge*.

### *1.1. 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 user reacts to tariffs and seeks to minimize his charges. For example, tariffs may make it economical for some users to shape their traffic, which would result in an increase of overall network performance. This is the key idea of *incentive compatibility*. Tariffs should guide the population of cost-minimizing users to select contracts and use the network in ways that are good for overall network performance, e.g., to maximize social welfare [11]. Tariffs that are not incentive compatible give the wrong signals and lead users to use the network in inefficient ways.

Well-designed tariffs should also be *fair*[1], i.e., charges should reflect a user's *relative* resource usage. As we will discuss, this is required for achieving economically efficient allocation of network resources. This point 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 resource usage.

The above remarks lead one to ask whether it possible to design tariffs that are sound, both in terms of incentive compatibility and fairness, but that are also not too complex, and their implementation does not require the network

[1] In the case of differential pricing and/or time-of-day pricing, fairness is considered for users of the same "class" that use network services within the same time period.

operator to make overly sophisticated or unrealistic measurements. Incentive compatibility will be hard to achieve if tariffs are too complex, since users will find it difficult to determine the effect the decisions under their control, such as traffic shaping, have on the charges they incur.

### *1.2. Contribution of the paper and related work*

In this paper we study a usage-based charging scheme that is a special case of the general model introduced in [3], where charges are linear functions of measurements of time and volume. Our consideration of such a charging scheme is motivated first by its technological feasibility, since current technology can readily support measurements of time and volume per connection or flow, and second for its simplicity for the users, since charges are simple linear functions of quantities they can easily understand. Consideration of more complex measurements would increase the performance at the expense of a large increase of implementation costs and complexity. Furthermore, the introduction of usage-based charging is itself a debatable issue, and in this paper we argue that effective charging schemes that require only two simple measurements (time and volume) can be created.

Our approach is based on the notion of effective bandwidth as a proxy for resource usage [9]. Both theory [6,9,3] and experimentation [5] has shown that a connection's resource usage depends on its context, i.e., the link resources and the composition and characteristics of the other traffic it is multiplexed with. In the effective bandwidth definition we consider, this dependence is only through a pair of parameters, the space and time parameters. The approach for creating usage-based charging schemes, whose mathematical foundation is developed in [3], transforms simple tariffs of the form $a_0T + a_1V$ into sound approximations of the effective bandwidth. The variables $T$ (duration) and $V$ (volume) are *dynamic* variables that are measured a posteriori, while the coefficients $a_0$ and $a_1$ are *static* parameters that depend on the traffic contract parameters and the operating point of the network. An important property of the approach is that it treats deterministic and statistical multiplexing in a unifying way.

We study the above charging approach for real broadband traffic consisting of Internet Wide Area Network traces and MPEG-1 compressed video, and for capacity and buffer sizes that will appear in broadband networks. In particular, we investigate the fairness of the charging approach, i.e., its ability to capture the relative amount of resources used by connections, and its incentive compatibility, i.e., the incentives it provides for guiding the system comprising of the network and its users to a stable and economically efficient operating point. Based on the conclusions of these investigations, we believe that our simple tariffs can serve their purpose well.

Our focus is on simple charging schemes that can accurately reflect the relative amount of resources used by connections. In this sense our work differs from [11] and [17], which investigate optimal pricing strategies assuming that

network resources (buffer and capacity) are charged separately, and [20], which also deals with optimal pricing but does not relate charges to the amount of resources a connection uses. Our approach can be applied to proposals such as expected capacity [1] and edge pricing [18], which address architectural issues of pricing network services, or can be complemented with approaches such as time-of-day pricing.

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 creating charging schemes that are linear in measurements of time and volume. In Section 3 we discuss the fairness of charging schemes, based on which we evaluate our approach for real broadband 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. We also investigate the effects of statistical multiplexing on the operating point of the network. In Section 5 we discuss the incentives for traffic shaping when our charging approach is used. Finally, in Section 6 we conclude the paper and identify some open issues.

## 2. A theory for usage-based charging

### *2.1. Effective bandwidths as a measure of resource usage*

Suppose the arrival process at a link is the superposition of independent sources of $J$ types: let $n_j$ be the number of sources 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 $t$ epochs. 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 [9]

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

where $s, t$ are system defined parameters that 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 [6,22]. The *space* parameter $s$ (measured in, e.g., $\text{kb}^{-1}$) 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 sources, 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 the half-space [9]

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

where $(s,t)$ is an extremizing pair in the following equation, called the *many sources asymptotic* ([6], also see [21] for the extension to networks):

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

The asymptotics behind this approximation assumes only stationarity of sources, and illustrative examples discussed in [9] and [16] include periodic streams, fractional Brownian input, policed and shaped sources, and deterministic multiplexing. Equation (2) expresses the local (at $s,t$) condition such that the QoS guarantee is met. The latter is encoded in the effective bandwidth definition through the value of $\gamma$ which influences the form of the acceptance region. Furthermore, investigations with real traffic [5] have shown that the above definition of the effective bandwidth can accurately quantify resource usage.

The above considered the buffer overflow probability as the measure of QoS. However, we could have also considered a bound on the maximum delay as the measure of QoS. Furthermore, we considered the case of a single First-Come-First-Served (FCFS) queue. The results can be extended to priority queueing [9], where each service class has its own pair of operating point parameters $s,t$.

We now stress the network engineering implications of the above results. For any given traffic stream, the effective bandwidth definition (1) is a measure of resource usage that depends on the link's operating point[2] through only two parameters $s,t$. Experimentation has revealed that the values of $s,t$ and the effective bandwidth are to a large extent insensitive to small variations of the traffic mix (percentage of different traffic types) [5]. Hence, particular pairs $s,t$ can be assigned to periods of the day during which the traffic mix remains relatively constant. These values can be computed off-line using (1) and (3),

[2] The operating point of a link is characterized by the combination and characteristics of the multiplexed traffic.

where the expectation in (1) is replaced by the empirical mean, which is computed from traffic traces taken for the particular period of the day.

