A HISTOGRAM BASED METHOD FOR VIDEO-NETWORK INTERFACES IN ATM NETWORKS

Yon Jun Chung, Chi Yuan Hsu, Antonio Ortega, and C.-C. Jay Kuo

Signal and Image Processing Institute and Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, California 90089-2564

ABSTRACT

In this paper, we introduce a source-network interface for variable bit rate (VBR) video sources. Our main goal is to develop a VBR video source-network interface with greater discrimination capability of bursty sources than the Leaky Bucket (LB) and which at the same time can exercise long term rate control. The real time policing mechanism utilized for our interface is based on comparing the worst case average/interval (WCA) between a constraint histogram and the content of a FIFO sliding window, monitoring the video source. Real video sequences and outputs from Markov based models are used in our experiments. We show that our method can selectively accept a subset of those bursty sequences which would be admitted by a LB of similar dimensions. This, in turn, will give the network greater control on the type of sequence to expect and result in a more informed decision in its call admission function.

1. INTRODUCTION: VIDEO-NETWORK INTERFACE

Numerous works have outlined the advantages of variable bit rate (VBR) transmission of video over standard constant bit rate (CBR) methods [1]. VBR transmission over asynchronous transfer mode (ATM) networks promises advantages in terms of both video quality and network utilization. Video coders operating at constant perceptual quality produce variable rate output; thus, VBR in principle offers the advantage of transmitting the video without requiring buffering at the encoder. While public and private ATM networks are being deployed, there are very few implementations of VBR video transmission.

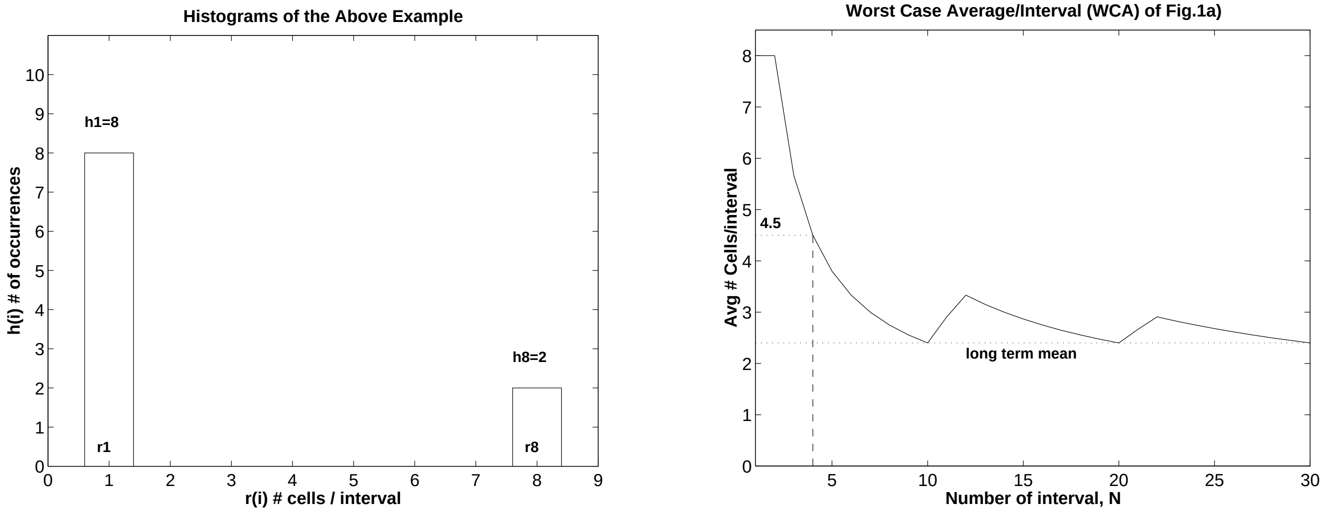
Estimates of video quality and multiplexing gain are usually based on simple scenarios, and in general have considered separately networking and source coding (e.g., assuming that constant quantizer bit rate traces are representative of VBR video transmission independently of the type of networking environment considered). However, the key to successful implementation of VBR video transmission lies in the interface between video and network, specifically in the rules used to determine the bit rate that can be allowed into the network from each source. Once we have specified the interface we can produce models for the video bit rate under the chosen set of rate constraints and use these models to estimate network utilization.

