Accepted at IEEE/IFIP IWQoS'99, UCL, London, UK, May 31 - June 4, 1999

Managing and Pricing Service Level Agreements for Differentiated Services*

Costas Courcoubetis[a,b] and Vasilios A. Siris[a]
[a]Institute of Computer Science (ICS)
Foundation for Research and Technology - Hellas (FORTH)
P.O. Box 1385, GR 711 10 Heraklion, Crete, Greece
E-mail: {courcou,vsiris}@ics.forth.gr
[b]Department of Computer Science, University of Crete, Heraklion, Greece

Abstract: We present an approach to manage and price service level agreements (SLAs) for differentiated services that uses a simple upper bound for the effective bandwidth of the conforming traffic as a proxy for resource usage. The bound depends on the user's traffic profile (peak rate and token bucket descriptor). Usage charges for a specific time period are proportional to this proxy, and their calculation requires only measurements of volume. We discuss and present experimental results regarding the incentives and fairness of the proxy, which is required in order to achieve economic efficiency. An important feature of our approach is the simplicity of the user's procedure for selecting optimal token bucket parameters. Our approach is quite generic and can be applied to scheduling disciplines that enable the provision of multiple service classes with different levels of performance. Finally, we present a case study for two service classes, real-time and non-real-time, with actual Internet traces.

1 Introduction

Recent advances in IP networks such as differentiated services (diffserv) [11] and integrated services (intserv), and architectures such as multiprotocol label switching (mpls) [10] support, similar to ATM, services that involve a *traffic contract* or *service level agreement*[1] (SLA) between the user and the network. According to such an agreement, the network provides some level of performance for the part of the user's traffic that is within a traffic profile. A widely used descriptor for a user's traffic profile consists of a peak rate and a token (or leaky) bucket.

The provision of service level agreements with some performance guarantees is also supported by current technology of network devices (routers and stand-alone devices) through mechanisms such as priority queueing, class based queueing (CBQ), weighted fair queueing (WFQ), etc. Such capabilities enable these devices to offer a different service to specific traffic flows based on, e.g., physical port, source/destination address, and protocol.

A main focus of the differentiated services work in the IETF is on the definition of *mechanisms* that are the building blocks for offering different levels of service to different users [11, 3]. Specific issues that are important include their simplicity, scalability, and deployment. Although mechanisms are used to provide services, these notions are separated, i.e., there is an effort not to embed a specific set of services in the internal mechanisms of the Internet, as in the case of integrated services and ATM. Such a separation allows the actual definition of services to evolve without modifying the internal mechanisms.

Work on pricing of differentiated services, such as [2, 17], focuses more on architectural issues such as where charges are computed and how multicast sessions and receivers can be charged, and not on how to compute usage charges. On the other hand, the proportionally fair pricing work of [15, 16] ([8] considers modifications of the TCP protocol to implement such an approach) investigates the problem of pricing services targeted for elastic applications, i.e., applications that can modify their traffic rate according to the available bandwidth inside the network.

In this paper we present and investigate a framework for managing and pricing differentiated services that offer some level of performance guarantees. Our goal is to quantify the amount of resources used by an SLA so that the network manager can decide how many such contracts can be offered simultaneously, and also by pricing certain aspects of the SLAs, provide users the incentive to select traffic contracts that reflect their actual needs. The framework is quite generic and can be used with a variety of mechanisms for implementing differentiated services.

There is a close relation with pricing and managing connections in ATM networks. An interesting difference is that SLAs for differentiated services are of a more static nature and the level of performance guarantees can be loose. Hence admission control is less strict, and has the goal of ensuring an average level of performance.

Our approach is based on using a bound to the effective bandwidth [14] as a "proxy" for quantifying resource usage. This bound, called the "simple" bound [5], considers an on-off approximation of the input traffic with a peak rate which depends on the traffic contract parameters (peak rate and token bucket parameters), while keeping

*The early stages of this work were supported by the European Commission under ACTS Project CASHMAN (AC-039).

[1] A service level agreement is typically more general than a traffic contract and can include such things as network availability, level of technical support, etc. In this paper we use the two terms interchangeably.

the same mean rate. Usage charges for specific time periods are proportional to this proxy, and their calculation requires only measurements of volume.

This rest of this paper is structured as follows. In Section 2 we describe our model and discuss the requirements for economic efficiency. In Section 3 we describe the proxy for resource usage, the information a network posts and how the user can select optimal token bucket parameters, which minimize his charge. We also discuss important incentive and fairness properties of the proxy, which are required for economic efficiency. In Section 4 we present a case study with two service classes, real-time and non-real-time service, and in Section 5 we conclude the paper and identify issues for further research.

2 The model

The model we consider is that of a single link with some amount of capacity and buffer, see Figure 1. The link is shared by a number of users, each with his own service level agreement (SLA) with the network provider. In practice, users can correspond to aggregations of individual traffic flows of the same class, such as that of large organizations (e.g., universities). SLAs are managed, through admission control enforced by the access link manager, so that some level of performance or *Quality of Service* (QoS) is ensured. An SLA includes traffic parameters, which describe the user's traffic profile (constrains the amount of traffic the user can send), and performance parameters, which characterize the level of performance that the network promises to provide to the conforming part of the user's traffic. In our framework the QoS is specified as a maximum queueing delay that is satisfied by some percentage of the conforming traffic.