### *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 (1), as determined by past connections of type $j$. A new connection of type $j$ would be charged at an amount per unit time proportional to $\alpha_j^*(s,t)$. This is the charging method adopted in an all-you-can-eat restaurant. At such a restaurant each user is charged not for his own food consumption, but rather for the average amount that similar users have eaten in the past. Under such a charging scheme, each user 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. Users 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 providers using a different (more competitive) charging method.

Similar incentive problems exist with a scheme where charges are based solely on the traffic contract parameters. In this case charges are proportional to the worst case traffic for a given traffic contract. However, because traffic contracts can be overly conservative, such a scheme will be unfair for users that use less resources than the maximum allowed by their contract.

At another extreme, one might charge a user 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. Incentive compatibility will be hard to achieve with such a complex scheme, since users will find it difficult to determine how the decisions under their control, such as traffic shaping, affect their charges. Furthermore, there is a conceptual flaw with the above approach that can be illustrated as follows. Suppose a user 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 user 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 extremes described above. We construct a charge that is based on the effective bandwidth, but which is a func-

tion of both *static* parameters which are part of the traffic contract (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 schemes. This leads to a charge with the right incentives for users, which also compensates the network operator for the amount of resources reserved. In the next section we describe our approach in detail.

### *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[3]

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

where $g(X)$ is the measured mean rate $(1/T)\sum_{i=1}^{T} X_i$ . 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. Consideration of such bounds is partly motivated by the remark that this is what we would charge to a user who makes maximal use of his traffic contract.

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 $m$. 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 $a_0[m, \mathbf{h}] = \bar{\alpha}(m, \mathbf{h}) - \lambda_m m$, $a_1[m, \mathbf{h}] = \lambda_m = \frac{\partial}{\partial m}\bar{\alpha}(m, \mathbf{h})$.

Our charging scheme works as follows. At connection setup and given his traffic contract $\mathbf{h}$ (chosen by the user), the user is offered a set of possible tariff

[3] As discussed in the introduction, we are concerned with the relative resource usage that should be attributed to each connection. Transformation of such relative charges to charges expressed in monetary units would be done using a multiplicative constant that depends on economic factors such as demand and competition.

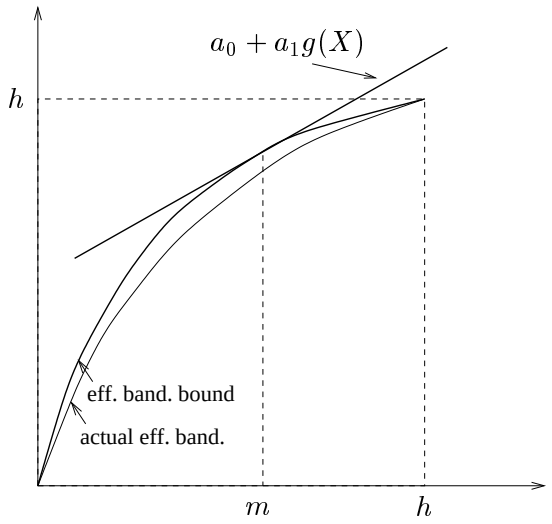



Figure 1. A user is charged according to a tangent to the effective bandwidth bound. Due to the concavity of the bound, a user's charge is minimized and becomes proportional to the bound if the user selects the tangent which corresponds to the a priori estimate of his mean rate.

pairs $(a_0, a_1)$ to choose from. These pairs correspond to tangents of the effective bandwidth bound $\bar{\alpha}(m, \mathbf{h})$ for different mean rates. Selection of a particular tariff pair $a_0, a_1$ defines the user's charging rate $f(X) = a_0 + a_1 g(X)$, where $g(X)$ is the user's mean rate (measured by the network).

By considering the concavity of $\bar{\alpha}(m, \mathbf{h})$ in $m$ [3], 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), Figure 1. Hence, a rational user that chooses the tariffs that minimize his charge will end up being charged in proportion to the maximal effective bandwidth that his connection could have, given all the available information at connection setup. Thus, the scheme offers the incentive for users to estimate their mean rates as accurately as possible, and reveal this estimate to the provider through their tariff selection. Furthermore, in addition to the static parameters of his traffic contract $\mathbf{h}$, the user's charge also depends on his actual mean rate $g(X)$; this provides the right incentives for avoiding the "all-you-can-eat restaurant" effect discussed in Section 2.2.

We note that although the tariff pairs $(a_0, a_1)$ are computed using sophisticated techniques (effective bandwidths and the link operating point parameters $s, t$), these are hidden from the user. The user is only required to select a tariff pair from a table.

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

In this section we consider approximations for $\bar{\alpha}(m, \mathbf{h})$, since $\bar{\alpha}(m, \mathbf{h})$ can be difficult to calculate and its value depends upon the operating point of the link.

We start with a simple approximation that shows the relation of the various time scales to buffer overflow. Suppose that a connection is policed by multiple leaky buckets with parameters $(\rho_k, \beta_k)$ for $k = 1, \ldots, K$ and let $m$ be the mean rate of the connection. The maximum amount of traffic $\bar{X}[0,t]$ produced in a time interval of length $t$ is

$$\bar{X}[0,t] \leq H(t) := \min_{k=1,\ldots,K} \{\rho_k t + \beta_k\}\,. \tag{6}$$

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

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

We call the right hand side of the last equation the *simple bound*. This equation is illuminating for the effects of leaky buckets on the amount of resource usage. Each leaky bucket $(\rho_k, \beta_k)$ constrains 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 $k$ that achieves the minimum in (6).

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

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

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 *on-off bound*. Charges based on this bound have been considered in [8].

