A hybrid VBR/ABR Service for Scalable MPEG2 Video Networking : A Simulation-based Analysis

Ahmed Mehaoua and Raouf Boutaba

Computer Science Research Institute of Montreal,
1801, Av. McGill College, Suite 800,
Montreal (Qc), H3H 2P2 - CANADA -
Email: amehaoua@crim.ca

Abstract

ABR will be one of the primary available ATM service for carrying data. Since it is based on the use of excess bandwidth in the network, it is expected to have a lower cost usage. Although, it is intended for non-real time applications, the inclusion of a minimum cell rate (MCR) makes it an attractive and economical candidate for the transmission of delay tolerant video applications. Therefore, in this paper we evaluate the effectiveness of a hybrid ATM service, composed of a dual VBR/ABR connection, for the transport of a 2-layer MPEG2 encoded application. We compare the ABR binary feedback and Explicit Rate mechanisms in situation of carrying the enhancement video sublayer for various data partitioning configurations. To carry out the comparison, we consider the variations of the switches buffer occupancy, the end-to-end cell transfer delay (CTD), and the 2-point Cell Delay Variation (CDV).We show that for one-way video applications, the mean end-to-end delay is below the recommended 1 second but at the cost of large buffer at the destination due to excessive jitter.

Keywords : ABR, MPEG2, Scalable coding, Delay, Jitter.

1. Introduction

An important set of emerging multimedia applications, such as live video-casts or television news, lie between the two extremes of interactive bi-directional and pre-recorded unidirectional video. In these applications, many users may be willing to tolerate a playback delay of several seconds and a graceful picture quality degradation in exchange for a smaller bandwidth and cost price requirements.

The question arises as to which class of service and associated congestion control schemes are suitable for such flexible and adaptive video services.

ATM networks provide multiple classes of service to transport video applications : Constant Bit Rate (CBR), Variable Bit Rate (VBR), Available Bit Rate (ABR), and Unspecified Bit Rate (UBR) [1].

In 1995, the ATM Forum has specified that video-on-demand (VOD) may be carried by Constant Bit Rate (CBR) service using AAL5 [2]. This requires the use of a local buffer at the source for smoothing but ease the admission control process. Besides, it's commonly accepted that for video encoded applications, the use of CBR coding can result in perceptible picture quality variation and bandwidth waste [3].

Therefore, ATM Forum's SAA sub-working group is currently addressing the transport of unrestricted (i.e. open loop) VBR MPEG2 video over VBR-rt and ATM Block Transfer with Delayed Transmission (ABT-DT) [4].

A hybrid of the CBR and VBR approaches, called Renegotiated CBR (RCBR), attempts to combine the simplicity of admission control for CBR with the greater statistical multiplexing gains of VBR-rt [5]. However, RCBR needs to predict an effective bandwidth over long intervals, and thus it requires an accurate model for the distribution of sources rates over long time scales [6].

In the other hand, best effort services, such as *Available Bit Rate (ABR), is* based on the excess bandwidth in the network with an expected lower usage cost. Moreover, Admission control for ABR is based on the minimum cell rate (MCR) negotiated at the connection setup. Consequently, it provides the same simplicity as CBR for admission control phase. In the explicit rate ABR schemes, in-band Ressource Management (RM) cells are periodically transmitted by each source to indicate the currently desired rate. If the network is unable to provide the requested bandwidth, it may adjust the Allowed Cell Rate (ACR) value before returning the RM cell back to the source. This is a kind of rate renegotiation between the source and the network, which is performed at a shorter time scale than RCBR and provide higher statistical multiplexing gain than CBR.

0-7803-4443-X/97/$10.00 © 1997 IEEE

Therefore, although ABR is intended for non-real time applications, it is an attractive and economical candidate for the transport of delay flexible video applications such live video broadcast and video-on-demand.

Consequently in this paper, we are considering ABR for the transmission of the enhancement sublayer of a hierarchically encoded MPEG2 application using data partitioning. We are assuming that the base layer is distinctively carried by a guaranteed ATM service class, which is in our case a VBR connection with PCR reservation.

The overall objective of this work is to assess the ATM layer end-to-end performance for supporting a hybrid VBR/ABR connection for transmitting respectively the base layer and the enhancement layer of a scalable one-way MPEG2 encoded video application.

We study the effect of various network/source parameters on the variance of the buffer occupancy, the cell transfer delay (CTD) and the cell delay variation (CDV) of a reference 2-layer VBR/ABR video connection Examples of these parameters include, the data partitioning ratio (i.e. load balancing factor) and the ABR rate-base control mode (i.e. binary or Explicit Rate).