We thus consider the following scenario which is compatible with the currently specified User Network Interface (UNI) [2, 3]. Video source and network agree upon a set of parameters (including the peak rate of the source) used to describe the connection and monitored by the network. Because monitoring at arbitrary points within the network is more difficult (delay jitter introduces errors in the measurements and the monitoring functions themselves have to be very simple), we concentrate on monitoring *at the point where the source enters the network*. This does not preclude other, potentially simpler, usage parameter control (UPC) or policing functions being used throughout the network.

Until now the most popular source-network interfaces for video sources have been based on the Leaky Bucket (LB) [4, 5]. The LB has many attractive features, mainly the rapid response time for small buffer sizes and the ease of implementation. It also provides an operational measure of how much the source bit rate exceeds a certain "average rate" (the output rate of the LB) over a time interval determined by the buffer size.

Assume we have a video source with known mean rate and peak rate for which we wish to select the LB parameters. If the monitoring of a video source is limited to a single LB, then the alternatives are (a) a LB with its output rate set to the mean rate of the source with a large buffer or (b) a LB with a higher output rate and a correspondingly smaller buffer. There are shortcomings with both scenarios. In case (a) the lower output rate performs long term rate monitoring, but at the expense of leaving the network susceptible to bursts proportional to the buffer size. Large bursts may degrade the Quality of Service (QOS) of neighboring sources sharing the ATM channel. Case (b) limits the size of the burst, but the higher output rate is a clear example of overdimensioning. This problem is addressed in [6] by *simultaneously* monitoring the video source with two LBs, a combination of (a) and (b), limiting the maximum burst size and at the same time performing long term rate control. While long term measures are not enforceable (it might be "too late" by the time the network detects a violation), they are useful in their potential contribution to the overall goal of increasing statistical multiplexing gain (SMG).

In this work, we introduce a histogram based constraint (HBC) as an alternative source-network interface to the LB. It extends the idea of limiting maximum burst rate and simultaneously peforming long term rate control, first presented in [6]. The HBC addresses the above mentioned shortcomings of the LB. It is more discriminating than the LB in the types of bursts admitted, and it performs long term rate control. We first describe HBC in greater de-



(a)

(b)

Figure 1: (a) Example of a histogram, 10 intervals long. (b) Corresponding WCA; it represents the running average of the descending order sort of Figure 1(a), repeated 3 times.

tail along with its policing mechanism in section 2. As in [6] we perform a "worst case" analysis, i.e., we compare the various strategies based on their performance when the maximum allowable number of bits is transmitted. In section 3, we describe our experiments with LB(a), LB(b), and HBC using real video sources and Markov model generated sources with varying degrees of burstiness.

2. HISTOGRAM BASED CONSTRAINT

2.1. Definition of HBC

A histogram is a graphical representation of the transmission rates of a sequence of finite duration. The duration of the sequence is divided into a set of fixed intervals. The duration can be expressed as $\sum h_i$, which equals 10 intervals in the example histogram in Fig. 1(a) . During these 10 intervals, there are $h_1 = 8$ intervals during which r_1 cells passed per interval, and $h_8 = 2$ intervals during which r_8 cells passed per interval, where $r_1 < r_8$. The observed transmission rate during one interval is recorded by incrementing the height of the bar of the corresponding rate. The mean of the histogram can be obtained via $\sum h_i r_i / \sum h_i$.

Previously, histograms of arrival rates were used in [7] to model the video sources and estimate ATM channel utilization, and discrete rates were used in [8] to represent states in a nonstationary finite state machine, whose transition probablility matrix determined the output of the source. The HBC mechanism we propose here is simpler to implement than other methods such as [8]. Even though histogram based models do not contain the temporal correlation information, they can still be used for allocation purposes [7].

