Bandwidth Allocation for VBR Video Service in High-speed Networks Based on Adaptive Filtering

Mladen Kos, Smiljan Pilipović and Alen Bažant

Department of Telecommunications
Faculty of Electrical Engineering and Computing
HR-10000, Zagreb, CROATIA
E-mail: {kos,smiljan,bazant}@tel.fer.hr

Abstract

Regarding video transmission in ATM networks big and yet unsolved problem is capacity allocation. During the setup of variable bit rate (VBR) video transmission service the negotiation of parameters tied with quality of service (QoS) takes place. Based on committed parameters it is necessary to determine the optimal capacity of the connection. According to some recent research, the statistics of information flow (amount of data for each frame) of VBR video traffic can be divided in low- and high-frequency component. Long lasting source activity, i.e. spectral components in the low-frequency region, is extremely important for the buffer overflow. Therefore it is satisfying to observe only low-frequency components for the description of the queue behavior in a buffer. In this article we propose an adaptive scheme for the variation of the cut-off frequency based on the buffer usage in a certain time slot.

1. Introduction

Regarding video transmission in ATM networks big and yet unsolved problem is capacity allocation while maintaining in the same time committed quality of service. During the setup of variable bit rate (VBR) video transmission service the negotiation of parameters tied with quality of service (QoS) takes place [1]. Based on committed parameters it is necessary to determine the optimal capacity of the connection, i.e. the maximum allowed speed at the point of entry of video traffic in the network. Multimedia traffic, especially its most important part, video streams, have unique features in comparison with other types of traffic: strong correlation between adjacent slices in each frame, between adjacent frames and groups of frames, and high burstiness. In the same time video services are time sensitive because the presentation time of each frame is defined in the moment of frame encoding. Every frame that comes after the specified time is of no significance for the presentation quality (i.e. it equals to lost and frames with errors).

Because of its inherited burstness the most appropriate service class for the transport is variable bit rate - real time (VBR-rt) [1]. Traffic parameters for VBR-rt service class are average bit rate, peak cell rate and maximum burstiness size. After the connection setup an usage parameters control (UPC) takes place. A network controls user's behavior and locates users that are not conforming to prearranged traffic parameters via UPC mechanisms. Noncomforming traffic is discarded or tagged for the discarding in neighboring nodes.

From the user's point of view, discarding of data leads to video quality deterioration and because of that, misbehavior is totally unwanted. In order to lessen the danger of data discarding, user could demand more bandwidth during the connection setup, but it leads to higher costs for the user and lower network utilization. Appropriate selection of traffic parameters has to accomodate these contrast demands.

This paper focuses on the link capacity allocation and the specification of encoder buffer output rate. We use a novel approach to this matter that has been proposed by San-qi Lee et al. in [2], [3] and [4]. Since the video traffic exhibit strong correlation the knowledge of first order statistics in capacity allocation has been shown inefficient and it is necessary to know the second and higher order statistics. Second order statistics can be expressed via autocorrelation function:

$$R(\tau) = E\big(a(t)\,a(t+\tau)\big) \qquad (1)$$

where $a(t)$ is a bitrate (number of bits in a frame). Equivalent representation of the same statistical properties

0-7803-3879-0 / 98 / $10.00

can be expressed by spectral transformation (Fourier transform) of (1) [2]:

$$P(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega\tau} d\tau \qquad (2)$$

$P(\omega)$ in this case is a power spectral function of the arrival process: $P(\omega) = |A(\omega)|^2$ where $A(\omega)$ is a Fourier transform of the arrival process $a(t)$.

In [4] authors discussed spectral characteristics of the bitrate generated in a video encoder and the impact of each frequency component on the queuing characteristics of the video stream. It has been shown that using low frequency components for driving the output bitrate gives good features in queuing characteristics (relatively low queue length and a fair link utilization. However the selection of the satisfactory cut-off frequency for the low - frequency components is left for further study.

In this paper we propose an adaptive mechanism where the cut-off frequency is not fixed at a certain value, but it can be adapted as a function of buffer occupancy.

The paper is organized as follows. In section 2 we define the problem and make an overview of link capacity assignment via filtered input rate. In the same section the framework is defined for analytical analysis of link capacity assignment techniques. In section 3 we define our proposal in defining cut-off frequency. Section 4 gives an experimental proof of our theory and the paper is concluded in section 5.

