MEASUREMENT-BASED USAGE CHARGES IN COMMUNICATIONS NETWORKS

COSTAS COURCOUBETIS

ICS-FORTH and Department of Computer Science, University of Crete, Greece

FRANK KELLY and RICHARD WEBER

University of Cambridge, Cambridge, U.K.

We study usage-sensitive charging schemes for broadband communications networks. We argue that a connection's 'effective bandwidth' is a good proxy for the quantity of network resource that the connection consumes and can be the basis for a usage charge. However, the determination of effective bandwidth can be problematic, since it involves the moment generating function of the cell arrival process, which may be difficult to model or measure. This paper describes methods of computing usage charges from simple measurements and relating these to bounds on the effective bandwidth. Thus we show that charging for usage on the basis of effective bandwidths can be well-approximated by charges based on simple measurements.

Charging and pricing are essential requirements in the operation of a communication network. They are needed not only to recover costs and make a profit. Even if a generous operator is willing to offer a network for free, there are still compelling reasons to charges for services in order to exercise control. The congestion that has plagued the Internet because it lacks any mechanism for charging and pricing highlights the fact that without charges it is difficult to control congestion or divide network resources amongst users in a workable and stable way.

Subject classifications: Communications: measurement-based charging.

Of course there are many considerations that influence the prices at which an operator will choose to sell network services. Marketing and regulation are certainly important, but these considerations are not unique to the operation of a communications network. Special considerations do, however, arise from the fact that a broadband communications network is intended simultaneously to carry a wide variety of traffic types.

Our conception of a broadband network is that of a collection of resources (links, buffers, switches, etc.) which can be used to provide a wide variety of communications services. These services are distinguished by traffic contracts, which specify parameters to which the traffic must adhere (a maximum peak rate, for example), and the quality of service which the network undertakes to guarantee (typically, cell loss or delay). These concepts are accepted as relevant to both the future development of the Internet (guaranteed and controlled-load services) and to ATM (ITU (1995), ATM Forum).

Given a set of offered services, along with traffic contracts, prices and methods of charging for those services, users will generate certain demands, fluctuating on various time scales, e.g., daily. Network operators will make decisions as to what quantities of differing services they wish to offer at various times. They will set and slowly change prices. Ultimately, in a competitive market, prices will reach an equilibrium. The market for services will be partitioned into various segments, characterised by different traffic contracts, qualities of service and charges. Some segments will be for high price, real-time services; others will be for low price, best-effort services.

There remains the important consideration that within each market segment there may be many types of traffic that are statistically different, purchased as different services, but which are actually indistinguishable at the cell level. They are close enough in their quality of service requirements that they may substitute for one another. For example, if a user desiring two units of service A realises that one unit of service B meets his needs just as well and costs less, then he will purchase one unit of B. Perhaps he will need to smooth his traffic to meet a slightly different peak rate restriction in the contract for type B traffic, or accept a slightly worse cell loss rate guarantee, but the cost of doing so might be relatively small. The implication is that charges for units of services A and B should be in the ratio of 1 to 2. From the network's point of view, this relative pricing of services A and B makes sense only if the network can carry a unit of service B with the same ease (in terms of maintaining guaranteed service to all customers) as two units of service A. This is one of the key ideas in the paper: that charges for services which are substitutable for one another must be in proportion to their resource usages.

This paper is concerned with substitutable connections of real-time traffic whose quality of service requirement is for small cell loss. Two examples are a video conferencing call between two sites of a company, and the delivery of a video film to a private residence. Both traffic types also have requirements on delay, but we suppose that this is engineered by the size of buffers and queueing disciplines, and that cell loss is the principal concern. We investigate the issue described in the above paragraph, namely, what are the implications for charging due to the fact that some services can substitute for others? We argue that charges for units of services A and B must be in proportion to their network resource usage, i.e., 'incentive compatible'. This begs the question as to how one might measure resource usage, to which our answer is the 'effective bandwidth' statistic described in Section 1. Subsequent sections of the paper are devoted to explaining the how the effective bandwidth statistic can be used to construct charges, or how one might use statistics which approximate to the effective bandwidth. This paper builds on work in Kelly (1994b), in which the idea of charges based upon effective bandwidths was introduced. What is new is the description of a general framework for a class of charging schemes that can be based on arbitrarily refined a posterior measurements and a priori information (such as that a connection is compliant with a given leaky-bucket policer).

Please note that this paper does not deal with explicit charges for fixed costs, network management, billing, maintenance, marketing, etc. These might be reflected within an overall charge by a fixed-charge component. Neither do we deal with charging for non-real-time, best-effort traffic. Such traffic is not subject to connection acceptance control nor given a strict quality of service guarantee, and effective bandwidth ideas do not directly apply. Ideas for pricing this traffic are developing, and envisage prices which dynamically adjust to the level of congestion in the network. See Courcoubetis, Siris and Stamoulis (1998b), Kelly (1997), Kelly, Maulloo and Tan (1998) and MacKie-Mason and Varian (1994).

Our interest is in that part of the charge which reflects usage of the shared physical network resources of bandwidth and buffering, where it is the finiteness of these resources that is the principal binding constraint. In terms of the example above, such a constraint can be expressed (at least locally) as a linear constraint $n_A\alpha_A + n_B\alpha_B \leq C$, where n_A, n_B are the number of connections of types A and B, and α_A, α_B and C depend upon the quantities of network resources and the statistical characteristics of type A and B connections. The coefficients α_A and α_B are the 'effective bandwidths'. In a competitive equilibrium, the overall social welfare, say $u(n_A, n_B)$, is maximized subject to this constraint. By social welfare we mean the sum of all user benefits. By formulating the Lagrangian optimization problem, we maximize without regard to the constraint, but with usage prices $\lambda\alpha_A$ and $\lambda\alpha_B$ incorporated into the objective function, i.e., by maximizing with respect to n_A and n_B an expression of the form

$$u(n_A, n_B) - \lambda\alpha_A n_A - \lambda\alpha_B n_B.$$

Here λ is the 'shadow price' for the constraint. If usage prices of $\lambda\alpha_A$ and $\lambda\alpha_B$ are posted then the social welfare optimum is obtained when n_A, n_B are chosen in a decentralised way and without regard to the constraint. If different usage prices were to be posted then the social welfare optimum would not be obtained; therefore such prices could not hold in a truly competitive setting. Thus a key fact that should be reflected in fixing these two usage charges is that they are proportional to α_A and α_B; this is what we capture in basing usage charges upon the effective bandwidths, or approximations to them. We do not attempt to state the absolute value of the usage charges; these depend on the value of λ, which is revealed by the market, via the mechanism of supply and demand for these two network services.

Charges not only generate income for the network, but also introduce feedback and control. For example, it may be economical for some customers to shape their traffic, and by their doing so the overall network performance may be enhanced. The key point is that after each user has minimized his own charges, the network should be left operating at an efficient point (e.g., with good utilisation and robustness). 'Incentive compatible' tariffs should guide the population of cost-minimizing customers to select contracts and to use the network in ways that are good for overall network performance. A closely related idea is that charges should have some fairness properties. Charges should reflect customers' relative network usages, so that a customer who makes less use of the network is charged less.

The concept of an effective bandwidth provides a notion of network resource usage appropriate for a multiservice broadband network, but this concept does not lend itself naturally to be used as a charging mechanism. This paper describes a methodology, based on the concept of an effective bandwidth, for developing families of charging schemes based on simple measurements, with the property that the expected charge of a connection bounds the effective bandwidth of the connection. These charges are sound, both in terms of incentive compatibility and fairness, but are not too complex. Their implementation does not require the network operator to make overly sophisticated or unrealistic measurements. They are also simple enough that users can determine the effects of decisions under their control, e.g., what effect a reduction in peak rate might have on incurred charges.

The paper is organised as follows. In Section 1 we summarise the notion of an effective bandwidth which underlies our entire approach. In Section 2 we use the concept of effective bandwidth to argue that charges should depend upon both *a priori* knowledge and *a posteriori* measurements of resource usage. We present a general method for charging, based on both static contract parameters, and measurements taken over the duration of a connection. In Section 3 we consider several special instances of our approach to charging. In Section 4 we give a numerical example, which compares a pricing approach described by Kelly (1994a) and a new approach that has similarities to the way personal income tax is charged. In Section 5 we discuss some related issues, such as connection acceptance control.

1 Effective bandwidths

1.1 Effective bandwidths as a basis for charging

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 $n = (n_1, \dots, n_J)$. We suppose that after taking into account all economic factors (such as sensitivity of demands to prices, competition, and so forth) the proportions of traffic of each of the J types remains close to that given by n, and we seek to understand the relative usage of network resource that should be attributed to each traffic type.

We take a discrete time model and let $X_{jk}[0,t]$ be the total load produced by the kth source of type j in epochs $1, \dots, t$. We assume that increments of $\{X_{jk}[0,t],\ t \geq 0\}$, such as $X_{jk}[0,t+s] - X_{jk}[0,t]$, have distributions which do not depend upon t (i.e, we have stationary increments), which may depend upon the type j, but not upon k. We do not require that sources be ergodic (i.e., that the distribution of $X_{jk}[0,t]$ can be found from a single sample path). The *effective bandwidth* of a source of type j is defined as

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

for some choice of a *time parameter* t and *space parameter* s. The effective bandwidth has the property that it increases from the mean to the peak value of $X_{jk}[0,t]/t$ as s increases from 0 to ∞.

