A Video Traffic Feedback Control Mechanism for ATM Networks

M. DOBRESCU, R. DOBRESCU, S. MOCANU
POLITEHNICA University of Bucharest
Splaiul Independentei 313
ROMANIA
radud@isis.pub.ro, http://isis.pub.ro

Abstract: The paper proposes a video traffic control mechanism that aims to solve the problem of the amount of feedback reduction, due to the increase of the number of destinations from the same source description, and the problem of the available bandwidth variation. For the successful use of this mechanism, two algorithms are developed: one to locate the most congested branch in the multicast tree, the other to observe the overall congestion state of this tree. In order to meet the target high priority cell rate a special encoding technique, based on an adjustment of the encoder’s quantization parameter is described. The performances of the proposed feedback mechanism, associated with the encoding algorithm were investigated on a series of simulations on a point-to-multipoint network model in comparison with other two known methods for rate control: Constant Bit Rate and Variable Type Rate. The simulation results showed the scalability with the number of destinations of the cell rate and of the video signal quality proving that video feedback mechanism offers better quality than a classical CBR service at the same amount of available bandwidth and can adapt to varying degrees of congestion in the network in order to improve video quality.

Keywords: ATM network, video services, available bit rate, adaptive encoding, congestion prevention

1 Introduction

In a common acceptation, video services refer to the transmission of moving images together with sound. Beside television, new challenges for video appear in the context of multimedia applications over packet-switched networks, such as those based on the Internet Protocol (IP) or on the Asynchronous Transfer Mode (ATM) protocol. It is commonly assumed that the main load in these multimedia networks is the video traffic, especially when multipoint applications appear as necessary. Such applications cover a very wide spectrum, including software distribution, replicated database update, command and control systems, audio/video conferencing, distributed interactive simulation. The proliferation of multimedia applications, especially real-time video services, associated with new high-speed networks, often based on ATM technology, is driving this need for reliable group communications mechanisms and protocols. Most of the early approaches to support the transmission of real-time video in ATM relied on preventive congestion control algorithms (like call admission - [1], [2]) and associated services, like Constant Bit Rate (CBR) service [3], that consists in maintaining the video output process constant by dynamically adjusting the quantization, but causes the fluctuation of the video quality and also need a large amount of bandwidth that implies the sacrifice of the multiplexing gain. Even a more developed service like Variable Bit Rate (VBR) – [4], [5], that require an accurate prediction of traffic predictors during call admission, requires statistical multiplexing gain and bandwidth renegotiation mechanisms. It was suggested that a better preventive congestion control technique must be a combination of preventive and reactive (feed-back) congestion control schemes, like in the Available Bit Rate (ABR) service recommended in [6]. Many researchers have studied feedback-based reactive congestion control algorithms for video ([7], [8], [9]), but only for point-to-point video transport. In this paper is propose a reactive congestion control algorithm that utilizes and adaptive encoding technique and an efficient multicast feedback mechanism to provide a scalable real-time video multicast service for ATM networks.

2 The Simulation Model of the Transmission System

Because the key of the optimization of the image quality is the coding technique, the performances of the proposed algorithm were tested in comparison
with those of CBR and VBR video coding. The model proposed for the test is similar with that proposed by Chen and Lin [10] in order to simulate both dependent or independent video coding techniques. Independent coding means that the distortion rate relation of a current video unit is independent of how an earlier video unit was coded (this is the case of motion JPEG schemes), while dependent coding (present in MPEG schemes) means that the distortion-rate relations of successive units are dependent and consequently inseparable.

The model of the transmission system is shown in fig. 1. Let assume a buffering and transmission delay of \( d \) video units (v.u.) between the encoder output and the decoder input of a point-to-point connection. The encoder performs a delayed coding with delay \( N \) v.u. At time \( n \) the encoder outputs \( b(n) \) bits to the buffer, which outputs \( t(n) \) bits to the ATM network. The finite buffer sizes of the encoder and of the decoder and the finite capacity of the output link \( W \) (bps) induce transmissions rate constraints on the coded video; then a specific transmission rate \( t(n) \) must be determined by the policing (congestion prevention) algorithm.

