FORTH-ICS / TR-214 February 1998

An Evaluation of Pricing Schemes that are based on Effective Usage

Costas Courcoubetis[1,2] *and Vasilios A. Siris*[1]

Abstract

Usage-based pricing enables network providers to recover costs from their customers in a fair manner. We evaluate usage-based pricing schemes that are based on bounds of the effective bandwidth. The schemes include ones that require simple measurements (time and volume) for the whole duration of a call, and two schemes that require measurements in distinct time intervals, smaller than the duration of a call: pricing with renegotiation and the virtual bucket scheme. Using MPEG-1 compressed video, the schemes are compared for different link capacities and buffer sizes. Our results show that measurements in distinct time intervals can greatly enhance *fairness*, i.e., the ability to capture the relative amount of resource usage by different connections. Furthermore, the virtual bucket scheme has performance close to the pricing with renegotiation scheme, without entailing the signaling overhead due to renegotiation.

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

An Evaluation of Pricing Schemes that are based on Effective Usage*

Costas Courcoubetis† and Vasilios A. Siris

Institute of Computer Science (ICS)
Foundation for Research and Technology - Hellas (FORTH)
P.O. Box 1385, GR 711 10 Heraklion, Crete, Greece
Email: {courcou,vsiris}@ics.forth.gr

Technical Report 214, February 1998

Abstract

Usage-based pricing enables network providers to recover costs from their customers in a fair manner. We evaluate usage-based pricing schemes that are based on bounds of the effective bandwidth. The schemes include ones that require simple measurements (time and volume) for the whole duration of a call, and two schemes that require measurements in distinct time intervals, smaller than the duration of a call: pricing with renegotiation and the virtual bucket scheme. Using MPEG-1 compressed video, the schemes are compared for different link capacities and buffer sizes. Our results show that measurements in distinct time intervals can greatly enhance *fairness*, i.e., the ability to capture the relative amount of resource usage by different connections. Furthermore, the virtual bucket scheme has performance close to the pricing with renegotiation scheme, without entailing the signaling overhead due to renegotiation.

Keywords: usage-based pricing, effective bandwidths, leaky buckets, broadband networks, MPEG-1

1 Introduction

The increasingly competitive nature of the telecommunications market and steps toward deregulation 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 the price 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 the access pipe which

*An abridged version of this paper appears in *Proc. of IEEE ICC'98*, Atlanta, GA, USA, June 7-11, 1998. This work was supported in part by the European Commission under ACTS Project CASHMAN (AC-039).

†Also with the Dept. of Computer Science, University of Crete.

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.

Unlike traditional telephone networks, which offer a single service that occupies a constant amount of bandwidth, broadband networks will support connections with different traffic characteristics and Quality of Service (QoS) requirements. An important first step to pricing is to find a workable and accurate way for quantifying resource usage which takes into account the above characteristics. The *effective bandwidth*, which is based on rigorous mathematical justification [7], is a scalar which summarizes the amount of resources required by a connection in order to preserve its own QoS requirements, and the requirements of the other connections it is multiplexed with. In addition to accurately capturing resource usage (in fact, as we will discuss in Section 3.1, accuracy alone is not an appropriate criteria for comparing usage-based pricing schemes), a pricing scheme must be efficient, in terms of signaling and accounting overhead, and must be based on parameters and measurements that are easily understood by customers.

In this paper we investigate usage-based pricing schemes that are based on bounds of the effective bandwidth. The schemes include ones that require simple measurements (time and volume) for the whole duration of a call, and two schemes that require measurements in distinct time intervals, smaller than the duration of a call: pricing with renegotiation and the virtual bucket scheme. The former, which we will refer to as "time-volume" pricing schemes, were initiated by the work of Kelly [6], and further extended in [3]. In [2] we investigated the incentive compatibility of these schemes and the effect of pricing on a network's equilibrium for Internet Wide Area Network (WAN) traffic.

As noted above, a user's charge 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 rest of this paper the term "price" or "charge" will refer to this component. Since the usage component is one part of the user's final charge, a desirable property of usage-based pricing schemes is their ability to capture the *relative* amount of resources used by connections. Based on this property, which we will refer to as *fairness*, we compare the pricing schemes for MPEG-1 compressed video, and investigate their robustness for different link capacities and buffer sizes, and the effect of traffic smoothing. Our results show that using measurements in time intervals smaller than the duration of a call increases, at the cost of increased accounting overhead, the fairness and robustness of usage-based pricing. Furthermore, the performance of the virtual bucket scheme is very close to the pricing with renegotiation scheme, without entailing the signaling overhead due to renegotiation.