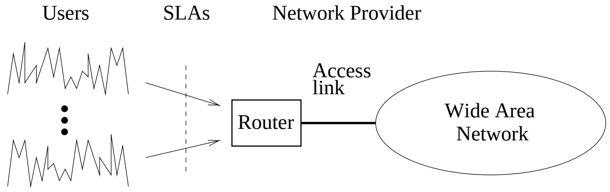


Figure 1: Access link of a Wide Area Network provider.

We continue with an informal discussion to define economic efficiency and motivate the use of effective bandwidths as a proxy for charging. SLAs with the same QoS are considered as being of the same type. For simplicity, we first consider that the link supports a single type of SLA, i.e., it supports a single service class with some target QoS.

Given an SLA with the vector of traffic contract parameters $\mathbf{x}_i$, user i's utility depends on these parameters and is a function $U_i(\mathbf{x}_i)$. The network's goal is to allocate SLAs to its users in a way that maximizes the sum of utilities (social welfare), while maintaining a given level of QoS. If i ranges over the set of contending users, this optimization can be written as

$$\max_{\{\mathbf{x}_i\}} \sum_i U_i(\mathbf{x}_i)$$

$$\text{subject to} \quad \sum_i \alpha(\mathbf{x}_i) \le K\,, \qquad (1)$$

where $\alpha(\mathbf{x}_i)$ is a measure of the resource usage, i.e., an effective bandwidth, consumed by the contract $\mathbf{x}_i$, and K is the effective capacity of the link (which depends on the capacity, buffer, and QoS). The above optimization problem is equivalent to

$$\max_{\{\mathbf{x}_i\}} \left\{ \sum_i U_i(\mathbf{x}_i) - p\left(\sum_i \alpha(\mathbf{x}_i) - K\right)\right\}\,,$$

where p is the shadow price of constraint (1). The first order conditions for this optimization are

$$\frac{\partial U_i(\mathbf{x}_i)}{\partial x_i^j} = p\frac{\partial \alpha(\mathbf{x}_i)}{\partial x_i^j}\,,$$

where j ranges over the components of the traffic contract. The last equation says that if the link is shared optimally, then there is a price p for which the user's marginal benefit of increasing some contract parameter x_i^j is equal to the price of the additional resources required. Hence, p represents a price per unit of effective usage $\alpha(\mathbf{x})$.

To achieve the social welfare optimum in a distributed manner, the network can post the price p and the function $\alpha(\mathbf{x})$, which is used to compute the amount of resources for contract $\mathbf{x}$. The usage charge per unit of time for a user with contract $\mathbf{x}$ will be $p\alpha(\mathbf{x})$. Observe that in the above formulation economic factors (e.g., demand, competition) are encoded in the price p, whereas technological factors (link resources, QoS, and service discipline) are encoded in the function $\alpha(\mathbf{x})$. Such an abstraction of technological factors is desirable since it allows the application of well-known economic results to networks employing statistical multiplexing and guaranteeing some level of performance.

A generalization for the case of a link offering two types of SLAs (two levels of QoS), each using dedicated resources, is the following[2]:

$$\max_{\{\mathbf{x}_i,\mathbf{y}_i\}} \sum_i U_i(\mathbf{x}_i,\mathbf{y}_i)$$

$$\begin{aligned}\text{subject to} \quad & \textstyle\sum_i \alpha_1(\mathbf{x}_i) \le K_1 \\ & \textstyle\sum_i \alpha_2(\mathbf{y}_i) \le K_2\,,\end{aligned}$$

where K_l for $l = 1,2$ is the effective capacity available for contracts of type l, and $\mathbf{x}, \mathbf{y}$ represent contracts of type 1 and 2, respectively. Here $\alpha_l(\mathbf{x})$ denotes the effective bandwidth of an SLA of type l with parameters $\mathbf{x}$. Similar to the single class case, at the optimum there exist prices p_1, p_2 that satisfy the following equations:

$$\frac{\partial U_i(\mathbf{x}_i)}{\partial x_{ij}} = p_1\frac{\partial \alpha_1(\mathbf{x}_i)}{\partial x_{ij}}, \qquad \frac{\partial U_i(\mathbf{y}_i)}{\partial y_{ij}} = p_2\frac{\partial \alpha_2(\mathbf{y}_i)}{\partial y_{ij}}\,.$$

[2] Simple arguments indicate that a similar approach can also be used for priority scheduling [14, 1].

In the above formulation, the prices p_l for $l = 1, 2$ are determined by the demand for services of type l. Note that if we can control the sharing of some fixed capacity $C = C_1 + C_2$ among the two service classes, then at the optimum we will have

$$p_1 \frac{\partial K_1}{\partial C} = p_2 \frac{\partial K_2}{\partial C} \, .$$

Hence if $\frac{\partial K_1}{\partial C} \approx \frac{\partial K_2}{\partial C}$ and the capacity is optimally shared between the two classes, then at the equilibrium the prices per unit of effective usage for the two classes will be close.

Interpretation of $\alpha(\mathbf{x})$

One can assign different interpretations for the effective bandwidth $\alpha(\mathbf{x})$. One alternative is to interpret $\alpha(\mathbf{x})$ as the effective bandwidth of the worst-case traffic subject to the traffic profile of the SLA. A disadvantage of such an approach is that resources are underutilized (the constraint (1) will be conservative), since it will typically be the case that the actual amount of resources used is much less than the maximum possible by the traffic contract.

A second alternative is to interpret $\alpha(\mathbf{x})$ as being the actual effective bandwidth. A disadvantage of this is that effective bandwidth expressions are in general complex functions requiring knowledge of detailed traffic statistics, unknown in most cases. Furthermore, a charge based directly on such a measure would be difficult for the network to construct and for the users to understand.