The paper is structured as follows. We briefly review, in section 2, the available scalable MPEG2 coding modes. The section 3 is devoted to the description of the statistical characteristics of the reference video connection and the hybrid VBR/ABR load balancing approach. In section 4, we present the simulation model, as well as the measured Quality of Service *(QoS)* parameters and the investigated scenarios. Finally, we discuss the results in section 5 and present our conclusion.

2. MPEG2 Hierarchical coding

Four scalable compression modes are defined in the MPEG-2 toolkit [7]. These coding techniques subdivide MPEG-2 video into numerous layers (base, middle, and high layers) mostly for prioritized transmissions [8]. At the destination, the lowest priority bitstreams, refered as enhancement layers, can be added to the base layer to display a higher quality picture. A brief summary of these different modes are presented below.

Spatial scalability : this mode codes a base layer at lower sampling dimensions (i.e. resolution) than the upper layers. This mode is useful in simulcasting, where a standard TV set needs only to decode the CCIT-601 720 x 480 base channel, and leave the higher HDTV 1440 x 960 data.

Temporal Scalability : the higher priority bitstream codes video at a lower frame rate (i.e. 15 *Hz*), and the intermediate frames are coded in a second bitstream to achieve a full frame rate (i.e. 30 Hz).

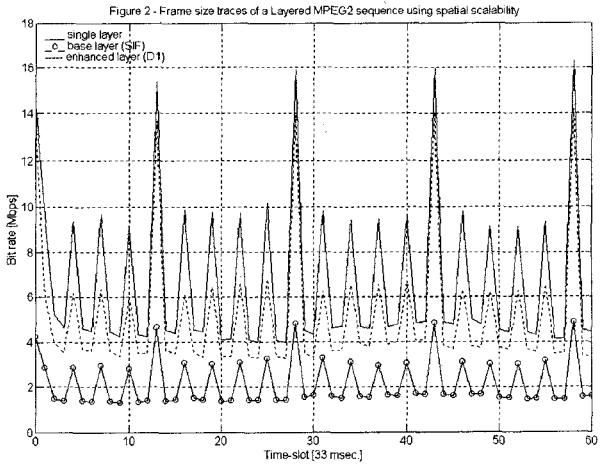
SNR Scalability : the layers are coded with differing picture quality by using different quantization step sizes (*Q*).

Data Partitioning : it is a frequency domain method that breaks the block of 64 quantized DCT coefficients into two bitstreams. The first, higher priority bitstream contains the lowest frequency coefficients and side information (such as motion vectors, macrobloc headers, ...). The second lower priority bitstream carries the remaining higher frequency AC coefficients.

3. The Hybrid VBR/ABR Service.

Due to implementation complexity, only few MPEG2 codecs provide these scalable coding modes. In addition, most encoders are implemented by software and are thus time consuming. For instance, the time needed to compute all the motion vectors of the following 61-frames ITU BT.601 704x480 (4:2:0) sequence is in the order of magnitude of days (3) using a DEC workstation. Its single, base and enhancement layers are presented in Figure 2. The data partitioning between the high priority base layer and the low priority enhancement layer is respectively of about 30% and 70%.

Figure 2 - Frame size traces of a Layered MPEG2 sequence using spatial scalability

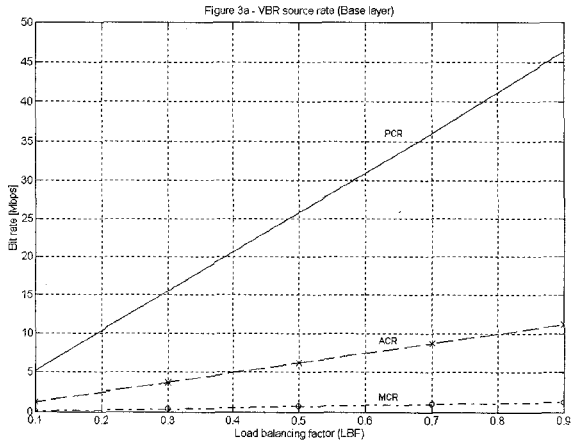


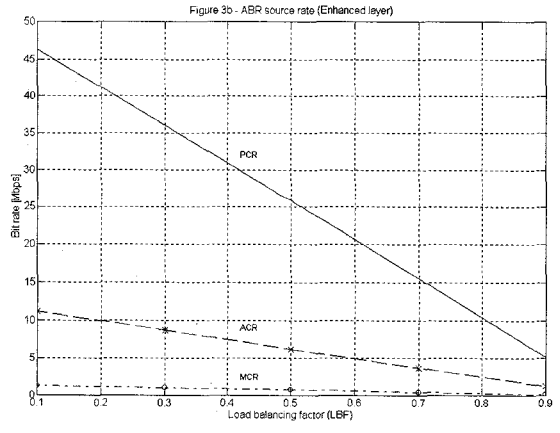
For our simulation analysis, we simulate static data partitioning to generate two complementary sub-streams from a single layer MPEG2 video sequence. The splitting approach used allows us to roughly approximate a real scalable codec. It precisely consists to divide the aggregate output bitstream by a ratio, named load balancing factor *(LBF)* ranging from 0.1 to 0.9, to generate two substreams transmitted over respectively a VBR connection with *LBF**ACR and a ABR connection with (1-*LBF*)*ACR. ACR corresponds

to the Average Cell rate of the reference stream. This partition is assumed static for every frame during the simulation run.