2. Link capacity allocation

Figure 1 explains the problem of link capacity allocation. Video encoder (or data storage system - if it is the off line presentation) is usually equipped with the buffer before the data are sent to the network. The aim of that buffer is to smooth the data stream before it reaches network. The amount of smoothing is subject of our research in this paper.

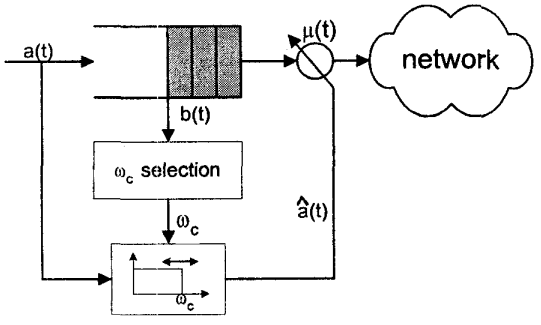

Fig. 1: Link capacity allocation

Video encoder produces cell stream $a(t)$. Here we put accent on the number of cells per frame, although the most accurate and instant values for this variable will be:

$$a(t) = \frac{1}{t(n) - t(n-1)} \text{ for } t \in [t(n\text{-}1), t(n)] \qquad (3)$$

where $t(n)$ is the arrival time of the nth cell in encoder buffer (figure 2).

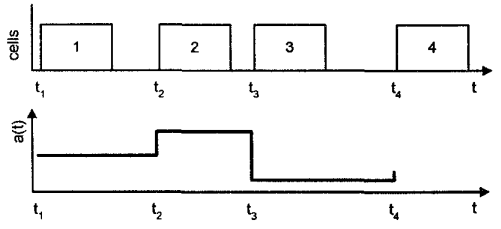

Fig. 2: Evaluation of cell rate

The cells from the encoder are stored in the encoding buffer whose occupancy is $b(t)$. The output rate from the buffer is $\mu(t)$. The output rate can be constant or it can be adaptively changed.

2.1 Static allocation

Output cell rate can be assigned in many different ways. If we specify:

$$\mu(t) = \max_t a(t) \qquad (4)$$

where link will be poorely utilized, but on the other hand, the need for the encoder buffer does not exist. Since the buffer outputs speed does not change during the time, an appropriate class of service for this case is constant bit rate (CBR). In frequency domain, (4) equals to $\omega_c = \eta$ where

$$\eta = \min_n [t(n) - t(n-1)] \qquad (5)$$

Another extreme case is link capacity allocation as:

$$\mu(t) = \max_t \hat{a}(t) \qquad (6)$$

where

$$\hat{a}(t) \equiv \frac{1}{T} \int_{t-T/2}^{t+T/2} a(t) dt \qquad (7)$$

and T is a duration of corresponding element in video coding (slice, frame, group of pictures, etc.). In this way we are approaching fluid flow models for queuing behaviour analysis. In frequency domain (6) and (7) are analogous to filtering out frequency components greater than $\omega_c = 2\pi / T$. In this case cells' arrival time during the observed period are neglected and there is a probability that all cells arrive in the same time at the beginning of the period T. On the contrary, $\mu(t)$ is specified for the case where the arrivals are evenly disposed through the period T. Because of that in this case a small buffer of the length $\max \hat{a}(t)$ should exist. The link utilization is greater than in (4) but essentially it is the case of CBR traffic with no statistical multiplexing.

The allocation can be performed with:

$$\mu(t) = \overline{a(t)} = \overline{\hat{a}(t)} \qquad (8)$$

where $\overline{(\,)}$ stands for the average value. In frequency domain it is analogous to filtering at the frequency of 0 Hz. The utilization is maximal and equals to 1, but this arrangement needs the buffer whose length should be infinite..

So far we have analized filtering at maximal and minimal frequency (from 0 to $2\pi/\eta$). Beside these frequencies, ω_C can be any other frequency from the interval $[0, 2\pi/\eta]$. If the cut-off frequency is changed, the link utilization and output cell rate are changed either. But, for the entire scheme of static allocation stands that its most appropriate transport scheme is constant bit rate and no statistical multiplexing takes place.

