Charging and Accounting for Bursty Connections

Frank P. Kelly

Abstract

Statistical sharing over several time-scales is a key feature of the Internet, and is likely to be an essential aspect of future ATM networks. In this chapter we describe how usage-sensitive pricing can encourage statistical sharing, and we provide a quantitative framework for the discussion of pricing issues in systems where statistical sharing is important.

In particular we describe a simple charging and accounting mechanism for real-time bursty connections, based on the concept of an effective bandwidth. The mechanism performs the dual role of conveying information to the network that allows more efficient statistical sharing, and information to the user about the resource implications of differently policed connection requests. The resulting tariff takes a strikingly simple form: a charge $a(x)$ per unit time, a charge $b(x)$ per unit volume of traffic carried, and a charge $c(x)$ per connection, where the triple $(a(x), b(x), c(x))$ are fixed at the time of connection acceptance as a function of the connection contract x. This form of tariff is also able to reveal user preferences concerning delay tolerant traffic, and thus promises to provide a unified pricing model over a wide range of quality of service classes.

1 Introduction

The definition of usage in a multi-service network is problematic, since the usage of a network resource may not be well assessed by a simple count of the number of bits carried. For example, to provide an acceptable performance to bursty sources with tight delay and loss requirements it may be necessary to keep the average utilization of a link below 10%, while for constant rate sources or sources able to accommodate substantial delays it may be possible to push the average utilization well above 90%.

What is needed is a measure of resource usage which adequately represents the trade-off between sources of different types, taking proper account of their varying characteristics and requirements. In recent years the notion of an effective bandwidth (Hui 1988, Gibbens and Hunt 1991, Guérin *et al* 1991, Kelly 1991) has been developed to provide such a measure. But while it is relatively easy to define quality of service requirements, the effective bandwidth of a source also depends sensitively upon the statistical properties of the source, and thus the issue becomes how much of the effort of statistical characterization should fall upon the network and how much upon the user responsible for a source.

Within the telecommunications and computer industries it is possible to discern two extreme approaches to this issue. One approach insists that a user provide the network with a full statistical characterization of its traffic source, which is then policed by the network. Another approach stresses the difficulty for a user of providing any information on traffic characteristics, and expects the network to cope nevertheless. These descriptions are, of course, caricatures. Note, though, that both approaches recognize the benefits of statistical sharing: they differ in how much characterization effort is required, and how this effort should be distributed over the combined system comprising users *and* network.

The correct balance will necessarily involve trade-offs between the user's uncertainty about traffic characteristics and the network's ability to statistically multiplex connections in an efficient manner. In this chapter we describe a charging mechanism that makes these trade-offs explicit, and show that it encourages the cooperative sharing of information and characterization effort between users and the network. An economist might describe the approach in terms of incentive-compatible tariffs in a stochastic environment. In circumstances where it may be inappropriate for the various charges to be converted into monetary units, the charges may instead be viewed as coordination signals that allow users to assess the impact of their actions on the network: thus a control engineer might describe the approach in terms of the hierarchical control through Lagrange multipliers of the entire system comprising users and network.

The organization of the chapter is as follows. In section 2 we review the concept of an effective bandwidth. In section 3 we consider the case of delay-sensitive connections, where statistical sharing is important over short time-scales, comparable or less than round-trip delay times across the network. We assume that such sources are policed, for example by leaky bucket regulators of the sort used to define the peak or sustainable cell rate parameters of an ATM traffic contract (ITU 1995). These policing parameters provide upper bounds on the behaviour of sources, but may not characterize sources well: for example sources may only occasionally need to burst at rates close to the bounds. If charges were based entirely on these bounds, then sources might as well fill the contract specification, and many of the benefits of statistical sharing over short time-scales would be endangered. In sections 3 and 4 we describe charging and connection acceptance mechanisms based on the traffic produced by a source, as well as any agreed policing parameters, a basis that can encourage statistical sharing. The advantages of statistical sharing are often lauded, but there is little recognition of just how effective it can be: it is quite possible to achieve high levels of statistical sharing *and* loss rates less than one in a billion, provided peak rates are policed and connection acceptance control makes use of both known peak rates and measured traffic volumes.

In section 6 we consider traffic less sensitive to delay. Statistical sharing is somewhat simpler for such traffic, since flows may be coordinated over time periods longer than round-trip delay times across the network, and since delay itself is available to convey control information. On the other hand the aims of pricing may be more complex, and may include providing feedback to users on network congestion and the revelation of user valuations (as discussed by MacKie-Mason *et al* 1996). We describe a simple scheme based on measurements of time and volume that can again achieve most of what could be expected of any scheme.

In section 7 we outline a unified pricing model suitable for a wide range of quality of service classes, described in terms of static parameters fixed for the duration of a connection and dynamic measurements of time and volume. This charging mechanism is currently undergoing trials, as part of the CA$hMAN project (Songhurst 1996a, 1996b).

2 Effective bandwidths

In this section we review the concept of an effective bandwidth, beginning with a simple model of statistical sharing at a transmission, switching or other scarce resource.