The focus of this paper is on effectively *measuring resource usage*. In this sense, it complements the work in [1, 11] which address architectural issues of pricing network services.

The rest of the paper is structured as follows. In Section 2 we briefly summarize important results from the theory of effective bandwidths and effective bandwidth bounds. In Section 3 we describe time-volume pricing schemes which require two measurements (time and volume) for the whole duration of a call, and in Section 4 we present two approaches to pricing which involve measurements in distinct time intervals, smaller than the duration of a call. In Section 5 we present experiments with MPEG-1 compressed video traffic. We conclude with Section 6 which summarizes the main points of the paper.

2 Effective bandwidths and effective bandwidth bounds

Next we present some basic results from the theory of effective bandwidths and effective bandwidth bounds[1] on which the pricing schemes investigated in this paper are based.

If $X_i[0,t]$ is the amount of workload produced by source i in an interval of length t, the effective bandwidth for that source is defined as [7]

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

where s,t are system parameters which depend on the characteristics of the multiplexed traffic and the link parameters (capacity and buffer size) [3]. Specifically, the *time* parameter t corresponds to the most probable duration of the busy periods of the buffer prior to 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 sources relative to the link capacity. In particular, for links with capacity much larger than the peak rate of the multiplexed sources, s tends to zero and $\alpha_i(s,t)$ approaches the mean rate of the source, while for links with capacity not much larger than the peak rate of the sources, s is large and $\alpha_i(s,t)$ approaches the maximum value of $X_i[0,t]/t$.

According to the theory of multiplexing developed in [7, 3], under conditions, for a target overflow probability $e^{-\gamma}$ and for local variations of the traffic mix, the constraint on the sum of effective bandwidths entering a link 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 source i and (s,t) is an extremizing pair of

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

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.

Experimentation with real traffic [4] has shown that the effective bandwidth can accurately quantify resource usage, and that the values of the system parameters 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 the day it is expected that the traffic mix will remain relatively constant. Hence, particular pairs (s,t) will correspond to periods of the day during which the traffic mix remains relatively constant. Such pairs can be precomputed off-line using traces of the actual traffic. In the rest of the paper, we will drop s,t from $\alpha_i(s,t)$.

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.

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

However, such an approach is costly due to the complexity of computing the logarithmic moment generating function, 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 effect 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" bound 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 of the effective bandwidth is the following [3]:

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

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 results from investigations in [3] which showed that in many cases, for a given pair (s,t), the worst case output of a leaky bucket consists of blocks of an inverted T pattern which repeats either periodically or with random gaps. Here, we consider the periodic pattern shown in Figure 1. Sources are assumed independent, hence their start is randomly distributed in the time interval $2t + t_{\text{off}}$.

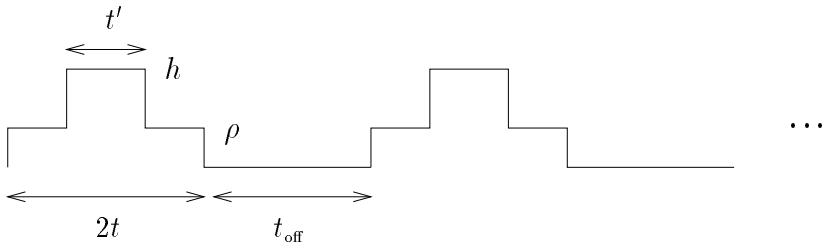


Figure 1: Periodic pattern for the inverted T approximation. $t' = \frac{\beta}{h-\rho}$, $t_{\text{off}} = \frac{(2t-t')\rho+t'h}{m} - 2t$

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

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

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

The simple bound and inverted T approximations were investigated for Internet WAN traffic in [2]. In the next section, we discuss how to construct simple pricing schemes, based on the above effective bandwidth approximations, that are linear in measurements of time and volume.