The tested reference MPEG2 video sequence has been encoded using the INRS-Telecom codec. The original *'TV News'* video sequence has been compressed using the following parameter set : GOP pattern : N=12 (24 fps) and M=3. Quantizer scales : 10 (I), 14 (P), 18 (B). Encoder input NTSC CCIR- 601 704x480 pixels with 24 bit color information. Number of frames : 240 (10 sec).

Figures 3a and 3b summarize the variation of the main bit rates for the two sub-layers as function of the load balancing factor *(LBF).* These traffic parameters are : the minimum cell rate *(MCR);* the average cell rate *(ACR)* and the peak cell rate *(PCR).* The corresponding Peak-to-mean ratio (B) is equal to 4.16 and gives us a measure of the burstiness level for the studied sequence. By way of comparison, the peak-to-mean ratio of a limited motion sequence like a *'TV talk-show'* is 2.8, whereas for a complex sequence like the well known *'Star-Wars'* is 5.0 [9]. Since the sub-layers are statically generated from the same bit stream using a complementary ratio, linear and symmetrical variations of the bit rates are visible in the figures.





4. Simulation Methodology

4.1 Quality of Service parameters

The evaluated performance parameters are the end-to-end cell transfer delay *(CTD)* and the *2-point Cell Delay Variation (CDV).* These quality of service *(QoS)* parameters are measured individually for the two sub-streams *(i.e. VBR, ABR)* of the reference connection. Emphasis is on the variation of these metrics at the cell level for each MPEG frame type sub-flow *(I, P, and B).* In order to accurately measure the impact of different switch control algorithms on the delay variation, the buffer queue lengths are also monitored and presented. We briefly remind in the following the definitions of these metrics.

The End-to-end queuing delay (CTD) is the time, $D_K = t_{0K} - t_{iK}$, between the departure of cell K from the source node (t_{iK}) and its arrival at the destination node (t_{0K}). The cells are assumed to be not lost or discarded during their transmission. This is achieved by using infinite buffer at the switches which accommodate all incoming cells.

As to the Cell delay variation (CDV), it is the variation of the periodic cell arrival process at the destination point. There are actually two performance parameters associated with CDV, the 1-point CDV and the 2-point CDV [10].

In this paper, we focus only on the 2-point CDV which is defined as the difference between the end-to-end delay experienced by the current cell K (D_K) and that experienced by cell 0 (D_0).

$$v_K = D_K - D_0$$

4.2 Network Simulation Model

We consider a simulation model consisting of two ATM switches (*SW1* and *SW2*), a bottleneck link (*L*) and a number of background connections crossing these switches as shown in the Figure 1. The distances between the source/destination and the switch nodes are constant and set to 0.125 miles (i.e. 0.2 km). The switch-to-switch link is equal to 625 miles (i.e. 1000 km). The background traffic consists of a number of ABR and VBR connections crossing each switch. In our model, all background streams exit at the output of each stage, and new traffic is added at the input of the next stage. We believe that this model representing the worst case traffic scenario is the most suitable for accurately evaluating the end-to-end delay variation. Indeed, when all the background connections originating from the source side sustain the path of the reference connection, all delay variations occur at the

first switch queue *(i.e. SW1)* since there is no contention between the cells at the subsequent node *(i.e. SW2).*

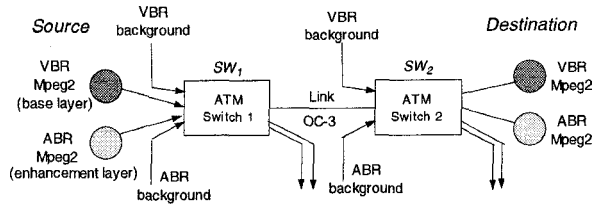


Figure 1 - The network model

The ABR background connections are characterized by an ON-OFF process with a mean burst length of 2 ms, a mean interval between bursts of 0.5 ms and a bit rate at which cells are generated during the ON periods of 100 Mbps. For the VBR background we generate 50 MPEG video connections using the *'Star-Wars'* MPEG-1 frame traces. The aggregate mean and peak cell rate of the background VP are respectively 18 Mbps and 212 Mbps. These 50 connections share the same VP but start transmitting at different instant in the range [0, 0.5 sec.].