Let $L(C,B,n)$ be the proportion of workload lost, through overflow of a buffer of size $B > 0$, when it is served at rate C and $n = (n_1, n_2, \dots, n_J)$. An important limiting regime, first considered in a key paper of Weiss (1986), is one in which the number of sources *and* the buffer size increase together. Courcoubetis and Weber (1996) have shown that

$$\lim_{N\to\infty} \frac{1}{N} \log L(CN, BN, nN) = \sup_t \inf_s \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - s(Ct+B) \right]. \tag{2}$$

The corresponding result has been proved in continuous time by Botvich and Duffield (1995) and for a special case by Simonian and Guilbert (1995). Courcoubetis, Fouskas and Weber (1995) and Montgomery and de Veciana (1996) have considered the accuracy of approximations for $L(C, B, N)$ based upon (2).

Let $A(\gamma, C, B)$ be the subset of $\mathbb{Z}_+^J$ such that $n \in A(\gamma, C, B)$ implies $\log L(C, B, n) \le -\gamma$. Such n are those for which the proportion of workload lost is below some predefined level and so expresses some quality of service (QoS) requirement. As $A(\gamma, C, B)$ is hard to compute we approximate it by using (2), from which it follows that

$$\lim_{N\to\infty} \frac{A(\gamma N, CN, BN)}{N} = A\,,$$

where

$$A = \bigcap_{0<t<\infty} A_t\,, \tag{3}$$

with

$$A_t = \left\{ n : \inf_s \left[st \sum_{j=1}^{J} n_j \alpha_j(s,t) - s(Ct+B) \right] \le -\gamma \right\}, \tag{4}$$

a region with convex complement in $\mathbb{Z}_+^J$ (Kelly (1996)). The set A is a scaled asymptotic approximation of the set $A(\gamma, C, B)$ and we refer to it as the 'acceptance region'.

Figure 1 illustrates an acceptance region when $J = 2$. The dotted lines mark the boundaries of A_t for three values of t. QoS is guaranteed when (n_1, n_2) lies within the region, A, bounded by thick lines.

If the boundary of the region A is differentiable at the point n, then the tangent plane determines a half-space

$$\sum_{j=1}^{J} n_j \alpha_j(s,t) \le C + \frac{1}{t}\left(B - \frac{\gamma}{s}\right) \tag{5}$$

where (s, t) is an extremizing pair in relation (2). Thus at points where the boundary of the region A is differentiable, the effective bandwidths $\alpha_j(s,t)$, $j = 1, \ldots, J$, determine the relative resource usages of traffic of different types, for local variations of the traffic mix

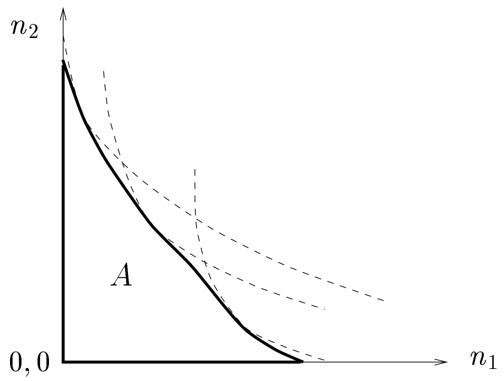

Figure 1: An acceptance region, A.

about the point n. This is the rationale for specifying usage charges that are proportional to effective bandwidths. At points where the boundary of the region A is not differentiable, two or more constraints of the form (5), with differing values of s and t, will be needed to characterise permissible local variations of the traffic mix. We illustrate this possibility in Sections 1.3 and 1.4; later, in Section 3.4, we describe how shadow prices may be associated with each constraint, and used to weight the various effective bandwidths arising from the different constraints.

Our approach is based on the asymptotic result (2) and implicitly assumes that the no connection occupies a large proportion of the link capacity. Therefore, when traffic fluctuates around the operating point, the changes in the proportions of traffic vary little and extremizing values of s and t are nearly constant. Experimental results for video traffic show that as the traffic mix varies there are points of discontinuity in t and in the effective bandwidths; for large links (150 Mbps and greater) the variations in the effective bandwidths are small as the traffic mix varies by 10-20%, whereas for small links the effect cannot be ignored. For more details see Courcoubetis, Siris and Stamoulis (1998a).

Note that because we have assumed a FCFS discipline all connections are offered the same QoS. In practice, switches may have facilities for weighted fair queueing, and methods of selective cell discard, whereby different connections can be offered different QoSs. Our results are directly applicable to some of these, such as the example of priority queues in Section 3.4, by which two QoS levels can be provided. We expect that other cases might be approached in a similar way, the key idea in any context being the substitutability of one source for another. However, we also expect that network operators will wish to provide only a small set of QoS-differentiated services and so it may be reasonable to suppose that all real-time traffic is offered no more than one or two QoS levels.

The effective bandwidth has been developed as a measure of a connection's resource usage at one switch. What justification is there for using this same measure of resource usage to assess the connection's resource usage along a route through the network, where it uses multiple buffers and switches? There are several things one can say.

Firstly, we might expect that it is often one bottleneck link which provides the binding constraint.

Secondly, effective bandwidths can sometimes 'decouple' at successive links on a route, i.e., that the effective bandwidth at a bottleneck link, in the midst of a route, is not affected by the smoothing that takes place in buffers upstream of that link. See Wischik (1998).

Thirdly, we emphasise that the principal attractive characteristics of the effective bandwidth statistic, $\alpha(s,t) = (1/st)\log E[\exp(-sX[0,t])]$. It is a univariate statistic, summarising the burstiness of a connection, and its two parameters, s and t, can be used to tune the statistic to the degree of statistical multiplexing that takes place in the network and the time scale over which an event which is guaranteed not to occur too often typically occurs. Although the effective bandwidth statistic is motivated by consideration of a constraint on cell loss rate in a single switch, the above characteristics make it an attractive candidate for wider use. In such use, s and t are to be set, not by solving equation (2), but adaptively or by experience. In fact, as we describe below, we do not advocate that one should try to make an on-line estimate of $\alpha(s,t)$, but rather use an approximation to it, leading to more intuitive charges, such as $aT + bV$, where T is the duration of a call, V the number of cells carried, and a, b replace s, t as parameters which the network operator can adjust.

1.2 Interpretations of space and time scales

The derivation of the large deviations approximation (2) in Courcoubetis and Weber (1996) gives a straightforward interpretation of the parameter t as the time for which the server has been busy preceding a buffer overflow. This interpretation has been experimentally verified by Courcoubetis, Siris and Stamoulis (1998a). The interpretation of s is less straightforward: over the busy period preceding a buffer overflow the amount of work produced by a source of type j has an exponentially tilted distribution, with with tilt parameter s. See Shwartz and Weiss (1995), page 13, for a description of the tilted distribution.

If we identify γ with the right hand side of the limit (2) then, by the envelope theorem

(Varian (1992)),

$$\frac{\partial \gamma}{\partial C} = st \quad \text{and} \quad \frac{\partial \gamma}{\partial B} = s \tag{6}$$

and these identities provide further interpretations of the space and time scales s and t respectively. We see that the parameter s has an interpretation as the derivative of the logarithm of the loss probability with respect to buffer size. A further interpretation of t is available from the deduction that

$$t = \frac{\partial \gamma}{\partial C} \Big/ \frac{\partial \gamma}{\partial B} \, .$$

Thus if B and C are chosen jointly to achieve a given γ, and if the optimal trade-off between B and C is made, e.g., to minimize a cost such as $f(B, C)$ subject to constraint (5), then

$$t = \frac{\partial f}{\partial C} \Big/ \frac{\partial f}{\partial B} = \frac{\text{marginal cost of unit capacity}}{\text{marginal cost of unit buffer}} \, .$$

The space and time scales s and t are defined as the extremizing pair in relation (2), and hence they depend upon the parameters B and C, and upon the aggregate traffic mix at the resource. Note that the traffic from a single source has, at least under the limit (2), no effect on s and t.

It is instructive to consider two special cases in which the acceptance region reduces to two linear constraints. In both these examples it is convenient to take time to be continuous rather than discrete.

1.3 Example: Leaky bucket policing models

Suppose that each source k, of type j, is guaranteed to satisfy the condition

$$X_{jk}[0, t] \le \rho_j t + \beta_j, \quad \text{for all } t \, . \tag{7}$$

A source which obeys this constraint is said to comply with leaky bucket policing, with token buffer of size β_j and leak rate ρ_j (see ITU Recommendation I371 (1995)). Leaky bucket constraints are one example of constraints that can arise as part of a traffic contract between the network and the user.