3 Pricing schemes linear in time and volume

The pricing schemes described in this section have the following simple form:

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

where T, V is the duration and transferred volume of a connection. Parameters a, b depend on the user's tariff selection: at connection setup, given his traffic contract, the user is offered a set of possible tariff pairs (a, b) to choose from. A rational user will select the pair which minimizes the a priori expected value of his charge. The tariff pairs (a, b) can be appropriately defined so that the expected charge for such a user is $\tilde{\alpha}T$, where $\tilde{\alpha}$ is some bound of the effective bandwidth. Specifically [6, 3], consider an effective bandwidth approximation $\tilde{\alpha}(m)$ which is concave in m (mean rate). If (a, b) are such that $a + bm$ are tangents of the bound $\tilde{\alpha}$, then the user will minimize his average charge if he selects[2] the tariff pair (a, b) which corresponds to the tangent of $\tilde{\alpha}$ at point M, where $M(= V/T)$ is the user's mean rate. In this case, his charge will be $(a + bM)T = \tilde{\alpha}T$, and he will be charged at a price per unit of time equal to $a + bM = \tilde{\alpha}$. In the above analysis, the only requirement on $\tilde{\alpha}(m)$ is that it is concave in m, hence both the simple bound (4) and the inverted `T` approximation (5) can be used [3].

Note that $\tilde{\alpha}T$ is a user's average charge provided he knows his mean rate. If he is inaccurate in determining his mean, then 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. For the rest of the paper, we assume that users are rational and that they know their mean rates. Hence, the comparison of the pricing schemes reduces to the comparison of the effective bandwidth bounds on which they are based.

3.1 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 $\tilde{\alpha}$ is a bound of the effective bandwidth α and let, with a slight abuse of notation, $\tilde{\alpha}(x)$ and $\alpha(x)$ be the corresponding charges for a connection x. For fairness, given any two connections x and y, we want to have $\tilde{\alpha}(y)/\tilde{\alpha}(x) \approx \alpha(y)/\alpha(x)$. As a measure of the *unfairness* of a pricing scheme over a set of connections we take the standard deviation of $\tilde{\alpha}(x)/(\mu\alpha(x))$, where μ is the average of $\tilde{\alpha}(x)/\alpha(x)$ as x ranges over the connection set.

4 Pricing schemes with measurements in distinct time intervals

In this section we describe two schemes that involve measurements in distinct time intervals, smaller than the duration of a call.

The first scheme applies to connections that support traffic parameter renegotiation[3] and is

[2] We assume that users can select from a continuum of tariff pairs (a, b). In practise, a small number of tariff pairs will be offered for each contract.

[3] For example, the Q.2963 signaling protocol [9] which is currently being developed for B-ISDN, and the Resource Reservation Protocol [12] being developed by the IETF support renegotiation.

an application of the simple pricing schemes described in the previous section for measuring resource usage in each renegotiation interval, rather than for the whole duration of the call. Note, however, that a straightforward without renegotiation extension of the time-volume schemes to include measurements of volume in fixed time intervals would give the same prices as the simple time-volume schemes, which involve measurements for the whole duration of a call; this occurs because (6) is linear in T, V and the pair (a, b) is the same for all the renegotiation intervals. It is the ability to renegotiate leaky bucket parameters (hence have different parameters a, b in different intervals) that can lead to a better approximation. Such an approach will be advantageous when the connection's traffic has different "modes" of operation, such as movie scenes with high and low action. In this case, leaky bucket parameters produce a tighter bound when the traffic they characterize has a single operating mode.

The second scheme, called the *virtual bucket* scheme, requires measurements in a priori fixed time intervals. However, the same leak rate of a "virtual" leaky bucket is used to measure the corresponding bucket size in each time interval. Hence, the scheme does not entail the renegotiation overhead of the previous approach.

4.1 Pricing services that support renegotiation

For services that support renegotiation, the schemes described in the Sections 2 and 3 can be used to measure resource usage for each renegotiation interval, where now the peak rate h_k and the leaky bucket parameters (ρ_k, β_k) are renegotiated in each interval k. In what follows, we will use the inverted `T` approximation since it is more accurate than the simple bound. Let $T_{\rm p}$ be the duration of the renegotiation intervals (for simplicity we assume they have equal length) and V_k the volume transferred during interval k. If K is the total number of renegotiation intervals and, as discussed in Section 3, a user can predict his mean rate in each interval, the average price per unit of time is[4]

$$\tilde{\alpha}_{\rm ren} = \frac{\sum_{k=1}^{K} \tilde{\alpha}_{\perp}(h_k, \rho_k, \beta_k, m_k)}{K}, \tag{7}$$

where $m_k = V_k/T_{\rm p}$.