In many cases [3], the worst case traffic consists of blocks of an inverted `T` pattern repeating periodically or with random gaps, with the size of the blocks and gaps depending on the values of $s, t$. In this paper we consider the periodic pattern shown in Figure 2, which gives the following approximation for the effective bandwidth bound, referred to as the *inverted `T` approximation*:

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

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

## 3. Performance evaluation

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 of the traffic contract, $g(x)$ is the measured mean rate of the connection, and $m$ is the user's anticipated value of this mean rate. The bound $\bar{\alpha}(m, \mathbf{h})$ can be approximated by (7), (8), or (9). In this section we evaluate the performance, in terms of fairness, of

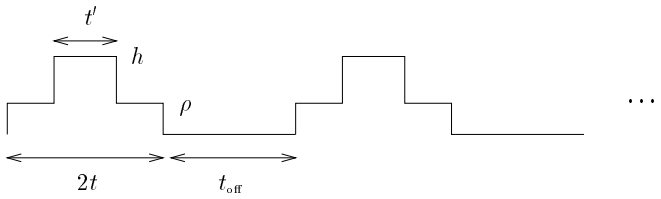



Figure 2. Periodic pattern for the inverted T approximation. $t' = \frac{\beta}{h \perp \rho}$, $t_{\text{off}} = \frac{(2t \perp t')\rho + t'h}{m} - 2t$

these approximations for real broadband traffic consisting of Internet Wide Area Network traces and MPEG-1 compressed video. We assume that a user knows his mean rate, hence his charge will be equal to the value of the approximation $\tilde{\alpha}(m, \mathbf{h})$ considered.

### *3.1. Fairness of charging schemes*

We first argue that charging in proportion to the actual effective bandwidth is required to achieve economic efficiency. Indeed, consider two types of connections, $A$ and $B$, and let $u_A(n_A)$ and $u_B(n_B)$ be the utility of accepting $n_A$ and $n_B$ connections of type $A$ and $B$, respectively. The constraint on the number of connections that can be accepted, while guaranteeing a target QoS, has the form of (2): $n_A\alpha_A + n_B\alpha_B \leq K$, for some value of $K$. The social welfare $u_A(n_A) + u_B(n_B)$ is maximized when $n_A, n_B$ are chosen in a decentralized way by imposing usage prices $\lambda\alpha_A, \lambda\alpha_B$ for connection types $A, B$ respectively, with $\lambda$ being the shadow price for the constraint $n_A\alpha_A + n_B\alpha_B \leq K$. On the other hand, if charges are not proportional to the effective bandwidth, then the optimal is not achieved. As an example, consider the case where both connection types $A$ and $B$ are charged at rate $g$. Then the number of connections of each type will be $n_A$ and $n_B$, with $u'_A(n_A) = u'_B(n_B) = g$. For such a traffic mix let $\alpha_A < \alpha_B$. Now assume that connections of type $B$ are charged slightly more than those of type $A$. The number of connections of type $A$ now becomes $n_A + \kappa$, while the number of connections of type $B$ becomes $n_A - \kappa\alpha_A/\alpha_B$, for some small value $\kappa$. Note that the new source combination also satisfies the acceptance constraint. Furthermore, the new combination results in an increase of the total utility by $\kappa(u'_A(n_A) - \alpha_A/\alpha_B u'_B(n_B)) = \kappa g(1 - \alpha_A/\alpha_B) > 0$, since $\alpha_A < \alpha_B$. The above example shows that charging both connection types the same does not achieve the optimal welfare.

An approximation $\tilde{\alpha}(m, \mathbf{h})$ of the actual effective bandwidth $\alpha(m, \mathbf{h}) = \alpha(s, t)$, given by equation (1), is defined as *fair* if the variance of the ratio $\tilde{\alpha}(m, \mathbf{h})/\alpha(m, \mathbf{h})$ is small, when $m, \mathbf{h}$ range over some interesting set of values. This implies that for this set we have $\tilde{\alpha}(m, \mathbf{h})/\alpha(m, \mathbf{h}) \approx k$, for some constant $k$, and pricing in proportion to $\tilde{\alpha}(m, \mathbf{h})$ is equivalent to pricing in proportion to $\alpha(m, \mathbf{h})$. Hence, pricing in proportion to a proxy $\tilde{\alpha}(m, \mathbf{h})$ that is fair can achieve economic efficiency. A reasonable measure of the *unfairness* of an approximation

for a set of connections is the standard deviation of $\tilde{\alpha}(m,\mathbf{h})/(\mu\alpha(m,\mathbf{h}))$, where $\mu$ is the average of $\tilde{\alpha}(m,\mathbf{h})/\alpha(m,\mathbf{h})$, as $m,\mathbf{h}$ take values corresponding to 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 multiplicative factor will have $\mathcal{U}=0$, hence would be preferred over some other approximation that, on the average, is closer to the true effective bandwidth, but whose ratio $\tilde{\alpha}(m,\mathbf{h})/\alpha(m,\mathbf{h})$ varies, hence $\mathcal{U}>0$.

Whereas fairness, or the *variability* of the approximation error, is important for achieving economic efficiency, the *absolute* approximation error of $\tilde{\alpha}(m,\mathbf{h})$, when that latter is used for acceptance control, is important for optimizing the use of resources, where optimality here refers to the maximal utilization of resources. In this paper we consider only the former, since our focus is on creating usage-based charging schemes that are incentive compatible and guide the network to an economically efficient operating point.

### *3.2. Experiments with real traffic*

Our experiments involved real broadband traffic, namely Internet Wide Area Network traces and MPEG-1 video. For the former we used the Bellcore Ethernet trace BC-Oct89Ext[4] [10], which has a total duration of approximately 34 hours. From the initial Bellcore trace we created a set of 17 non-overlapping trace segments, each with a duration of approximately 116 minutes. Our model consists of a link with capacity $C$ and buffer $B$ that multiplexes a number of connections. The traffic of each connection is given by one trace segment. Hence there are a total of 17 different connection types. Furthermore, for each connection the start frame within the trace segment is randomly chosen. The effective bandwidth for each such connection is estimated from (1) with the expectation replaced by the empirical mean. Hence, if $T$ is the duration of a trace segment $j$ then the effective bandwidth $a_j$ for a connection whose traffic is given by this segment is computed from

$$\alpha_j = \frac{1}{st}\log\left[\frac{1}{T/t}\sum_{i=1}^{T/t} e^{sX_j[(i-1)t,it]}\right], \tag{10}$$

where $X_j[(i-1)t,it]$ is the load produced in the interval $[(i-1)t,it)$. For the parameters $s,t$ we consider "typical" values that correspond to a target overflow probability equal to $10^{-6}$, link capacities 34, 155, and 622 Mbps, and buffer sizes that are anticipated for broadband networks. Such typical values[5] are computed using equations (1), (3), and (10). In the experiments with Internet traffic that follow, the values of $s,t$ were obtained for a traffic mix containing the same num-

[4] Obtained from The Internet Traffic Archive, http://www.acm.org/sigcomm/ITA/

[5] Typical values of parameters $s,t$ for Internet WAN traffic and other traffic types and mixes are available at http://www.ics.forth.gr/netgroup/msa

ber of connections for each connection type, where the traffic for each connection type is given by one of the 17 trace segments.