We have noticed that the periodic properties of the background video traffic, i.e. GOP pattern, have an impact on the buffer occupancy and thus on the delay variations which confirms the conclusions in [11]. Besides, we also assume that the MPEG codec output rate are constant (i.e. piecewise-CBR fashion) during the transmission of each frame. The bit rate corresponds to *(frame_size*fps)* for each video connection.

We assume shared output FIFO buffers with two congestion thresholds : When the buffer occupancy exceeds the high threshold *(HT),* a congestion flag is set and the EFCI field of outgoing data cells is marked. When the buffer occupancy drops below the Low Threshold *(LT),* the congestion flag is cleared.

We also assume that there are two priority queues at each switching node, namely the VBR queue *(VBR Q)* and the ABR queue *(ABR Q).* Depending on their type of service, arriving cells are stored in corresponding queues. Cells in the VBR queue have priority over cells in the ABR queue, i.e., only when the VBR queue is empty ABR traffic is sent.

For all proposed scenarios, the processing delay at the ATM layer is not explicitly modeled. We assume that its contribution to the end-to-end delay experienced by each cell is relatively constant, and thus it can be omitted..

4.3 ABR rate-based Congestion Control

To evaluate the efficiency of the rate-based approach in carrying a layered video connection, we compare the binary feedback mechanism (EFCI) and the rate-based explicit rate (ER) mode. In this paper, the ER scheme performed by the switches is EPRCA [12].

The EPRCA scheme computes a mean allowed cell rate *(MACR)* using exponential weighted averaging and a Fair share.

$$MACR = (1 - EAF) * MACR + EAF * CCR$$
$$Fairshare = SW\_DPF * MACR$$

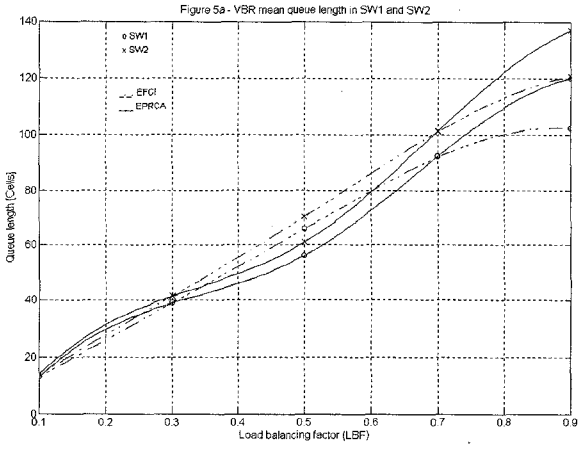
Here, EAF is the exponential averaging factor, CCR is the current cell rate of the connection indicated in the RM cells and SW_DPF is a multiplier called switch down pressure factor set close to but below 1. Various simulation studies [13] suggest that the appropriate values for EAF and SW_DPF are respectively 1/6 and 7/8. The initial MACR and the queue threshold *(QT)* are respectively set to PCR and 15.

We define five network scenarios by varying the Load Balancing Factor. In the first scenario, the source performs a static data partitioning using a *LBF* of 0.9, which corresponds to 90 % of the original video data transmitted through the VBR connection. The remaining 10% is transmitted over a distinct ABR virtual channel. The scenario 2 is similar to the first but with a *LBF* equal to 0.7. Finally, scenario 3, 4 and 5 have respectively *LBF* equal to 0.5, 0.3 and 0.1.