4.2 The virtual bucket scheme

Unlike the previous scheme, the *virtual bucket* scheme does not require renegotiation. Let r be the leak rate of a "virtual" leaky bucket and h the peak rate, both declared at call setup, e.g., indirectly through the tariff selection. In fixed size intervals of length $T_{\rm p}$ (pricing interval), two measurement are made for each connection: the volume V_k and the minimum bucket size b_k such that the connection's traffic in interval k conforms to the leaky bucket (r, b_k). The quantity b_k can be measured using two counters: B_1 and B_2. Counter B_1 decreases at rate r (measured in cells per second), and increases by one for every cell received. At each update of B_1, its value is compared to B_2. If $B_1 > B_2$, then B_2 is set to B_1. At the end of the pricing interval, B_2 will contain the bucket size b_k we are looking for. If K is the total number of pricing intervals, the

[4]Equations (7) and (8) assume that the duration of a call is an integer multiple of the renegotiation or pricing interval. They can be extended to cover the general case.

average price per unit of time is

$$\tilde{\alpha}_{\mathrm{vb}}(h,r) = \frac{\sum_{k=1}^{K} \tilde{\alpha}_{\perp}(h,r,b_k,m_k)}{K}\,, \tag{8}$$

where $m_k = V_k/T_p$.

5 Experiments with real traffic

In this section we investigate the four pricing schemes (or equivalently, the effective bandwidth bounds on which they are based) we presented above: the simple bound (4), the inverted `T` approximation (5), the pricing with renegotiation scheme (7), and the virtual bucket scheme (8). The specific issues we address are the following:

- Effect of the pricing/renegotiation interval on the accuracy of the pricing schemes.
- Effect of the duration of a call on the accuracy of the pricing schemes.
- Comparison of the schemes in terms of *fairness* for different link capacities and buffer sizes.
- Effect of traffic smoothing.

Our investigations involve real MPEG-1 compressed video[5] (Table 1). In our experiments, we assume that cells belonging to the same frame are evenly spaced throughout the duration of a frame (40 msec).

We assume that users are rational and that they know their traffic profile, hence they can compute their actual mean rate and select the optimal leaky bucket parameters for the whole duration of the call (in the case of the simple bound and inverted `T` approximations) and for each renegotiation interval (in the case of the pricing with renegotiation scheme). For the virtual bucket scheme, we assume that users choose the leak rate that minimizes their charge. Furthermore, in the pricing with renegotiation scheme we assume that renegotiations take place in fixed time intervals.

Our results are for s, t values that are typical for the specific capacity and buffer size. These are computed using equations (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, particular pairs (s,t), which are computed off-line, will characterize different periods of the day.

Effect of the pricing/renegotiation interval

Figure 2 shows the effect of the pricing interval for the virtual bucket scheme and the renegotiation interval for the pricing with renegotiation scheme. The curve labeled "effective bandwidth" is estimated by taking the average of the effective bandwidth estimates in time intervals with length equal to the pricing/renegotiation interval. These estimates are computed using (1). The reason for this is that both the pricing with renegotiation and virtual bucket schemes measure resource usage in time intervals with length equal to the pricing/renegotiation interval.

[5]Made available by O. Rose [10], at <ftp://ftp-info3.informatik.uni-wuerzburg.de/pub/>.

As expected, the accuracy of the pricing with renegotiation and virtual bucket schemes is much better than both the simple bound and inverted `T` approximation. Furthermore, the accuracy of the virtual bucket scheme is close to that of the pricing with renegotiation scheme.

Figure 4(a) shows the effect of the pricing and renegotiation interval on the average of the ratio of the effective bandwidth bound over the actual effective bandwidth. The latter is computed as before, by taking the average of the effective bandwidth estimates in time intervals with length equal to the pricing/renegotiation interval. This average is computed over a set of 54 non-overlapping trace segments, each with a duration of approximately 3 minutes, that were created from the original MPEG-1 traces. This figure reaffirms our observations regarding Figure 2. However, note that the simple bound and inverted `T` approximations are more accurate; this is because the length of the segments (call duration) used in Figure 4(a) is 3 minutes, which is smaller than the length (30 minutes) of the *Star Wars* trace used in Figure 2. Also note that the ratio for the time-volume schemes decreases as the pricing/renegotiation interval increases. This is because the average of effective bandwidth estimates over small time intervals is smaller than the effective bandwidth estimate over the sum of the intervals.