We assume that connections are policed by two leaky buckets $\mathbf{h} = \{(h, 0), (\rho, \beta)\}$, and that the traffic is shaped in a buffer of size $d$ (measured in msec). Traffic shaping is performed by averaging the amount of traffic in windows of length $d$.[6] For the experiments with Internet traffic, we assume that traffic is shaped with shaping delay $d = 20$ msec. The pairs $(\rho, \beta)$ for which all the traffic is conforming form an *indifference curve* $G$ (Figure 3). Finally, we assume that all users know their mean rates and are "rational", i.e., they select the pair $(\rho, \beta)$ that minimizes their charge.

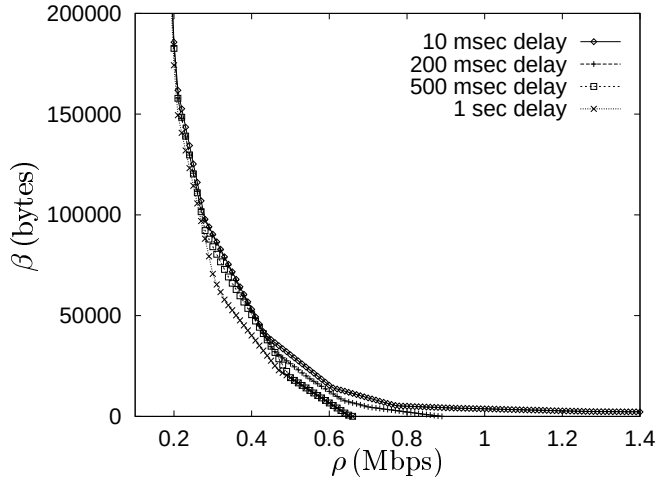


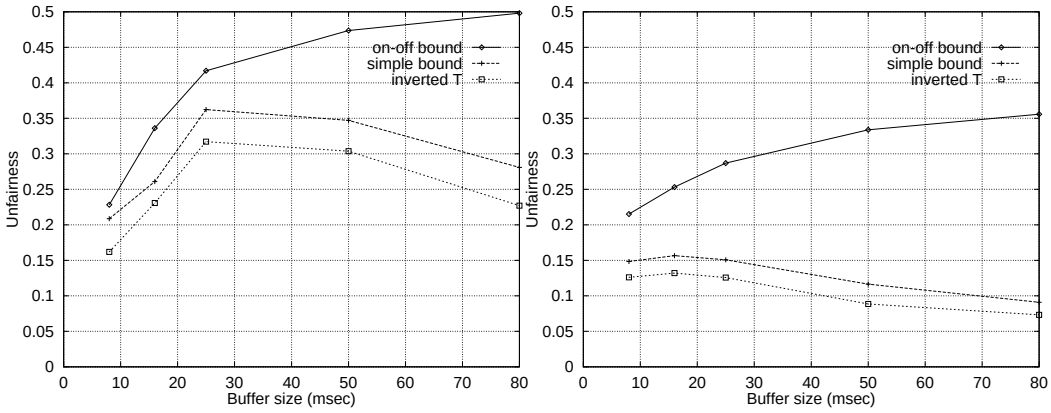

Figure 3. Indifference curve $G(d)$ for various shaping delays $d$. [ Bellcore Internet WAN traffic ]

The experimental results that we present next were obtained as follows. For each connection type, we compute the effective bandwidth using (10) and the effective bandwidth approximations (7), (8), and (9). Having done this for all 17 connection types, we then compute, for each approximation, the unfairness index $\mathcal{U}$ as defined in Section 3.1.

Figures 4(a) and 4(b) show, for two link capacities, the dependence of the unfairness index on the buffer size. We can make the following observations:

- The unfairness of the simple bound and inverted `T` approximation is close, with the inverted `T` approximation being strictly fairer. This result holds for other capacities and buffer sizes we have investigated, and occurs because the inverted `T` approximation is tighter than the simple bound, which for the

[6] This is *one* way for performing traffic shaping; we are not assuming that it is the best. Furthermore, the value $d$ represents an upper bound on the maximum packet delay. For example, for $d = 200$ msec, the actual maximum packet delay is 45 msec, whereas the average delay is less than 1 msec.




(a) $C = 34$ Mbps (b) $C = 155$ Mbps

Figure 4. Unfairness of the three effective bandwidth bounds for Internet WAN traffic.

traffic used in our experiments leads to a less variable approximation error, i.e., ratio $\tilde{\alpha}(m, \mathbf{h})/\alpha(m, \mathbf{h})$. The inverted `T` approximation (9) is tighter than the simple bound (7) due to the convexity of the exponential and the constraint $X_-(t) \leq H(t)$, where $H(t)$ is defined in (6), with the strict inequality occuring for some intervals of length $t$.

- Whereas for small buffer sizes the on-off bound is fair, for large buffer sizes its unfairness increases. This can be explained as follows: For small buffer sizes the value of $t$ is small and the minimization in (6) occurs for $k = 1$, i.e, $H(t)$ is given by the peak rate $\rho_0 = h$. Due to this, the simple bound (7) coincides with the on-off bound (8). On the other hand, for large buffer sizes the value of $t$ is large and the minimization in (6) occurs for $k = 2$, i.e, $H(t)$ is determined by the leaky bucket parameters and is independent of the peak rate. As a result, for large buffer sizes the on-off bound has an approximation error with higher variability.
- The unfairness of the simple bound and inverted `T` approximation initially increases as the buffer size increases. However, this occurs up to some buffer size, after which the unfairness decreases as the buffer size increases. This behavior can be explained as follows: For small buffer sizes, parameter $s$ tends to be large. For large values of $s$, the effective bandwidth (1) approaches the peak rate measured over an interval $t$. In such a case both the simple bound and inverted `T` approximation are tight and have an approximation error with small variability. On the other hand, for large buffer sizes, $s$ tends to be small and $t$ tends to be large. In the case of small values of $s$, the effective bandwidth is determined by the mean, the variance, and higher moments of $X[0, t]$ [9]. Because the value of $t$ is large, the variance and higher moments of $X[0, t]$ are small, hence the effective bandwidth approaches the mean. In such a case, as in the case of small buffers, both the simple bound and inverted `T` approximation

are tight and have an approximation error with small variability.

