In Proc. of Packet Video Workshop '94, Portland, Sept. 1994

MULTIPLE LEAKY BUCKETS FOR INCREASED STATISTICAL MULTIPLEXING OF ATM VIDEO

*Antonio Ortega**
Dept. of EE-Sys. and SIPI,
Univ. of Southern California,
Los Angeles, CA 90089,

Martin Vetterli
Dept. of EECS,
Univ. of California,
Berkeley, CA 94720

ABSTRACT

We examine the difference between the short term traffic descriptors that are relevant to the network and the long term averaging that is common in video coding. We propose to combine short and long term rate constraints, e.g. several leaky buckets with different window sizes, so as to accomodate the network requirements, by restricting short term bursts, while preventing "greedy" video encoder operation, by enforcing very long term rate constraints.

1. INTRODUCTION

Recent implementations of packet video transmission have been reported for video over local area networks [1] or video multicasting over the Internet [2]. Both cases have in common the lack of Quality of Service (QOS) requirements for the network performance. The user can only expect "Best-effort" performance from the network and therefore a rate control at the encoder is needed in order to change the video frame rate and/or frame quality depending on the network conditions. (If the rate was not changed, the information might sometimes, e.g. when congestion occurs, be received too late to be usable by the receiver.) In this paper we will argue that rate control may not only be needed even in a guaranteed environment such as that offered by ATM networks [3] but may actually be *beneficial* for the overall performance of these networks, if an appropriate type of rate constraints are set up.

Thus far much of the work on packet video has involved little interaction between the network and video coding communities. Analyses of network performance have tended to assume that a video source could be characterized by a more or less elaborate probabilistic model [4, 5, 6], while work on encoding schemes for packet video coders has tended to see the network as a black box, as determined by the source policing interface [6]. The announced statistical multiplexing gain (SMG), i.e. increased network efficiency and better video quality, could then be achieved provided that the real world sources behaved as predicted by the model. This type of analysis could be misleading in that (i) it may be hard to characterize the sources when more than a few seconds of encoded video are considered [7] and, more importantly, (ii) the models do not take into account that for a given network constraint the source might be using some kind of rate control. Fig. 1 illustrates the idea of "self-regulating coders" [8, 9]. While, typically, models tend to characterize sources operating with a "constant quantizer" mode (Fig. 1(a)) the rate control needed to keep a source within the constraints set by the policing function (Fig. 1(b)) will require that different quantizers be used as is also the case in CBR encoders (Fig 1(c)). In [10] it was argued that the "constant quantizer" assumption is likely to be incorrect, as video encoders are designed to maximize the quality for the available transmission resources. We termed this approach "greedy" and proposed an alternative "non-greedy" approach which involved encoders using just enough bit rate to provide "sufficient" quality. The non-greedy approach (see also [9]) had the additional advantage that it provided performance equal to that of greedy encoders for the most difficult scenes while using less total rate, thus making possible increased SMG.

In this paper, our initial assumption is that encoders *do* perform rate control and our goal is to consider the problem of *designing* rate constraints that will be beneficial for the network while enabling good quality for the encoded video. Our work differs from previously reported work in two ways: (1) we do not fix the video bit rate and instead assume that it will be close to the largest allowable rate within the constraint, i.e. we pessimistically assume video encoding algorithms to be greedy. (2) Our goal is thus not to "match" the rate constraint to some previously determined bit rate sequence, as in [11] for instance, but rather to design it based on assumption (1) and the need for efficient network utilization.

Section 2 examines possible alternatives for rate constraints and notes the difference in the time scales that are relevant at the encoder and the network. In Section 3 we show that the danger posed by greedy source coding can be limited by resorting to schemes where the policing function constraints the bit rate at several time scales, e.g. using multiple leaky buckets.