Suppose that J sources share a single resource of capacity C, and let A_j be the load produced by source j. Assume that $A_j, j = 1 \ldots, J$, are independent random variables, with possibly different distributions. Can the resource cope with the superposition of the J sources? More precisely, can we impose a condition on the distributions of $A_1, A_2, \ldots, A_J$ which ensures that

$$P\left\{\sum_{j=1}^{J} A_j > C\right\} \leq e^{-\gamma} \tag{2.1}$$

for a given value of γ? The answer to this question is, by now, fairly well understood. There are constants s, K (depending on γ and C) such that if

$$\sum_{j=1}^{J} B(A_j) \leq K\,, \tag{2.2}$$

where

$$B(A_j) = s^{-1} \log E\left[\exp(sA_j)\right]\,, \tag{2.3}$$

then condition (2.1) is satisfied. The expression (2.3) is called the effective bandwidth of source j. This result, originally due to Hui (1988), was extended (Kelly 1991) to show that if the resource has a buffer, and if the load produced by source j in successive time periods is a sequence of independent bursts each distributed as A_j, then the probability the delay at the resource exceeds b time periods will be held below $e^{-\gamma}$ provided inequality (2.2) is satisfied, with B again given by equation (2.3), where $K = C$ and $s = \gamma/(bC)$. It is by now known that for quite general models of sources and resources it is possible to associate an effective bandwidth with each source such that, provided the sum of the effective bandwidths of the sources using a resource is less than a certain level, then the resource can deliver a performance guarantee (see de Veciana and Walrand 1995, Courcoubetis and Weber 1996, and Kelly 1996 for recent results and reviews). Often the relevant definition is of the form

$$B(A_j) = (st)^{-1} \log E\left[\exp(sA_j[t])\right] \tag{2.4}$$

for particular choices of s and t where $A_j[t]$ is the arriving workload from source j over a random interval of length t. There may be several constraints of the form (2.2) corresponding to different physical or logical resources within a network.

For example, suppose a single resource gives strict priority to sources $j \in J_1$ which have a strict delay requirement, but also serves sources $j \in J_2$ which have a much less stringent delay requirement. Then two constraints of the form

$$\sum_{j \in J_1} B_1(A_j) \leq K_1\,, \qquad \sum_{j \in J_1 \cup J_2} B_2(A_j) \leq K_2 \tag{2.5}$$

will generally be needed to ensure that both sets of requirements are met, where B_1 and B_2 are calculated using different values of the space and time scales s and t appearing in expression (2.4) (cf. Bean 1994, de Veciana and Walrand 1995, Elwalid and Mitra 1995, Kelly 1996). If the less stringent delay requirement becomes very weak, corresponding to a very large buffer and almost no sensitivity to delay, then $s \to 0$ in (2.3), and

$$B_2(A_j) \to E(A_j)\,, \tag{2.6}$$

the mean load produced by source j. The second constraint of (2.5) then becomes the simple constraint that the mean loads of all sources should not exceed the capacity of the resource, the minimal constraint necessary for the queue to be stable.

Under specific modeling assumptions on sources it is often possible to refine the constraints (2.2), (2.5) by numerical computations (Gibbens and Hunt 1991, Elwalid and Mitra 1995). The effective bandwidths defined by the more refined constraints may no longer have the simple analytical forms (2.3) and (2.4), but share a qualitatively similar dependence on the statistical properties and performance requirements of the sources.

To develop some understanding of this dependence, let us consider the very simple case of an on/off source of peak rate h and mean rate m, for which

$$P\{A=0\} = 1 - \frac{m}{h}\,, \qquad P\{A=h\} = \frac{m}{h}\,. \tag{2.7}$$

The effective bandwidth (2.3) of such a source is then

$$B(h,m) = \frac{1}{s}\log\left[1 + \frac{m}{h}\left(e^{sh}-1\right)\right]\,, \tag{2.8}$$

and, with s replaced by st, this expression provides a bound on the effective bandwidth (2.4) of any source with peak rate h and mean rate m. For fixed h the function (2.7) is increasing and concave in m, while for fixed m it is increasing and convex in h. As $s \to 0$ (corresponding to a very large capacity C in relation to the peak h), the effective bandwidth approaches m, the mean rate of the source. However as s increases (corresponding to a larger peak h in relation to the capacity C) the effective bandwidth increases to the peak rate h of the source.

Next we consider how an effective bandwidth might be used for charging and connection acceptance control. One possible charging mechanism might assess the effective bandwidth of a connection, using an empirical average to replace the expectation (2.4), and then charge according to the assessment. Apart from the difficulty of interpreting this 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 (2.4) 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 are primarily concerned with expectations of *future* quality of service, the distinction matters.

An alternative charging mechanism might calculate the largest possible effective bandwidth subject to the agreed policing parameters, and charge accordingly. For example, the expression (2.7) might be evaluated with h and m

set equal to a peak and sustainable cell rate respectively. This tariff is certainly easier to explain, but severely penalizes users whose mean traffic may be unpredictable and not easily characterized by policing parameters.

Instead our approach is to regard the effective bandwidth as a function of both static parameters (such as the parameters of leaky bucket regulators) and dynamic parameters (such as duration and volume); to police the static parameters and measure the dynamic parameters; to bound the effective bandwidth by a linear function of the measured parameters, with coefficients that depend on the static parameters; and to use such linear functions as the basis for simple charging and connection acceptance mechanisms. In sections 3 and 4 we illustrate this approach.

3 Charging mechanisms

In this section we consider the case of delay-sensitive connections, where statistical sharing is important over short time-scales, comparable or less than round-trip delay times across the network. Examples include the variable bit rate service class of ATM standards, and some of the proposals for real-time services over the Internet.

3.1 Known peak rate, unknown mean rate

Consider first the case of an on/off source with a known (and perhaps policed) peak rate h, but with a mean rate that may not be known with certainty, even to the user responsible for the source. Assume, however, that the user has a prior distribution G for the mean rate M of the connection. The distribution G may represent very vague information, or might be constructed by recording past observed mean rates. Then the expected mean rate of the connection is

$$E_G M = \int_0^h x dG(x) .$$

If the network knew the prior distribution G for the mean rate M, then the network would determine the effective bandwidth of the connection, from equations (2.3) and (2.7), as

$$\begin{aligned} \frac{1}{s} \log E e^{sA} &= \frac{1}{s} \log E_G E\left(e^{sA} | M\right) = \frac{1}{s} \log E_G \left[1 + \frac{M}{h}\left(e^{sh} - 1\right)\right] \\ &= \frac{1}{s} \log \left[1 + \frac{E_G M}{h}\left(e^{sh} - 1\right)\right] . \qquad (3.1) \end{aligned}$$

But expression (3.1) is just the effective bandwidth if M is not random, but identical to its mean value under G. We see that since the source is on/off with known peak rate the network need only know $E_G M$, the user's expected mean rate; further detail about the distribution G does not influence the effective bandwidth, and would be superfluous for the network to even request.