Histograms provide more information than mechanisms such as the LB. For example, many different histograms may share the same mean, but they can be further characterized based on their worst case analysis. The worst case can be achieved by sorting the histogram contents in descending order, e.g., $r_8, r_8, r_1, \ldots, r_1$, because this would cause the greatest amount of congestion in the network amongst all possible permutations of sequences that are 10 intervals long and contain 8 r_1 and 2 r_8. The cumulative sum of this descending order sort is divided by the number of intervals $N = 1, 2 \ldots$ to yield a running average. We call this running average the worst case average/interval (WCA). Fig. 1(b) represents the WCA for the histogram in Fig. 1(a) repeated for three periods. The WCA for our example equals r_8, $(r_8 + r_8)/2$, $(r_8 + r_8 + r_2)/3, \ldots$ so on.

The WCA serves as an upper bound on the maximum number of ATM cells that can be expected within a given number of intervals. In Fig. 1(b), the maximum average bound is 4.5 cells/interval for sequences 4 intervals long. This means that if the first three intervals of a given input sequence contained 7, 6, and 3 cells respectively, then in order to maintain the maximun average of 4.5 cells/interval, the fourth input of this sequence is limited to 2 cells. Also notice that for durations of 1, 2, 3 intervals, this input sequence has an average/interval of 7, 6.5, 5.3 cells/interval, which are less than the respective maximum averages of 8, 8, 5.7 cells/interval in Fig. 1(b). In general, the HBC can control the WCA of any input sequence at every point along the WCA graph.

In a similar manner, the WCA of a constraint histogram can serve as the threshold for policing a VBR video source. The source and the network can agree on a constraint histogram that suits the need of the source. The line dividing the two shaded regions in Fig. 2 represents the WCA of a constraint histogram. A FIFO sliding window of equal length to the constraint histogram can track the video source "on the fly", as in Fig. 3. The WCA of the constraint histogram is the threshold of the policing algorithm. If the WCA of the FIFO sliding window's content lies above the threshold at one or more points, i.e., in the rejected region, then sufficient number of cells must be discarded or tagged

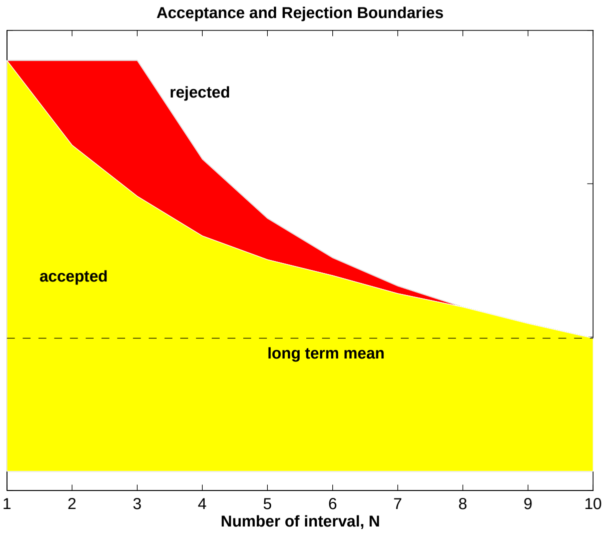


Figure 2: The acceptance and rejection regions for a worst case average/interval (WCA) of a constraint histogram

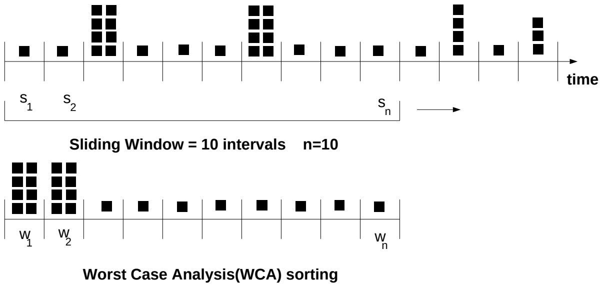

time
s_1
s_2
s_n
Sliding Window = 10 intervals n=10
w_1
w_2
w_n
Worst Case Analysis(WCA) sorting