Suppose we desire that there be no cell loss, so that $\gamma = \infty$. Assume that there is positive probability of equality in (7). Then it is possible to show that

$$\mathbb{P}(\text{buffer overflow}) = 0 \iff \sum_j n_j \rho_j \le C \text{ and } \sum_j n_j \beta_j \le B. \tag{8}$$

In this case the region A is completely defined by two linear constraints, corresponding to the limit of the set A_t as $t \to \infty$ and $t \to 0$ respectively. Note that $\alpha_j(s,t)$ is correspondingly equal to either ρ_j or β_j, depending upon whether n is such that the first or second constraint on the right hand side of (8) holds with equality. More generally, if each source is policed by a positive but finite number of leaky buckets, then the acceptance region is completely defined by a finite number of linear constraints (see Cruz (1991) and Kelly (1996), Section 3.4). For example, ITU Recommendation I371 (1995) discusses the use of two leaky buckets to police a source, one with a small value of β to bound the peak rate, and one with a much larger value of β and a smaller value of ρ to bound the sustainable cell rate.

1.4 Example: Brownian bridge model

To motivate the second example, consider several independent sources, where each behaves as a periodic source, producing a burst of size ρ at unit spaced times $\{U+n, n=0,1,\dots\}$, where U is uniformly distributed on the interval $[0,1]$. The superposition of such a collection can be well approximated by a Brownian bridge (for a recent review see Hajek (1994)). This motivates study of the source

$$X_j[0,t] = \rho_j t + \rho_j Z(t - \lfloor t \rfloor)$$

where $Z(t), 0 \le t \le 1$, is a standard Brownian bridge. The resulting functions $\alpha_j(s,t)$ produce a region A defined by the two constraints

$$\sum_j n_j \rho_j \le C, \tag{9}$$

$$\sum_j n_j \left(\rho_j + \rho_j^2 \frac{\gamma}{2B} \right) \le B + C. \tag{10}$$

(Indeed this acceptance region is exact for a simple queue fed by Brownian bridge inputs - Kelly (1996), Section 3.5). Constraint (9) is of the canonical form (5) with $t \to \infty$. Constraint (10) may be thrown into the form (5), with for example the choice $(s,t) = (2\gamma/B, 1/2)$.

2 Charging schemes

2.1 Combining prior information with measurements

We have argued above that effective bandwidths can provide a way to assess resource usage and that usage charges should be proportional to effective bandwidths. However, there are subtleties in the conversion of an effective bandwidth into a charge, arising from whether we estimate the effective bandwidth of a connection of a given type from a priori or a posteriori information.

A priori information which might be available for connections of type j could include the fact that all connections of this type are subject to a common traffic contract, possibly defined in terms of leaky bucket parameters, but might also include information gleaned from historical data on past connections of type j. For example, one might estimate the effective bandwidth of connections of type j in the following way. For each connection k that we see of type j we could compute

$$\frac{1}{T/t}\sum_{i=1}^{T/t} e^{sX_{jk}[(i-1)t,it]}, \tag{11}$$

average such estimates over all connections we have seen to date of type j, form an empirical estimate of the expectation appearing in the formula (1), and hence make an estimate, $\tilde{\alpha}_j(s,t)$, say, of the effective bandwidth of a connection of type j. Note that we must average over many connections of type j, since because we have not assumed ergodicity of sources of type j the evaluation of (11) may differ significantly between two connections of this type.

We could now simply charge each newly admitted connection of type j an amount per unit time equal to the empirical estimate $\tilde{\alpha}_j(s,t)$, as determined by past connections of type j. This is the charging method which is adopted by 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. (There is only one customer type, except that some such restaurants have a lower price for children or different prices depending on the time of day.) The existence of all-you-can-eat restaurants demonstrates that this charging scheme is viable. This is analogous to the charging scheme used when local telephone calls are unmetered, or when the only cost a student pays to browse the WWW is the cost of waiting for a free seat in the computer room. But all-you-can-eat restaurants are not for everyone. They encourage diners to over-eat; they tend to serve only the lower quality part of the market. Customers with small appetites are likely to feel they are over-charged in such restaurants. Others are put off by the bare-bones, help-yourself, no-frills ambience.

The problem with adopting a charging scheme in which a connection is charged at a rate per unit time which is determined wholly in terms of parameters that are known at call setup, is that users are not penalised for using more than the typical amount of resources used by others of their type. Supposing that connections of a given type are subject to the same traffic contract, (e.g., a leaky bucket constraint), each user of that type may as well use the maximum of network resources that the contract allows. This results in a situation where the operator calculates the largest effective bandwidth that is possible subject to the agreed policing parameters and charges for it. 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 though as they did and will seek network service providers using a different charging method.

For a concrete example, consider the case of a source with peak rate h and mean rate m, (perhaps policed as the peak and sustainable cell rate, ITU Recommendation I371 (1995)). Then, as we show in Subsection 3.3, the effective bandwidth (1) is bounded above by the expression

$$G(m,h) = \frac{1}{st} \log \left[1 + \frac{m}{h} \left(e^{sht} - 1 \right) \right] . \tag{12}$$

But a charge based on the bound (12), evaluated with h and m replaced by policed peak and sustainable cell rates respectively, severely penalises users whose mean traffic may be unpredictable and not easily characterised by policing parameters such as a sustainable cell rate.

At the other extreme, one might charge a user wholly on the basis of a posteriori measurements that are made for his connection, e.g., charge

$$\hat{\alpha}_{jk}(s,t) = \frac{1}{st} \log \left(\frac{1}{T/t} \sum_{i=1}^{T/t} e^{sX_{jk}[(i-1)t,it]} \right) . \tag{13}$$

as measured for this connection. Apart from the difficulty of interpreting this complicated tariff to users, there is a conceptual flaw, which can be illustrated as follows. Suppose a user requests a connection policed by a high peak rate, but then happens to transmit very little traffic over the connection. Then an *a posteriori* estimate of quantity (1) will be near zero and the charge near zero, even though the *a priori* expectation may be much larger, as assessed by either the user or the network. Then too much of the risk associated with the uncertainty of a user's traffic is borne by the network, since the network may have to allocate at least some resources on the basis of a priori information about the connection.

Our approach attempts to deal with the difficulties illustrated in the discussion of the above two charging methods. We construct a charge based on the effective bandwidth, which is a function of both static parameters (known a priori, such as parameters of leaky bucket constraints) and dynamic parameters (known a posteriori, such as the duration and volume of the connection); the static parameters might arise from traffic contracts, while the dynamic parameters arise from measurements of the connections. 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.

2.2 Charges that are linear in chosen measurements

In this section we investigate a charging scheme in which the per unit time charge for a connection of type j can be expressed as a linear function of the form

$$f(X) = a_0 + a_1 g_1(X) + \cdots + a_L g_L(X) = a_0 + a^\top g(X)\,, \tag{14}$$

where $g_1(X), \ldots, g_L(X)$ are measurements taken from the observation of $X = (X_1, \ldots, X_T)$, or some functions of those measurements. Here X and $a_0, \ldots, a_L$ depend on j, and hence perhaps on policing parameters for sources of type j, but we suppress the dependence on j for convenience.

For example, with $L = 1$ we could take $g_1(X)$ equal to expression (11). Or we could take $L = 1$ and

$$g_1(X) = (1/T)\sum_{i=1}^{T} X_i\,. \tag{15}$$

In the first case the total charge is quite complicated to compute. In the second case the total charge is just a function of the total number of cells carried, and, through a_0, the duration of the connection. These are of course practically the simplest measurements we could take and lead to a total charge of $a_0 \cdot \text{time} + a_1 \cdot \text{volume}$. Below we consider other possibilities.

Note that by making the charge a linear function of specified statistics the user's expected charge is a function only of the expected values of those statistics. The expected value of the charge per unit time is just

$$\mathbb{E}f(X) = a_0 + a^\top \mathbb{E}g(X)\,.$$

Next we describe how linear functions of the form (14) may be constructed so that the expected charge bounds the effective bandwidth.

Suppose that $X \in \mathcal{X}(\mathbf{h})$, for a given set $\mathcal{X}(\mathbf{h})$ parameterised by some vector $\mathbf{h}$, and the measurements satisfy $\mathbb{E}g(X) = \mathbf{m}$. Let $\bar{\alpha}(\mathbf{m}, \mathbf{h})$ be an upper bound on the effective bandwidth subject to the above constraints, i.e.,

$$\bar{\alpha}(\mathbf{m}, \mathbf{h}) := \sup_{X:\mathbb{E}g(X)=\mathbf{m};\, X\in\mathcal{X}(\mathbf{h})} \left\{ \frac{1}{st} \log \mathbb{E}\left[e^{sX[0,t]}\right] \right\}\,. \tag{16}$$

The supremum is taken over X having stationary increments, $\mathbb{E}g(X) = \mathbf{m}$ and $X \in \mathcal{X}(\mathbf{h})$. Note that s and t are fixed here by the system wide operating parameters n, B, C. Consideration of $\bar{\alpha}(\mathbf{m}, \mathbf{h})$ is partly motivated by the remarks at the start of this section, that this is what we would should charge a user who makes maximal use of his service contract. Later we develop examples where $\mathbf{m}$ might be a mean rate and $\mathbf{h}$ might be peak rate. An important property of $\bar{\alpha}(\mathbf{m}, \mathbf{h})$ is that it is concave in $\mathbf{m}$.