2.2 Dynamic allocation

If output bitrate from the encoding buffer is changed on regular basis or periodicaly than we speak about dynamic link allocation case. The cut-off frequency can be taken from the same set as in previous section, but the output bitrate is calculated not from the maximal filtered rate, but from:

$$\mu(t) = C\,\hat{a}(t) \qquad (9)$$

since C is a constant and

$$E[\mu(t)] = C\,E[\hat{a}(t)] \qquad (10)$$

$$C = \frac{E[\mu(t)]}{E[\hat{a}(t)]} = \frac{E[\mu(t)]}{E[a(t)]} = \rho^{-1} \qquad (11)$$

where ρ is an average link utilization.

In [4] it is shown that dynamic approach has smaller necessary buffer, while it maintains the same link utilization.

3. Novel proposal for the adaptive dynamic allocation

In this case the output from the buffer is variable and the most appropriate class of service is variable bit rate (VBR). As it was mentioned in the introduction, VBR service has three traffic parameters that are negotiated during the connection setup. They are: average bit rate, peak bit rate and maximum burstiness size. During the duration of the connection the commited parameters are controled via GCRA algorithm or leaky bucket [1]. In [5] authors made an analysis of buffered CBR and VBR connection with leaky bucket control. If the leaky bucket has parameters: drain rate $\overline{C}$ and the leaky bucket size LB_{max}, we can write buffer constraint as follows:

$$0 \le b(t) = \int_0^t a(\tau)d\tau - \int_0^t \mu(\tau)d\tau \le b_{\max} \qquad (12)$$

$$0 \le \int_0^t \mu(\tau)d\tau - t \times \overline{C} \le LB_{\max} \qquad (13)$$

(12) is a constraint for encoder buffer, while (13) is a similar constraint for the state of leaky bucket. Adding (12) and (13) we obtain:

$$0 \le \int_0^t a(\tau)d\tau - t \times \overline{C} \le b_{\max} + LB_{\max} \qquad (14)$$

In that way VBR connection can be treated as an CBR connection with the buffer length of $LB_{max}+b_{max}$ and the bitrate of $\overline{C}$. The key questions are how to distribute buffer occupancy among virtual buffer of leaky bucket and the real encoding buffer and why do we have to use VBR traffic if the same results are obtained with CBR and parameters specifed above. The answer to the second question is that by reducing the capacity of encoding buffer we reduce service delay which is very important in the case of interactive services. Concerning occupancy distribution and taking the network's benefits into accout we think that the fair algorithm would be to use proper buffer as much as possible and using the leaky bucket possibilities only in the moments of extreme needs. In that way burstiness of the video traffic (observed by the network) would be significantly reduced while the service delay is in acceptable bounds.

From the previous discussion it is clear that the most appropriate selection of buffer output rate will be such that keeps buffer occupancy large. But the selection of cut-off frequency with this constraint depends on character of the video sequence, coding patern (I, P and B pictures in MPEG standard [6]), selection of quantizers, etc.

Because of that we suggest following algorithm which is depicted in fig. 1. The information of buffer occupancy $b(t)$ is transmitted to the block for ω_C selection. As a signal processing (filtering) is conveyed by digital filters, it is possible to estimate several sets of filter coefficients for different cut-off frequencies and, depending on the selected cut-off frequency, it is possible to use one of the pre-estimated sets. In the following section we are going to analize the number of useful sets of parameters and the associated buffer occupancy.

Doing that way the accent is put on the maintenance of requested quality of service (delay is bounded by the length of the buffer, number of discarded cells due to the fact that unconforming behaviour is lessen because the fluctuation of the bitrate is as low as possible). The fluctuations of information flow in the buffer are partially compensated, and part of them is transferred to the network.

4. Experimental results

The starting point in the analyses of a link capacity allocation mechanisms is the input traffic. Regarding the fact that analytical models of multimedia traffic are to complex and they still do not adequately describe input traffic, all the analyses in our paper are based on the finite data sequence obtained from the movie Mr. Bean which has been published by the courtesy of Institute of Computer Science, University of Würzburg and can be obtained by ftp://www-info3.informatik.uni-wuerzburg.de /pub/MPEG/traces/MrBean.tar.gz. The sequence from the above mentioned film is coded using an MPEG-1 coder [7]. The file containes 40,000 integers each of them representing number of bits per frame (each frame consists of one slice). The recording frequency was 25 frames per second and the GOP pattern was IBBPBBPBBPBB. The entire trace and first 20 data with GOP pattern visible are shown in fig. 3 and fig. 4, respectively. The values are presented in bits per frame originaly. Since we are interested in the behaviour of video sequence in ATM network, the values had to be converted in cells/frame. The conversion depends on the AAL layer that is used, the length of the PDU in that layer, padding bytes in the last cell, etc. We made an assumtion that padding bytes can be neglected because of the large quantity of cells for each frame and an AAL5 layer is used for the coding. In that way we get the converted values by simple multiplication of original values with 1/(8×48) where 8 stands for 8 bits/byte and 48 stands for 48 bytes of payload/one cell. The major data for observed stream is: mean value = 43.9 cells/frame, maximum of 597 cells / frame and minimal value = 1 cell / frame. The standard deviation in this proces is 52.12.