Figure 3: Sliding Window and its WCA Sorting

to force the sliding window's WCA below the threshold. The comparison between the two WCAs, those of the constraint histogram and sliding window, is akin to the comparison between the cumulative sums of their descending order sort. Let $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$ represent the descending order sort, i.e., $w_i \geq w_j,\ j > i$, of the sliding window rates $\mathbf{s} = \{s_1, s_2, \ldots, s_n\}$. Likewise, let $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$ be the descending order sort of the constraint histogram. Their corresponding cumulative sums are $W_k = \sum_{m=1}^{k} w_m$ and $C_k = \sum_{m=1}^{k} c_m$. A fraction of the cells to arrive in the sliding window during the latest interval, s_n, is rejected or tagged for rejection if the cumulative sum of the sliding window is greater than that of the constraint histogram at one or more points. Note that our method imposes a constraint on both long term average and short term bursts.

2.2. Policing Function

The policing function uses a compliance criterion to compare the WCA of the sliding window to that of the constaint histogram. It calculates the number of arriving cells to reject or tag for rejection. We state that the content of the sliding window is compliant if $C_k \geq W_k\ \forall\ k \in [1, n]$, in which case no cells are rejected. If it is non-compliant, i.e., $C_k < W_k$ for some k, then the policing function rejects q *cells* out of the x *cells*, latest input in slot s_n. Therefore, the actual number of cells admitted will be $x^* = x - q$, forcing the resulting sliding window into compliance.

Assume that the sliding window is in a state of compliance at time t. At $t+1$, the content of s_1 is no longer part of the sliding window and a new number of packets arrives at $s_n = x = w'_f$. If $\{w_1, w_2, \ldots, w_n\}$ is compliant and w'_f creates a violation, then none of the rates larger than w'_f may have changed, otherwise there would be a net loss in rate within the sliding window. Now the descending order sort of the sliding window becomes $\mathbf{w}' = \{w_1, w_2, \ldots, w_{f-1}, w'_f, \ldots, w'_n\}$. Define $W'_k = \sum_{m=1}^{k} w'_m$, therefore $W_k = W'_k \leq C_k,\ k \in [1, f-1]$. Let $q \equiv \max_{k \in [f,n]}\{W'_k - C_k\}$, be the number of cells to tag or reject. Subtract q from the new input w'_f, $x^* \equiv w'_f - q$ and let $w'_f - q = w^*_p$, i.e., the maximum, q, occurs at index p and therefore $q = \{W'_p - C_p\}$. If the maximum, q, occured at f, i.e. $p = f$ and $q = \{W'_f - C_f\}$, then the only difference between $\mathbf{w}$ and $\mathbf{w}^*$, the descending order sort of the modified sliding window is w^*_f. $\mathbf{w}^* = \{w_1, \ldots, w_{f-1}, w^*_f, w_{f+1}, \ldots, w_n\}$ if $C_f \geq W^*_f$ holds, the modified content of the sliding window will be compliant,

$$
\begin{aligned}
C_f &\overset{?}{\geq} W^*_f \\
W^*_f &= W_{f-1} + w^*_f = W_{f-1} + w'_f - q \\
W^*_f &= W_{f-1} + w'_f - [W'_f - C_f] \\
W^*_f &= C_f,\ \text{since } W'_f = W_{f-1} + w'_f
\end{aligned}
$$

and therefore, $C_k \geq W^*_k\ \forall\ k \in [1, n]$. If $p > f$, the maximum, q, occurs at an index after f , then $q = \{W'_p - C_p\}$ and $\mathbf{w}^* = \{w_1, \ldots, w_f, \ldots, w_{p-1}, w^*_p, \ldots, w^*_n\}$.

$$
\begin{aligned}
C_k &\geq W^*_k = W_k,\ k \in [1, p-1] \\
W^*_p &= W'_p - q = W'_p - [W'_p - C_p] = C_p \\
W^*_j &= W'_j - q,\ j \in [p+1, n] \\
&= W'_j - \max_{k \in [1,n]}\{W'_k - C_k\},\ j \in [p+1, n]
\end{aligned}
$$

therefore

$$
C_j \geq W^*_j\ j \in [p+1, n]
$$

Last equation holds because there can be only one single maximum value for q. Combining the first, second and last equations leads to $C_l \geq W^*_l = W_l,\ l \in [1, n]$ and as a result the modified sliding window is compliant.