Effect of the call duration

In Figure 3 we observe that the accuracy of the simple bound and inverted `T` approximation decreases when the duration of the call increases, hence these schemes are less fair to long duration connections. On the other hand, both the pricing with renegotiation and virtual bucket schemes are more robust to the call duration. These observations can be explained by noting that the leaky bucket forms a less tighter traffic bound when the duration of a call is long (hence includes a large number of scenes) compared to when the duration is small (hence includes a smaller number of scenes).

Comparison in terms of fairness

Figure 4(b) shows that for a pricing interval of a few tens of seconds, the fairness of the pricing with renegotiation and virtual bucket schemes is better than the time-volume pricing schemes. Furthermore, the fairness of the virtual bucket scheme is close to that of the pricing with renegotiation scheme.

For various buffer sizes, Figures 5(a), 6(a), and 7(a) compare the pricing schemes in terms of fairness for link capacities 34 Mbps, 155 Mbps, and 622 Mbps, respectively. Three values of the pricing/renegotiation interval are considered: 10, 20, and 30 seconds. The results pertain to the same set of 54 non-overlapping trace segments we considered above. Observe that the pricing with renegotiation and virtual bucket schemes, compared to the simple pricing schemes, are fairer and more robust, i.e., their fairness is less affected by the link capacity and buffer size. Furthermore, the fairness for the time-volume 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 6(a)), the fairness of the time-volume pricing schemes does not change significantly when the buffer increases 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 time scales, i.e., 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.

Effect of traffic smoothing on unfairness

In the results shown in Figures 5(b), 6(b), and 7(b) the traces are smoothed such that cells produced during two consecutive frames are evenly spaced throughout the duration of the two frames. Comparing these figures with Figures 5(a), 6(a), and 7(a), we observe that traffic smoothing has an effect for the whole range (up to 16 msec) of buffer sizes shown for capacity 34 Mbps, and up to a certain buffer size for 155 Mbps and 622 Mbps: 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 Conclusion

We investigated time-volume pricing schemes which require two simple measurements (time and volume) for the whole duration of a call, and two pricing schemes that involve measurements in distinct time intervals, smaller than the duration of a call: pricing with renegotiation and the virtual bucket scheme. Our investigations involved MPEG-1 compressed video traffic, and their results are summarized as follows: 1) pricing schemes with measurements in distinct time intervals can achieve, at the cost of higher accounting overhead, greater performance in terms of fairness and robustness compared to pricing schemes with measurements for the whole duration of the call, 2) the performance of the virtual bucket scheme is very close to the pricing with renegotiation scheme, without entailing the signaling overhead due renegotiation, and 3) in general, the fairness of the time-volume pricing schemes is higher for larger link capacity and buffer sizes, for smaller call durations, and for smoother traffic; the exact behavior depends on the time scale that is important for buffer overflow (expressed by the time parameter t).

The simple pricing schemes discussed in Section 3 assumed that the peak rate h of a connection is known by the user and policed by the network. If this is not the case, we can again have a pricing formula similar to (6). As an example, consider the "on-off" approximation $\tilde{\alpha} = \frac{1}{st}\log z$, where $z = 1 + \frac{m}{h}(e^{sht} - 1)$. The function $\tilde{\alpha}(z)$ is concave in z, hence similar to the approach described in Section 3, we can define the tariff pairs (a, b) to correspond to tangents of $\tilde{\alpha}(z)$. A user who chooses the tariff (a, b) will be charged the amount

$$\begin{aligned}(a + bz)T &= \left(a + b\left[1 + \frac{V}{Th}(e^{sht} - 1)\right]\right)T \\ &= (a + b)T + bV\frac{e^{sht} - 1}{h} \\ &= (a + b)T + bV',\end{aligned}$$

where T, V is the duration and volume, and V' is the "scaled" volume, which depends on the measured volume V and the measured peak rate h. If we assume that the peak rate is independent of the transferred volume, we get that the expected value of the price per unit of

time is $(a+b)+bm\sigma$, where m is the mean rate and $\sigma = E\left[\frac{e^{sht}-1}{h}\right]$. Hence, in order to choose the optimal tariff, a user needs to know his "scaled" mean $m\sigma$.