2. TRAFFIC CONTRACTS, SHAPING AND RATE CONTROL

Current versions of the ATM user-network interface (UNI) specifications [3] call for a negotiation process, prior to connection set up, where the user and the network decide on a series of parameters that define the connection. One of the parameters that are agreed upon is a choice of one or several traffic descriptors, as well as their values. The considered traffic parameters, such as peak cell rate (PCR) or sustainable cell rate (SCR), are in fact operational, rather than statistical. That is, they are defined in terms of counters analogous to Leaky Buckets (LB) [11], which the network monitors in order to determine whether each cell of a given source is compliant. More precisely, a SCR traffic descriptor will be determined by a leak rate and a window size. The leak rate will be the maximum admissible cell rate averaged

*Work supported in part by a scholarship from the Fulbright Commission and the Ministry of Education and Science of Spain. This work was performed while at Dept. of Elec. Eng. and Center for Telecom. Research, Columbia University, NYC, NY.

Figure 1: Three configurations for transmission of VBR coded video. Note that the control box sets a quantization parameter Q. (a) Typical configuration for studying the statistical behavior of video source and modeling the output bit rate. (b) Self policing for transmission over a packet network. (c) Transmission over a CBR link.

over the window size. We will thus refer to high/low rate and short/long window SCR as appropriate depending on the chosen values for those parameters. A certain cell will be compliant if the LB or counter measuring the running average does not exceed the agreed maximum rate.

Note that determining whether the traffic offered by a certain source is compliant with that specified in the contract is only one of the tasks performed by the network as a part of its usage parameter control (UPC) or "policing" function. Thus, the networks may have other means of controlling the traffic and therefore a non-compliant cell may not necessarily be removed from the network.

It is important to note that the network and the users look at video traffic at different time scales.

On the network side the time scale corresponds to the cell level so that, for instance, the PCR corresponds to the minimum interval in between consecutive cells corresponding to a given video connection. Similarly, the SCR is measured on relatively short time intervals. The justification for this approach is that traffic descriptors are assumed to allow the network to make decisions on allocation and admission control [12]. From this perspective, parameters such as SCR with long time windows are useless because, in order to guarantee the QOS for those sources, the network would have to either use very long internal buffers or operate at low utilization. Each of these two possibilities is highly undesirable and thus traffic descriptors, even those measuring "averages" tend to be used over relatively short time windows.

Conversely, video encoders often use relatively small blocks (typically 8 by 8 pixels) as their basic coding units, but video encoding algorithms tend to average the rate over a frame, or even across frames. A good resource allocation within a frame may yield rather large variations in short term bit rate, thus making a short window SCR too restrictive unless its rate is sufficiently high. Moreover, in typical video coding algorithms such as MPEG [13], the bit rates per frame can cover a wide range (e.g. well over and below average for I and B frames respectively). In this situation, short term traffic descriptors are all the more meaningless.

To avoid this type of short term constraint, the encoder can resort to traffic shaping [8]. The idea is to store the encoded bits corresponding to a frame and then packetize them and transmit the corresponding cells at equally spaced time intervals during the duration of the next frame. This approach has the advantage of being transparent to the encoding algorithm, in the sense that traffic shaping is done after the frame has been compressed.

Another possibility would be to include the constraint defined by the traffic descriptor(s) within the encoding loop at the encoder. By doing this, the encoder would itself reduce locally the number of bits it spends so as to not violate the traffic contract. Rate control differs from shaping in that here the actual encoding (how many bits are used for each block) can vary depending on the constraints imposed by the network. There are two disadvantages if this approach is used with usual (i.e. short memory) traffic descriptors. For a low SCR (close to the average rate in a frame) the short memory may not be sufficient to average the large variations in rate within a frame so that quality may not be acceptable (e.g. runs of busy, high rate areas, will be blurred). Conversely, given a high rate SCR the encoder algorithm may adopt a greedy rate control so as to produce a rate always close to SCR, thus reducing SMG [10].

3. MULTIPLE LEAKY BUCKETS

The previous section has examined the different time scales considered in video encoding algorithms and network traffic descriptors. Long window traffic descriptors are useless to the network in that they cannot prevent high rate bursts and thus a compliant source may still cause congestion by being bursty. Conversely, short window traffic descriptors may either be too restrictive for typical video coders or are bound to be abused by malicious coders which will transmit at close to the monitored rate all the time.

