In Proc. of IEE Colloquium on "Charging for ATM - the Reality Arrives", London, UK, November 1997

COMPARING USAGE-BASED PRICING SCHEMES FOR BROADBAND NETWORKS[1]

C. Courcoubetis[2], V. A. Siris[2], and G. D. Stamoulis[2]

Tariff structures where charges depend on resource usage allow network operators to fairly recover costs from their customers. The notion of effective bandwidth is an important mathematical abstraction to the complex problem of quantifying resource usage. We consider usage-based pricing schemes which are based on some bound of the effective bandwidth and involve two measurements: the duration of a connection and the total volume transferred. The schemes are compared according to their fairness, i.e., their ability to capture the relative amount of resources used by connections. Our experiments involve a large set of MPEG-1 compressed video with different contents, and show that for links with a high degree of multiplexing, simple approximations can yield good performance.

1 Introduction

The increasingly competitive nature of the telecommunications market and steps toward deregulation, along with Internet's intense congestion problems, are pushing towards prices which take some account of actual resource usage. Usage-based pricing enables network providers to recover the costs of service provision from their customers in a fair manner. Of course prices for network services will be affected, to a large degree, by economic, regulatory, and marketing issues. However, it is unlikely that a pricing scheme that does not take into account resource usage is an effective mechanism for controlling congestion and for providing the right incentives for efficient and stable network operation. A prime example is that of the Internet, which faces intense congestion problems. Many engineers and economists believe that the Internet's congestion problems are due to its ineffective pricing structure, namely *flat rate* pricing, where prices depend only on the rate of

[1]This work was supported in part by the European Commission under ACTS Project CASHMAN (AC-039)

[2]ICS-FORTH and Department of Computer Science, University of Crete, P.O. Box 1385 GR 711 10 Heraklion, Greece. Email: {courcou, vsiris, gstamoul}@ics.forth.gr

the access pipe which connects a customer to his provider [8, 5]. Such a pricing scheme provides no incentives for users to use less bandwidth than the rate of their access pipe.

In this paper we consider usage-based pricing schemes that are based on bounds of the effective bandwidth. These schemes are linear in two measurements (the duration of a connection and the total volume transferred) and were initiated by the work of Kelly [6], and further extended in [2] to include both a priori parameters (such as leaky bucket parameters) and a posteriori parameters (such as the total volume) and for an arbitrary number of measurements. In [1] we investigated the incentive compatibility of these schemes and the effects of pricing on a network's equilibrium for Internet Wide Area Network (WAN) traffic.

As noted above, the final price a customer is charged will depend on other issues (e.g., economic, regulatory, marketing), in addition to the amount of resources used. Our focus is solely on the usage part of the price, and in the the rest of this paper the term "price" or "charge" will refer to this component. Since the usage component of a price is one part of the amount a customer is charged, an important property of usage-based pricing schemes is their ability to capture the *relative* amount of resources used by connections. Based on the above property, which we will refer to as *fairness*, we compare the pricing schemes for MPEG-1 (Motion Picture Experts Group) compressed video traffic, investigate their robustness for different link capacities and buffer sizes, and investigate the effects of traffic smoothing. Our experiments show that for links with a high degree of multiplexing, simple approximations can give good performance.

The rest of this paper is structured as follows. In Section 2 we briefly summarize important results from the theory of effective bandwidths and effective bandwidth approximations on which the investigated pricing schemes are based. In Section 3 we describe simple pricing schemes which require measurements of time and volume for the whole duration of a call. In Section 5 we compare the schemes in terms of their fairness for MPEG-1 compressed video traffic. We conclude with Section 6, which summarizes the main points of the paper.

2 Effective Bandwidths

Next, we present some basic results from the theory of effective bandwidths and effective bandwidth approximations[3] on which the pricing schemes investigated in this paper are based.

If $X[0,t]$ is the amount of workload produced by a traffic stream in an interval of length t, the

[3] For a more rigorous mathematical treatment of the subject, the reader is referred to [7, 2].

effective bandwidth for that stream is defined as [7]

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

where s, t are system parameters which depend on the characteristics of the multiplexed traffic and the link parameters (capacity and buffer size), see [2]. Specifically, the *time* parameter t corresponds to the most probable duration of the busy periods prior to buffer overflow. The *space* parameter s corresponds to the degree of multiplexing and depends, among others, on the size of the peak rate of the multiplexed streams relative to the link capacity. In particular, for links with capacity much larger than the peak rate of the multiplexed streams, s tends to zero and the $\alpha(s,t)$ approaches the mean rate of the stream, while for links with capacity not much larger than the peak rate of the streams, s is large and $\alpha(s,t)$ approaches the maximum value of $X[0,t]/t$.