Lemma 1 *$\bar{\alpha}(\mathbf{m}, \mathbf{h})$ is concave in $\mathbf{m}$.*

Proof. Suppose $X, Y \in \mathcal{X}(\mathbf{h})$, and $\mathbb{E}g(X) = \mathbf{m}_1$, $\mathbb{E}g(Y) = \mathbf{m}_2$ where $0 < \theta < 1$. Let Z be X or Y with probabilities θ and $1 - \theta$ respectively. This corresponds to the practical circumstance of being unsure of the type of a connection. Then

$$\mathbb{E}g(Z) = \theta \mathbb{E}g(X) + (1-\theta)\mathbb{E}g(Y) = \theta \mathbf{m}_1 + (1-\theta)\mathbf{m}_2 \,.$$

So

$$\begin{aligned}
\bar{\alpha}(\theta \mathbf{m}_1 + (1-\theta)\mathbf{m}_2, \mathbf{h}) &\geq \frac{1}{st} \log \mathbb{E}\left[e^{sZ[0,t]}\right] \\
&= \frac{1}{st} \log \left[\theta \mathbb{E}e^{sX[0,t]} + (1-\theta)\mathbb{E}e^{sY[0,t]}\right] . \\
&\geq \theta \frac{1}{st} \log \left[\mathbb{E}e^{sX[0,t]}\right] + (1-\theta)\frac{1}{st} \log \left[\mathbb{E}e^{sY[0,t]}\right]
\end{aligned}$$

where the first inequality is by definition of $\bar{\alpha}(\cdot,\cdot)$ and the second by concavity of $\log(\cdot)$. Since this holds for all $X[0,t]$ and $Y[0,t]$ satisfying the constraints, we have after maximizing the right hand side

$$\bar{\alpha}(\theta \mathbf{m}_1 + (1-\theta)\mathbf{m}_2, \mathbf{h}) \geq \theta \bar{\alpha}(\mathbf{m}_1, \mathbf{h}) + (1-\theta)\bar{\alpha}(\mathbf{m}_2, \mathbf{h}) \,.$$

■

The fact that $\bar{\alpha}(\mathbf{m}, \mathbf{h})$ is concave in $\mathbf{m}$ means that there is a tangent hyperplane to $\bar{\alpha}(\mathbf{m}, \mathbf{h})$ at $\mathbf{m}$ and Lagrangian methods apply. So there exists $\lambda_{\mathbf{m}}$ such that

$$\begin{aligned}
\bar{\alpha}(\mathbf{m}, \mathbf{h}) &= \max_X \left\{ \frac{1}{st} \log \mathbb{E}\left[e^{sX[0,t]}\right] - \lambda_{\mathbf{m}}^\top (\mathbb{E}g(X) - \mathbf{m}) \right\} \\
&= \min_\lambda \max_X \left\{ \frac{1}{st} \log \mathbb{E}\left[e^{sX[0,t]}\right] - \lambda^\top (\mathbb{E}g(X) - \mathbf{m}) \right\} .
\end{aligned}$$

We are now able to define a family of charging functions of the form,

$$f_{\mathbf{m},\mathbf{h}}(X) := \max_Y \left\{ \frac{1}{st} \log \mathbb{E}\left[e^{sY[0,t]}\right] - \lambda_{\mathbf{m}}^\top (\mathbb{E}g(Y) - g(X)) \right\} = \bar{\alpha}(\mathbf{m}, \mathbf{h}) + \lambda_{\mathbf{m}}^\top (g(X) - \mathbf{m}) \tag{17}$$

parameterised by $\mathbf{m}$ and $\mathbf{h}$. Here $\mathbf{h}$ is fixed by the type of connection, but the user is permitted to choose $\mathbf{m}$. These charging functions are of the form $a_0 + a^\top g(X)$, where

$$a_0[\mathbf{m},\mathbf{h}] = \bar{\alpha}(\mathbf{m},\mathbf{h}) - \lambda_{\mathbf{m}}^\top \mathbf{m}\,; \quad (a_1[\mathbf{m},\mathbf{h}],\dots,a_k[\mathbf{m}]) = \lambda_{\mathbf{m}}^\top = \left(\frac{\partial}{\partial \mathbf{m}_1}\bar{\alpha}(\mathbf{m},\mathbf{h}),\dots,\frac{\partial}{\partial \mathbf{m}_k}\bar{\alpha}(\mathbf{m},\mathbf{h})\right).$$

Observe that, for any given choice of $\mathbf{m}$, the expected value of the charging rate satisfies

$$\begin{aligned}
\mathbb{E}f_{\mathbf{m},\mathbf{h}}(X) &= \max_Y \left\{\frac{1}{st}\log \mathbb{E}\left[e^{sY[0,t]}\right] - \lambda_{\mathbf{m}}^\top(\mathbb{E}g(Y) - \mathbb{E}g(X))\right\} \\
&\geq \min_\lambda \max_Y \left\{\frac{1}{st}\log \mathbb{E}\left[e^{sY[0,t]}\right] - \lambda^\top(\mathbb{E}g(Y) - \mathbb{E}g(X))\right\} \\
&= \max_Y \left\{\frac{1}{st}\log \mathbb{E}\left[e^{sY[0,t]}\right] - \lambda_{\mathbb{E}g(X)}^\top(\mathbb{E}g(Y) - \mathbb{E}g(X))\right\} \\
&= \bar{\alpha}(\mathbb{E}g(X),\mathbf{h})\,,
\end{aligned}$$

with equality if $\mathbf{m} = \mathbb{E}g(X)$. As we intended, the coefficients $a_i[\mathbf{m},\mathbf{h}]$ depend upon both static information, such as knowledge of a policed peak rate, as well as the user's expectations about measurements that will be taken during the duration of the call.

When $g_1(X)$ is equal to expression (11) we find $\bar{\alpha}(\mathbf{m},\mathbf{h}) = \alpha(s,t)$, and thus $\mathbb{E}f(X) = \alpha(s,t)$ if the user chooses the tariff indexed by $m = \mathbb{E}g_1(X)$. The expected charge is then *equal* to the effective bandwidth. But as we have noted this charge is difficult to compute and interpret to the user. It is likely that we will wish to measure something less complicated. It is clear that some measurements are more useful than others in terms of constructing a sensible charge. The measurements should be informative about the effective bandwidth. But whatever the measurement the charge we have described has the following desirable properties.

1. *It is a simple linear function of measured statistics, $g_1(X),\dots,g_L(X)$. The coefficients depend on static parameters which can reflect the network resources, policing parameters and QoS guarantees.*

2. *The user minimizes the expected charge for his connection if he chooses the charging function $f_{\mathbf{m},\mathbf{h}}(X)$ parameterised by $\mathbf{m} = \mathbb{E}g(x)$.*

3. *The expected charging rate for a connection, $\bar{\alpha}(\mathbb{E}g(X),\mathbf{h})$, is conservative, in the sense that it has the maximum effective bandwidth possible amongst connections having the same value of $\mathbb{E}g(X)$ and which are parametrised by the same static parameters.*

Figure 2 shows how the effective bandwidth might be bounded by a linear function of a measured parameter, m_1.

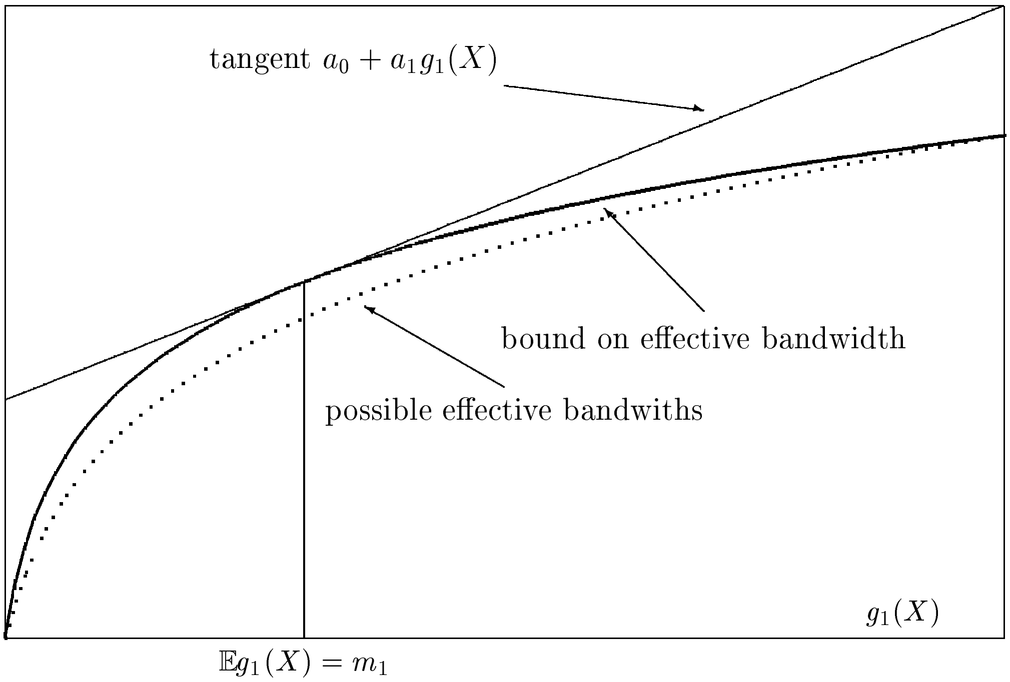

Figure 2: Effective bandwidth and bounds as a function of m_1

3 Some instances of the approach

In this section we consider some special instances of this approach. The first two subsections treat cases in which the upper bound in (16) can found explicitly, either by solving a linear programming or a Markov decision problem. The third subsection explains a simple upper bound which holds generally. The final subsection illustrates that multiple constraints can arise when it is desired to give priority to one type of traffic.