Since the goal of transmission through ATM networks is to reap benefits to both video quality and network efficiency, we now propose a solution which tends to fulfill that objective by considering *both* short term and long term traffic descriptors. In what follows we will concentrate on LB traffic descriptors. Note that throughout this discussion we assume that the applicable cell-level, very short term traffic constraints are met through shaping at the encoder.

In [10] it was shown how encouraging the use of non-greedy coding techniques can provide appropriate multiplexing gain while maintaining the video quality for the most difficult scenes. Our point of view in this section is to assume that it may not be always realistic to assume sources behaving in a non-greedy fashion. Therefore we study the trade-off involved in choosing the leaky bucket by looking at the maximum average rate that can be generated while still abiding by the given LB constraint. We will call these the worst case bursts and they will be measured as the maximum average that can be used over a window of i frames when a $LB(N_b, R)$ is used. We here denote $LB(N_b, R)$ a LB of window size N_b and leak rate R, where the window size is given in units of frame intervals, i.e. 1 corresponds to one frame interval, and R corresponds therefore to the "average" rate per frame.

For a given single $LB(N_b, R)$ constraint the admissible sequences can be very different. As an example, a sequence where every frame uses R bits is admissible, as will be one that uses $N_b \cdot R$ bits for every N_b-th frame and zero bits for those in between. Obviously, these are extreme cases but indicate that sources with varying degrees of "peakiness" can be admissible.

To better understand the trade-offs involved we define a curve that can describe the "worst case" performance of LB policing mechanism. We plot $R_{MAX}(i)$ which we

Figure 3: Motivation for using a double leaky bucket. The worst case short term behavior is determined by the short bucket, while the long term average is set by the long bucket. As before the range of admissible average rates is represented by the area under whichever curve is closer to the x-axis, for a given window size.

We now show an example of how a double LB can produce the desired result of allowing a certain sequence to be transmitted but preventing worst case scenarios (that could result when using only one LB). We use the coding examples of [10], which include a greedy and a non-greedy sequence, to choose the appropriate parameters for the LB.

We consider two separate single LB schemes. First a short window LB, LB(3,60), is chosen, see Fig 4. Here the maximum allowable peaks are small (180kbit/frame) whereas the long term average is 60kbit/frame. Thus the danger is that a greedy source could use continuously 60kbit/frame. Indeed the greedy source of [10] would be admissible under these constraints. Conversely one could choose a longer window LB, LB(60,55), see Fig 5 where the long term average would be kept lower (55kbit/frame). However the danger here is that a source could be admissible while generating a peak rate of up to 3300kbit for one frame.

When the two LB are combined, see Fig 6, we observe that the unwanted properties of each of the single LB schemes are avoided. Thus the maximum short term

peak is kept small as is the long term average.

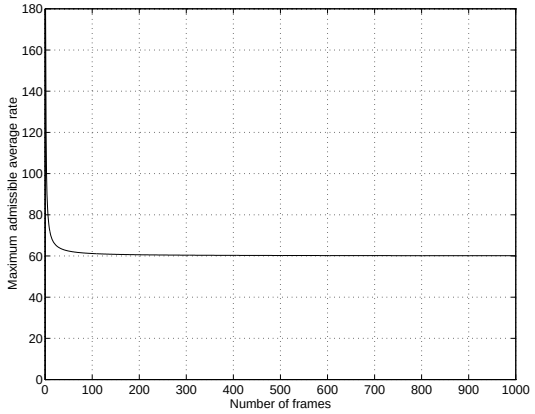

Figure 4: Worst case burst curve for a LB(3,60) that has been chosen for the non-greedy source of [10]. The window is short ($N = 3$) and thus the leak rate has to be large enough to permit the larger frames to be sent. The drawback is that the long term average is 60kbit/frame, while the actual sequence's average was 46.3 kbit/frame. The greedy sequence of [10] would also be admissible.