References

[1] D. D. Clark. A model for cost allocation and pricing in the Internet. In L. W. McKnight and J. P. Bailey, editors, *Internet Economics*. MIT Press, MA, 1996.

[2] C. Courcoubetis, F. P. Kelly, V. A. Siris, and R. Weber. A study of simple usage-based charging schemes for broadband networks. In *Proceedings of IFIP International Conference on Broadband Communications (BC'98)*, Stuttgart, Germany, April 1-3, 1998. To appear.

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

[4] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Application and evaluation of large deviation techniques for traffic engineering in broadband networks. In *Proc. of ACM SIGMETRICS'98*, June 1998. To appear.

[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] International Telecommunication Union Telecommunications Sector (ITU-T) Draft Recommendation Q.2963. Geneva, Switzerland, 1991.

[10] 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.

[11] 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.

[12] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala. RSVP: A Resource ReSerVation Protocol. *IEEE Network Mag.*, pages 8–91, July 1993.

Name	Mean (Mbps)	Peak (Mbps)	Peak/Mean	Std. dev. (Mbps)
Jurassic Park	0.37	3.31	9.03	0.71
The Silence of the Lambs	0.21	3.71	17.92	0.31
Terminator 2	0.31	2.20	7.20	0.28
James Bond: Goldfinger	0.68	6.75	9.99	0.71
Star Wars	0.26	3.46	13.18	0.36
Movie Preview	0.40	4.77	11.94	0.52

Table 1: MPEG-1 traces. The duration of each trace was approximately 30 minutes and the frame pattern was IBBPBBPBBPBB with 25 frames per second. The above statistics refer to the streams created after the MPEG-1 traces were segmented into cells with a 48 bytes payload.

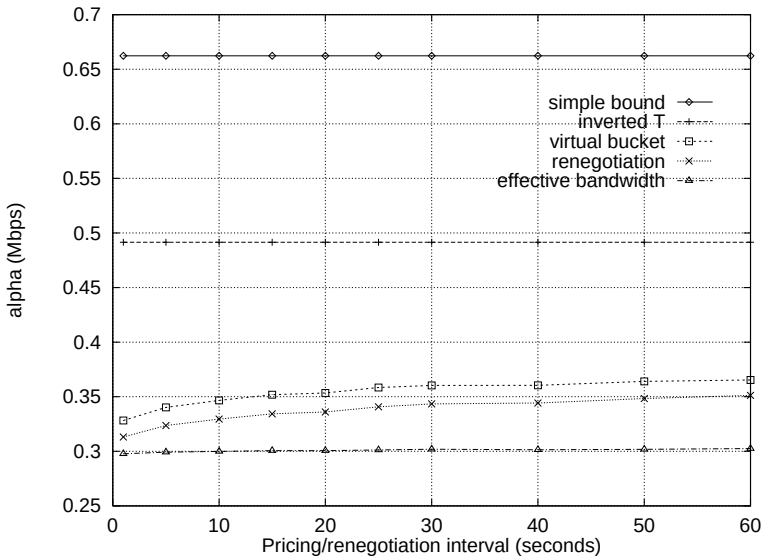


Figure 2: Effect of the pricing/renegotiation interval. $C = 155$ Mbps, $B = 8$ msec (≈ 2811 cells), *Star Wars* trace (duration ≈ 30 minutes).