- Finally, comparison of Figures 4(a) and 4(b) shows that the unfairness of all bounds decreases when the capacity increases. This occurs because for a larger degree of multiplexing, which occurs on higher capacity links, the bounds become tighter and have a less variable approximation error.

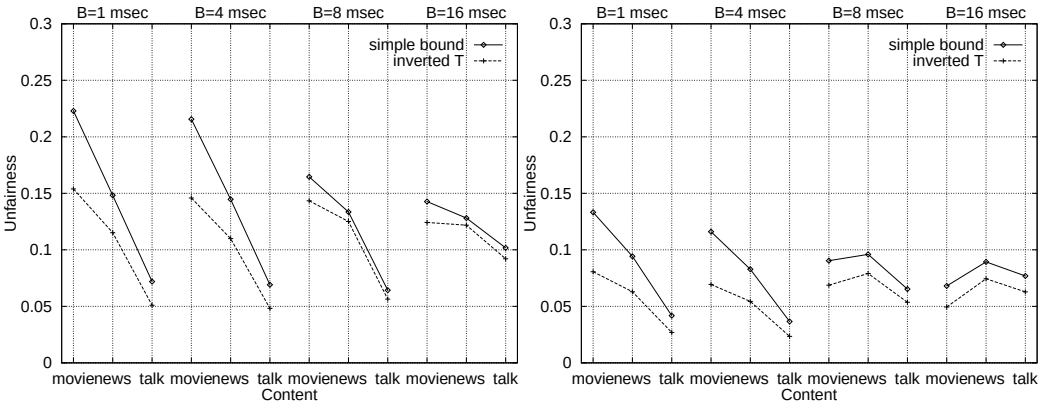
Next, we investigate the fairness of the effective bandwidth bounds for MPEG-1 compressed video with various contents. Because video will typically have stricter delay requirements than general Internet traffic, the buffer sizes we consider in the experiments that follow are smaller, hence correspond to smaller queueing delays, than those we considered previously for Internet traffic. Three sets of video traffic[7] were used in our experiments: *movies*, *news* and *talk shows*. These were created by breaking the cell streams containing the MPEG-1 traffic into non-overlapping segments, each with a duration of approximately 3 minutes (4500 frames). The resulting *movies* set contained 54 segments, the *news* set contained 16 segments, and the *talk show* set contained 18 segments. Because the on-off bound is less accurate (for large buffers) or coincides (for small buffers) with the simple bound, we do not consider it further.

The results shown in Figures 5(a) and 5(b) were obtained for values of $s, t$ when the majority of the multiplexed MPEG-1 traffic was of type movie. We can make the following observations regarding the above figures:

- As we observed for Internet WAN traffic, unfairness is not necessarily monotonous with the buffer size. Rather, there may be an initial increase of unfairness as the buffer size increases, after which unfairness decreases with increasing buffer.
- The unfairness of the two bounds depends on the traffic content. In particular, the fairness of the bounds is higher for talk shows compared to action movies. The fairness for news is somewhere in the middle. This can be explained by noting that talk show traffic is less variable than news traffic, which in turn is less variable than action movie traffic. Furthermore, the leaky bucket characterization is tighter for less variable traffic, hence the bounds we consider, which are based on the leaky bucket, produce a less variable approximation error for less variable traffic.
- As can be seen in Figures 4 (b) and 5 (a), for the same capacity and buffer ($C = 155$ Mbps and $B \approx 10$ msec), both the Internet and the MPEG-1 traffic segments we considered displayed similar unfairness.

The general conclusions of this section hold for other traffic types, traffic mixes, and overflow probabilities that we have investigated. Indeed for higher overflow probabilities, unfairness tends to be smaller. This occurs because the effective bandwidth approaches the mean rate, and all the bounds become tighter.

[7] Made available by O. Rose [15], at ftp://ftp-info3.informatik.uni-wuerzburg.de/pub/




(a) $C = 155$ Mbps (b) $C = 622$ Mbps

Figure 5. Unfairness of the simple bound and inverted `T` approximation for MPEG-1 video with various contents.

An important issue for further investigation includes the effects of uncertainty in the value of the expected mean rate.

Our focus in this section was on the fairness of charging schemes. As discussed in Section 3.1, such a property is required for achieving economic efficiency. We considered charging schemes based on the bounds (7), (8), and (9), which depend on measurements of time and volume. It was found that the inverted `T` approximation has smaller unfairness than the other two bounds. On the other hand, it is more complex than the other two bounds. Indeed, for the simple bound, if parameter $t$ is known then a user can compute the leaky bucket parameters that minimize his charge using a simple procedure [4]. Hence, for the simple bound and the inverted `T` approximation there is a tradeoff between simplicity and fairness. The on-off bound, for large buffer sizes, has higher unfairness than the other two bounds. One the other hand, the unfairness for the simple bound and inverted `T` approximation is small for very small or large buffer sizes. Finally, the unfairness of all bounds decreases with the link capacity, which results in a larger degree of multiplexing, and for smoother traffic.

How do the above schemes compare to other charging schemes, such as the ones mentioned in Section 2.2? We have found that charging based solely on traffic contract parameters has much higher unfairness than the above schemes. This is expected since such contracts, in many cases, provide a very crude and overly conservative characterization of the actual effective bandwidth. We note, nevertheless, that there can be cases where such an approach is desirable either because no traffic measurements are possible or because the traffic of users is close to the maximum allowed by their contract, e.g., see [19] which contains a comparison of the tariffing approach presented in this paper with real tariffs published by a network operator. On the other hand, charging based on the

actual effective bandwidth would be fair, but has other problems, as discussed in Section 2.2: complexity for the user and high implementation costs, in the case that the effective bandwidth is measured for each connection, or the incentive to "over-eat", in the case that users are charged based on measurements of the effective bandwidth for previous connections of the same type.