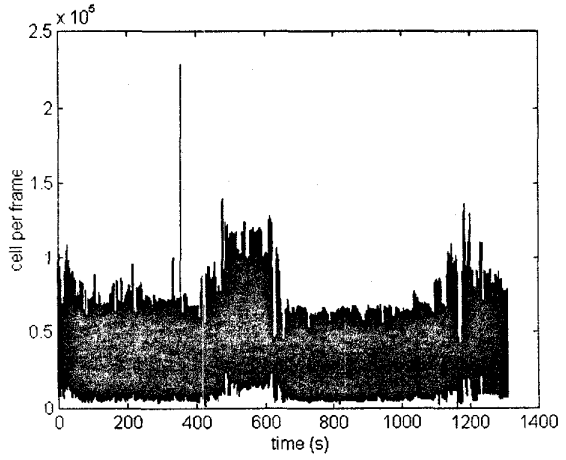

Fig. 3: Trace from *Mr. Bean*

Fig. 5 depicts power spectral density for the same video trace. The frame frequency (frame rate) is 25 Hz, GOPs are 12 frames long, so their frequency is around 2.1 Hz which is visible in the spectra.

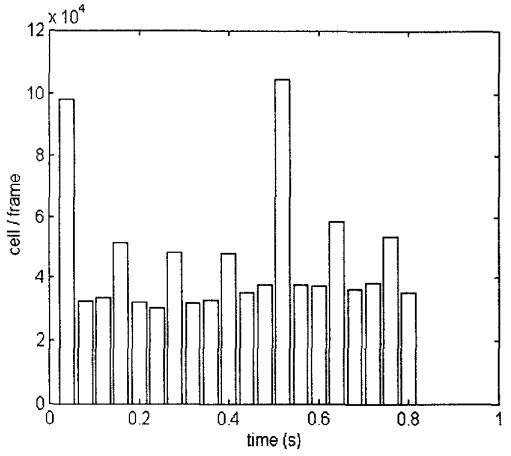

Fig. 4: GOP pattern

The highest peak is at 8.3 Hz and it comes from the patter XBB where X stands either for I or P picture. The highest power concentration is in lowest frequencies which is more visible in Fig. 6. On that picture we gave the power distribution in spectral domain, i.e. the power quantity ratio which is NOT in the band specified by frequency on the abscissa. It can be seen that more than 65% of power is in very narrow band less than 1 Hz.

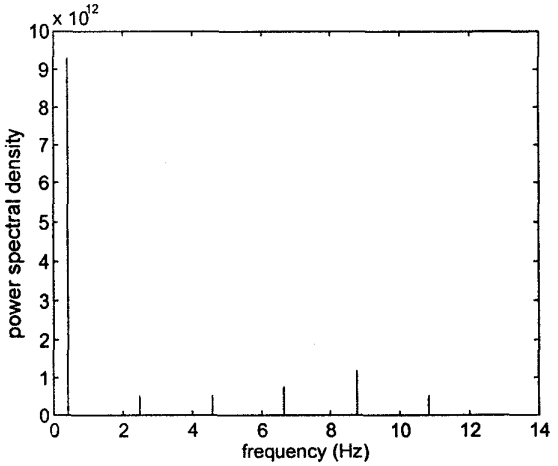

Fig. 5: Spectral characteristics of the observed trace

According to above observation we have selected cut-off frequencies of 1 Hz and 3 Hz that collect arround 70% of the signal power. The filter coefficients should be calculated in advance and they should alternate according to the buffer occupancy.

In experimental conditions we were using another approach, FFT. FFT can not be implemented in real time systems because it has infinite response, but since our aim is just to proove the correctness of our proposal we have

made filtering via removing spectral components from the spectral transform and than moving the signal back to time domain. The controlling signals, i.e. filtered components of original signal are depicted in fig. 7.