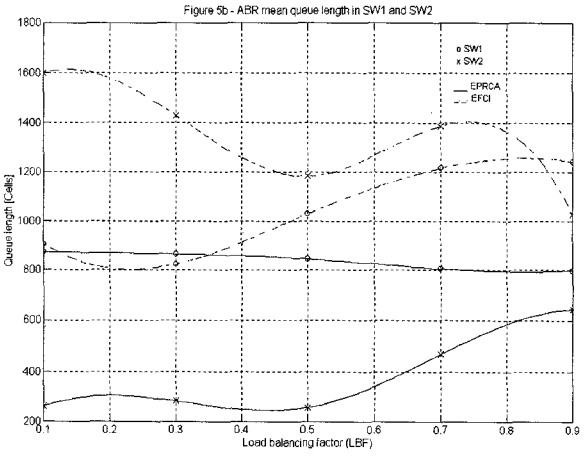
5. Simulation Results

5.1 Impact on Queue Length

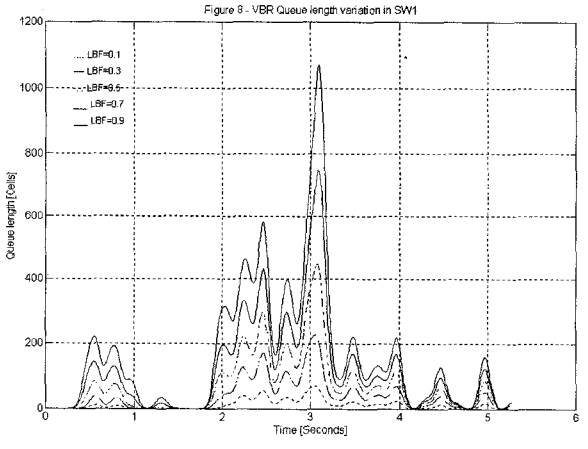
Figure 5a, 5b and 8 present the variation of the buffer queue lengths as function of the Load Balancing Factor. The first observation concerns the proportional increase of the average number of cells in the queue. Indeed, by comparing the five curves of Figure 8., it can be seen that the plot shapes are similar. Alternatively figures 5a and 5b show the advantage of EPRCA over the EFCI scheme. The mean occupation ranges from 14 cells for *LBF*equal to 0.1 to a maximum of 138 cells for *LBF* equal to 0.9. It is also interesting to note that this variation does not depend on the ABR control mode *(EFCI, EPRCA).* One should expect a correlation between the ABR and VBR queues, but this is not the case here. Indeed, due to the separation in two distinct queues and the used service discipline which favor high priority VBR traffic over ABR one, we can consider that the average transit delays are closely depending on the number of VBR background load. This mean transit delay is bounded to 64.4 ms in the worst case scenario, i.e., of 50 simultaneous background video connections.



The CTD experienced by I-frames are the most important in comparison to the other frames. In the worst case corresponding to the enhancement layer with EPRCA, it reaches 782 ms, while for P and B sub-streams is respectively to 668 and 778 ms. In unidirectional video services, the tolerated CTD is 1 second [2].



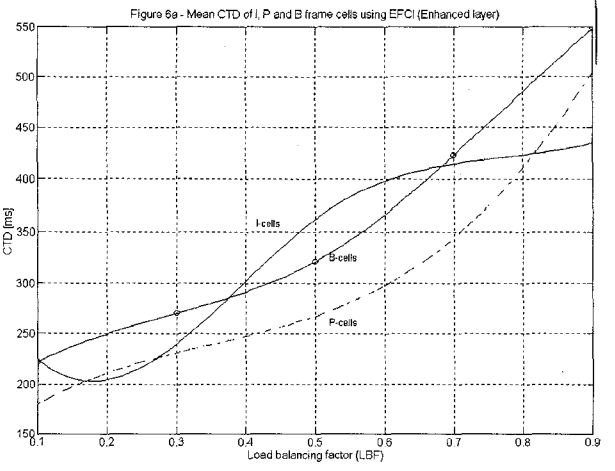
We note that the average occupation for the ABR queues is increasing linearly for the two ABR control modes with values ranging from 259 to 1601 cells. In all cases the ER mode minimizes buffer oscillations.



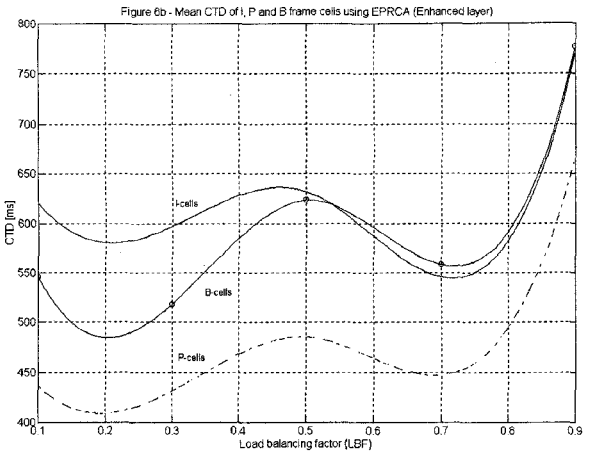
The last set of curves (Figures 10.a, b, c, d) show the effect of varying the *LBF* ratio on the instantaneous ABR switch buffer occupancy for EFCI and EPRCA.

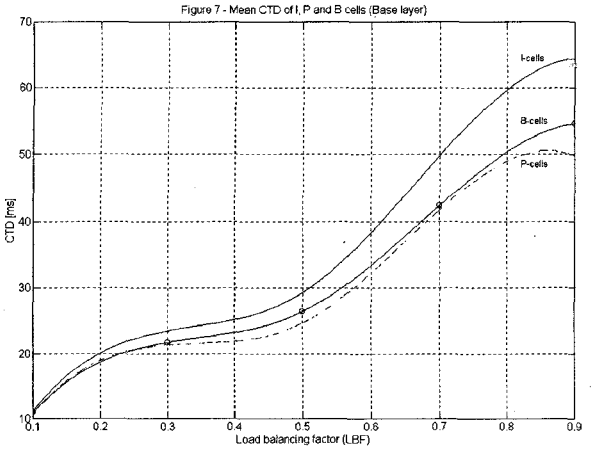
5.2 *Impact on Cell Transfer Delay*

In Figures 6.a, 6.b and 7 the average end-to-end cell transfer delay for each cell class (I, P and B) is illustrated for various *LBF* values. It is interesting to notice that for both ABR and VBR connections, the experienced mean CTD of I-frame cells are greater to those of other frame type cells. This can be explained by the impact of intra coding on the output frame size and subsequently on the source cell rate and switch buffer occupancy.