3.1 A linear program

For a stationary sequence of discrete random variables $X_1, X_2, \ldots$ let

$$p(x_1, x_2, \ldots, x_\tau) = P\{X_1 = x_1, X_2 = x_2, \ldots, X_\tau = x_\tau\}$$

and let

$$\mathbf{p} = (p(x_1, x_2, \ldots, x_\tau), \tau = 1, 2, \ldots; x_i = 0, 1, 2, \ldots,\ 1 \le i \le \tau)\,.$$

Then

$$p(x_1, x_2, \ldots, x_\tau) \ge 0, \qquad \sum_{x_1, x_2, \ldots, x_\tau} p(x_1, x_2, \ldots, x_\tau) = 1, \tag{18}$$

and stationarity provides the further condition

$$p(x_{\tau+1}, x_{\tau+2}, \ldots, x_{\tau+t}) = \sum_{x_1, x_2, \ldots, x_\tau} p(x_1, x_2, \ldots, x_{\tau+t}). \tag{19}$$

The probability $P\left\{\sum_{i=1}^t X_i = x\right\}$ may be written

$$p_t(x) = \sum_{x_1, x_2, \ldots, x_t : \sum_1^t x_i = x} p(x_1, x_2, \ldots, x_t). \tag{20}$$

Then $\bar{\alpha}(\mathbf{m}, \mathbf{h}) = (1/st) \log Y$ where Y is the optimal value attained in the linear program

$$\text{maximize} \quad \sum_x e^{sx} p_t(x) \quad \text{subject to} \quad \mathbb{E}g(X) = \mathbf{m} \quad \text{and} \quad \mathbf{p} \in \mathcal{P}(\mathbf{h}),$$

where $\mathcal{P}(\mathbf{h})$ is the set of vectors $\mathbf{p}$ satisfying (18), (19), (20) and additional conditions parameterised by $\mathbf{h}$.

For example, in the case where sources are policed by K leaky buckets we might have $\mathbf{m} = (m)$, $\mathbf{h} = (\beta_k, \rho_k, k = 1, 2, \ldots, K)$ and the additional conditions

$$p_\tau(x) = 0, \quad x > \beta_k + \rho_k \tau, \quad \tau = 1, 2, \ldots, T; \quad k = 1, 2, \ldots, K.$$

Note that the constants $a(\mathbf{m}, \mathbf{h})$ are provided by the dual variables corresponding to the constraint $\mathbb{E}g(X) = \mathbf{m}$, through the relation

$$a(\mathbf{m}, \mathbf{h}) = \frac{\partial}{\partial \mathbf{m}} \bar{\alpha}(\mathbf{m}, \mathbf{h}) = \frac{1}{st} \frac{\partial}{\partial \mathbf{m}} \log Y = \frac{1}{stY} \frac{\partial Y}{\partial \mathbf{m}}.$$

Example: the case $t = 1$

The case $t = 1$ allows several simplifications, and provides some helpful examples. In practical terms it is the case that is appropriate when the buffer is small. Suppose the source is policed, so that $X_i \le h$. Define

$$\theta_j(X) = \frac{1}{T} \mathbb{E}\left[\sum_{i=1}^T 1\{X_i = j\}\right], \quad j = 1, \ldots, h,$$

i.e., θ_j is the expected proportion of epochs in which the load is j. The charging formula becomes quite simple in the case that $\mathbb{E}g(X)$ is linear in θ. Suppose $\mathbb{E}g(X) = A\theta$, where $\theta = (\theta_0, \ldots, \theta_h)^\top$ and A is a $k \times (h+1)$ matrix. Then

$$\begin{aligned} \bar{\alpha}(\mathbf{m}) &= \frac{1}{s} \max_{\theta_0, \ldots, \theta_h} \left\{ \log \frac{1}{T} \mathbb{E}\left[\sum_{i=1}^T e^{sX_i}\right] : \sum_{j=1}^h \theta_j = 1, A\theta = \mathbf{m} \right\} \\ &= \frac{1}{s} \max_{\theta_0, \ldots, \theta_h} \left\{ \log \sum_{j=1}^h \theta_j e^{sj} : \sum_{j=1}^h \theta_j = 1, A\theta = \mathbf{m} \right\} \end{aligned}$$

The maximization problem is equivalent to one in which the objective function is replaced by $\sum_j \theta_j e^{sj}$. This gives a linear program (LP):

$$\text{maximize } \sum_{j=0}^{h} \theta_j e^{sj}\,, \qquad \text{subject to } \sum_{j=0}^{h} \theta_j = 1,\ A\theta = \mathbf{m}.$$

In accordance with the usual theory of linear programming, the solution will have $\theta_j \neq 0$ for at most $k+1$ different j. So we see that the worst case distribution is concentrated on $k+1$ points.

Let us take $L = 1$, $g_1(X) = (1/T)\sum_{i=1}^{T} X_i$. We adopt the notation that $\mathbf{m}_1 = m$, so that m is the expected number of cells carried per period. Then the LP described in the previous subsection is equivalent to

$$\text{maximize } \sum_{j=0}^{h} \theta_j e^{sj}$$

$$\text{subject to } \sum_{j=0}^{h} \theta_j = 1,\ \sum_{j=0}^{h} j\theta_j = m.$$

which clearly has the solution $\theta_0 = (1 - m/h)$, $\theta_{ht} = m/h$ and $\theta_j = 0$ otherwise. Then as in (12)

$$\bar{\alpha}(m,h) = \frac{1}{st}\log\left[\left(1 - \frac{m}{h}\right) + \frac{m}{h}e^{sh}\right] = G(m,h)$$

This gives a total charge of

$$\begin{aligned} f(X) = {} & \frac{1}{s}\left\{\log\left[\left(1 - \frac{m}{h}\right) + \frac{m}{h}e^{sh}\right] - \left[\frac{e^{sh}-1}{me^{sh} + (h-m)}\right] m\right\} T \\ & + \frac{1}{s}\left[\frac{e^{sh}-1}{me^{sh}+(h-m)}\right] V \end{aligned} \tag{21}$$

where $V = \sum_{i=1}^{T} X_i$ is the total volume of cells carried. This is the charge that has been described by Kelly (1994a).

3.2 Markov decision process formulation

Suppose a connection is policed by leaky buckets with parameters $(\rho_k, \beta_k), k = 1, 2, \ldots, K$. Then the linear program of Section 3.1 may be formulated as a Markov decision problem, as follows. Let

$$Z_k(\tau) = \min_{i<\tau}\left\{\rho_k(\tau - i) + \beta_k - \sum_{j=i}^{\tau-1} X_j\right\}.$$

Thus $0 \leq Z_k(\tau) \leq \beta_k, k = 1, 2, \ldots, K$. Let the state of the Markov decision process at time τ be

$$(Z_k(\tau), k = 1, 2, \ldots, K; X_{\tau-1}, X_{\tau-2}, \ldots, X_{\tau-t+1}).$$

The action allowed at time τ is to choose X_τ in the region

$$0 \leq X_\tau \leq \min_k \{Z_k(\tau) + \rho_k\},$$

with associated reward $\exp[s(X_\tau + X_{\tau-1} + \cdots + X_{\tau-t+1})]$. Note that if $t, \beta_k, \rho_k, k = 1, 2, \ldots, k$, are all rational, and X_j is integral, then the Markov decision process has a finite state and action space, and can hence be written as a finite linear program (Ross (1983), Chapter 5.3; Whittle (1983), Chapter 32.3).

For example, if $(s, t) = (1, 2)$ and a source is policed by a single leaky bucket with parameters $(\rho, \beta) = (1, 2)$, then the worst case traffic takes the form of a periodic sequence with repeated blocks of the form, 00131. If $(s, t) = (1, 4)$ and a source is policed by two leaky buckets with parameters $(\rho_1, \beta_1) = (3, 0)$ and $(\rho_2, \beta_2) = (2, 2)$, then the worst case traffic takes the form of a periodic sequence with repeated blocks of the form, 0223322.

Note that the blocks have the shape of an inverted `T`. Doshi (1994), see also Worster (1994), has previously noted that an `L` shape may, for sources policed by a leaky bucket, be more difficult to multiplex than a source taking two levels.

More complex optima may arise, but fortunately a simple bound exists.

3.3 A simple bound

A constraint $0 \leq X[0, t] \leq \bar{X}[0, t]$ together with the convexity of the exponential function implies that

$$\begin{aligned} \alpha(s, t) &\leq \frac{1}{st} \log \left[1 + \frac{tm}{\bar{X}[0, t]} \left(e^{s\bar{X}[0,t]} - 1\right)\right] \\ &= G(m, \bar{X}[0, t]/t), \end{aligned}$$

where $G(m, h)$ is defined in equation (12) and m is the mean rate. If a source is policed by leaky buckets with parameters (ρ_k, β_k), $k = 1, 2, \ldots, K$, then we have

$$X[0, t] \leq \min_k \{\rho_k t + \beta_k\},$$

and so

$$\bar{\alpha}(m, \mathbf{h}) \leq G(m, \min_k \{\rho_k + \beta_k / t\}).$$

3.4 Multiple constraints

The paradigm described in Section 2 applies in some interesting circumstances in which the acceptance region is described by multiple constraints.