2.3. Versatility of the HBC

In this subsection, we address the discrimination versatility of our histogram based constraint. The WCA of the LB can be used in the manner just presented, and its performance would be nearly identical to that of the same LB used in its standard fashion. The advantage of the HBC is that, unlike the LB, it allows for a given mean, the selection of different constraints on rates over shorter windows. A LB has two parameters at its disposal, the buffer size and the output, or long term rate. Such is not the case with histograms.

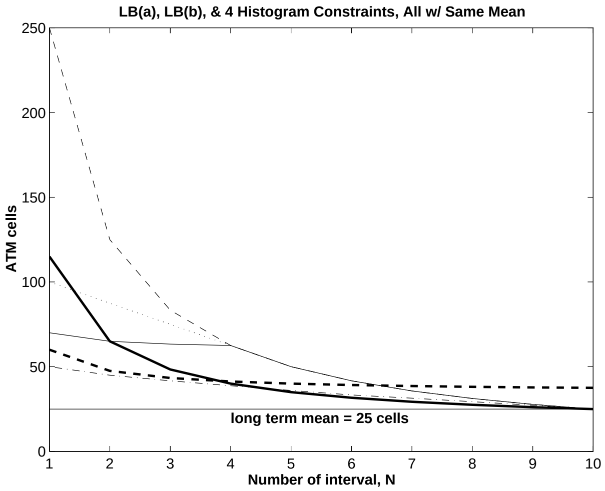


Figure 4: The WCA plots for LB(a) (thick) and LB(b) (thick&dashed) and four Histograms that have the same long term mean as LB(a), equal to 25 cells/interval.

The WCA for the constraint histogram can be shaped by choice.

The buffer size, the long term rate, and all the points in between can be tailored to suit the needs of a particular video source. Figure 4 shows the WCA for four different histograms with the same mean. Even though they share a common mean they have different shapes. The *thick&solid* line corresponds to the WCA of LB(a) and the *thick&dashed* line to that of LB(b). As stated earlier in the introduction, LB(a) is susceptible to large bursts whereas LB(b) is a case of overdimensioning. If the LB is modified to remedy the situation, i.e., if we decrease the output rate of LB(b) or the buffer size of LB(a), the resulting LB degenerates into a form of CBR interface. Herein lies LB's weakness.

Given the histogram's versatility to take on different WCA plots, the maximum burst rate and long term rate can be controlled simultaneously. This ability implies that not only can the HBC limit the size of the burst in any one time interval, it can also discriminate sources according to the proximity in the occurrences of bursts. This gives the HBC greater ability to differentiate among sequences with equal long term rates and consequently, admit only a subset of those allowed using the LB. More importantly, the HBC can give the network greater control on the types of sequences the network can expect than the LB. With more information about the video source, the network can make more informed decisions in its call admission function. This is the topic of our future research. While we developed the HBC interface with ATM primarily in our mind, this work could equally be relevant to other applications allowing capacity reservations.

3. EXPERIMENTAL RESULTS

In the following experiments we generate the source traffic with different traffic patterns from model, and apply different policing functions on the source traffic. We show that the admissible traffic patterns that can enter the networks are fewer than these monitored by LB policing function, because histogram based policing function is more discriminating on the burst type traffic than LB. In this section units of ATM cells, 1 cell equaling 384 bits, are used. The experiments used video source inputs corresponding to 25 frames/second and each frame duration, 1/25 second, is further divided into 10 intervals. Each interval is 1/250 second long.