Based on the theory of multiplexing developed in [4, 2], for a target overflow probability $e^{-\gamma}$, under conditions, the constraint on the sum of the effective bandwidths of the multiplexed streams has the following linear form

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

where C, B is the link capacity and buffer, $\alpha_i(s,t)$ is the effective bandwidth of connection i and (s,t) is an extremizing pair in

$$\sup_t \inf_s \left[st \sum_i \alpha_i(s,t) - s(Ct + B) \right] . \tag{3}$$

An important property of effective bandwidths, expressed by equation (2), is that they reflect effective usage: if source A has twice as much effective bandwidth as source B (measured at a particular operating point s, t), then source A uses twice as much resources as source B.

Relation (3) is used to compute the values of s, t. Experimentation with broadband traffic [3], has revealed that the effective bandwidth sufficiently captures the amount of resources used by a connection, and has provided evidence that the values of s, t are, to a large extent, insensitive to small variations of the traffic mix (i.e., percentage of different traffic types). The latter has important engineering implications: during different times of day it is expected that the traffic mix will remain relatively constant, hence specific pairs (s,t) will correspond to different periods of the day. Such values can be precomputed off-line based on traces of the actual traffic. In the rest of the paper, we will drop s, t from $\alpha(s,t)$.

Turning to the problem of pricing, one can price a connection based on the direct estimate of the effective bandwidth using (1), where the expectation is replaced by the empirical mean. However, such an approach is costly due to the complexity of computing the logarithmic moment generating function in (1), a task which needs to be performed for every connection. An additional disadvantage of this approach is that it will be difficult for users to determine the effects of their decisions (e.g., the reduction of their peak rate) on prices, making it difficult for them to behave "rationally"; the latter is a requirement for economically efficient and stable network operation.

For the above reasons, conservative approximations of (1) which involve easy to measure quantities are desirable. One such approximation considered in [6] is the "on-off" approximation, which depends solely on a connection's peak rate h, which in the rest of the paper we assume is policed, and its mean rate m. When a connection, in addition to having his peak rate policed, is policed by a leaky bucket with leak rate ρ and bucket size β, a tighter bound is the following [2]

$$\hat{\alpha}_{\text{sb}}(m,h,\rho,\beta) = \frac{1}{st}\log\left[1+\frac{tm}{H(t)}(e^{sH(t)}-1)\right], \tag{4}$$

where $H(t) = \min\{ht, \rho t + \beta\}$, i.e., $H(t)$ is the maximum amount of workload that the source can produce in a window of time t. We will refer to approximation (4) as the "simple bound".

A third and more accurate approximation, which is motivated from investigations in [2] which showed that in many cases, for a given pair (s,t), the worst case traffic from the output of a leaky bucket policer consists of blocks of an inverted T shape being either periodic or having random gaps. Here, we consider the periodic inverted T pattern shown in Figure 1 [1]. Sources are assumed independent, hence their corresponding patterns have random phases, i.e., their start is randomly distributed in the time interval $2t + t_{\text{off}}$.

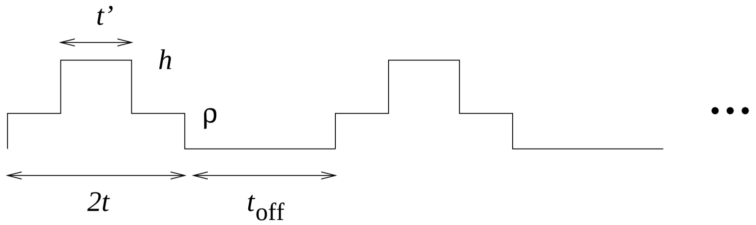


Figure 1: Periodic pattern for the inverted T approximation. $t' = \frac{\beta}{h-\rho}$, $t_{\text{off}} = \frac{(2t-t')\rho + t'h}{m} - 2t$

Based on the periodic pattern in Figure 1, we have the following effective bandwidth approximation:

$$\hat{\alpha}_{\perp}(m,h,\rho,\beta) = \frac{1}{st}\log E\left[e^{sX_{\perp}[0,t]}\right], \tag{5}$$

where $X_\perp[0,t]$ denotes the amount of workload produced by the inverted `T` pattern in an interval of length t. The expected value in the right-hand side of (5) is computed analytically.

In the next section, we discuss how to construct simple pricing schemes, based on the above effective bandwidth bounds, that are linear in measurements of time and volume.