Example: priority queueing

It is sometimes the case that one will want to give different qualities of service to different classes of traffic. One way to do this is by priority queueing.

Suppose traffic classes are partitioned into two sets, J_1 and J_2. Service is FCFS, except $i \in J_1$ is always given priority over $j \in J_2$.

For $i \in J_1$ there is a QoS guarantee on delay of the form:

$$\mathbb{P}(\text{delay} > B_1/C) \leq e^{-\gamma_1} .$$

For all sources there is a QoS guarantee on cell loss rate:

$$\mathbb{P}(\text{buffer overflow}) \leq e^{-\gamma_2} .$$

This gives two constraints:

$$\sum_{j \in J_1} n_j \alpha_j(s_1, t_1) \leq K_1 \tag{22}$$

$$\sum_{j \in J_1 \cup J_2} n_j \alpha_j(s_2, t_2) \leq K_2 \tag{23}$$

where

$$K_1 := C + \frac{1}{t_1}\left(B_1 - \frac{\gamma_1}{s_1}\right) \qquad K_2 := C + \frac{1}{t_2}\left(B - \frac{\gamma_2}{s_2}\right) .$$

and the s_i, t_i are the appropriate extremising values.

For example, suppose $J_1 = \{1\}$, $J_2 = \{2\}$. Then we have

$$n_1 \alpha_1(s_1, t_1) \leq K_1$$

$$n_1 \alpha_1(s_2, t_2) + n_2 \alpha_2(s_2, t_2) \leq K_2$$

If $K_1/\alpha_1(s_1, t_1) < K_2/\alpha_2(s_2, t_2)$ then the region in which the network provider can expect to operate is illustrated in Figure 3.

Suppose a network operator charges f_i per unit time for a connection of type i, $i = 1, 2$. The revenue $f_1 n_1 + f_2 n_2$ is maximized by operating, if possible, at some point on the boundary

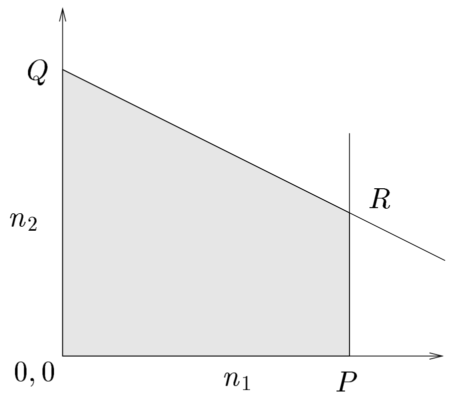

Figure 3: An acceptance region defined by two constraints

QRP. The operating point will be determined by issues such as the price sensitivity of the demand for the two types of traffic.

However, whatever the nature of the demand, there will be shadow prices λ_1 and λ_2 associated with relaxation of the constraints (22) and (23) respectively. If (22) is active (i.e., the network operates at capacity constrained by PR), then it will be appropriate to charge type 1 connections an amount which bounds $\lambda_1\alpha_1(s_1,t_1)$. If (23) is active then it will be appropriate to charge both type 1 and type 2 connections, at prices which bound $\lambda_2\alpha_1(s_2,t_2)$ and $\lambda_2\alpha_2(s_2,t_2)$ respectively. If operation is at point R then type 1 connections should incur a total charge which bounds $\lambda_1\alpha_1(s_1,t_1)+\lambda_2\alpha_1(s_2,t_2)$.

In all these cases the charge is a linear function of the measurements $g_1(X),\dots,g_L(X)$. For example, suppose we charge simply for time and volume, as in Subsection 3.1, and operation occurs at R. The charge for a type 1 call should be

$$\begin{aligned} f_1(\mathbf{m},\mathbf{h}) &= \lambda_1\Big(a_1[\dots]T + b_1[\dots]V\Big) + \lambda_2\Big(a_2[\dots]T + b_2[\dots]V\Big) \\ &= a[\dots]T + b[\dots]V\,, \end{aligned}$$

where a_1, b_1 depend on m_1, h_1, s_1, t_1, and a_2, b_2 depend on $m_1, h_1, m_2, h_2, s_2, t_2$, and a, b are simply the appropriate linear combination of these coefficients. The point is that the form of the charge remains the same. The user can be offered a number of tariffs, that differ only in the weights placed on T and V.

As a further extension of this example, suppose that traffic from a source of type j consists of correlated streams of high and low priority traffic. Let

$$\alpha_j^{(k)}(s,t) = \frac{1}{st}\log\mathbb{E}\Big[\exp\big(sX_j^{(k)}[0,t]\big)\Big]$$

where $X_j^{(k)}[0,t]$ is the workload of priority k or higher produced by a source of type j over the interval $[0,t]$. Then the acceptance region is given by two constraints

$$\sum_j n_j \alpha_j^{(1)}(s_1, t_1) \leq K_1,$$
$$\sum_j n_j \alpha_j^{(2)}(s_2, t_2) \leq K_2.$$

Note that these reduce to constraints (22) and (23) respectively in the case where a source produces either high or low priority traffic, but not both.

Example: Brownian bridge model

Two constraints may arise in even simpler circumstances. For example, consider again the Brownian bridge model of Subsection 1.4. Then a charge based on the two effective bandwidths arising in the constraints (9) and (10) takes the form

$$\lambda_1 \rho_j + \lambda_2 \left(\rho_j + \rho_j^2 \frac{\gamma}{2B} \right)$$

for a connection of type j. Within our framework this might correspond to a case where there are no measurements, and each type j connection is policed by a constraint

$$X_j[t, t+1] \leq \rho_j \text{ for all } t.$$

(The worst case traffic subject to this form of sliding window constraint is a periodic stream, with bursts of size ρ_j at unit spaced times.)

4 A comparison of two charging schemes

An important issue for a charging scheme is its complexity. We expect that by taking more measurements the charge can be made to more faithfully reflect the effective bandwidth. In this section we consider a family of charging schemes and quantify their performance for a particular form of traffic.

4.1 Tax band charging schemes

We look at a family of schemes corresponding to choices of the matrix A in the example of Section 3.1. The simplest member of this family is is the one in which the charge is (21), i.e., based upon only time and volume. Other members of this family are refinements in which separate time and volume measurements are made for periods during which the source rate lies in one of ℓ bands.

We divide the time interval $[0,T]$ into T/t intervals of length t. For notational ease let $X_i = X[(i-1)t, it]$, $i = 1, \dots, T/t$. Choose thresholds, $h_0, \dots, h_t$, such that $-1 = h_0 < h_1 < \cdots < h_\ell = ht$, and define,

$$I_k = \{i : h_{k-1} < X_i \le h_k\}, \quad g_k(X) = \frac{1}{T/t}\sum_{i\in I_k} 1, \quad g_{\ell+k}(X) = \frac{1}{T/t}\sum_{i\in I_k} X_i\,,$$

for $k = 1, \dots, \ell$. So

$$\mathbb{E}g_k(X) = \sum_{j\in I_k}\theta_j, \quad \mathbb{E}g_{\ell+k}(X) = \sum_{j\in I_k} j\theta_j\,, \quad k = 1, \dots, \ell.$$

The LP is then

$$\text{maximize} \sum_{j=0}^{h}\theta_j e^{sj}$$

$$\text{subject to} \quad \sum_{j=0}^{ht}\theta_j = 1, \quad \sum_{j\in I_k}\theta_j = m_k\,, \quad \sum_{j\in I_k} j\theta_j = m_{\ell+k}\,, \quad k = 1, \dots, \ell.$$

This is easily solved to give

$$\bar\alpha(\mathbf{m}, h) = \frac{1}{st}\log\sum_{k=1}^{\ell}\left[\frac{m_{\ell+k} - m_k h_{k-1} - m_k}{h_k - h_{k-1} - 1}e^{sh_k} + \frac{m_k h_k - m_{\ell+k}}{h_k - h_{k-1} - 1}e^{s(h_{k-1}+1)}\right]$$

It follows from $a_k = \partial\bar\alpha(\mathbf{m}, h)/\partial m_k$ that

$$a_1 > a_2 > \cdots > a_\ell \qquad \text{and} \qquad a_{\ell+1} < a_{\ell+2} \cdots < a_{2\ell}\,.$$

Furthermore

$$a_k + (h_{k-1}+1)a_{\ell+k} = \beta e^{s(h_{k-1}+1)} \qquad \text{and} \qquad a_k + h_k a_{\ell+k} = \beta e^{sh_k}\,,$$

where $\beta = \frac{1}{st}e^{-st\bar\alpha(\mathbf{m},h)}$. The total charge takes the form

$$a_0 T + \sum_{i=1}^{T/t}\max_k\{a_k + a_{\ell+k}X_i\}\ .$$

One can check that the maximum on the r.h.s. occurs for k if $i \in I_k$, i.e., $h_{k-1} < X_i \le h_k$. This corresponds to a 'tax-band charging scheme', in that cells incur a cost per cell that depends upon in which of ℓ bands X_i falls. If $h_{k-1} < X_i \le h_k$, then the charge for those cells is

$$a_k + (h_{k-1}+1)a_{\ell+k} + (X_i - h_{k-1} - 1)a_{\ell+k} = \beta e^{sh_{k-1}} + (X_i - h_{k-1} - 1)a_{\ell+k}\,.$$

The case $\ell = 1$ corresponds to the charge of Section 3.1, in which the charge is based on only T and V. When we take $\ell > 1$ we are maximizing over a more restricted set and the expected charge is less. This is at the price of measuring more.