A possible solution is to use an approximation $\bar{\alpha}(\mathbf{x})$ of the actual effective bandwidth $\alpha(\mathbf{x})$. Such an approximation can depend on the traffic contract parameters (a priori information) and on simple measurements (a posteriori information), such as the mean rate. In this case we denote the above approximation by $\bar{\alpha}(\mathbf{x}, m)$, where m is the actual mean rate of the contract $\mathbf{x}$. Now the usage charge for a time period of duration T is $p\bar{\alpha}(\mathbf{x}, V/T)T$, where V is the volume transferred in that period, a quantity which can be easily measured.

2.1 Requirements for economic efficiency

An issue with the approach described in the last paragraph is that pricing in proportion to some arbitrary function $g(\mathbf{x}, m)$ for a contract $\mathbf{x}$ and mean rate m does not necessarily guide the system (network and users) to the economically optimal operating point that is achieved when pricing in proportion to the actual effective bandwidth. A proxy $\bar{\alpha}(\mathbf{x}, m)$ is *fair* if the variance of the ratio $\bar{\alpha}(\mathbf{x}, m)/\alpha(\mathbf{x}, m)$ is small, when $\mathbf{x}, m$ range over some interesting set of services (here $\alpha(\mathbf{x}, m)$ is the actual effective bandwidth). This implies that for such services $\bar{\alpha}(\mathbf{x}, m)/\alpha(\mathbf{x}, m) \approx k$, for some constant k. Pricing in proportion to $\bar{\alpha}(\mathbf{x}, m)$ is equivalent to pricing in proportion to $\alpha(\mathbf{x}, m)$, if we set the price per unit of $\bar{\alpha}(\mathbf{x}, m)$ equal to p/k. Hence pricing in proportion to a proxy $\bar{\alpha}(\mathbf{x}, m)$ that is fair can achieve economic efficiency.

A proxy for resource usage may be fair for *typical* users, as defined in the previous paragraph, but pricing based on it might not give users the incentive to remain typical. For example, pricing based solely on the traffic contract parameters does not discourage users from sending the worst-case traffic allowed by their contract. Hence, on a possibly long time scale, the users' traffic will change from typical to worst-case. Since users send more traffic than their actual needs, economic efficiency is not achieved. A way to remedy this is to account for actual usage instead of only the worst-case. Hence, when volume measurements can be obtained, $\bar{\alpha}(\mathbf{x}, m)$ should be defined as the worst-case effective bandwidth of the traffic resulting from contract $\mathbf{x}$, and having mean m. In this case, users are discouraged from increasing their mean rate, since this would increase their charge.

A final observation is that charges based on a subset of the traffic contract parameters can create substantial problems, by providing the wrong incentives to users to request unjustifiably "large" contracts. In addition to creating problems concerning the management of large contracts, in such cases users will be tempted to no longer remain typical. In Section 5 we describe such a pricing scheme.

In conclusion, for a given set of typical users, economic efficiency is achieved when the price per unit of actual effective bandwidth, which is proportional to the ratio $\bar{\alpha}(\mathbf{x}, m)/\alpha(\mathbf{x}, m)$, does not vary much. On the other hand, such a set of users will be unstable on a long time scale if there are non-typical flow and contract combinations that achieve a much lower price per unit of effective bandwidth. Since in such cases users end up sending more traffic than they actually need to, economic efficiency is not achieved.

3 Pricing and managing SLAs

In this section we discuss the basic components of our pricing and management scheme, namely the effective bandwidth bound that we use as a proxy for resource usage (Section 3.1) and its fairness (Section 3.4), the information posted by the network, and the pricing and management of SLAs (Section 3.2), and the user selection of traffic contract parameters (Section 3.3). We assume that the user's traffic contract $\mathbf{x}$ includes a peak rate h and a token bucket (ρ, β), where ρ is the token rate and β is the bucket depth.

3.1 A proxy for resource usage

Much research has been done on how to quantify resource usage in broadband networks. This research has shown that a stream's resource usage cannot be accurately quantified if the context of the stream (the link and the multiplexed traffic) is not taken into account. [14, 5] propose an effective bandwidth definition where the stream's context is encoded in just two parameters, the *space* and *time* parameters s, t, which depend on the link resources (capacity and buffer) and the characteristics of the multiplexed traffic. Specifically, the space parameter s (measured in, e.g., Mbit^{-1}) indicates the degree of multiplexing and depends, among others, on the size of the peak rate of the multiplexed streams relative to the link capacity: For links with capacity much larger than the peak rate of the multiplexed streams, s tends to zero and the effective bandwidth approaches the mean rate, while for links with capacity not much larger than the peak of the streams, s is large and the effective bandwidth approaches the peak rate measured over an interval of duration t. On the other hand, the time parameter t (measured in, e.g., seconds) corresponds to the most probable duration of the buffer busy period prior to overflow.

Investigations with real broadband traffic [7] have shown that the above effective bandwidth definition is quite accurate. These investigations have also shown that the parameters s, t are to a large extent insensitive to small variations of the traffic mix. Hence for given link resources and service discipline, pairs of s, t can be assigned to periods of the day during which the traffic mix remains relatively constant. The parameters can be computed offline from actual traffic traces. Related software and typical values of these parameters for various link capacities, buffer sizes, and types of traffic can be found in [18].