How, then, should the network encourage the user to assess and to declare the user's expected mean rate? We next investigate whether the charging mechanism might be used to provide the appropriate amount of encouragement.

Suppose that, before a connection's admission, the network requires the user to announce a value m and then charges for the connection an amount $f(m;M)$ per unit time, where M is the measured mean rate for the connection. We suppose that the user is risk-neutral and attempts to select m so as to minimize $E_G f(m;M)$, the expected cost per unit time: call a minimizing choice of m, $\widehat{m}$ say, the best declaration for the user. What properties would the network like the best declaration $\widehat{m}$ to have? Well, first of all the network would like to be able to deduce from $\widehat{m}$ the user's expected mean rate $E_G M$. A second desirable property would be that the expected cost per unit time under the best declaration $\widehat{m}$ be proportional to the effective bandwidth of the connection (or, equivalently, equal to the effective bandwidth under a choice of units). In Kelly (1994a) it is shown that these two requirements essentially characterize the tariff $f(m;M)$ as

$$f(m;M) = a(m) + b(m)M\,, \tag{3.2}$$

defined as the tangent to the curve $B(h,M)$ at the point $M = m$ (see Figure 1). By a simple differentiation of the function (2.7), the coefficients in expression (3.2) are given by

$$b(h,m) = \frac{e^{sh}-1}{s\left[h + m\left(e^{sh}-1\right)\right]}\,, \qquad a(h,m) = B(h,m) - mb(h,m) \tag{3.3}$$

where we now make explicit the dependence of the coefficients on the peak rate h.

Observe that the choice of m simply labels the choice of a linear function (3.2), and that the presentation of tariff choices for a given peak rate h may be entirely couched in terms of pairs $(a(h,m), b(h,m))$, giving a charge per unit time and a charge per unit volume respectively, with no mention of the word "mean". It is not essential to provide the user with a continuum of tariff choices: the relevant functions may be well approximated by a small number of tangents, especially if the capacity C is large in relation to the peak rate h. Later, in section 4, we shall see that connection acceptance control is also primarily concerned with tangents, rather than their points of contact with effective bandwidth functions, and that any inaccuracies in the choice of tangent by a user cause the connection acceptance control to accept less, just as they cause the user to be charged more.

Thus, under the tariff f, the user has no incentive to "cheat", by choosing a tangent other than the tangent that corresponds to the user's expected mean rate. The property that the expected cost per unit time under the best declaration is equal to the effective bandwidth has several further incentive compatible properties: the benefit to a user in reduced charges of either shaping traffic to have a different peak or mean or of better characterizing traffic through improved prediction of statistical properties is exactly the expected reduction in the effective bandwidth required from the network. Thus users are not encouraged to do more work determining the statistical characteristics of their connections than is justified by the benefit to the network of better characterization.

If we do not insist on these further incentive compatible properties, but require only that the best declaration for the user be the user's expected mean rate then many tariff structures are possible: it would be sufficient for the family (3.2), as m varies, to form the envelope of a strictly concave function

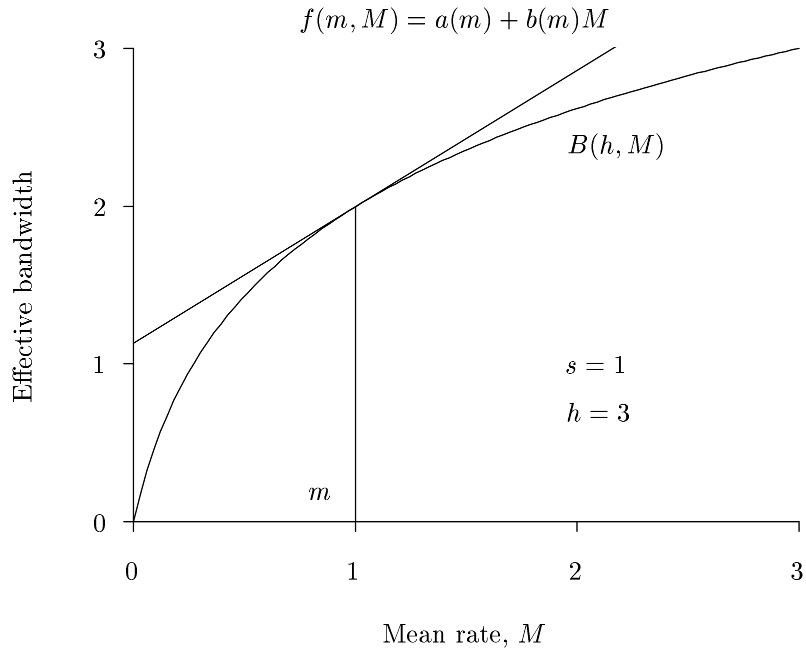

Figure 1: Implicit pricing of an effective bandwidth. The effective bandwidth is shown as a function of the mean rate, M, for a given peak rate h. The user is free to choose a tangent to this curve, and is then charged an amount $a(m)$ per unit time, and an amount $b(m)$ per unit volume.

	Rate (Mbps)		Charge	
Service type	Peak	Mean	fixed (s^{-1})	variable (Mbit^{-1})
1	0.1	0.04	2.7×10^{-4}	1.0
2	2.0	0.02	1.3×10^{-4}	1.4
3	10.0	0.01	1.1×10^{-3}	7.9
	h	m	$a(h,m)$	$b(h,m)$

Table 1: Typical charges for traffic with low mean rate

of M. We shall see later, in sections 4 and 5, that while connection acceptance control places a special emphasis on the particular concave function $B(h,M)$, yet further incentive issues may justify the simultaneous use of a modified function for tariffing.