4.2 Discussion of the tax band schemes

Charging schemes which result from taking $\ell = 1$ and $\ell = 2$ are equivalent to bounding the effective bandwidth by using $\mathbb{E}e^{sX[0,t]} \leq \mathbb{E}\phi(sX[0,t])$, where $\phi = \phi_1$ is either single chord lying above e^{sx} (the case $\ell = 1$), or two chords $\phi = \phi_2$ lying above e^{sx} (the case $\ell = 2$), graphed against x in the three plots in Figure 4.

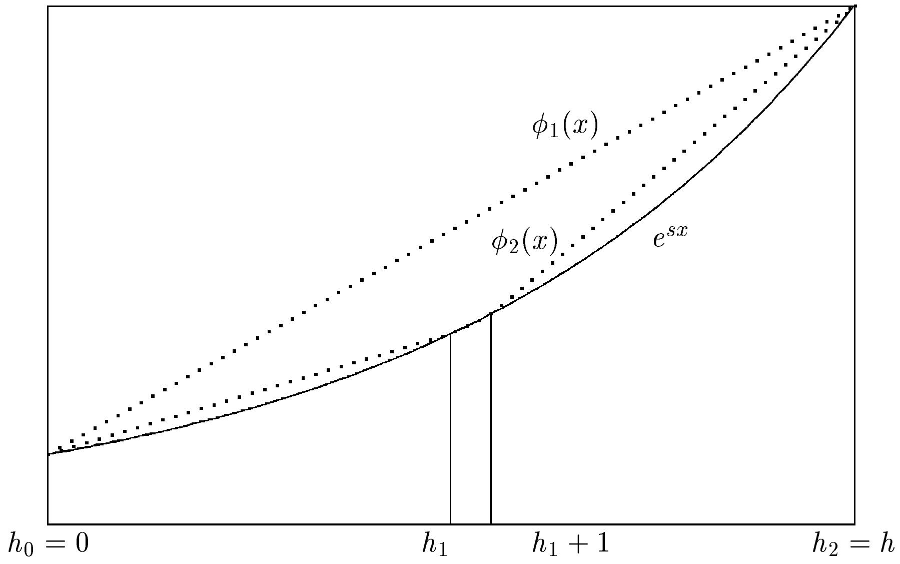

Figure 4: Bounds for tax band schemes with 1 and 2 bands

We have called these 'tax-band schemes' because of their similarity to the system of graduated (or banded) income tax which is operated in most countries. In such countries there is a charging period for income tax calculation, t, which is usually one year. Suppose for simplicity that there are only two tax bands, and that the tax for a year's income I is computed from the following table, in which $0 < a_3 < a_4 < 1$ and $h_1 > 0$.

income band	tax
$0 \leq I \leq h_1$	$a_0 + Ia_3$
$h_1 < I$	$a_0 + h_1 a_3 + (I - h_1)a_4$

One can think of a_0 as a fixed 'poll' tax which is paid equally by all taxpayers, irrespective of income. This leads to total tax over T years that can be written in the form

$$a_0 T + a_1 T_1 + a_2 T_2 + a_3 I_1 + a_4 I_2,$$

where T_1 and T_2 are respectively the numbers of these years in which the taxpayer does not or does pay higher rate tax, I_1 and I_2 are respectively his total income, summed over years of these types, and $a_1 = 0$, $a_2 = a_0 + h_1(a_3 - a_4)$.

Notice that a taxpayer who has a constant income in each of 40 years will pay less tax than a taxpayer who has the same total income over those 40 years but whose income varies above and below h_1 from year to year. While this anomaly is not desirable in a tax system (and indeed some countries allow taxpayers to average their income over several years to reduce this effect), it is precisely what we want, in the context of charging for variable rate traffic, so that a source that is more bursty will pay a greater charge.

Our analogous charging scheme for a connection is to divide the duration of the connection, T, into T/t intervals of length t, where its value is determined from the parameters of the system. In the examples we consider, the critical parameter t ranges from about 50ms to 1500ms. Each such interval will be our *charging* interval for the traffic stream.

Let us classify a charging interval as being of type I or II as the total volume of cells the source produces during that interval is either $\leq h_1$ or $> h_1$. the charging function takes the form

$$a_1[\ldots]T_1 + a_2[\ldots]T_2 + a_3[\ldots]V_1 + a_4[\ldots]V_2. \tag{24}$$

Recall that T_1 and T_2 are the total durations of intervals of types I and II respectively; so $T_1 + T_2 = T$. Similarly, V_1 and V_2 are the total volumes of cells generated by the source during intervals of types I and II respectively; so $V_1 + V_2 = V$. We have, as above, $a_1 > a_2$, $a_3 < a_4$. Then the charge is equivalent to

$$\sum_{i=1}^{T/t} \max\{a_1 t + a_3 v_i \, , \, a_2 t + a_4 v_i\},$$

where v_i is the number of cells which the source generates during the ith charging interval. The maximum is achieved by the first or second term as $v_i \leq h_1$ or $v_i > h_1$. Notice that the charge is an increasing and convex function of v_i.

4.3 Numerical comparison of the schemes

To illustrate these ideas, we give some numerical results for a theoretical model of a source that has been a popular testbed in other studies. This source alternates between on and off states according to a two-state Markov process. The mean duration of an on phase is 350 ms and the mean duration of an off phase is 650 ms. In the on phase the source produces cells at a constant rate of 64 Kbps, and in the off phase it is silent. Thus these sources have mean rate $.35(64) = 22.4$ kbps. Suppose the total bandwidth available to a number of such calls is 155 Mbps. On the basis of peak rate allocation, the switch could carry just $155,000/64 = 2,420$ connections, (or perhaps slightly less, due to cell scale effects; although these effects are very small for this model once the buffer is more than about 50 cells.) If there were an infinite buffer, and delay is not an issue, then it would be possible to carry $155,000/22.4 = 6,920$ connections.

Suppose the switch has a buffer of 200 ATM cells (= 84.4k bits, since a cell is 53 bytes.) Solving so that (2) gives about 17.75 (which we take as a target value as $e^{-17.75} \doteqdot 2 \times 10^{-8}$), we find $n = 6,350$ and the extremizing values in equation (2) of $t = 95$ms and $s = 0.027$.

Table 1 displays the results of similar calculations when the bandwidth which is available to these sources is reduced to 50% and 25% of the maximum (perhaps because some of the bandwidth is being used by constant bit rate sources). Other lines in the table show results for buffers of smaller and larger sizes.

B (cells)	C (Kbps)	n	γ	t ms	s (Kb)$^{-1}$
50	155,000	6,315	17.6	46	0.054
200	155,000	6,350	17.6	95	0.027
	77,500	3,075	17.6	116	0.032
	37,750	1,430	17.7	143	0.039
2000	155,000	6,505	17.5	359	0.008
10000	155,000	6,705	17.9	1332	0.003

Table 1: Values of s and t for various B and C.

Table 2 shows values of

$$(a)\ \frac{1}{st}\log \mathbb{E}e^{sX[0,t]} \quad (b)\ \frac{1}{st}\log \mathbb{E}\phi(sX[0,t]) \quad (c)\ \frac{1}{st}\log\left[\left(1-\frac{m}{h}\right)+\frac{m}{h}e^{sht}\right]$$

for $t = 150$ ms and $t = 1000$ ms. Here $\phi(x)$ denotes the piecewise linear function that bounds e^x for a tax-band scheme with two bands, (as shown in Figure 4), where the point at which the two bands are divided is chosen to minimize (b). The data for $t = 150$ demonstrates that there is not much difference for typical values of t and s. The data for $t = 1000$ demonstrates that the tax-band scheme can gives a substantially better approximation to the effective bandwidth when t is large. However, as we have seen, the circumstance in which it is appropriate to take t large (relative to the mean durations of on and off phases of a source) is when there is a large buffer; in this case s is small and there is not much difference between the two charging schemes.