As a proxy for resource usage we consider a bound, the so-called "simple" bound in [5], which is a function of the mean rate m and traffic contract $\mathbf{x} = \{h, (\rho, \beta)\}$. This bound is given by

$$\bar{\alpha}(\mathbf{x}, m) = \frac{1}{st} \log \left[1 + \frac{m}{H(t)} \left(e^{stH(t)} - 1 \right) \right] , \qquad (2)$$

where $H(t) := \min\{h, \rho + \beta/t\}$ and m is the mean rate. $H(t)t = \min\{ht, \rho t + \beta\}$ is the maximum amount of traffic that can be sent in time interval t. Note that (2) corresponds to the effective bandwidth of an on-off fluid with peak rate $H(t)$ and mean rate m, for which the changes of the state occur much slower than the time scale t of buffer overflow. For the above reason we will refer to $H(t)$ as the *effective peak* and denote it simply by H. Finally, observe that $\bar{\alpha}(\mathbf{x}, m)$ is increasing in H and increasing and concave in m. An important issue is whether the above bound has the fairness property discussed in the previous section. We investigate this issue in Section 3.4.

3.2 Network functions for pricing and managing SLAs

The network posts the value of parameter t and a family of pricing curves $f_H(m)$ parameterized by H. The time parameter t, as we discuss in Section 3.3, can be used to simplify the selection of optimal token bucket parameters, i.e., parameters that minimize a user's charge. The pricing curves are given by $f_H(m) = p\bar{\alpha}(\mathbf{x}, m)$, where p is the price per unit of effective bandwidth and $\bar{\alpha}(\mathbf{x}, m)$ is given by (2) for particular values of H, s, t. If the user selects a contract $\mathbf{x} = \{h, (\rho, \beta)\}$, then he will be charged according to the curve $f_H(m)$ with $H = \min\{h, \rho + \beta/t\}$, and his charge for a time period of duration T will be $f_H(V/T)T$, where V is the volume transferred in that period.

There is an alternative charging scheme through which users provide the network provider with an estimate of their mean rate, which the provider can use to perform more effective admission control [12]. According to the scheme [13, 5], users select a tariff pair (a, b) from some set offered by the provider, and are charged using the simple formula $aT + bV$, where T is the duration of the charging period and V is the transferred volume. The tariffs (a, b) correspond to tangents to some bound of the effective bandwidth, for different values of the mean rate. A rational user will select the pair which minimizes the a priori expected value of his charge. Because the bound is concave in the mean rate, this value is minimized for the pair (a, b) which corresponds to the user's expected mean rate.

For differentiated services, as currently being defined by the IETF, the guarantees offered by the network can be loose. Hence, admission control can be performed in a less strict manner. Furthermore, connection setup for differentiated services is performed at the management level (service provisioning is on a much longer time scale), and admission control is performed on a longer time scale compared to admission control in a switched connection environment. In such environments, the information regarding a user's expected mean rate might be less important for the provider to know prior to accepting a user (since the provider can always measure it after admitting the user, and take it into account in future actions) compared to switched connection environments with stricter guarantees, such as ATM networks. In what follows, we will assume that the network posts pricing curves f_H and not a set of tariff pairs from which users select the pair according to which they will be charged. We note, nevertheless, that the two approaches correspond to a trade-off between simplicity of the charge computation and simplicity of the tariff negotiation.

Setting prices

The network sets the price p to reflect the demand for effective bandwidth. Since p corresponds to the shadow price of constraint (1), a direct approach is to measure the sum $\sum_i \alpha_i$ (for simplicity we use α_i to denote $\alpha(\mathbf{x}_i, m_i)$) and decrease the price p if the sum is smaller than the effective capacity K or increase the price p if the sum is larger than K. A practical alternative would be to directly measure the offered performance Q_m, and compare it with the target value Q_t. If $Q_m < Q_t$ then the price p is decreased, whereas if $Q_m > Q_t$ then the price p is increased.

We note that the above price adjustment occurs in very long time scales (months/years), hence prices are fixed for the whole duration of the service level agreement.

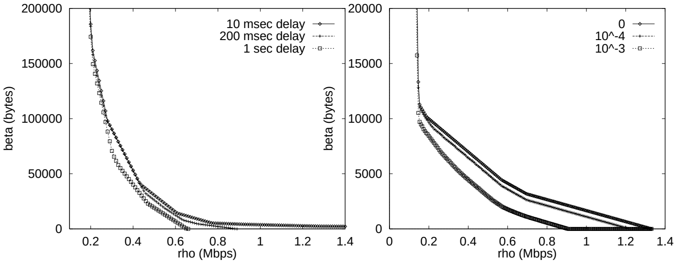
Managing SLAs

The above approach for setting the price p does not require the knowledge of the multiplicative factor k ($\approx \bar{\alpha}_i/\alpha_i$). Nevertheless, this factor can be estimated from measurements of $\bar{\alpha}_i$ and α_i, and its value can be used in admission control since the QoS constraint is satisfied if $\sum_i \bar{\alpha}_i \leq kK$. In this sense, k represents an *oversubscription factor*. The applicability of such an approach also requires that the proxy for resource usage is fair.

3.3 User functions for selecting traffic contract parameters

Given the information posted by the network, the user must select[3] his traffic contract $\mathbf{x} = \{h, (\rho, \beta)\}$, where h is the peak rate and (ρ, β) are the token bucket parameters. This will determine the value of the effective peak H, hence the pricing curve f_H according to which he will be charged. The selection of traffic contract parameters can be based on past measurements collected by the user. Furthermore, if it is possible to modify these parameters, a user can adjust them to better fit his traffic requirements as he collects new traffic measurements.