3.2 A numerical example

We now illustrate the formulae (3.3) with a numerical example. Suppose that the predominant traffic offered to a link of capacity 100 megabits per second falls into three categories, with peak and mean rates as described in Table 1. Then the choice $s = 0.333$ in expression (2.3) is reasonable (Kelly 1994b). Note that almost all of the charge for these three service types arises from the variable charge $b(h,m)$.

While the predominant traffic may be of types 1, 2 and 3, connections are not constrained to just these types. For example, a connection with a known peak rate of 2 megabits per second could select any pair $(a(2,m), b(2,m))$ from Figure 2, or a connection with a known peak rate of 10 megabits per second could select any pair $(a(10,m), b(10,m))$ from Figure 3.

Similarly tariffs may be calculated for sources with other peak rates. For a peak rate of 0.1 megabits per second the bandwidth $B(h,M)$ is almost linear in M, producing a variable charge $b(h,m)$ per unit of traffic that is almost constant in m. Since statistical multiplexing is efficient for sources with such low peak rates, very little incentive need be given to determine mean rates accurately. Peak rates above 2 megabits per second produce more concave effective bandwidths and hence more incentive to accurately estimate the mean. The various charges shown in Tables 1 and 2 are expressed in the same units (of resource usage per second or per megabit) and are directly comparable with each other.

Observe that the total charge for service type 1 is higher than that for service type 2: for these service types at this resource statistical sharing is relatively easy, and the advantage of a lower mean rate outweighs the disadvantage of a higher peak rate. The total charge for service type 3 is, however, more than twice as high as for service types 1 and 2: statistical sharing becomes more difficult with a peak rate as high as 10% of the capacity of the resource. Observe that for the three service types shown in Table 1 almost all of the total cost to the user arises from the variable charge. For the service types shown in Table 2 much more of the total cost arises from the fixed charge, more than half in the case of service type 6.

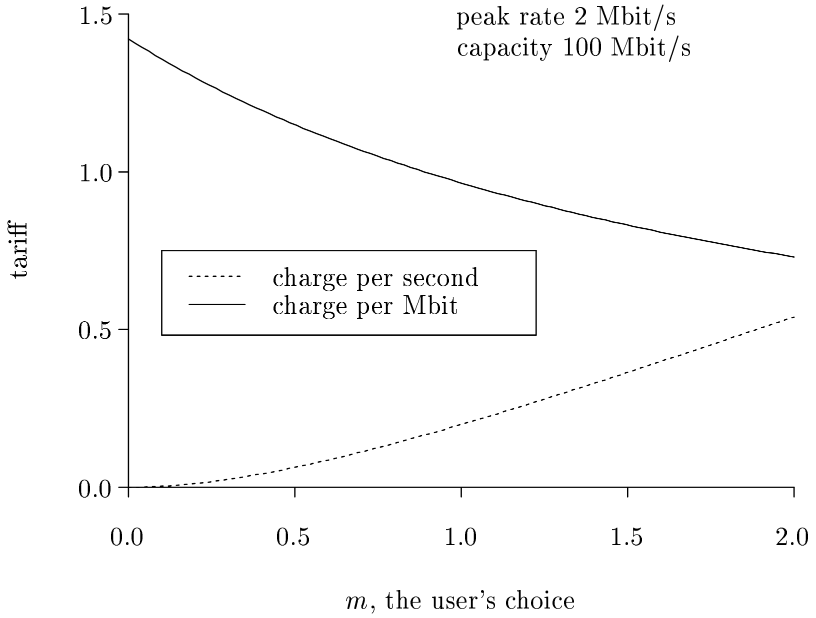

Figure 2: Tariff choices, for a peak rate of 2 Mbps. The user can choose a lower charge per megabit, with a higher charge per second.

	Rate (Mbps)		Charge	
Service type	Peak	Mean	fixed (s^{-1})	variable (Mbit^{-1})
4	2.0	1.0	0.2	1.0
5	10.0	1.0	1.7	2.2
6	10.0	2.0	3.0	1.3
	h	m	$a(h,m)$	$b(h,m)$

Table 2: Charges for traffic with higher mean rates.

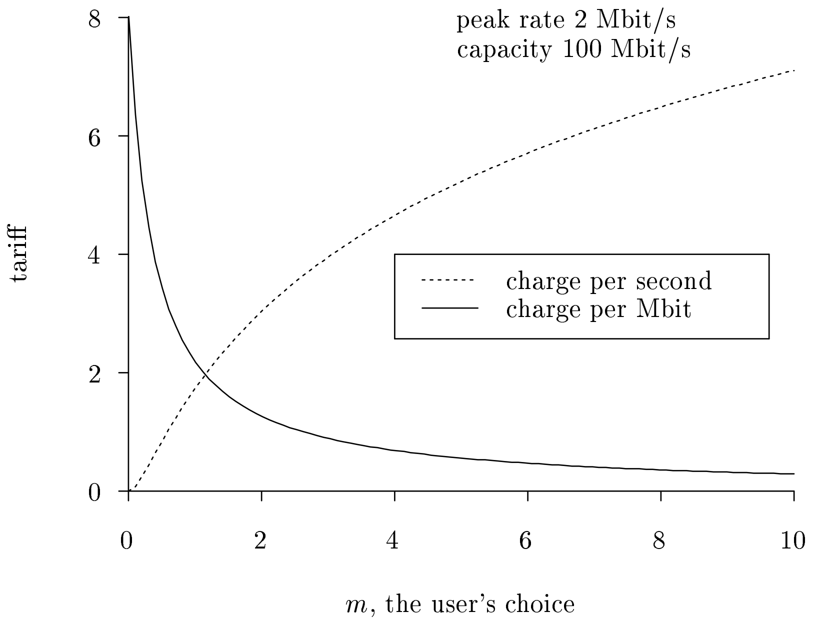

Figure 3: Tariff choices, for a peak rate of 10 megabits per second. The charge per second is typically higher with a peak rate of 10 megabits per second than with a peak rate of 2 megabits per second, since statistical sharing of the resource is harder.