In order to minimize the delay due to packet formation, the streaming mode used when forming the AAL-PDU’s is considered such that as soon as a cell payload is delivered to the AAL, the payload is placed into a cell and transmitted.

Three video generation rules were used in simulations.

- The first, defined as Constant Rate Generation (CRG) use the CBR control and generate a cell at regular interval of \( T_{fix} \) (seconds), i.e. the first cell in a stream is formed \( T_{fix} \) seconds after the encoder starts putting the bits into the output buffer, such that \( V \cdot T_{fix} = 48 \) bytes, with \( V \) (bps) the bit rate of the CBR video stream.

- The second, defined as Variable Rate Generation (VRG) use the VBR control and generate a cell as soon as the encoder generates 48 bytes of data.

- The third, defined as Adaptive Rate Generation (ARG) use a new policing mechanism with feedback derived from the ABR control that allows adjusting the cell rate according to the traffic status in order to avoid congestion. This mechanism will be detailed in the next section.

In the same time, three different coding methods were associated with the cell generation procedures, as follows in order: CBR encoding, constant video quality CVQ-VBR encoding and adaptive video quality AVQ encoding. In CBR encoding, a hypothetical rate control buffer of size \( B \) is considered at the output of the encoder, which is emptied at the target rate \( V \). The short term variations produced by the encoder per time interval are compensated by the rate control buffer. In the CVQ-VBR encoding, the video quality level is used as feedback to control the quantizer scale in order to maintain the desired level of quality. In the AVQ encoding a special algorithm that will be presented in section IV ensures the adjustment of the quantization parameter according to the total output rate from the video encoder, maintaining the quality in admissible limits.
For traffic scenarios involving one type of video the same video source was used to generate all streams for a given run of simulation. As quality parameter the loss of information (in macroblocks) of the aggregate traffic generated by independent and identically distributed (i.i.d) $N$, sources was used.

3 The Feedback Rate Control Mechanism

In ATM networks, the closed-loop feedback is utilized to control the overall cell rate of the video source. The source periodically transmits a resource management cell (i.e. a forward feedback cell, $FFC$), that is returned from the destination as a backward feedback cell, $BFC$. Usually, a single bit of this cell indicates the congestion. This binary congestion indication is considered in this approach.

The feedback mechanism provide guaranteed service to all video traffic under the Minimum Cell Rate ($MCR$) and also try to rise up the speed to the Peak Cell Rate ($PCR$). The goal of the proposed mechanism, denoted Adaptive Rate Generation ($ARG$) is to adjust the high $PCR$ according to the state of the most congested branch of the multipoint connection, making use of the unutilized bandwidth. The principle of the control mechanism is the following: for each $N = 16$ video cells transmitted, the video source transmits a single $FFC$, by which the source probes two destinations in order to obtain two information: the location of the most congested link in the multipoint connection (useful to adjust the high $PCR$) and the congestion state of the entire connection, useful to adjust the overall cell rate ($OCR$). Each destination is assigned a unique identification that the video source uses while polling. When the source generates the $FFC$, it indicates the targets for both probes. Quite every destination participating in the point-to-multipoint video calls receives the probe, because the $FFC$’s are multicast, only two destinations being polled respond with a $BFC$. Since each $FFC$ contains two probe targets, in the return results two $BFC$, depending of the type of probe the destination is receiving. The format of the feedback cell is the same for both forward and backward feedback and corresponds to the indications of the Enhanced Proportional Rate Control Algorithm devised by the ATM Forum for $ABR$ service.