3 Pricing schemes linear in time and volume

The pricing schemes we consider have the simple form

$$\text{Price} = aT + bV\,, \tag{6}$$

where T, V are the duration and transferred volume of a connection, respectively. Parameters a, b correspond to the user's tariff selection: at connection setup, given the user's traffic contract, the user is offered a set of possible tariff pairs (a, b) to choose from. A rational user will select the pair (a, b) which minimizes the a priori expected value of his price. According to the theory developed in [6, 2], the tariff pairs (a, b) can be appropriately defined so that the expected price for a rational user is $\hat{\alpha}T$, where $\hat{\alpha}$ is some bound of the effective bandwidth. Let $\hat{\alpha}(m, \mathbf{h})$ be an upper bound on the effective bandwidth subject to the mean rate being m and the traffic being within the description contained in the traffic contract $\mathbf{h}$. The function $\hat{\alpha}(m, \mathbf{h})$ is concave in m [2]. If (a, b) are such that $a + bm$ are tangents of the bound $\hat{\alpha}(m, \mathbf{h})$, then the user will minimize his average price if he selects the tariff pair[4] (a, b) which corresponds to the tangent of $\hat{\alpha}$ at point M, where $M = V/T$ is the user's mean rate. In this case, his price will be $(a + bM)T = \hat{\alpha}(M, \mathbf{h})T$, and his price per unit of time will be $a + bM = \hat{\alpha}(M, \mathbf{h})$.

Note that $\hat{\alpha}T$ is a user's average price only if he knows his mean rate. If he is inaccurate in determining his mean, his price will be higher. Hence, constructing tariffs using the above approach provides the user with the incentive to know his mean rate as accurately as possible. The user reveals to the network the estimate of his mean rate indirectly through his tariff selection. In the rest of this chapter we assume that users are rational and that they know the exact value of their mean rate. Under these assumptions, the comparison of the pricing schemes involves the comparison of the bounds of the effective bandwidth on which they are based. The comparison of pricing schemes is discussed in the next section.

[4]We assume that users can select from a continuum of tariff pairs (a, b). In practice there will be a small set of tariff pairs, e.g., three pairs corresponding to a small, medium, and large mean rate.

4 Fairness of pricing schemes

As discussed in the introduction, the fairness property of a pricing scheme is its ability to capture the *relative* amount of resources used by connections. Suppose $\hat{\alpha}$ is an approximation of α and let, with a slight abuse of notation, $\alpha(x)$ and $\hat{\alpha}(x)$ be the corresponding charges for a connection x. For fairness, given any two connections x and y, we want to have $\hat{\alpha}(y)/\hat{\alpha}(x) \approx \alpha(y)/\alpha(x)$. As a measure of the *unfairness* of a pricing scheme over a set of connections we take the standard deviation of $\hat{\alpha}(x)/(\mu\alpha(x))$, where μ is the average of $\hat{\alpha}(x)/\alpha(x)$ as x ranges over the connection set. A small value for this indicator means that pricing schemes based on approximation $\hat{\alpha}$ will tend to provide users with the right incentives, i.e., a user who chooses a tariff which results in a smaller price is actually making less use of the network.

5 Experiments with Real Traffic

In this section we investigate the pricing schemes presented in Sections 2, 3: the simple bound (4) and the inverted `T` approximation (5). Among the issues we wish to investigate is how the fairness of the pricing schemes is affected by the link parameters (capacity and buffer size) and traffic smoothing.

Our investigations involve MPEG-1 compressed video[5] whose characteristics are shown in Table 1. Three sets of video traffic are used in our experiments: *movies*, *news* and *talk shows*. These were created by breaking the cell streams containing the MPEG-1 traffic into non-overlapping segments, each with a duration of approximately 3 minutes (4500 frames). The resulting *movies* set contained 54 segments, the *news* set contained 16 segments, and the *talk show* set contained 18 segments.

We assume that users are rational and that they know their traffic profile, hence they can compute their actual mean rates and select the optimal leaky bucket parameters for the whole duration of the call.

Finally, our results are for s, t values that are typical for the specific capacity and buffer size. These are computed using relations (1) and (3), where the expectation in (1) is replaced by the empirical mean, which is estimated from the traces. As discussed in Section 2, we assume that specific pairs (s, t), which are computed off-line, will characterize different periods of day.