The above charges can be compared with those that might be appropriate if the link were entirely loaded by delay-insensitive traffic (that is, traffic whose effective bandwidth approaches its mean rate, (2.6), and for which the average utilization could reasonably exceed 90%). A variable charge of about 0.5 per megabit would recover from such traffic about the same total revenue as can be recovered from a link loaded with high-priority traffic from the categories shown in Table 1. Note that delay-insensitive traffic does not directly substitute for high-priority traffic until its volume exceeds a certain level, about 50 megabits per second for the numerical example of this sub-section. There are two distinct constraints in (2.5): when the first constraint is tight and the second is not then the link is unable to accept any more high priority traffic and yet is capable of accepting more delay-insensitive traffic.

3.3 Unknown peak and mean rate

Next consider the case of an on/off source where the peak rate as well as the mean rate may not be known with certainty. Let G be the joint prior distribution for the peak rate H and the mean rate M. If the network knew the prior distribution then it would determine the effective bandwidth of the call, from equations (2.3) and (2.7), as

$$\frac{1}{s}\log E_G e^{sA} = \frac{1}{s}\log E_G E\left(e^{sA}|H,M\right) = \frac{1}{s}\log E_G\left[1+\frac{M}{H}\left(e^{sH}-1\right)\right] .$$

The network thus needs to know just the quantity $E_G Z$ where

$$Z = 1+\frac{M}{H}\left(e^{sH}-1\right) . \tag{3.4}$$

Suppose the network charges an amount $f(z; M, H)$ per unit time, where z is chosen by the user, and M and H are subsequent measurements of the mean and peak rates respectively. In Kelly (1994a) it is shown that the optimal choice for the user is $z = E_G Z$, and with this choice the expected cost per unit time is equal to the effective bandwidth of the call, if and only if

$$f(z; M, H) = a_z + b_z Z \tag{3.5}$$

where the right hand side of equation (3.5) is the tangent to the curve $B(Z) = s^{-1}\log Z$ at the point $Z = z$.

We may rewrite the form (3.5), using the definition (3.4), as

$$f(z; M, H) = a_z + M b_z(H) \tag{3.6}$$

where

$$a_z = s^{-1}\left(\log z - z^{-1}\right) \qquad b_z(H) = \frac{e^{sH}-1}{szH} . \tag{3.7}$$

The form (3.6) is linear in the measured mean M, with a peak dependent charge of $b_z(H)$ per unit of carried traffic. The function $b_z(\cdot)$ is increasing and convex: hence a connection that is more uncertain about its peak rate will prefer to choose a higher value z, thus incurring a higher charge per unit time and a lower charge per unit of carried traffic.

The tariff structure (3.6)-(3.7) subsumes the simpler structure (3.2)-(3.3): if a user knows its peak rate H, but not the mean M, then its optimal choice of z under the structure (3.6)-(3.7) incurs charges identical to those incurred with an optimal choice of m under the structure (3.2)-(3.3). Note that although the function (3.5) is linear in Z, the function (3.6) is non-linear in H and M. Difficulties with the structure (3.6)-(3.7) become apparent for networks comprising multiple heterogeneous resources, where a distinct choice of z may be required for each resource: for the simpler structure (3.2)-(3.3) a single choice of m suffices.

4 Connection acceptance control

In this section we outline how the framework of sections 2 and 3 allows the development of connection acceptance control mechanisms sympathetic to statistical sharing. Again our emphasis will be on the case where statistical sharing is important over short time-scales, comparable or less than round-trip delay times across the network.

Suppose that a resource has accepted connections $1, 2, \ldots, J$, and write (a_j, b_j) for the coefficients $(a(h,m), b(h,m))$ describing the choice of a tangent (3.2), (3.3) to the effective bandwidth function of connection j (see Figure 1). Suppose that the resource measures the arriving workload $A_j[t]$ from connection j over a period of length t (the same length t as appears in the definition (2.4)), and let $M_j = A_j[t]/t$. Define the *effective load* on the resource to be

$$\sum_{j=1}^{J} (a_j + b_j M_j) \ . \tag{4.1}$$

Then a connection acceptance control may be defined as follows. A new request for a connection should be accepted or rejected according as the most recently calculated effective load lies below or above a threshold value, with the proviso that if a request is rejected then later requests are also rejected until either a short interval has elapsed or an existing connection has terminated.

If admitted calls have high peak-to-mean ratios, as for the service types in Table 1, then the above control may amount to comparison of $\sum_j b_j M_j$ with a threshold, where b_j is the variable charge for call j. It may seem that such a call admission control is naively straightforward: after all, the measurements M_j will be highly variable, according to whether the source is on, off or somewhere in between. In Gibbens *et al* (1995) it is shown that such a simple call admission control can be both robust and efficient: an example is analyzed where a 1 in 10^9 condition on loss rates is combined with a utilization at least 97% of that achievable when mean rates are known.

Both the charging mechanisms of section 2 and the connection acceptance control described above use bounding tangents to the effective bandwidth function. If the same tangents are used for both purposes then the effective load has a natural interpretation as an aggregate charge at the resource over a recent short period. But there is no necessity for identical tangents to be used for charging and for connection acceptance. Thus users choosing a small peak rate might be offered no further choice of tariff, so that for charging purposes the effective bandwidth function is bounded by a single tangent. In contrast, the resource might choose its tangent to the effective bandwidth function at the point where the mean rate is the long-term observed average for traffic with that peak rate. Or, if distinct effective bandwidth functions are used for charging and connection acceptance control, then the resource might still choose its tangent according to user's declaration of expected mean rate.