$t = 150$ ms

s	(a)	(b)	(c)
0	22.4	22.4	22.4
0.001	22.46	22.46	22.47
0.01	22.97	23.00	23.11
0.05	25.34	25.51	26.03
0.1	28.42	28.75	29.82
0.2	34.48	35.09	37.03
0.5	47.04	47.74	50.20
1	54.79	55.18	57.00
2	59.23	59.38	60.50
5	62.05	62.08	62.60
∞	64	64	64

$t = 1000$ ms

s	(a)	(b)	(c)
0	22.4	22.4	22.4
0.001	22.56	22.63	22.87
0.01	24.04	24.73	27.29
0.05	31.05	34.34	44.46
0.1	38.80	42.75	53.53
0.2	47.94	50.49	58.75
0.5	56.61	57.36	61.90
1	60.15	60.36	62.95
2	62.04	62.09	63.48
5	63.21	63.21	63.79
∞	64	64	64

Table 2: Comparison of bounds (b) and (c) to the effective bandwidth (a)

However, we do expect to see more marked differences when there is a mixture of sources of different types and whose burstiness differs widely. For example, if there are also a small number of sources for which the mean on and off phases are 35 and 65 ms then the extremizing values of t and s in equation (2) will be much as before (since the proportion of these sources is small.) For these sources (and also for a source with even shorter on and off phases) the values of (a)–(c) at $t = 95$ and $s = 0.027$ are given in Table 3.

mean on, off	(a)	(b)	(c)
350, 650	23.47	23.51 (+ 0.2%)	23.62 (+ 0.6%)
35, 65	22.84	23.01 (+ 0.7%)	24.62 (+ 3.4%)
3.5, 6.5	22.44	22.57 (+ 0.6%)	24.62 (+ 5.3%)

Table 3: Over estimates of the effective bandwidth by tax-band schemes

Thus to charge for these sources simply on the basis of T and V results in 3.4% over-charging relative to their effective bandwidths, even when the most favourable tariff of this type is used. Users whose sources are of this type will prefer to be charged according to the tax-band scheme or to smooth their traffic in a small buffer so that it has a smaller peak rate.

For this model of Markov modulated on-off sources we can conclude that

1. typical values of t are 100–200 ms;

2. typical values of s are 0.01–0.04 per Kb;

3. the quantities of interest change slowly in s, t, so it is not important that these be determined very accurately;

4. as buffer or bandwidth increases, t increases and s decreases;

5. the charge based on T and V over-charges, as compared to a charge which is simply the effective bandwidth, by about 0.6%;

6. the charge based on a tax-band scheme using two bands over-charges by about 0.1%.

5 Discussion

It is important to appreciate the interface between usage charges and connection acceptance control (CAC). Given charges, demand rates, and a CAC policy, the network will carry various numbers of different services; these numbers will fluctuate, in response to randomly offered traffic, around some point on the boundary of an acceptance region, which is itself a function of the CAC policy. There will be a shadow price for the binding constraint at this point, and effective bandwidths for each type of traffic. For example, if the CAC is to accept a call if this leaves $\sum_i n_i h_i < C$, where h_i is the peak rate, then the effective bandwidth for traffic type i is h_i and the usage charge for traffic type i should be of the form λh_i. Note that operation takes place around a point at the boundary of the acceptance region; if it were taking place at an interior point then prices should come down, so as to admit more traffic.

In the case of real-time bursty sources subject to a non-zero cell loss rate guarantee, an efficient network operator will wish to take advantage of statistical multiplexing and use an acceptance region close to that described in this paper, i.e., $A = \{n_i : \sum_i n_i \alpha_i(s,t) < K\}$.

In our approach the CAC which is consistent with our charge is the acceptance region $\bar{A} = \{n_i : \sum_i n_i \bar{\alpha}_i(\mathbf{m}, \mathbf{h}) < K\}$, where $\bar{\alpha}_i(\mathbf{m}, \mathbf{h})$ is an upper bound on the effective bandwidth, subject to knowledge of static parameters and a priori estimates of measured parameters. The

CAC treats two sources as equivalent if, conditional on the a priori information available, their values of $\bar{\alpha}_i(\mathbf{m},\mathbf{h})$ are the same. A dynamic CAC that is also responsive to on-line measurement might conceivably make even more efficient use of network resources. However, within the class of CAC that we have described, the most efficient CAC is when users make accurate a priori estimates of the expected value of the measured parameter (11), so that $\bar{\alpha}_i(\mathbf{m},\mathbf{h}) = \alpha_i(s,t)$ and the acceptance region is A. Generally, users know less, we measure less, and $\bar{A}$ lies strictly within A. However, as our numerical examples have shown, fairly simple measurements can be enough to ensure that $\bar{A}$ is almost as large as A.

We have developed results in the context in which the traffic mix is fixed at some operating point. In practice, the traffic mix will fluctuate around this point, which may itself be different at different times of day. Different constraints will be binding at different operating points, and so usage charges will differ. If the binding constraint is a single link then the usage charge at that time of day is to be computed from the effective bandwidths that hold at that link. If the traffic passes along two links, each of which is filled to capacity, then the usage charge should be found from a weighted sum of the effective bandwidths, where the weights are the shadow prices of the link constraints. Sometimes traffic passes through two or more networks (e.g., a local one and a wide area one). The local network might be uncongested, resulting in a charge with only a fixed-charge component; the wide area network might be congested, resulting in a charge with both fixed and usage charge components. The user should be charged for the networks he uses, and therefore to choose from a menu of tariffs will need to know which parts of his traffic is carried on one or other of the networks, or both.

Acknowledgement

The partial support of the Commission of the European Communities ACTS project AC039, entitled Charging and Accounting Schemes in Multiservice ATM Networks (CA$hMAN), is acknowledged.

We are grateful to Peter Austing and John de Sa for solving the MDP examples in Section 3.2, to Meena Lakshmanan for help with the calculation of the tables of Section 4.3, and to George Stamoulis and Vasilios Siris for helpful discussions.

References

Botvich, D. D. and Duffield, N. (1995) Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems* **20**, 293–320.

Courcoubetis, C., Fouskas, G. and Weber, R. R. (1995) An on-line estimation procedure for cell-loss probabilities in ATM links. In *Proceedings of the 3rd IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, UK.*

Courcoubetis, C., Siris, V. A. and Stamoulis, G. D. (1998a) Application and evaluation of large deviation techniques for traffic engineering in broadband networks. In *ACM Sigmetrics '98*. Madison, Wisconsin.

Courcoubetis, C., Siris, V. A. and Stamoulis, G. D. (1998b) Integration of pricing and flow control for available bit rate services in ATM networks. In *Proceedings IEEE Globecom '96*, pp. 644–648. London, UK.

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

Cruz, R. L. (1991) A calculus for network delay. *IEEE Trans. Information Theory* **37**, 114–141.

Doshi, B. T. (1994) Deterministic rule based traffic descriptors for broadband isdn: worst case behavior and connection acceptance control. In J. Labetoulle and J. W. Roberts, eds., *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proceedings of the 14th International Teletraffic Congress — ITC 14*, volume 1a of *Teletraffic Science and Engineering*, pp. 591–600. Elsevier Science B.V. Antibes Juan-les-Pins.

Hajek, B. (1994) A queue with periodic arrivals and constant service rate. In F. P. Kelly, ed., *Probability Statistics and Optimization, a Tribute to Peter Whittle*, pp. 147–157. Wiley.

ITU Recommendation I371 (1995) Traffic control and congestion control in B–ISDN. Geneva, Switzerland.

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

Kelly, F. P. (1994b) Tariffs and effective bandwidths in multiservice networks. In J. Labetoulle and J. W. Roberts, eds., *The Fundamental Role of Teletraffic in the Evolu-*

tion of Telecommunications Networks, Proceedings of the 14th International Teletraffic Congress — ITC 14, volume 1a of *Teletraffic Science and Engineering*, pp. 401–410. Elsevier Science B.V., Amsterdam. Antibes Juan-les-Pins, France.

Kelly, F. P. (1996) Notes on effective bandwidths. In F. Kelly, S. Zachary and I. Ziedins, eds., *Stochastic Networks: Theory and Applications Telecommunications Networks*, volume 4 of *Royal Statistical Society Lecture Notes Series*, pp. 141–168, Oxford. Oxford University Press.

Kelly, F. P. (1997) Charging and rate control for elastic traffic. *European Transactions on Telecommunications* **8**, 33–37.

Kelly, F. P., Maulloo, A. and Tan, D. (1998) Rate control in communications networks: shadow prices, proportional fairness and stability. *J. Operational Res. Soc.* **49**. to appear.

MacKie-Mason, J. K. and Varian, H. R. (1994) Pricing the Internet. In B. Kahin and J. Keller, eds., *Public Access to the Internet.* Prentice-Hall, Englewood Cliffs, New Jersey.

Montgomery, M. and de Veciana, G. (1996) On the relevance of time scales in performance oriented traffic characteristics. In *Proceedings of IEEE INFOCOM '96*, pp. 513–520.

Ross, S. (1983) *Introduction to Stochastic Dynamic Programming.* Academic Press.

Shwartz, A. and Weiss, A. (1995) *Large Deviations for Performance Analysis.* Chapman and Hall, London.

Simonian, A. and Guilbert, J. (1995) Large deviations approximation for fluid queues fed by a large number of on-off sources. *IEEE J. Selected Areas Commun.* **13**, 1017–1027.

Varian, H. (1992) *Microeconomic Analysis.* Norton, New York, third edition.

Weiss, A. (1986) A new technique for analyzing large traffic systems. *Adv. Appl. Prob.* **18**, 506–532.

Whittle, P. (1983) *Optimization Over Time*, volume 2. Wiley, Chichester.

Wischik, D. (1998) The output of a switch, or, effective bandwidths for networks. Submitted for publication.

WORSTER, T. (1994) Modelling deterministic queues: the leaky bucket as an arrival processlarge deviations approximation for fluid queues fed by a large number of on-off sources. In J. Labetoulle and J. W. Roberts, eds., *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proceedings of the 14th International Teletraffic Congress — ITC 14*, volume 1a of *Teletraffic Science and Engineering*, pp. 581–590. Elsevier Science B.V. Antibes Juan-les-Pins.