3.1. Video Traffic Model

A traffic model which can approximately simulate the cell arrival process of the real video traffic is used in the experiments. In [8] and [7], the real video traffic is modeled as an eight-state Markov chain. Each state in the Markov chain represents a transmission rate. From the experiment in [8], this Markov chain model can simulate video traffic arrival with reasonable degree of accuracy. Therefore, we adopt this finite-state Markov chain as the model of the arrival process of video traffic in our experiment. In [8] the transition probabilities of the Markov chain obtained from the real video source shows that the probabilities of the transitions between two adjacent states are much greater compared to that of jumping two or more states. Therefore we define the model as an eight-state Markov chain with transitions only between two adjacent states. We set the the transition probabilities to the next higher transmission-rate state all equal to α, and the transition probabilities to the next lower transmission-rate state equal to β. We define R_i as the the transmission rate when the source is in state i, where $R_0, R_1, ...R_7$ are in ascending order. This source traffic model is depicted in Figure 5:

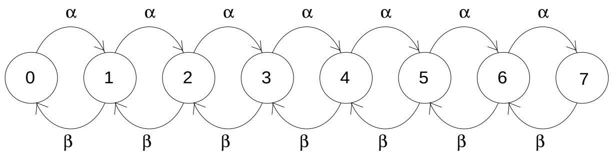


Figure 5: Video traffic model

The model can generate source traffic with different behavior by changing the values of α and β. The long term mean rate of this source traffic is:

$$\bar{R} = \left(\frac{1-(\frac{\alpha}{\beta})}{1-(\frac{\alpha}{\beta})^8}\right) \cdot \left(\sum_{i=0}^{7} (\tfrac{\alpha}{\beta})^i \cdot R_i\right) \qquad (1)$$

$$\text{where } 0 \le \alpha \le 1, \quad 0 \le \beta \le 1, \quad \alpha + \beta \le 1$$

By adjusting the parameters of the model to generate various kinds of simulated video traffic, we can compare our newly proposed histogram based policing function to other approaches, observing how each policing function works on the simulated traffic.

Generally, the long term mean traffic rate is related to the ratio $\frac{\alpha}{\beta}$. Model with higher $\frac{\alpha}{\beta}$ value will generate traffic with high mean rate. For the models with the same mean rate, i.e. same $\frac{\alpha}{\beta}$ ratio, the model with with small α or β value will tend to generate source traffic with longer burst duration, because the source is more likely to stay at the same state for a long time.

3.2. Simulation of Traffic Enforcement with Traffic Model

In this experiment, we assume the eight possible transmission rates of the source traffic are 15, 20, 25, 30, 35, 40, 45 and 50 cells/interval, where each cell contains 384 bits of data. With the source rates generated by the model, we apply the following three different policing functions on the source traffic:

(i) *Histogram based constraint:* with window size equal to 10 intervals and long term average rate equal to 32 cells/interval.

(ii) *Leaky bucket (large bucket):* with drain rate equal to 28 cells/interval and bucket size equal 310 cells.

(iii) *Leaky bucket (small bucket):* with drain rate equal 31 cells/interval and bucket size equal 90 cells.

We have selected the 3 constraints by selecting a source ($\alpha = 0.136$, $\beta = 0.2$) and choosing the constrain parameters so that the source is admissible. The worst cast average rates of the above three policing functions are depicted in Figure 6.

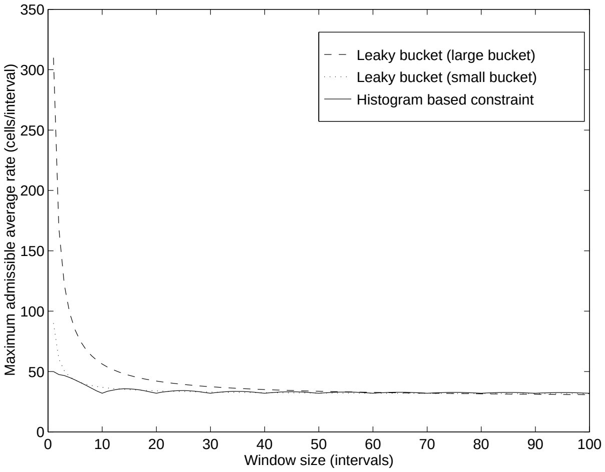


Figure 6: WCA curves for (i) Histogram based, (ii) LB (large bucket) and (iii) LB (small bucket).