Distinct effective bandwidth functions might be appropriate for charging and connection acceptance control, since the two areas have quite different time-scales and requirements for precision. Connection acceptance control must use accurately calculated effective bandwidths, based on the buffer sizes, port speeds and other features of current hardware to make decisions on connections as they request connection, otherwise quality of service guarantees on loss rates may be compromised. While charges need to be precisely defined, they influence users' behaviour and software application design over much longer time-scales, where features of hardware may evolve. Thus tariff design might include consideration of the possible effective bandwidth functions appropriate to future hardware and network scale.

A fuller discussion of connection acceptance control would consider multiple constraints of the form (2.5), network routing, and the shadow prices associated with different physical and logical resources (cf. Kelly 1988). The analysis and implementation of dynamic routing schemes often use Lagrange multipliers or shadow prices for each of the internal resources of the network, but only certain aggregates of this network detail might usefully influence charges to users. For example, competition between network providers might appear to users as a choice between routes, and averaged congestion measures might motivate a predictable time-of-day element to charges for delay-sensitive connections. We shall see in Section 6 that there are other ways of providing dynamic feedback on congestion, for users able to respond to such feedback.

5 Further incentive issues

5.1 Adaptation

Until now we have supposed that the choice of tariff is made once, at the start of a connection. Might it be worthwhile for a connection to vary this choice over its duration? Changes in the tariff regime might be made over time intervals longer than round-trip delay times, but shorter than a connection. Such changes would incur additional control and complexity overhead, but might allow more efficient statistical sharing, through a more precise indication of short-term statistical characteristics. Where statistical sharing is easy we might expect there to be little benefit, but there may be circumstances where a connection can make good short-term predictions that might be helpful to the network.

It is relatively easy to describe a mechanism which allows these various trade-offs to be assessed. Suppose that a connection may change its choice of tariff regime at any time, but must pay a fixed charge c every time this choice is exercised. Equivalently we allow a connection to become a sequence of shorter duration connections, but charge c for the set-up costs incurred by the network for each connection. Then a user is able to make its own evaluation of the benefit of changing the choice of tariff, and to assess whether the benefits of short-term prediction of its load outweigh the set-up charge c.

5.2 Incentives to split traffic

The independence of loads produced by different sources is an essential feature of the analysis leading to the constraints (2.2), (2.5), and to the concept of an effective bandwidth. But might a user split a connection with a high effective bandwidth into multiple connections, such that the sum of the effective bandwidths is smaller, in order to pay less? This is an important issue: users should *not* be encouraged to produce correlated sources. Fortunately there seem to be several mitigating factors.

To understand why an incentive to split traffic may exist, consider an effective bandwidth of the form (2.7) for a resource with a fixed capacity C. This expression is more than doubled if the peak h and mean m are both doubled, reflecting the fact that statistical sharing becomes harder as the peak-to-capacity ratio increases. Conversely, if capacity were to increase, while peak and mean remained constant, then effective bandwidth would decrease, illustrating the economies of scale of statistical sharing itself. Finally, if all three of capacity C, peak h and mean m were to double, then the effective bandwidth would be exactly doubled (as is clear from scaling arguments for very many resource models: in expression (2.7) the space scale s would be halved.)

For connection acceptance control it is important to use effective bandwidths calculated for current resources, where the effective bandwidth function increases faster than linearly with the peak, for a given mean-to-peak ratio. But, as discussed in section 4, prices influencing longer-term user behaviour may need to consider the future network scale. In particular, larger users should not be deterred if their attraction to a network allows the capacity of the network to increase proportionately. Indeed if the cost of capacity were to rise *less* than linearly with capacity, there would be a clear rationale for having the effective bandwidth function used for pricing increase *less* than linearly with the peak, for a given mean-to-peak ratio. A simple compromise might be constructed as follows: calculate time and volume charges for a single "typical" peak-to-capacity ratio, and then scale these charges to be proportional to the peak for a given mean-to-peak ratio (so that $a(h,m) = ha(m/h)$, $b(h,m) = b(m)$.) Such charges lessen the incentive to split traffic and simplify the presentation of tariffs, while retaining a (weakened) incentive for a user to lower the peak rate of a connection.

As well as transmission and buffering resources, a connection also uses other resources of the network. We have already discussed, in section 5.1, set-up costs incurred per connection. Additionally, in an ATM network, a connection holds a virtual circuit indicator (VCI) for its duration. If a rent is charged for use of a VCI then the charge per unit time of the connection is increased, and this further lessens the incentive to split traffic.

Finally we note that there may be occasions when it is to the joint advantage of network and users for traffic to be split, for example if the network is able to route correlated traffic streams over disjoint paths.

5.3 Discussion

We have discussed several incentive compatibility issues in sections 3, 4 and 5, and seen that compromises may be necessary between the incentive given to users to lower peak rates and to split traffic, as well as between the effective bandwidth functions used for charging and connection acceptance control. Simplicity and predictability of tariffs are other important issues. For example, suppose that users choosing a small peak rate relative to capacity are offered no further choice of tariff, corresponding to the bounding of the effective bandwidth function by a single tangent. Then this tangent might be chosen with predictability in mind: thus the tangent to the effective bandwidth function at the point where the parameter M is equal to the peak rate or a sustainable cell rate has the property that the charge to a user is bounded above in terms of the peak rate or the sustainable cell rate respectively, without overly penalizing users whose mean traffic may be not easily characterized by policing parameters. Quite where the various compromises should be drawn may not yet be clear, but the theory of sections 2 and 3 does at least provide a framework for analysis of the several tradeoffs involved.