The general trend of these curves is, the more the load balancing factor increases, the more the mean delay increases. The optimum *LBF* value which minimize the mean CTD of the reference connection is 0.2 regardless to the rate-base mechanism. However, for the reference VBR connection the CTD of any sub-flow is bounded by a mean value of 64 ms, whereas for ABR stream the average end-to-end transit delay seems to be out to acceptable bounds. Indeed, for most load balancing configuration the experienced ABR cell delays are above the recommended upper bound of 150 ms for video applications [14]. It appears that for a heavy loaded wide area networks (i.e. 85-90 % bandwidth utilization and 1000 km link length), the ABR service with the presented control schemes is not suitable for transmitting interactive two-ways video applications with stringent temporal requirements such as videoconferencing.



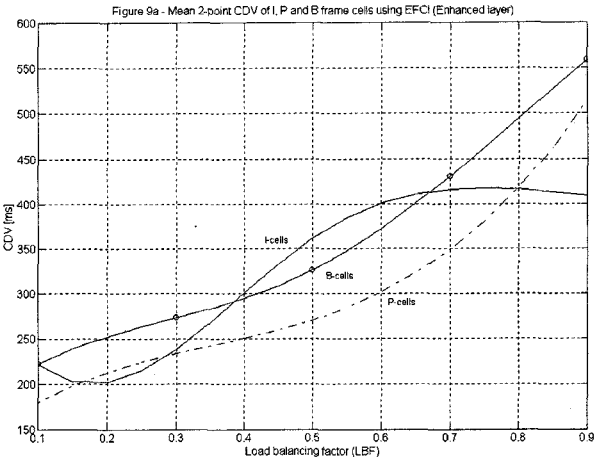


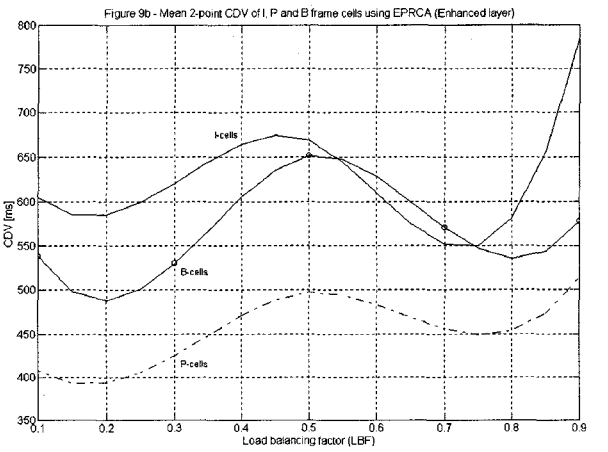
Despite of these excessive measured delays lighter network configuration, such as LANs associated with more efficient buffer cell scheduling algorithms, will provide better network performance regarding to unidirectional video retrieval services.

5.3 Impact on Cell Delay Variation

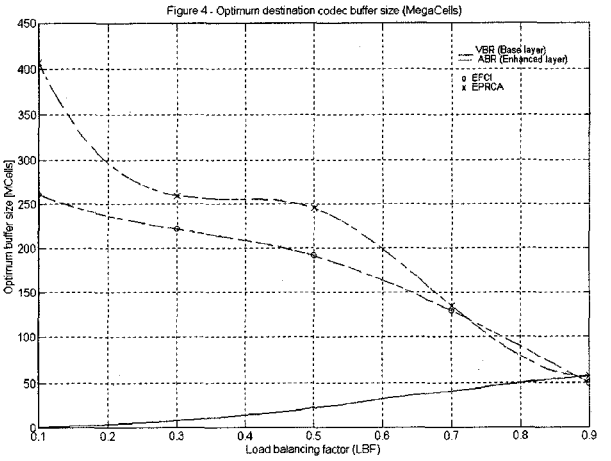
From the mean 2-point CDV curves (Figure 9a and 9b) and the Figures 6.a and 6.b (mean CTD) the following observations can be done.

Firstly, the distribution of the mean 2-point CDV is identical to the distribution of the mean end-to-end delay, with a difference of a D_0 time-shift. By definition, the 2-point CDV strongly depends on the random variable D_0. This means that from one simulation run to the other the collected 2-point CVD statistics may be very different.