The choice of peak rate h depends on the amount of shaping that the user performs: A smaller peak rate results from a larger amount of shaping, hence the user's traffic incurs a larger delay before entering the network. We will assume that the user performs the largest amount of shaping, corresponding to some maximum shaping delay d that he can tolerate. Given a maximum shaping delay[4], which corresponds to some peak rate h, there will be pairs of (ρ, β) for which all of the user's traffic is conforming. These pairs form the indifference curve $G(h)$. Examples of such curves are shown in Figure 2(a).



(a) Indifference curves for various shaping delays.

(b) Indifference curves for various percentages of non-conforming traffic, and shaping delay 40 msec.

Figure 2: Indifference curves for various shaping delays (each corresponding to some peak rate) and percentages of non-conforming traffic ($0 - 10^{-3}$). Bellcore Internet WAN traffic[5].

Users can also select (ρ, β) so that some percentage of their traffic can be non-conforming. Examples of such indifference curves are shown in Figure 2(b). If the network offers a single packet loss rate, a user will typically select the percentage of non-conforming traffic to be of the same order of magnitude as the loss rate inside the network, since there are no gains in selecting a smaller percentage and a higher percentage results in worst performance.

For a particular shaping delay and percentage of non-conforming traffic, the indifference curve determines the set of pairs (ρ, β) from which a user can choose from. The specific choice will depend on how the network charges: A rational user will select the token bucket pair (ρ, β) that minimizes his charge. The structure of the simple bound (2) allows this selection to be performed *without* having to explicitly compute charges. Specifically, observe that (2) is increasing in H. Hence, the user can simply select the pair (ρ, β) that minimizes $H = \min\{h, \rho + \beta/t\}$. If the minimizer of the last expression is h, then the token bucket selection does not affect the charge. On the other hand, if the minimizer is $\rho + \beta/t$ then the pair (ρ^*, β^*) that minimizes H is given by the tangent to the indifference curve with slope $-t$, Figure 3.

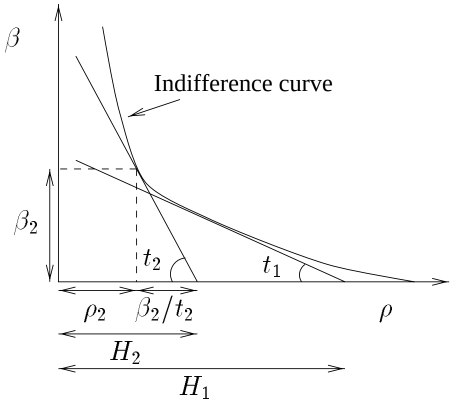


Figure 3: The optimal token bucket is given by the tangent to the indifference curve with slope $-t$. Observe that $H_1 > H_2$ for $t_1 < t_2$, where t_1 corresponds to a smaller buffer than t_2.

It is interesting to note that the above approach for determining optimal token bucket parameters is related to the interpretation of the time parameter t as the ratio of the marginal cost per unit capacity over the marginal cost per unit of buffer [5]. This interpretation has also been considered in [9] to guide users, or flows, to select the same ratio of token rate and bucket depth values. Such a selection makes it simpler to determine the total amount of resources required to simultaneously carry all the flows.

The network and user functions are collectively shown in Table 1. As noted previously, we assume that the user has selected the traffic he will send, hence his mean rate.

3.4 Incentives and fairness

Note that, in general, the unfairness of our effective bandwidth bound increases when users choose in an arbitrary way their token bucket parameters (ρ, β); a user's only requirement is that (ρ, β) is on or above his indifference curve. However, under the incentives provided through charging, although any choice of (ρ, β) does not affect the resulting traffic hence resource usage, choosing (ρ, β) rationally specifies a tighter traffic profile, i.e., a profile closer to the actual effective bandwidth. For such a selection

[3] We implicitly assume that the user has already selected the traffic he will send, hence his mean rate. In general, an organization can control its quantity of traffic (set of flows) using policy rules that specify the treatment of individual end-user and application flows.

[4] We assume that shaping is performed by averaging the amount of traffic in intervals of length d (shaping delay). This is one way of performing shaping; we are not assuming it is the best. Furthermore, d represents an upper bound on the maximum packet delay. The actual maximum and average delay is smaller. For example, when $d = 100$ msec the actual maximum delay is 47 msec while the average delay is less than 1 msec.

[5] Available from the Internet Traffic Archive at `http://www.acm.org/sigcomm/ITA/`

Network functions

1. Posts a family of curves $f_H(m)$ parameterized by H.
3. Posts the time parameter t.

User functions

1. Selects the maximum shaping delay; this determines his peak rate h.
2. Selects the percentage of non-conforming traffic; 1 & 2 determine the user's indifference curve G.
3. Selects the token bucket parameters (ρ, β) by considering the tangent to G with slope $-t$; this determines the value of $H = \min\{h, \rho + \beta/t\}$, hence the price curve f_H. His charge for a time period T will be $f_H(V/T)T$, where V is the volume (measured) transferred in that period.

Table 1: Network and user functions. We assume the user has traces of typical traffic. On a slower time scale, the network adjusts the price p based on the demand. The price curves are given by $f_H(m) = p\bar{\alpha}(\mathbf{x}, m)$, with $\bar{\alpha}(\mathbf{x}, m)$ given by (2).

of traffic profiles, experimental results indicate that the simple bound is fairer.