6 Delay insensitive traffic

Sections 3 and 4 have concerned delay-sensitive traffic, and we have seen that charging schemes based on measurements of time and volume can encourage the coordination of users and the network, and allow statistical sharing over short time-scales. In this section we consider traffic less sensitive to delay, arising from applications that might be termed elastic (Shenker 1995). Statistical sharing is somewhat simpler for such traffic, since flows may be coordinated over time periods longer than round-trip delay times across the network, and since delay itself is available to convey control information. On the other hand the aims of pricing may be more complex, and may include providing feedback to users on network congestion and the revelation of user valuations. We shall describe a simple scheme based on measurements of time and volume that can again achieve almost all of what could be expected of any scheme.

We shall describe the scheme in terms of the available bit rate (ABR) service class of ATM standards. This service class assumes that a connection responds to rate control messages from the network, but that a user may choose a minimum cell rate (MCR) below which the connection will not be asked to fall. The essence of the scheme is that traffic up to the MCR is charged at one rate, while traffic above the MCR is charged at a lower rate. If a resource within the network has spare capacity beyond that required for high-priority and MCR traffic, then it may be shared amongst ABR connections, in proportion to their MCRs. Thus the choice of MCR by a user buys a share of spare capacity, as well as providing a minimum cell rate.

More precisely, we suppose that there is a charge of a times the chosen MCR per unit time, and additionally a charge of b per unit volume, where b may be zero. Thus traffic above the chosen MCR is charged at a lower rate, possibly a substantially lower rate. To illustrate the properties of this scheme, consider a typical ABR application, such as a file transfer of a given size. By choice of MCR a user can obtain an upper bound on the time taken to transfer the file, although the user would expect a much faster transfer if the network were lightly loaded. Note the important feature that both the time taken to transfer the file *and* the total charge to transfer the file will be larger when the network is congested, since the higher charge a applies to a larger volume of the transfer. Users may of course complain that they are charged more for a slower service, but this is the key characteristic of any incentive-compatible scheme designed to ease congestion. At times of congestion the user can speed up file transfers by increasing the chosen MCR for connections: each user is able to act according to its own trade-off between delay and cost.

Note that users and network can achieve a coordinated response to congestion without the need for the charges a and b to depend upon the level of congestion or even upon factors such as the time of day. The key point is that for traffic that is not highly delay sensitive, both price and delay are available as coordination signals (cf. Shenker *et al* 1996). The scheme described here allows delay to carry feedback on network congestion to users; the charges a and b turn the delay signal into a price signal with many of the attractive properties of the "smart market" of MacKie-Mason and Varian (1995). In particular, the price signal encourages the revelation of user preferences.

There are several other ways to describe the above scheme, both practically and theoretically. The scheme closely resembles some of the existing tariffs for frame relay, where the committed information rate plays a similar role to the minimum cell rate. It can be described in terms of the effective bandwidths of section 2, where the charges a and b play the role of the respective shadow prices for the two constraints (2.7). It would be interesting to explore further relationships with the expected capacity service of Clark (1996), where packets tagged *out* might correspond naturally with traffic charged at the rate $b = 0$.

Ihe approach of sections 2-5 could be readily integrated with the above scheme to price traffic below the MCR, and this may be worthwhile if the statistical properties of users' traffic are such that MCR s are frequently not filled. Similarly the approach outlined in section 4 will have relevance for the statistical multiplexing over short time-scales of delay-tolerant traffic that is bursty within its rate control envelope. A fuller discussion of rate control for delay-tolerant traffic would depend upon implementation details, and particularly upon the allocation of buffering across the network, but it seems unlikely that this level of network detail could usefully influence the structure of tariffs.

7 A unified pricing model

The schemes of sections 3 and 6 are both based on time and volume measurements, and this allows a simple unified description. The essence of the pricing model is as follows. The cost of a connection is given by the expression

$$a(x)T + b(x)V + c(x) \tag{7.1}$$

where T is the duration of the connection (measured in seconds, hours or months), V is the volume of the connection (measured in megabits or gigacells), and x describes tariff choices allowed to the user by the network at the time of connection acceptance. The tariff choice x includes the service class (for example variable bit rate or available bit rate), the traffic contract parameters (such as the peak cell rate or minimum cell rate), and further choices which might allow a user to lower the "per unit time" rate $a(x)$ at the cost of raising the "per unit volume" rate $b(x)$, as described in section 3. The charge per connection is $c(x)$.

The above scheme thus makes a distinction between static and dynamic parameters of a connection. The static parameters contained in x may be many and varied, but are fixed for the duration of the connection: the dynamic parameters T and V are measured in real-time. The cost of a connection to the customer, and the connection acceptance decisions of the network, are based on both static and dynamic parameters, but in a very restricted manner: through linear functions of the dynamic parameters, where the coefficients in the linear functions are themselves functions the static parameters. Mathematically, linear functions are natural approximants to the effective bandwidth function: practically, linear functions are easier for users to interpret as per unit time and per unit volume charges, and are easier to implement in real-time connection acceptance control schemes.

For expositional convenience we have called a, b and c charges, and expression (7.1) the cost, but these are really measured in units of resource usage. The conversion of such units into monetary units and a final bill is likely involve many other factors, for example customer discounts and marketing promotions, connection and subscription charges, whose discussion seems not so inextricably linked with statistical sharing.