Secondly, the 2-point CDV defined in section 3 can be related to the calculation of the optimum sizes of the destination codec buffer. Indeed in [15], it has been shown that the optimum buffer size must be at least equal to twice the value $(D_{\max} - D_{\min})/T_{\min}$. The numerator is the maximum of the 2-point CDV calculated at the minimum possible value of $D_0 = D_{\min}$.



In our case study, this minimum (D_{min}) is equal to a one-way trip to the destination with empty buffer along the path for the reference VBR connection and the smallest occurring cell delay for the reference ABR connection. The denominator corresponds to the inverse of the Peak Cell Rate *(PCR)* of the targeted connection. In VOD, video streams are previously encoded and stored, which allow us to set the value T_{min} to the largest encoded frame size P_{max} .

As illustrated in Figure 4., the best results are obtained when *LBF* is superior to 0.6.

6. Conclusion

Available bit Rate (ABR) service has been initially developed to carry non real-time bursty data. Because MPEG2 video sources are rate adaptive and scalable, the ABR service has the promise of supporting compressed video as well. Indeed, the ABR protocol allows the network to control the source rate directly, ABR schemes can then be adapted to control compression ratio at the codec sources, in addition to controlling cell-loss rates. Given this, and if the Sustained Cell Rate (SCR) selected for the VBR connection is chosen to correspond to the minimum rate required by the video encoder, the Available Cell Rate (ACR) determined by the ABR scheme may be used to determine the Current Cell Rate (CCR) of the enhancement ABR connection at the video encoder.

Therefore, in this paper we have studied through intensive simulation the suitability of a hybrid VBR/ABR connection for the transport of a 2-layer MPEG2 encoded streams. We evaluated the effectiveness of the ABR binary feedback and Explicit Rate control modes in situation of carrying the enhancement MPEG2 sub-stream with various static load balancing configurations. To carry out the comparison, we used actual MPEG2 video traces and frequency domain data partitioning.

For most practical cases, the impact of changing some network parameters, such as the load balancing factor and the ABR switch control algorithm, on the network performance have in the given order the most significant effect on the buffer occupancy, the mean Cell Transfer Delay, and the Jitter. From the evaluation results, we noticed that the I-frame cells experienced the highest end-to-end delay but they never exceeded the 1 seconde required for one-way video (Video on Demand or TV broadcast).

Finally, if such compressed video retrieval applications with flexible temporal constraints will be widely provided over ATM networks, the use of scalable MPEG2 coding associated with best effort services may maintain high throughput during overload in ATM switches with expected picture quality enhancements.

7. References

[1] ATM Forum, Traffic Management Sub-Working Group, "Traffic Management Specification 4.0", at-tm-0056.000, April 1996.

[2] ATM Forum, "Audio visual Multimedia Services : Video on Demand v1.0", Service Aspects and Applications Sub-Working Group , af-saa-0049, December 1995.

[3] D. Reininger, B. Melamed and D. Raychaudhury, "Variable bitrate MPEG video : characteristics, modeling and multiplexing", International Teletraffic Congress, June 1994, pp 314-319.

[4] ATM Forum, Service Aspects and Applications Sub-Working Group, "VBR MPEG2 specification", draft, January 1997.

[5] M. Grossglauser, S. Kershav and D. Tse, 'RCBR : A Simple and Efficient Service for Multiple Time-scale Trafic', ACM SIGCOMM '95, Cambridge, USA, Sept. 1995.

[6] T.V. Lakshman, P.P. Mishra, K.K. Ramakrishnan, 'Transporting Compressed Video over ATM Networks with Explicit Rate Feedback Control', IEEE INFOCOM '97, Kobe, Japan, March 1997.

[7] ISO/IEC 13818-2 MPEG-2, "Information Technology - Generic Coding of Moving Pictures and Associated audio", 1995.

[8] P. Pancha and M. El Zarki, 'Prioritized Transmission of Variable Bit Rate MPEG Video', IEEE GLOBECOM'92, Orlando, FL, December 1992, pp. 1135-1139.

[9] M. W. Garrett, "Contribution toward real-time services on packet switched networks". PhD thesis, Columbia University, 1993.

[10] ITU-T recommendation I.356, "B-ISDN ATM layer cell transfer performance", Geneva, Nov. 1993.

[11] O. Rose and M. Frater, 'Impact of MPEG Video Trafic on an ATM multiplexer', IFIP High Performance Networking '95 (HPN'95), Palma, Spain, Setp. 1995, p.157-168.

[12] L. Roberts, "Enhanced Proportional Rate Control Algorithm (EPRCA)", ATM Forum, 94-0735R1, August 1994.

[13] Y. Chang, N. Golmie, and al., "Simulation study on the new rate-based PRCA traffic managment mechanism", ATM Forum, af-tm-94-0809, Sept. 1994.