The first four bytes of a Feedback Cell contain two 16-bit identifiers of the destination that should respond with a $BFC$ for a probe of type 1 ($P1$), respectively for a probe of type 2 ($P2$). The next 5 bit fields in the five byte are traffic indicators with the following significance:

- $CI$ – Congestion Indication – indicates whether the source-to-destination path has experienced congestion
- $LPCI$ – Low-Priority Congestion Indication - indicates whether the source-to-destination path is in danger of losing a high priority cell
- $PRI$ – Probe Response 1– one-bit field set by the destination responding to a probe of type 1
- $PR2$ – Probe Response 2– one-bit field set by the destination responding to a probe of type 2
- $FD$ – Feedback Direction - one-bit field set in 0 in $FFC$ and set in 1 in $BFC$
- The feedback polling algorithm requires two algorithms, one to locate the most congested branch (LB) in the multicast tree, the other to observe the overall congestion state (OC) of this tree

3.1 The description of the LB algorithm

The algorithm has two states: the TEST state and the ADJUST state. In the TEST state, the source probes each destination, one at a time. Whenever a $FBC$ arrives at a destination, the $P1$ field is examined. If the identifier contained in $P1$ matches with the destination’s identification, then the destination examines its low priority traffic congestion status, generates a $BFC$ and sets the $LPCI$ bit accordingly. When the source receives a $BFC$ with $LPCI=1$ it enters in the ADJUST state. In this state, the source probes the destination that generated the $LCI$ and decrements its high $PCR$ by an amount proportional to its current value. Only when the probed destination returns a $BFC$ with $LPCI=0$, the source returns to the TEST state. It is obvious that the LB algorithm reduces rapidly the high $PCR$ whenever this cell is in danger of being lost, and increases the high $PCR$ only after every destination has indicated a lack of low priority traffic congestion. So, LB algorithm produces a high $PCR$ that approximate the available bandwidth on the multipoint connection’s most congested link.

3.2 The description of the OC algorithm

Usually an ATM switch considers a buffer congested from the moment when an output port’s buffer occupancy exceeds an upper threshold and
until its occupancy drops below a lower threshold. When a buffer is congested, the switch marks the Explicit Forward Congestion Indicator (EFCI) bit in the header of the cell being served to indicate congestion to destination and stores the congestion state of the forward path. When the destination is probed, it generates a BFC and marks its CI bit if the last cell arrived to the destination had a marked EFCI. Upon receiving the BFC the source stores the cell’s congestion indication CI bit in a congestion stack. This stack maintains the last N congestion indications received by the source. Then the source calculates a Target Overall Cell Rate (TOCR) as:

\[ TOCR = MCR + (N_0/N) \times (PCR - MCR) \]  

(1)

where \(N_0\) is the number of zeros in the congestion stack. If the current overall cell rate is less than the target overall cell rate, it is increased additively. In the alternate situation, when the current overall cell rate is greater than the target overall cell rate, it is decreased proportionally. In this way TOCR is determined by the degree of congestion on all the source-to destination paths. The minimum value, when all these paths are congested, i.e. \(N_0 = 0\), \(TOCR=MCR\). At the other extreme, when all the paths are uncongested, i.e. \(N_0 = N\), \(TOCR=PCR\). For congestion levels between these two extremes, TOCR is set at an intermediate value corresponding linearly to the degree of congestion.

4 The Adaptive Encoding Mechanism

In order to meet the target high priority a special encoding technique denoted Adaptive Video Quality AVQ was adopted. The control of the overall output rate of the video encoder requires an adjustment of the encoder’s quantization parameter \(Q\). The video buffer is filled by the encoder and served at an output rate equal to the overall cell rate. When the video buffer’s occupancy changes, \(Q\) is adjusted to prevent buffer underflow (\(Q\) is increased) or underflow (\(Q\) is decreased). It is obvious that a decrease of \(Q\) leads to the increase of the combined output rate and image quality.