In this simulation, the model generates the source traffic with all possible choices of parameters α and β (chosen on a discrete grid), and the policing functions are used to enforce these source traffic generated by the model. The probability that the source traffic violates each policing function can be empirically measured. For each policing function, we find out the set of admissible parameters α and β, with which the traffic generated can go through the policing function with violation probability smaller than 5% and 1%. From the corresponding admissible region of α and β, we can compare the discrimination versatility of each policing method. From the discussion about the source model, we know that $\frac{\alpha}{\beta}$ is related to the long term mean rate as eq (1), and the the value of α or β with a same $\frac{\alpha}{\beta}$ value can generate source traffic with different burstiness but same mean rate. Instead of plotting the admissible region in terms of α and β, we plot this region in terms of long term average rate v.s. the value of α. The admissible region of α and β of the HBC is smaller than that of other constraints because HBC is more discriminating on the burst type traffic.

We can get a general idea of how each policing function constrains the source in terms of mean rate and burstiness. Figure 7 shows the admissible region of each policing function.

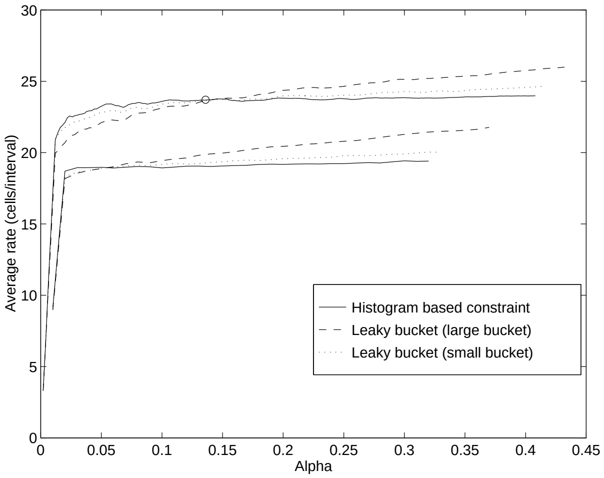


Figure 7: Admissible subset of the source traffic. The region below each curve is the admissible region of α and average rates such that the traffic generated by the model can go through the policing function with violation probability $\leq$ 5% (upper 3 curves) and 1% (lower 3 curves).

3.3. Real Video Trace

In this experiment, the input to the three interface schemes is the MPEG Football sequence 150 frames long filmed at 25 frames/second. The original format is in YUV, and the dimensions are 360x240 pixels. It was encoded using MPEG-2 standard with software provided by [10]. There are 12 frames in the group of pictures (GOP) and the I/P frame distance is 3, i.e., IBBP. The footage was encoded at 46020 bits/frame. The experiment is conducted at the interval level time frame (1/250 second) to observe the intraframe as well as interframe interactions.

Once again, we refer to the two LBs mentioned earlier in the introduction. The parameters of the HBC and the two LBs are shown in Table 1. The size of these parameters were chosen as small as possible, but still sufficiently large to admit the MPEG video source into the network without tagging any cells. Figure 8 shows the three WCAs (HBC, LB(a), LB(b)). Notice the large buffer size of LB(a). A video source using LB(a) could potentially transmit 180 cells in a single interval without any cells being tagged or rejected. A burst of such magnitude could degrade the QOS for other sources sharing a common ATM channel.