Currently the CA$hMAN project, a consortium of network operators, equipment manufacturers and theoreticians, is examining the implementation of the above and other approaches to the pricing of ATM networks. Key issues explored in the early experiments include the following (Songhurst 1996a, 1996b). Is it feasible for the time, T, and volume, V, measurements to be made in hardware? Where should the cost be calculated, and how can it be transferred around the network, both for accounting purposes and to provide feedback to the user? What is the response of customers to the tariff structure, and can a user or application designer make good use of the feedback on resource usage provided by charges? Can the information on statistical characteristics conveyed through tariff choices be of assistance to the connection acceptance mechanisms of the network, and should the intelligence required to make tariff choices be provided by a user, an application, or a device elsewhere in the network that is knowledgeable about the traffic patterns usually generated?

8 Conclusion

The major issues for usage-sensitive pricing may be categorized under four headings.

Incentive compatibility. Users of a multiservice network will make quality of service choices, and the network will enforce priorities. Pricing is the major method of communication between users and network concerning the consequences of these actions. For example, prices will generally give some incentive to shape traffic: it is important that this incentive be compatible with the efficient operation of the *entire* system, comprising users and network.

Accounting and transaction costs. Such costs are a major deterrent to usage-sensitive pricing schemes. A less obvious, but ultimately more significant, cost may be that poorly designed tariffs impede innovation or network usage, by making uneconomic potentially important classes of application. The Internet has allowed the organic development of applications in a highly distributed manner: well designed charging schemes should encourage similarly decentralized decisions about resource allocation.

Commercial reality. Pricing schemes must be stable under competition, and under resale. The various economies of scale and scope present in communication networks make this a challenging issue, where more work is clearly needed. The issue interacts with the major topic of this paper, since important economies of scale and scope are provided by statistical sharing.

Technical feasibility. Any usage-sensitive pricing scheme must be capable of implementation in hardware and systems: for example the time-scales imposed by speed of light constraints on round-trip delay times must be respected.

In this chapter we have described how usage-sensitive pricing can encourage the efficient statistical sharing of scarce network resources required by many users, and have outlined a time and volume based charging mechanism which is both incentive compatible and simple enough to integrate with network control. The theoretical basis for the mechanism provides some prospect that the disparate issues outlined above may be treated in a coherent way.

References

[1] Bean, N. 1994. Effective bandwidths with different quality of service requirements. In *IFIP Transactions, Integrated Broadband Communication Networks and Services* , V.B. Iverson (editor), Elsevier, Amsterdam, 241–252.

[2] Clark, D.D. 1996. A model for cost allocation and pricing in the Internet. *This volume.*

[3] Courcoubetis, C. and Weber, R. 1996. Buffer overflow asymptotics for a switch handling many traffic sources. *Journal of Applied Probability* 33.

[4] Elwalid, A. and Mitra, D. 1995. Analysis, approximations and admission control of a multi-service multiplexing system with priorities. In *Proc. IEEE INFOCOM*, 463–472.

[5] de Veciana, G. and Walrand, J. 1995. Effective bandwidths: call admission, traffic policing and filtering for ATM networks. *Queueing Systems* 20, 37–59.

[6] Gibbens, R.J. and Hunt, P.J. 1991. Effective bandwidths for the multi-type UAS channel. *Queuing Systems* 9, 17–28.

[7] Gibbens, R.J., Kelly, F.P. and Key, P.B. 1995. A decision-theoretic approach to call admission control in ATM networks. *IEEE J. Selected Areas Communication* 13, 1101–1114.

[8] Guérin, R., Ahmadi, H. and Naghshineh, M. 1991. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Selected Areas Communication.* 9, 968–981.

[9] Hui, J.Y. 1988. Resource allocation for broadband networks. *IEEE J. Selected Areas Communication* 6, 1598–1608.

[10] ITU Recommendation I371, 1995. Traffic control and congestion control in B-ISDN. Geneva, Switzerland.

[11] Kelly, F.P. 1988. Routing in circuit-switched networks: optimization, shadow prices and decentralization. *Advances in Applied Probability* 20, 112–144.

[12] Kelly, F.P. 1991. Effective bandwidths at multi-class queues. *Queueing Systems* 9, 5–16.

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

[14] Kelly, F.P. 1994b. Tariffs and effective bandwidths in multiservice networks. In *The Fundamental Role of Teletraffic in the Evolution of Telecommunication Networks*, J. Labetoulle and Roberts J.W. (editors). Elsevier, Amsterdam, 401–410.

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

[16] MacKie-Mason, J.K. and Varian, H. 1995. Pricing the Internet. In *Public Access to the Internet*, B. Kahin and J. Keller (editors). MIT Press. 269-314.

[17] MacKie-Mason, J.K., Murphy, L. and Murphy, J. 1996. The role of responsive pricing in the Internet. *This volume.*

[18] Shenker, S. 1995. Service models and pricing policies for an integrated services Internet. In *Public Access to the Internet*, B. Kahin and J. Keller (editors). MIT Press. 315-337.

[19] Shenker, S., Clark, D., Estrin, D. and Herzog, S. 1996. Pricing in computer networks: reshaping the research agenda.

[20] Songhurst, D. 1996a. Charging schemes for multiservice networks. In *Proc. 13th UK Teletraffic Symposium.* IEE, London.

[21] Songhurst, D. (editor) 1996b. Experiment design for round I. The CA$hMAN Consortium (http://www.isoft.intranet.gr/cashman/).

Acknowledgements

The support of the Commission of the European Communities ACTS project AC039, entitled Charging and Accounting Schemes in Multiservice ATM Networks (CA$hMAN), is acknowledged. This chapter has benefited from the questions and criticisms of participants at the Internet Economics Workshop, of members of the COST 242 project, and especially of colleagues in the CA$hMAN project, to all of whom I am grateful.

Author Information

Frank P. Kelly (f.p.kelly@statslab.cam.ac.uk) is with the Statistical Laboratory, University of Cambridge, and can be reached at 16 Mill Lane, Cambridge CB2 lSB, England.

Publication

In *Internet Economics*, Lee W. McKnight and Joseph P. Bailey (editors). MIT Press, 1996.