Another important parameter that affects a user's charge is the amount of traffic shaping he performs, which determines his peak rate. In many cases shaping does not affect the actual effective bandwidth. For example, observe in Table 2(a) that the actual effective bandwidth is not affected much when the peak rate decreases: A decrease of the peak rate from 2.34 Kbps to 0.28 Kbps results in a 5.8% decrease of the effective bandwidth (7.37 Kbps to 6.94 Kbps). On the other hand, the same decrease of the peak rate results in a 50.7% decrease of the simple bound (21.1 Kbps to 10.4 Kbps). Hence, one might ask why charges, such as the ones we propose, should provide the incentive to decrease the peak rate, even when the peak rate has a small effect on actual usage. The answer goes back to the requirements for economic efficiency that were discussed in Section 2.1: Providing such an incentive guides users to select traffic contracts for which the simple bound is fairer and smaller, hence corresponds to a tighter worst-case traffic bound.

The last argument is also supported by Tables 2(a), 2(b), and 2(c), which show the effective bandwidth and simple bound for different Internet traces. The bottom rows of these tables correspond to an average shaping delay less than 4 msec, which is acceptable for delay insensitive Internet traffic. Observe that for such shaping delays $k(\approx \bar{\alpha}_i/\alpha_i) \in [1.5, 2.3]$ which is smaller than if users did not have the incentive to decrease their peak rate, namely $[1.6, 2.9]$, $[1.6, 2.7]$, and $[1.6, 2.5]$ for the first three lines in the tables. Additional results regarding fairness are presented in [4], where we investigate the variance of k for various link capacities and buffer sizes, in the case of Internet traffic from the same source and MPEG-1 video traffic with various content.

peak(Mbps)	α(Kbps)	$\bar{\alpha}$(Kbps)
2.34	7.37	21.1
1.33	7.37	20.0
0.76	7.37	18.0
0.28	6.94	10.4

(a) Bellcore Internet WAN traffic

peak(Mbps)	α(Mbps)	$\bar{\alpha}$(Mbps)
8.89	1.45	3.65
5.84	1.45	3.53
4.75	1.45	3.42
3.39	1.13	2.64

(b) LBL TCP WAN traffic[5]

peak(Mbps)	α(Mbps)	$\bar{\alpha}$(Mbps)
11.88	3.16	4.90
9.16	3.16	4.90
6.69	3.16	4.90
4.48	1.83	3.48

(c) SDSC FDDI traffic[6]

Table 2: Effective bandwidth and simple bound. C = 34 Mbps, B = 63 × 10^3 Bytes (maximum queueing delay is approximately 15 msec) ($s = 17$ Mbit^{-1}, $t = 0.2$ s).

4 Case study: real-time and non-real-time services

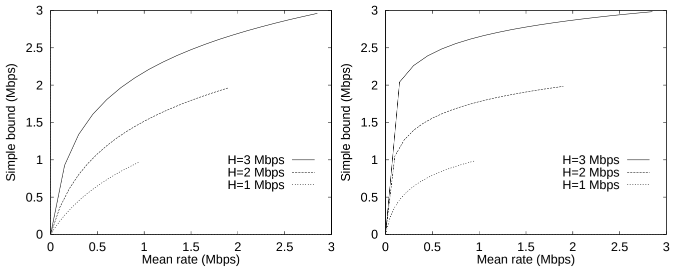
In this section we present a case study for real-time and non-real-time services, demonstrating the application of the pricing approach discussed in Section 3 for real network traffic. In addition to showing a typical family of curves for each of these services and making some observations regarding the application of our approach, we also present and discuss experimental results on the effects of the traffic mix and the scheduling discipline.

Figures 4(a) and 4(b) show a family of simple bound curves for real-time and non-real-time services, which ensure with probability 10^{-6} a maximum queueing delay D_1 = 4 msec and D_2 = 16 msec, respectively. Observe that as the effective peak H increases, the value and convexity of the simple bound increases. This is expected since a bursty stream (high peak) requires more resources than a less bursty stream (low peak).

Comparison of Figures 4(a) and 4(b) also shows that the values and convexity of the simple bound curve for the same effective peak H are *higher* for non-real-time service than for real-time service. Although this might at first seem counterintuitive, it can be explained as follows: For large buffers, the value of the time parameter t and product st increases[7]. This suggests, see [7], that for the overflow phenomena the traffic appears smoother.

[6] Available from NLANR at `http://moat.nlanr.net/Traces/Traces/`

[7] Formally, the higher values and convexity of the simple bound is due to the higher value of the product st which appears in (2).



(a) Real-time. $D_1 = 4$ msec, $s = 64$ Mbit^{-1}, $t = 0.02$ s.

(b) Non-real-time. $D_2 = 16$ msec, $s = 13$ Mbit^{-1}, $t = 0.24$ s.

Figure 4: Simple bound curves for real-time and non-real-time services in a link with $C = 34$ Mbps multiplexing Bellcore Internet WAN traffic. The probability of violating the respective delay is 10^{-6}. The curves are indexed by the effective peak $H = \min\{h, \rho + \beta/t\}$.

For this reason the aggregation of streams in a large buffer appears as a smooth (constant rate) stream. Hence, multiplexing these with a bursty stream, whose time scales of burstiness are slower than the time scale t of the buffer overflow, requires more additional capacity than in the case of a small buffer, where the aggregate traffic does not appear as smooth, hence there are gains due to statistical multiplexing.