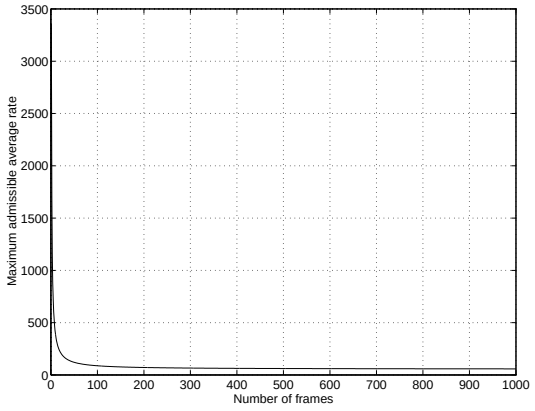

Figure 5: Worst case burst curve for a LB(60,55). The non-greedy sequence of [10] is also admissible under this LB. Note that the longer window $N_b = 60$ enforces a lower long term average. However, there is the danger that a compliant source may generate burst of up to $3000 kbit/frame$!.

4. CONCLUSIONS

We have examined the problem of designing rate constraints for video transmission over ATM networks. We have proposed using both short term and long term traffic descriptors as a way of accommodating the differing time-scales of video encoders and network managers. An example of this approach involving the use of two leaky buckets has also been presented.

REFERENCES

[1] A. Eleftheriadis, S. Pejhan, and D. Anastassiou, "Algorithms and performance evaluation of the Xphone multimedia communication system," *ACM Multimedia 93 Conf.*, (Anaheim, CA), pp. 311–320, Aug. 1993.

[2] J.-C. Bolot and T. Turletti, "A rate control mechanism for packet video in the internet," *Infocom'94*, (Toronto), pp. 1216–1223, Jun. 1994.

[3] ATM Forum, *ATM User-Network Interface Specification, Version 3.0.* Prentice-Hall, 1993.

[4] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. on Comm.*, vol. 36, pp. 834–843, Jul. 1988.

[5] W. Verbiest, L. Pinnoo, and B. Voeten, "The impact of the ATM concept on video coding," *IEEE JSAC.*, vol. 6, pp. 1623–1632, Dec. 1988.

[6] "Special issue on packet video." IEEE Trans. on CAS for Video Tech., Jun. 1993.

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

[8] G. Rigolio, L. Verri, and L. Fratta, "Source control and shaping in ATM networks," *GLOBECOM'91, Phoenix*, 1991.

[9] A. R. Reibman and B. G. Haskell, "Constraints on variable bit-rate video for ATM networks," *IEEE Trans. on CAS for video tech.*, vol. 2, pp. 361–372, Dec. 1992.

[10] A. Ortega, M. W. Garrett, and M. Vetterli, "Toward joint optimization of VBR video coding and packet network traffic control," *Packet Video Workshop*, (Berlin), Mar. 1993.

[11] E. P. Rathgeb, "Modeling and performance comparison of policing mechanisms for ATM networks," *IEEE JSAC*, vol. 9, pp. 325–334, Apr. 1991.

[12] J. W. Roberts, "Variable-bit-rate traffic control in B-ISDN," *IEEE Comm. Mag.*, pp. 50–56, Sept. 1991.

[13] D. LeGall, "MPEG: a video compression standard for multimedia applications," *Comm. of the ACM*, vol. 34, pp. 46–58, Apr. 1991.

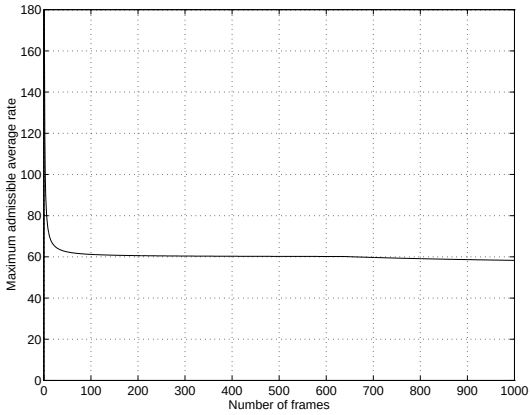

Figure 6: Effect of combining two LB's. The resulting worst case burst curve shows both the lower long term average and smaller short term burst.