## 4. Incentive compatibility and user-network interaction

As we have mentioned in the introduction, the operating point of the link and the posted tariffs are inter-related in a circular fashion. The network operator posts tariffs that have been computed for the current operating point of the link, which corresponds to some values of the parameters $s, t$. These tariffs provide *incentives* for the users to change their contracts in order to minimize their anticipated charges. Under these new contracts, the link's operating point will move, since the network operator must guarantee the performance requirements of these new contracts. Hence, the operator will calculate new tariffs for the new operating point. This interaction between the network and the users will continue until an equilibrium is reached. Tariffs are incentive compatible, if they provide the incentives to users, which act to maximize their individual utility, so that the decentralized allocation of network resources coincides with the economically optimal allocation that would have been made centrally by the network operator, if he knew the utilities of all users.

We show below, for a simple example, that if the network operator uses our charging approach then an equilibrium exists and maximizes the social welfare, which we take to be the number of users admitted to the system. For simplicity, we assume that all users are policed by a single leaky bucket. Furthermore, we assume that the indifference curve $G$ is convex, tends to infinity when $\rho$ goes to the mean rate $m$, and is zero for $\rho = h$. The network consists of a shared link with capacity $C$ and buffer $B$.

Consider first the case of deterministic multiplexing (zero overflow probability). For deterministic multiplexing, our effective bandwidth theory suggests that $s = \infty$ (this follows from (2) when $\gamma = \infty$), and that the effective bandwidth of a connection policed with $(\rho, \beta)$ is $\alpha_j(\infty, t) = \bar{X}[0,t]/t = \rho_j + \frac{\beta_j}{t}$ for $t > 0$ and $\alpha_j(\infty, 0) = \beta_j$. The acceptance region $A$ (one-dimensional in our case) is defined by the constraints $\sum_j \rho_j \leq C$ and $\sum_j \beta_j \leq B$. For each of these constraints, the effective bandwidth is defined for $t = \infty$ and $t = 0$ and is $\rho_j$ and $\beta_j$, respectively.

A rational user will seek to minimize his charge, hence performs the following optimization:

$$\text{USER:} \quad \min_{(\rho,\beta)} \alpha(\infty, t; m, \mathbf{h}) \quad \text{such that} \quad (\rho, \beta) \in G\,, \tag{11}$$

where $t = \infty, 0$ for which the effective bandwidth becomes $\rho_j, \beta_j$ respectively.

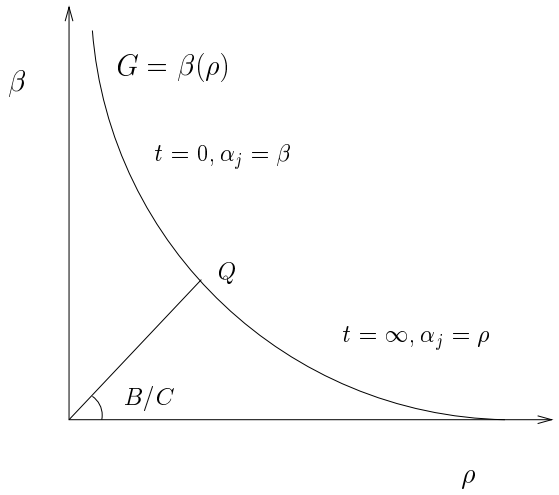



Figure 6. Indifference curve $G$ and effective bandwidth for deterministic multiplexing. For points above $Q$ users tend to decrease $\beta$, since their charge is proportional to $\beta$. On the other hand, for points below $Q$ they tend to decrease $\rho$, since their charge is proportional to $\rho$.

The network tries to maximize the number of users it accepts, hence performs the following:

$$\text{NETWORK:} \quad \max n \quad \text{such that} \quad \sum_j \rho_j \leq C \ \text{ and } \ \sum_j \beta_j \leq B\,. \tag{12}$$

We assume that the network and the users update their parameters in discrete steps as follows: (i) the network computes the values of $s, t$ based on the current traffic, (ii) the users synchronously select their leaky bucket parameters, and (iii) the network recomputes the values of $s, t$ for the new operating point. For simplicity, we assume that users have identical requirements, hence their choice for $(\rho, \beta)$ coincide. Due to this, the constraints in (12) become $n\rho \leq C$ and $n\beta \leq B$, where $n$ is the total number of users.

Consider the point $Q \in G$ where the two constraints are both active, and let $\rho^*, \beta^*$ be the leaky bucket parameters at point $Q$, Figure 6. Hence, $n^* = C/\rho^* = B/\beta^*$. The value $n^*$ is the maximum number of users (welfare optimum), since moving away from $Q$ decreases the number of users that can be accepted. Note also that the point $Q$ is the intersection of the indifference curve $G$ with the line of slope $B/C$ that passes from the origin.

Next we show that under our charging approach the network operating point will move towards point $Q$ and that point $Q$ is an equilibrium.

If the users choose a point $M$ below $Q$, then the first constraint of (12) will be active ($t = \infty$) and the charge will be proportional to $\rho$; this will guide users to reduce $\rho$ and move towards $Q$. On the other hand, if the users choose

a point $M$ above $Q$, then the second constraint will be active and the charge will be proportional to $\beta$; this will guide users to reduce $\beta$ and move towards $Q$. Assuming that, in order to avoid oscillations, users are allowed to make small changes to their traffic contracts, point $Q$ will be eventually reached.