It is important to note that the same values of H for real-time and non-real-time services do not necessarily correspond to the same traffic contract. Indeed, for the buffer sizes considered, the minimizer of $H = \min\{h, \rho + \beta/t\}$ is $\rho + \beta/t$, which depends on t. Hence, for a larger buffer, which corresponds to a larger value of t, the same traffic contract has a smaller effective peak $H = \rho + \beta/t$. Adding this last point to that of the previous paragraph, we see that larger buffer sizes give rise to two effects: (i) larger values of st (which push towards larger values of the simple bound (2)), and (ii) larger values of t, hence smaller values of the effective peak $H = \rho + \beta/t$ (which push towards lower values of the simple bound). Experiments for the buffer sizes considered show that the combination of these two effects results in lower values of the simple bound for larger buffers. This is shown in Figure 5. Of course, the charges will also depend on the prices, hence on the demand, for real-time and non-real-time services. However, as discussed in Section 2, if the capacity of a link is optimally shared between the two service classes, then the prices per unit of effective bandwidth will be approximately the same for both classes. In this case, the same contract will cost less for the non-real-time service than for the real-time service due to the smaller effective bandwidth, hence resource usage, for the former.

Figure 5 shows the simple bound for the same traffic contract for real-time and non-real-time services. However, as discussed in Section 3.3, the optimal token bucket parameters, hence the traffic contract, are not the same for different values of the time parameter t. In particular, as shown in Figure 3, larger values of t (which correspond to larger buffers) lead to smaller values of the effective peak H. Hence, for the same indifference curve one expects a smaller effective peak for a non-real-time service than for a a real-time service. This is shown in the first and second line of Table 3.

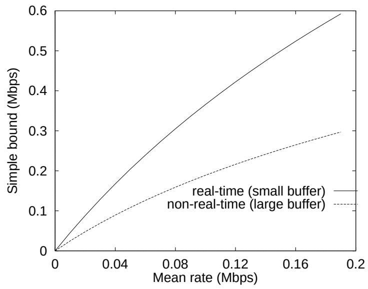


Figure 5: Simple bound for real-time and non-real-time services for the same traffic contract $h = 2$ Mbps, $(\rho, \beta) = (0.2 \text{ Mbps}, 10000 \text{ Bytes})$.

shap.	service	ρ(Mbps)	β(bytes)	H(Mbps)	$\bar{\alpha}$(Mbps)
20	r-t	1.43	1000	1.82	0.025
20	non-rt	1.43	1000	1.46	0.110
20	non-rt	0.16	11700	0.54	0.016
200	non-rt	0.15	9900	0.48	0.014

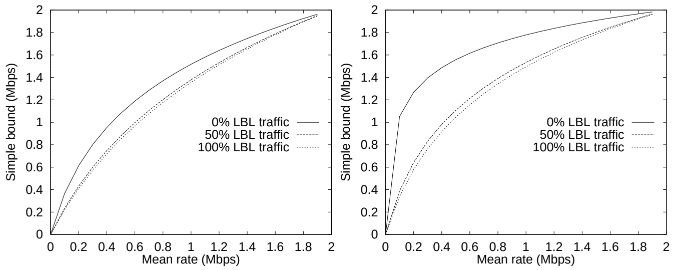
Table 3: Token bucket selection for real-time and non-real-time services. Comparison of the second and third line shows that the optimal selection of token bucket parameters has a large effect on the charge. (We assume that the price per unit of effective bandwidth is the same for both classes.) Comparison of the first and third line shows that transferring the same traffic with lower quality would incur a smaller charge. Finally, comparison of the third and fourth line shows the effect of shaping (the first column shows the shaping delay in msec. Bellcore Internet WAN traffic.

4.1 Effects of the traffic mix

Figure 6 shows the simple bound curves for effective peak $H = 2$ Mbps and various traffic mixes of Bellcore and LBL traffic. These figures show that the simple bound is not uniformly affected when the traffic mix varies. In practice, see Figure 6, traffic mixes with over 50% LBL traffic can be characterized by the same simple bound curve.

4.2 Effects of the scheduling discipline

Figure 4 shows the family of simple bound curves when each service class (real-time and non-real-time) has dedicated capacity 34 Mbps. Now we consider the case when the two service classes share capacity $C = 2 \times 34 = 68$ Mbps, with class 1 (real-time service) having priority over class 2 (non-real-time service). A minimum capacity $C_2 = 34$ Mbps is guaranteed for the non-real-time services (see [14, 5] for details on how our approach can be applied to priority queueing). Each service class guarantees the same QoS as before, namely maximum delay $D_1 = 4$ msec (real-time service) and $D_2 = 16$ msec (non-real-time service) with probability 10^{-6}.



(a) Real-time. $D_1 = 4$ msec. (b) Non-real-time. $D_2 = 16$ msec.

Figure 6: Simple bound curves for real-time and non-real-time services in a link with $C = 34$ Mbps multiplexing a mix of Bellcore and LBL traffic. The probability of violating the respective delay bound is 10^{-6}. $H = 2$ Mbps.

The family of simple bound curves for the real-time service are identical to those shown in Figure 4(a), while those for the non-real-time service are different, and are shown in Figure 7. Observe that the simple bound is lower in the case of shared capacity than in the case of dedicated capacity. This is due to the more efficient statistical multiplexing which results from sharing the total link capacity between the two service classes.

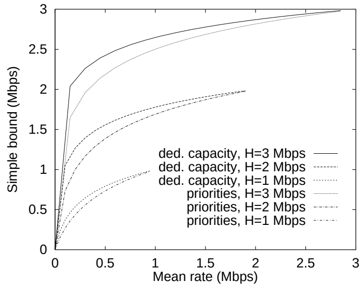


Figure 7: Non-real-time services for classes with dedicated capacity and for priority classes with shared capacity. A maximum queueing delay $D_2 = 16$ msec with probability 10^{-6} is ensured. For dedicated capacity $s = 13$ Mbit$^{-1}, t = 0.24$ s, and for shared capacity $s = 11$ Mbit$^{-1}, t = 0.2$ s.