[5]Made available by O. Rose [9], at ftp://ftp-info3.informatik.uni-wuerzburg.de/pub/

Unfairness for various link capacities and buffer sizes

Figure 2 compares the fairness of the pricing schemes for various link capacities and buffer sizes. Observe that the fairness for the simple bound and inverted `T` pricing schemes, in general, increases when the link capacity and buffer size increases. However, this depends on the time scale t of the system (link and multiplexed traffic). For example, at a link with capacity 155 Mbps (Figure 3(b)), the fairness of the simple charging schemes does not change significantly when the buffer is increased from 1 msec ($\approx$ 381 cells) to 8 msec ($\approx$ 2811 cells). However, it does increase when the buffer increases to 16 msec ($\approx$ 5622 cells). This can be explained as follows. For small values of t, which are typical for buffer sizes up to 8 msec, the leaky bucket does not produce a tight bound. On the other hand, for the larger time scales which are typical for a 16 msec buffer, the leaky bucket produces a tighter bound.

Effects of traffic smoothing on unfairness

In the results shown in Figure 3, the traces are smoothed such that cells produced during two consecutive frames are evenly spaced throughout the duration of the two frames. Comparing Figure 3 with Figure 2, we observe that traffic smoothing has an effect on pricing up to a certain buffer size: 8 msec ($\approx$ 2811 cells) for capacity 155 Mbps, and 4 msec ($\approx$ 5660 cells) for capacity 622 Mbps. Again, this is related to the time scale of the system. For large buffer sizes, the time scale t is large, and smoothing traffic in a smaller time scale has no effect.

6 Conclusions

In this paper we have considered simple pricing schemes that are based on bounds of the effective bandwidth and involve two measurements, the duration of a connection and the total volume transferred. The schemes are compared according to their fairness, i.e., their ability to capture the relative amount of resources used by connections. Our experiments involved MPEG-1 compressed video and were performed for link capacities and buffer sizes that are anticipated for broadband networks. Our results show that for a high degree of multiplexing, simple approximations of the effective bandwidth can give good performance.

References

[1] C. Courcoubetis, F. P. Kelly, V. A. Siris, and R. Weber. A study of simple usage-based charging schemes for broadband networks. Technical Report No. 193, ICS-FORTH, May 1997. Submitted for publication.

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

[3] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Many sources asymptotic and effective bandwidths: Investigation with MPEG traffic. Presented at the 2nd IFIP Workshop on Traffic Management and Synthesis of ATM Networks. Extended version submitted for publication., September 24-26, 1997.

[4] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources. *Journal of Applied Probability*, 33:886–903, 1996.

[5] E. Firdman. Rx for the Internet: Usage-Based Pricing. *Data Communications*, January 1997. Available at <http://www.data.com/business_case/rx_internet.html>.

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

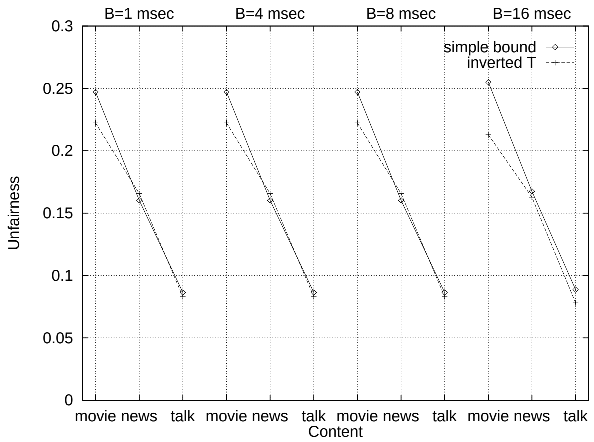
[7] F. P. Kelly. Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. Zeidins, editors, *Stochastic Networks: Theory and Applications*, pages 141–168. Oxford University Press, 1996.

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

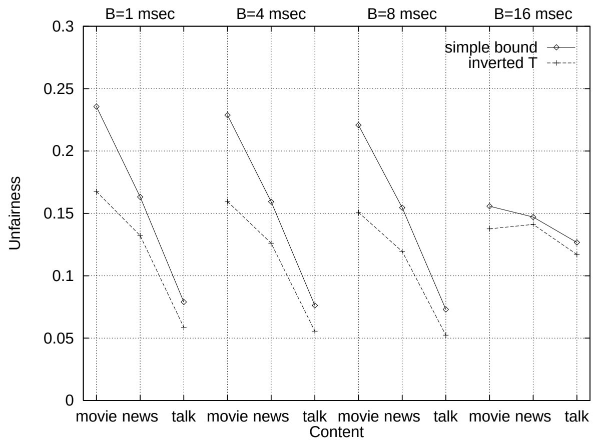
[9] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. Technical Report 101, University of Wuerzburg, February 1995.