At $Q$, both constraints in (12) are active and charges will be proportional to a linear combination $\lambda_1\rho + \lambda_2\beta$ of the effective bandwidths $\rho$ and $\beta$ that correspond to the two constraints, where $\lambda_1, \lambda_2$ are the shadow prices of the optimization problem in (12). Furthermore, point $Q$ is an equilibrium since users minimize their charge at that point. Indeed, if we consider the Lagrangian of (12) with the constraints replaced by $n\rho \leq C$ and $n\beta \leq B$, then from the necessary conditions for optimality we can show that the shadow prices $\lambda_1, \lambda_2$ satisfy the equation $\lambda_1\rho^* + \lambda_2\beta^* = 1$, which corresponds to the tangent of the indifference curve $G$ at point $Q$. Assume now that a user moves away from point $Q$ by decreasing his leak rate to $\rho^* - \Delta\rho$, while still staying on the indifference curve. Due to the convexity of the indifference curve, the user's bucket size becomes $\beta^* + \lambda_1/\lambda_2\Delta\rho^* + \delta$, where $\delta > 0$. His charge will then become $\lambda_1\rho^* + \lambda_2\beta^* + \lambda_2\delta$, hence it increases. Thus, the point $Q$ on the indifference curve is an equilibrium.

We now turn to the case of statistical multiplexing. Assume that the network charges using the simple bound $\tilde{\alpha}$ in (7). Since a rational user seeks to minimize his charge, he will select the pair $(\rho, \beta)$ that minimizes (7). Hence, the user performs the following optimization:

$$\text{USER:} \quad \min_{(\rho,\beta)} \tilde{\alpha}(s, t; m, \mathbf{h}) \quad \text{such that} \quad (\rho, \beta) \in G\,, \tag{13}$$

where $m$ is the user's mean rate and $\mathbf{h}$ the policing constraints of his contract, which include the peak rate $h$ and leaky bucket $(\rho, \beta)$.

On the other hand, the network tries to maximize the number of sources that it can accept while satisfying the QoS constraint $P(\text{overflow}) \leq e^{-\gamma}$. This can be written as follows:

$$\text{NETWORK:} \ \max n \ \text{ such that } \ \sup_t \inf_s [stn\tilde{\alpha}(s, t; m, \mathbf{h}) - s(Ct + B)] \leq -\gamma\,. \tag{14}$$

By performing the optimization in (14), the network computes a new pair $(s, t)$, i.e., the link's operating point moves. This new pair $(s, t)$ will affect the users' charges, hence will guide them to perform the optimization in (13) and select a new leaky bucket pair $(\rho, \beta)$, which in turn will affect the link's operating point, hence the tariffs, and so on. Our experimental results indicate that, for the traffic considered, the above user-network interaction leads to an equilibrium. Table 1 shows such equilibria for a range of buffer sizes and for target overflow probability $10^{-6}$. As in the case of deterministic multiplexing, for statistical multiplexing our results show that in the equilibrium the number of sources is maximized. This is expected since the users' objective to minimize their charge, which is proportional to their effective bandwidth, coincides with the network's

Table 1
Equilibrium under deterministic and statistical multiplexing. [ $C = 34$ Mbps, $P(\text{overflow}) \leq 10^{\perp 6}$ (for statistical multiplexing), Bellcore Internet WAN traffic. ]

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

objective to maximize the number of users accepted, since the latter is achieved when the effective bandwidth per user is minimized.

For both cases discussed in this section, we have assumed that the network has only partial information regarding the user traffic, namely the mean rate $m$ and traffic contract $\mathbf{h}$. Our incentive compatibility results indicate that the number of connections accepted under decentralized control with each user selecting his own $(\rho, \beta)$ pair is the same as the number of users that would have been accepted had the network provider, knowing the mean rates and indifference curves of all users, centrally chosen $(\rho, \beta)$ for all users. If the network had a priori the full statistical information of all users, then it would have loaded the link while satisfying the constraint $\sup_t \inf_s [stn\alpha(s,t;m,\mathbf{h}) - s(Ct+B)] \leq -\gamma$. Since typically $\alpha(s,t;m,\mathbf{h}) < \tilde{\alpha}(s,t;m,\mathbf{h})$, the number of users accepted in the case of full statistical information would be larger than the number accepted in the cases considered, where the network has only partial information.

## 5. Incentives for traffic shaping

Our final investigation deals with the incentives for traffic shaping that are provided by our charging scheme.

We first begin with a discussion of how traffic shaping affects the maximum link utilization. Figure 7 shows the maximum link utilization, i.e., the aggregate mean rate divided by the link capacity, for different buffer sizes and shaping delays, when $C = 34$ Mbps and a target overflow probability $10^{-6}$ is met. Observe that for a moderate buffer ($75 \times 10^3$ bytes), shaping traffic in a buffer with delay less than 500 msec does not affect the maximum utilization, thus the amount of resources used by each connection. Hence, for such buffer sizes the charging scheme need not provide the incentives for traffic shaping.

Of course a user can use shaping to make a contract with a lower peak rate. However, contrary to intuition, this might not affect his effective bandwidth as seen by the network. For example, Table 2 shows that for a buffer larger than $100 \times 10^3$ bytes, reducing the peak rate from 2.3 Mbps to 0.28 Mbps does not

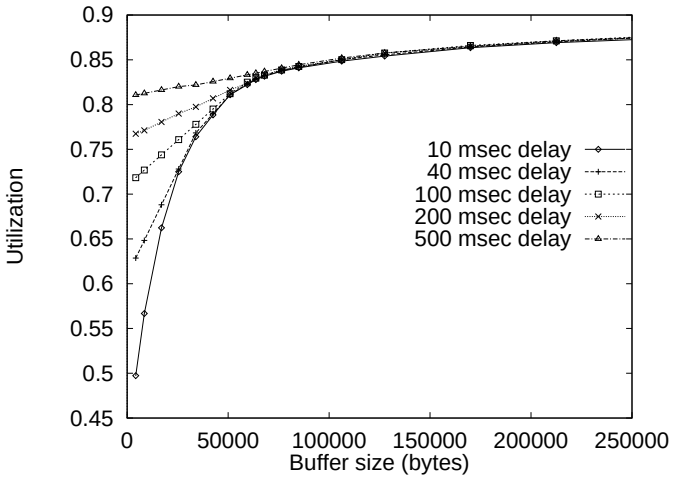



Figure 7. Effects of traffic shaping on the maximum link utilization. For a relatively small buffer ($100 \times 10^3$ bytes), shaping in a buffer with delay less than 500 msec does not affect the maximum utilization. [ $C = 34$ Mbps, $P(\text{overflow}) = 10^{\perp 6}$, Internet WAN traffic ]

change the effective bandwidth, which is equal to 7 Kbps.