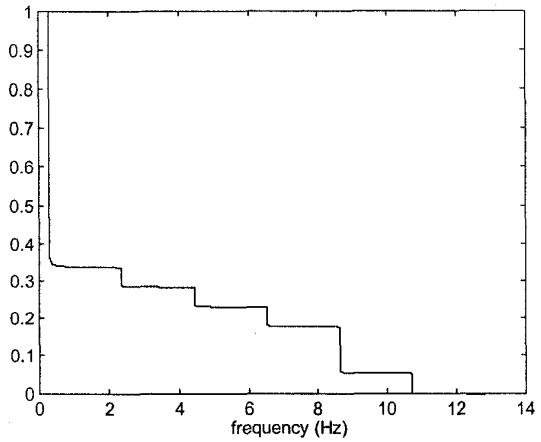

Fig. 6: Power distribution in frequency domain

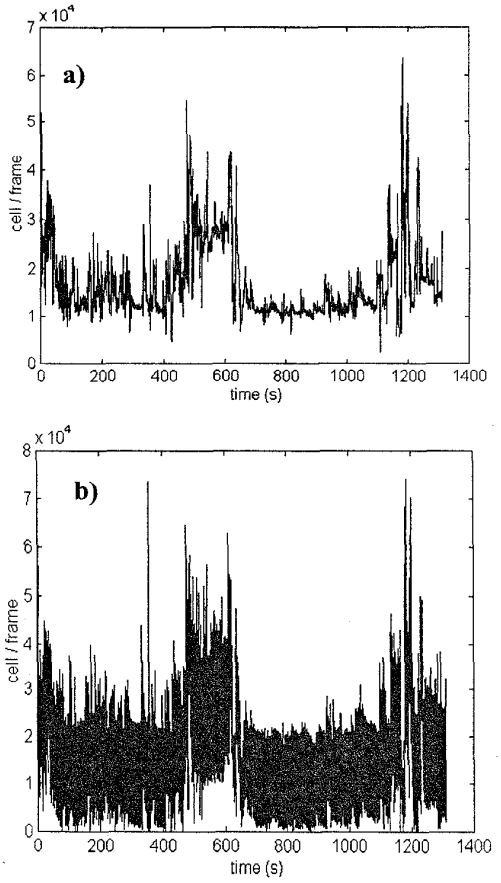

Fig. 7: Service rates for selected cut-off frequences

5. Conclusion

In this paper we have discussed the problem of allocation buffer output rate in video applications. From the network point of view it is desirable to send as little bursty traffic as possible. In order to respect this constraint we proposed technique that adaptively changes output bitrate as a function of the buffer occupancy. This technique is based on the filtering statistical data about cell arrival into the encoder buffer.

Acknowledgments

This research is supported by Croatian Ministry of Science and Technology (Project 036-004) and Croatian Post and Telecommunications (HPT) (contract 0/14-24/87-97).

References

[1] ATM Forum, *ATM User-Network Interface (UNI) Specification, Version 3.1*, Prentice Hall, 1995.

[2] S.-Q. Li, C.-L. Hwang, "Queue Response to Input Correlation Functions: Discrete Spectral Analysis", *IEEE/ACM Transactions on Networking*, Vol. 1, No. 5, 17, pp. 522-533, October 1993.

[3] S.-Q. Li, C.-L. Hwang, "Queue Response to Input Correlation Functions: Continuous Spectral Analysis", *IEEE/ACM Transactions on Networking*, Vol. 1, No. 6, 17, pp. 678-692, December 1993.

[4] S.-Q. Li, S. Chong and C.L. Hwang, "Link Capacity Allocation and Network Control by Filtered Input Rate in High-Speed Networks", *IEEE/ACM Transactions on Networking*, Vol. 3, No. 1, pp. 10-25, February 1995.

[5] C.-Y. Hsu, A. Ortega, A. R. Reibman, "Joint Selection of Source and Channel Rate for VBR Video Transmission Under ATM Policing Constraints", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 6, pp. 1016-1028, August 1997.

[6] MPEG1 Draft International Standard, ISO/IEC JTC 1, 1992

[7] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems", *Proc. of the 20th Annual Conference on Local Computer Networks*, pp.397-406, Minneapolis, MN, 1995.