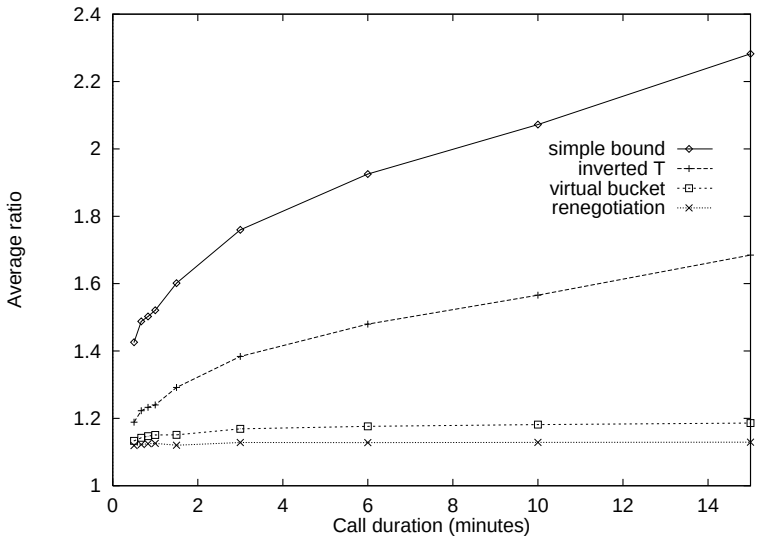
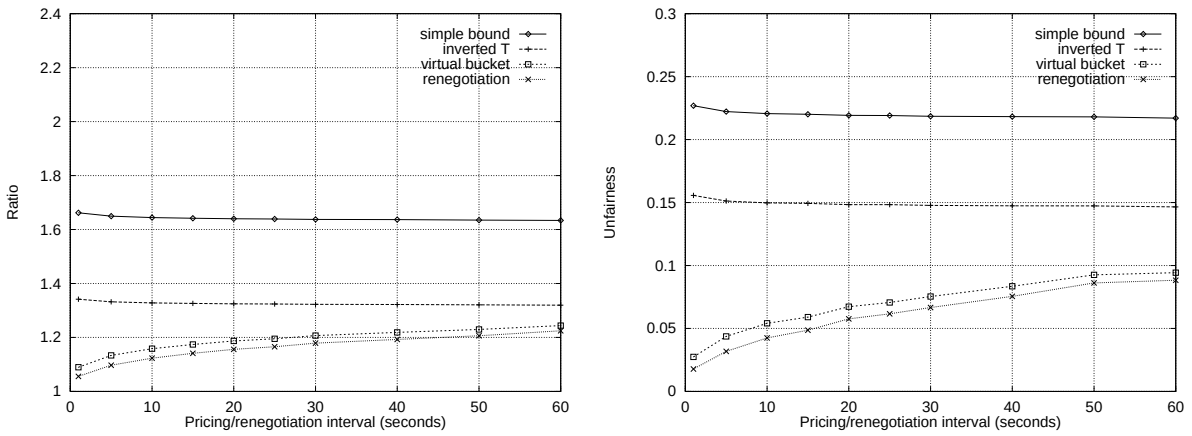


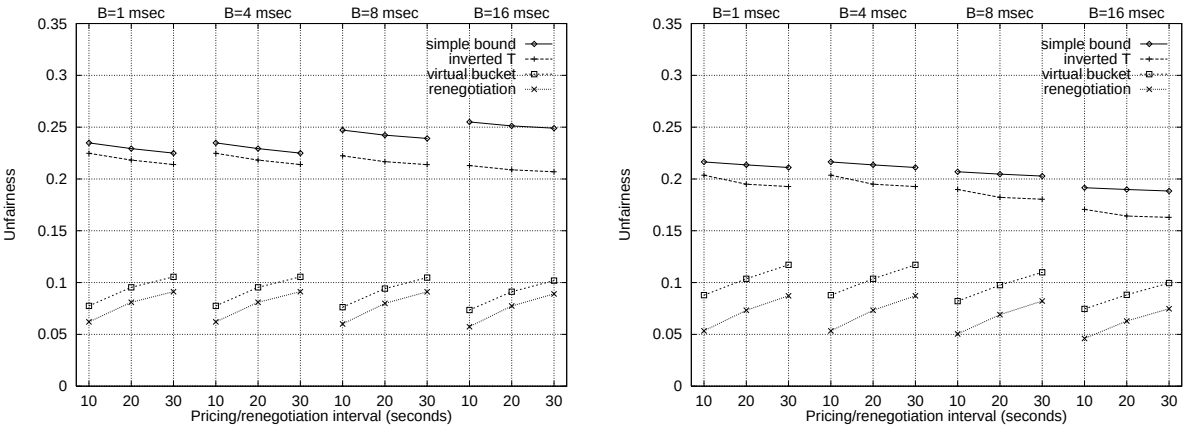
Figure 3: Effect of the call duration. $C = 155$ Mbps, $B = 8$ msec (≈ 2811 cells), $T_p = 10$ seconds.