The above discussion demonstrates how the theory of effective bandwidths described in Section 2 captures the effects of the various traffic and network parameters, such as the relevant time scales, on the amount of resources used by connections. Hence our charging approach, which is based on effective bandwidths, correctly takes into account the effects of the above parameters on resource usage and provides incentives for users to shape their traffic only when shaping can increase the maximum link utilization.

## 6. Conclusions

This paper has dealt with one important part of the charging activity: the part that aims to access a connection's network resource usage. In this direction,

Table 2
Effective bandwidth for different peak rates $h$. For buffers larger than $100 \times 10^3$ bytes, reducing $h$ from 2.3 Mbps to 0.28 Mbps does not change the effective bandwidth, which is equal to 7 Kbps. [ $C = 34$ Mbps, $P(\text{overflow}) = 10^{\perp 6}$, Internet WAN traffic ]

| | Effective bandwidth (Kbps) | | |
|---|---|---|---|
| Buffer ($10^3$ bytes) | $h = 2.30$ Mbps | $h = 0.76$ Mbps | $h = 0.28$ Mbps |
| 17 | 8.47 | 7.51 | 7.08 |
| 25 | 8.12 | 7.49 | 7.07 |
| 65 | 7.20 | 7.20 | 7.07 |
| 106 | 7.00 | 7.00 | 7.00 |

we have provided a framework for constructing incentive compatible charges that reflect effective resource usage. Our charging schemes are based on bounds of 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 users. 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). 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.

We have presented numerical results, with real broadband traffic, that display the fairness of the three effective bandwidth bounds that we considered, namely, the on-off bound, the simple bound, and the inverted `T` approximation, and how fairness depends on the link parameters (capacity and buffer). Furthermore, we have displayed the incentive compatibility and the user-network interaction of the proposed scheme for the cases of deterministic and statistical multiplexing. Based on the results of these investigations we believe that our approach for constructing tariffs can be used to fairly recover costs from users and lead to efficient and stable network operation.

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.

## Acknowledgements

The authors thank the anonymous referee for his thorough review and detailed comments, which have improved the presentation of the paper.

## References

[1] D. Clark. Internet cost allocation and pricing. In L. W. McKnight and J. P. Bailey, editors, *Internet Economics*. MIT Press, Cambridge, MA, 1997.

[2] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang. Pricing in computer networks: Motivation, formulation, and examples. *IEEE/ACM Trans. on Networking*, 1(6):614–627, November 1993.

[3] C. Courcoubetis, F. P. Kelly, and R. Weber. Measurement-based usage charges in communications networks. Technical Report 1997-19, Statistical Laboratory, University of Cambridge, 1997. To appear in *Operations Research*.

[4] C. Courcoubetis and V. A. Siris. Managing and pricing service level agreements for differentiated services. In *Proc. of 6th IEEE/IFIP International Conference of Quality of Service (IWQoS'99)*, London, UK, May-June 1999.

[5] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Application of the many sources asymptotic and effective bandwidths to traffic engineering. *Telecommunication Systems*, 12:167–191, December 1999. A shorter version appeared in *Proc. of ACM SIGMETRICS'98/PERFORMANCE'98*.

[6] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources. *J. Appl. Prob.*, 33:886–903, 1996.

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

[8] F. P. Kelly. On tariffs, policing and admission control for multiservice networks. *Operations Research Letters*, 15:1–9, 1994.

[9] F. P. Kelly. Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. Zeidins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

[10] W. E. Leland and D. V. Wilson. High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection. In *Proc. of IEEE INFOCOM'91*, pages 1360–1366, Bal Harbour, FL, USA, April 1991.

[11] S. H. Low and P. P. Varaiya. A new approach to service provisioning in ATM networks. *IEEE/ACM Trans. on Networking*, 1(3):547–553, October 1993.

[12] J. K. Mackie-Mason and H. R. Varian. Pricing congestible network resources. *IEEE J. Select. Areas in Commun.*, 13(7):1141–1149, September 1995.

[13] J. K. Mackie-Mason and H. R. Varian. Pricing the Internet. In B. Kahin and J. Keller, editors, *Public Access to the Internet*. Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[14] C. Parris, S. Keshav, and D. Ferrari. A framework for the study of pricing in integrated networks. Technical Report TR-92-016, International Computer Science Institute, Berkeley, CA, March 1992.

[15] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. Technical Report 101, University of Wuerzburg, February 1995.

[16] B. K. Ryu and A. Elwalid. The importance of the long-range dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *Proc. of ACM SIGCOMM'96*, pages 3–14, Stanford, CA, USA, August 1996.

[17] J. Sairamesh, D. F. Ferguson, and Y. Yemini. An approach to pricing, optimal allocation and quality of service provisioning in high-speed packet networks. In *Proc. of IEEE INFOCOM'95*, Boston, MA, USA, April 1995.

[18] S. Shenker, D. Clark, D. Estrin, and S. Herzog. Pricing in computer networks: Reshaping the research agenda. *ACM Computer Communication Review*, pages 19–43, 1996.

[19] V. A. Siris, D. J. Songhurst, G. D. Stamoulis, and M. Stoer. Usage-based charging using effective bandwidths: studies and reality. In *Proc. of the 16th Int. Teletraffic Congress (ITC - 16)*, North Holland, 1999. Elsevier Science B. V.

[20] Q. Wang, J. M. Peha, and M. A. Sirbu. The design of an optimal pricing scheme for ATM integrated-services networks. In J. P. Bailey and L. Mcknight, editors, *Internet Economics*, Massachusetts, 1996. MIT Press.

[21] D. Wischik. The output of a switch, or, effective bandwidths for networks. To appear in *Queueing Systems*, 1999.

[22] D. Wischik. Sample path large deviations for queues with many inputs. To appear in *Annals of Applied Probability*, 1999.