[14] Raif O. Onvural, "Asynchronous transfer Mode Networks : Performance Issues", Artech House, 1994.

[15] H. Naser and A. Leon-Garcia, "A simulation study of Delay and Delay Variation in ATM Networks, Part 1 : CBR Traffic", IEEE INFOCOM'96, San Francisco, March 1996.

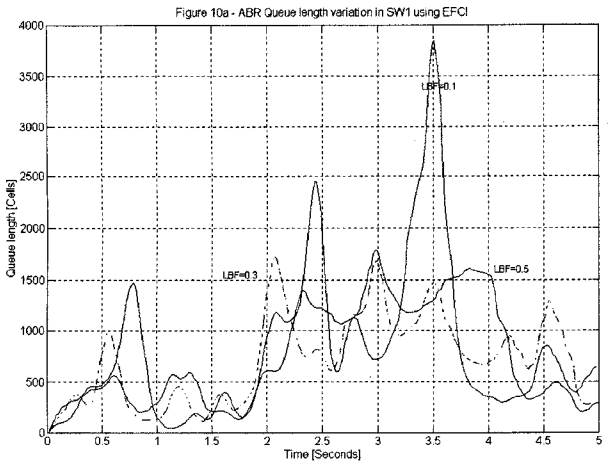


Figure 10a - ABR Queue length variation in SW1 using EFCI
Queue length [Cells]
Time [Seconds]
LBF=0.1
LBF=0.3
LBF=0.5

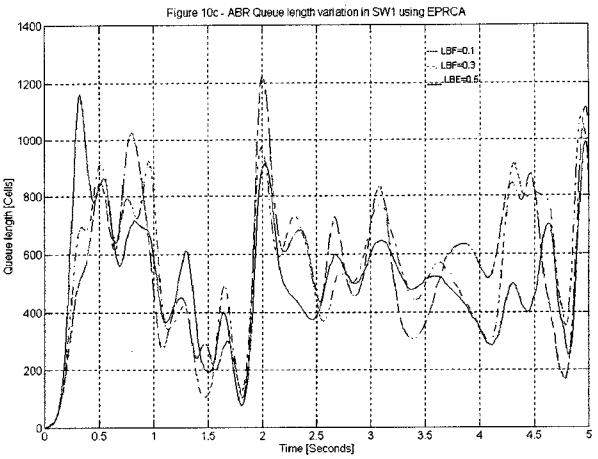


Figure 10c - ABR Queue length variation in SW1 using EPRCA
Queue length [Cells]
Time [Seconds]
LBF=0.1
LBF=0.3
LBF=0.5

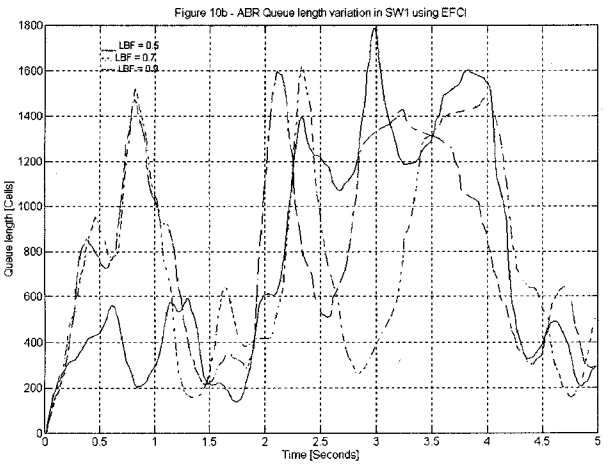


Figure 10b - ABR Queue length variation in SW1 using EFCI
Queue length [Cells]
Time [Seconds]
LBF = 0.5
LBF = 0.7
LBF = 0.9

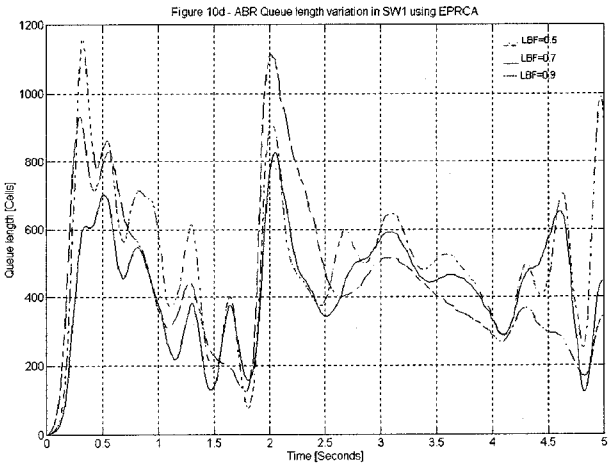


Figure 10d - ABR Queue length variation in SW1 using EPRCA
Queue length [Cells]
Time [Seconds]
LBF=0.5
LBF=0.7
LBF=0.9