While LB(b) is significantly less susceptible to burst than LB(a), its long term rate is twice the mean rate of the MPEG encoded Football sequence. This is done to accomodate the occurence of bursts in close proximity. LB(b) is a costly constraint to implement in terms of bandwidth and

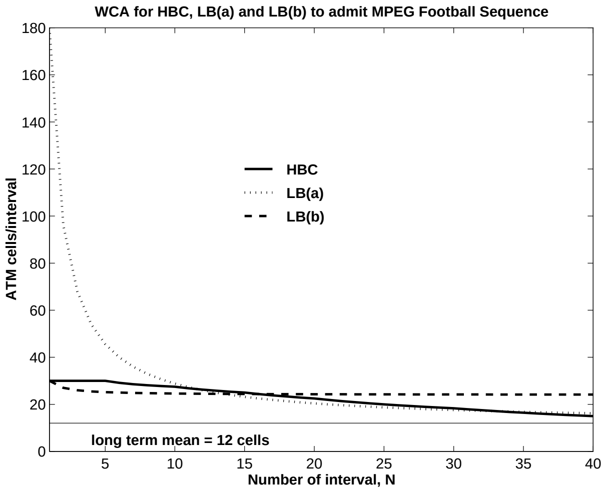


Figure 8: The WCAs for HBC, LB(a), and LB(b) necessary to admit MPEG Football sequence. The dimensions of the three constraints are given in Table 1.

an example of overdimensioning. Using LB(b) will decrease the overall SMG of ATM channels.

On the other hand, the HBC can perform long term rate control as well as make the network impervious to large bursts. More importantly, HBC will be able to accomodate other MPEG sources with the same mean and maximum burst rate as our MPEG Football sequence, but with different proximity in the burst occurrence, just by altering the intermediate points (points other than max burst rate and long term rate) of its WCA.

MPEG at 46020 bits/frame = 12 cells/interval							
HBC						LB	
i	r_i	h_i	i	r_i	h_i	(a)	(b)
1	5	10	5	25	5	leak rt	leak rt
2	10	10	6	30	5	12	24
3	15	5	7	40	0	buf size	buf size
4	20	5	8	50	0	180	30

Table 1: Dimensions of the HBC and Leaky Buckets (a) & (b), required to admit MPEG Football sequence encoded at 46020 bits/frame. Units for r_i and h_i are (# cells/interval) and (# intervals) respectively.

4. CONCLUSIONS

In this paper, we have introduced a source-network interface for a VBR video that is compatible with the currently specified UNI and has some inherent advantages over the LB. The advantages are a greater discriminating capability than the LB and a more versatile policing mechanism to better suit the need of the video source according to its burstiness. A possible future research direction is to build a call admission function based on the HBC and try to improve the impermeability of the QOS of neighboring sessions sharing the same ATM channel.

5. REFERENCES

[1] W. Verbiest, L. Pinnoo, and B. Voeten, "The impact of the ATM concept on video coding," *The IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 1623–1632, December 1988.

[2] "The ATM Forum." Prentice-Hall, 1993. ATM user-network interface specification, V 3.0.

[3] "Traffic control and congestion control in B-ISDN." Intl. Telecomm. Union, March 1993.

[4] J. S. Turner, "New directions in communications (or which way to the information age?)," *IEEE Commun. Mag.*, vol. 24, pp. 8–15, October 1986.

[5] E. P. Rathgeb, "Modeling and performance comparison of policing mechanisms for ATM networks," *IEEE J. on Sel. Areas in Comm.*, vol. 9, pp. 325–334, April 1991.

[6] A. Ortega and M. Vetterli, "Multiple leaky buckets for increased statistical multiplexing of ATM video," *Packet Video Workshop, Portland, OR*, September 1994.

[7] P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 446–458, August 1993.

[8] H. Heeke, "A traffic-control algorithm for ATM networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, pp. 182–189, June 1993.

[9] M. W. Garrett, *Contributions Toward Real-Time Services on Packet Switched Networks.* PhD thesis, Dept. of Electrical Eng., Columbia Univ., 1993.

[10] MPEG Sofware Simulation Group `ftp://ftp.netcom.com/pub/cf/cfogg/mpeg2/`.