(a) Accuracy

(b) Unfairness

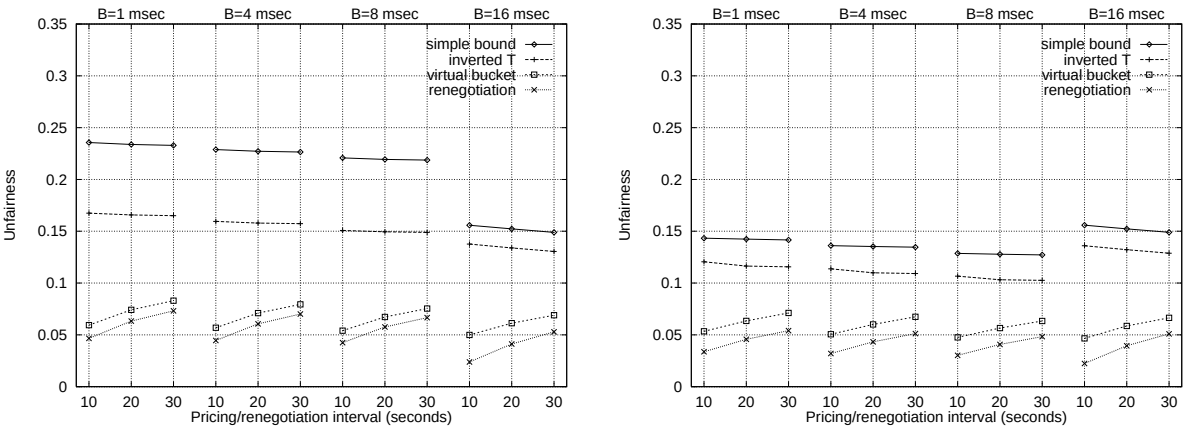
Figure 4: Effect of the pricing/renegotiation interval. $C = 155$ Mbps, $B = 8$ msec (≈ 2811 cells), call duration ≈ 3 minutes.



(a) no smoothing

(b) with smoothing

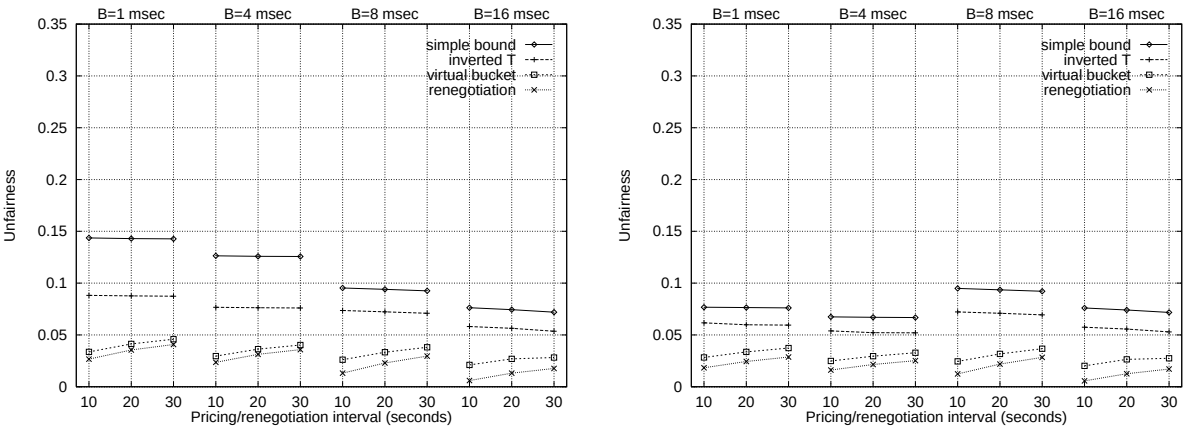
Figure 5: Unfairness for link capacity 34 Mbps. Call duration $\approx$ 3 minutes.



(a) no smoothing

(b) with smoothing

Figure 6: Unfairness for link capacity 155 Mbps. Call duration $\approx$ 3 minutes.



(a) no smoothing

(b) with smoothing

Figure 7: Unfairness for link capacity 622 Mbps. Call duration $\approx$ 3 minutes.