5 Conclusions and further research

This paper has presented a framework for managing and pricing service level agreements (SLAs) for differentiated services that uses a simple upper bound for the effective bandwidth as a proxy for resource usage. This bound depends on the traffic parameters of the SLA and the mean rate of the traffic sent. Usage charges for a specific time period are proportional to the proxy, and their calculation requires measurements of the transferred volume. We have discussed and investigated the incentive and fairness properties of the proxy and how the network can set prices for various services. An important feature of our approach is the simplicity of the procedure for selecting optimal token bucket parameters. This procedure, along with the computation of indifference curves, can be performed on behalf of the user by an intelligent agent.

Charging in proportion to the mean rate is a special case of our approach when there is a large degree of multiplexing, for example due to large capacities or relaxed performance guarantees. In such cases, the effective bandwidth of typical users approaches their mean rate. For this reason, we expect that the mean rate may be fairer than a bound of the effective bandwidth (which accounts for worst-case users), hence more preferable for use as a proxy of resource usage. Furthermore, usage charges become *proportional* to the measured volume. Of course, for very large degrees of multiplexing, effective bandwidth bounds also tend to become linear in the mean rate, hence will be as fair as the mean rate. Charging according to the mean rate, however, does not discourage users from selecting large traffic contracts, which create problems for network management. The provider must use other means to limit the size of contracts, for example by setting upper limits for the values of the traffic contract parameters.

It is interesting to consider the properties of charging in proportion to the rate ρ of the token bucket, and not take into account the other contract parameters. Under such a pricing scheme, users will have the incentive to select a small value for ρ, hence ρ will be very close to the mean rate. Since in the case of large buffers the effective bandwidth of typical users is close to their mean rate, pricing in proportion to the token rate will be fair. However, the scheme does not discourage users from requesting large values for the peak rate h and bucket depth β, nor from sending the maximum traffic allowed by their traffic contract (such wrong incentives were discussed in Section 2.1). Since the contract for typical users has effective bandwidth much larger than their mean rate, this scheme cannot lead to economically efficient operation of the network. With the presence of increasing intelligence at the user end, such issues should be carefully considered.

Issues for further investigation include the extension of our framework to the case of a network. One approach can be to separately consider the national traffic (which typically traverses lightly loaded links) and international traffic (which typically traverses highly congested and expensive international links) of a large organization. Another issue is how to provide incentives to avoid traffic splitting. This is important because the underlying theory of statistical multiplexing assumes that the multiplexed traffic streams are independent.

We are currently looking into the application of our charging approach to a real networking environment, and in particular the link connecting a large organization (university) to a Wide Area Network provider (e.g., see [6]). Specific issues we plan to investigate include the following:

- Selection of token bucket parameters for different periods of the day. One would expect that these parameters change throughout the day, but are similar for the same periods of different days.
- Investigation of the oversubscription factor ($k \approx \bar{\alpha}_i/\alpha_i$) and fairness for different periods of the day,

and application of the oversubscription factor to admission control.

- Investigation of the link parameters s, t for different periods of the day, and creation of a library of simple bound curves for various link resources, scheduling disciplines, and traffic mixes. [18] is a step in this direction and contains related software.

References

[1] A. W. Berger and W. Whitt. Extending the effective bandwidth concept to networks with priority classes. *IEEE Commun. Mag.*, pages 78–83, August 1998.

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

[3] D. Clark and J. Wroclawski. An approach to service allocation in the Internet. Internet draft: draft-clark-diff-svc-alloc-00.txt, July 1997.

[4] C. Courcoubetis, F. P. Kelly, V. A. Siris, and R. Weber. A study of simple usage-based charging schemes for broadband networks. In *Proc. of IFIP Int. Conference on Broadband Communications (BC'98)*, Stuttgart, Germany, April 1998.

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

[6] C. Courcoubetis and V. A. Siris. Measurement and analysis of real network traffic. Technical Report No. 252, ICS-FORTH, March 1999.

[7] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Application of the many sources asymptotic and effective bandwidths to traffic engineering. To appear in *Telecommunication Systems*, 1999. A shorter version appeared in *Proc. of ACM SIGMETRICS'98/PERFORMANCE'98.*

[8] J. Crowcroft and P. Oechslin. Differentiated end-to-end Internet services using a weighted proportional fair sharing TCP. *Computer Communications Review*, 28(3):53–67, July 1998.

[9] N. G. Duffield and S. H. Low. The cost of quality in networks of aggregate traffic. In *Proc. of IEEE INFOCOM'98*, 1998.

[10] D. O. Awduche *et al.* Requirements for traffic engineering over MPLS. Internet draft: draft-ietf-mpls-traffic-end-00.txt, October 1998.

[11] Y. Bernet *et al.* A framework for differentiated services. Internet draft: draft-ietf-diffserv-framework-02.txt, February 1999.

[12] R. J. Gibbens and F. P. Kelly. Measurement-based connection admission control. In *Proc. of the 15th Int. Teletraffic Congress (ITC - 15)*, North Holland, 1997. Elsevier Science B. V.

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

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

[15] F. P. Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8:33–37, January 1997.

[16] F. P. Kelly, A. Maulloo, and D. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49, 1998.

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

[18] V. A. Siris. Large deviation techniques for traffic engineering.
`http://www.ics.forth.gr/netgroup/msa/`.