Movies	
bond	James Bond: Goldfinger
dino	Jurassic Park
lambs	The Silence of the Lambs
movieprev	Movie Preview
star	Star Wars
terminator	Terminator II
News	
news1	News
news2	News
Talk Shows	
talk1	Talk show
talk2	Talk show

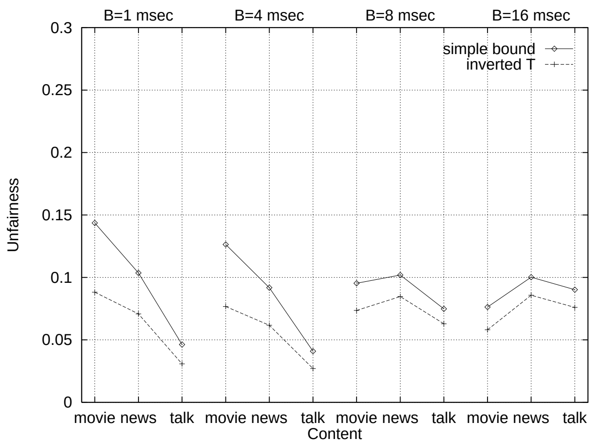
Table 1: MPEG-1 sequences. The frame pattern of the sequences is IBBPBBPBBPBB and their frame rate is 25 frames per second.



(a) $C = 34$ Mbps

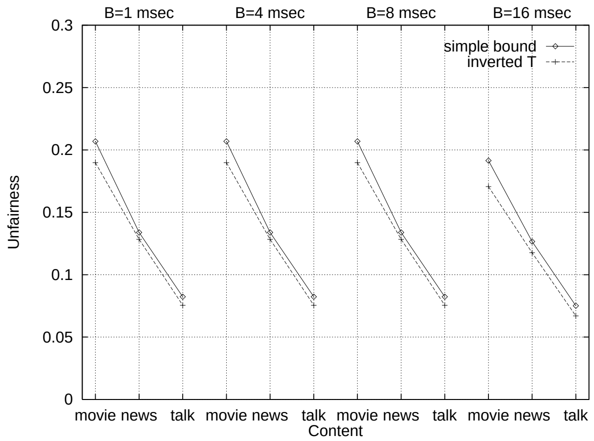


(b) $C = 155$ Mbps

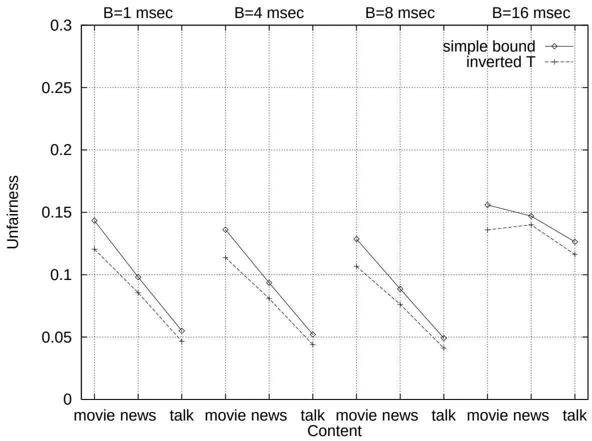


(c) $C = 622$ Mbps

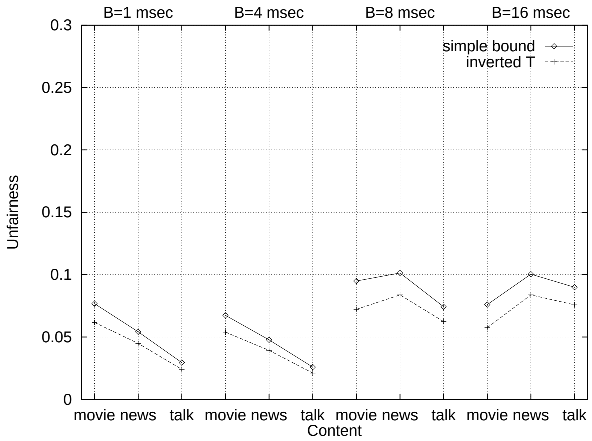
Figure 2: The fairness of the time-volume schemes increases as the link capacity and buffer size increases. [MPEG-1 video traffic (each segment has duration ≈ 3 minutes)]



(a) $C = 34$ Mbps



(b) $C = 155$ Mbps



(c) $C = 622$ Mbps

Figure 3: Comparison of this figure with Figure 2 shows that traffic smoothing increases the fairness of the time-volume pricing scheme up to a certain buffer size, which depends on the link capacity. The buffer sizes for which smoothing has a very small effect are such that the time parameter t is larger than the smoothing interval which is two frame times (=80 msec). [MPEG-1 video traffic (each segment has duration $\approx$ 3 minutes)]