The adjustment of the quantization parameter achieves control of the total output rate from the video encoder. In addition, data partitioning is utilized to control the output rate of low and high priority information. Data partitioning consists in the division of the encoded video bit stream into two layers with different priorities. Because the video bit stream results according a MPEG procedure after applying a run-length coding one the two-dimensional DCT coefficients appears like a sequence of run-level pairs. To split this stream into two, a priority breakpoint is introduced. This priority breakpoint is an integer specifying the number of run-level pairs per block to place into the high priority stream. All remaining run-level pairs are placed into the low priority stream. This encoding algorithm ensures for every macroblock that is encoded and packed into ATM cells, the priority breakpoint is adjusted to produce the desired high priority cell rate.

On the other side, the encoding mechanism must operate under the constraints on the transmission rate derived from the limits of the decoder buffer. Referring to the scheme in fig.1, let \(R(n)\) denote the decoder buffer size and let \(r(n)\) be the decoder buffer level after extraction of \(b(n-d)\) data for the \((n-d-N)\)-th v.u. for decoding. Cumulated over time, it yields:

\[ r(n+\lambda) = r(n-1) + \sum_{i=n-\lambda}^{i=n+\lambda} t(i) - \sum_{i=n-d}^{i=n-\lambda} b(i), \lambda \geq 0 \]  

(2)

In such conditions, the decoder underflow \(U_d(n,\lambda)\) and the decoder overflow \(O_d(n,\lambda)\) must satisfy the following series of inequalities:

\[ U_d(n,\lambda) = \sum_{i=n-d}^{i=n+d+\lambda} b(i) - r(n-1) \leq \sum_{i=n}^{i=n+d+\lambda} t(i) \]  

(3)

\[ \leq \sum_{i=n-d}^{i=n+d+\lambda} b(i) - r(n-1) + sR = O_d(n,\lambda) \]

where \(s\) is a scaling factor. Let now check the constraints from the network congestion prevention mechanism. If \(N\) is the dimension of the congestion stack, then

\[ N(n) = N(n-1) + t(n) - R \]  

(4)

and in a similar matter cumulated over time the resulted underflow \(U_n(n,\lambda)\) and overflow \(O_n(n,\lambda)\) must satisfy the following series of inequalities:

\[ U_n(n,\lambda) \leq (\lambda+1)R - N(n-1) \leq \sum_{i=n}^{i=n+d+\lambda} t(i) \leq (\lambda+1)R - N(n-1) + sN = O_n(n,\lambda) \]  

(5)

It results that after encoding of \([n-N, n]\) v.u., \(t(i)\) for \(i = n, \ldots, n+N\) is determined according to:
\[
\max \{0, (U_D, U_N)(\eta, \lambda)\} \leq \sum_{i=1}^{\eta+\lambda} f(i) \leq \min \{(O_D, O_N)(\eta, \lambda)\},
\]

with \(\kappa=0,1, \ldots, N\).

The AVQ algorithm can be now summarized as:

- **Step 1.** Grow the coding tree recursively, one level at time, subject to the rate constraints derived above.
- **Step 2.** On the path determined from the LB algorithm, quantize and encode \(v_u n\) using the scaling factor \(s\) and determine the transmission rate \(t(n)\) subject to (6).
- **Step 3.** Update the buffer levels, increment \(n\) by 1 and return to step 1.

### 5 Experimental Results

Two kinds of experiments were developed on the simulation model. First, a comparison of the performances for the three packet generation rules was realised on a point-to-point link. Secondly, an analysis of the new encoding and rate control algorithms was made one the point-to-multipoint ATM network model. The tests involved two types of video sequences of about one minute length: a videoconferencing-type (VC) sequence that shows a person sitting and talking and a piece of a commercial advertisement (CA) that contains fast movements and animated scenes. In order to realise a realistic comparison between CBR transmission and the other two variable-rate methods, the video quality was measured by the scale factor \(s\) which was maintained at about the same value above a minimum value \(s_{min}\). The quality factor was measured over one second interval, considered appropriate to the rate of changes in the scene and to the response time of the human visual system.

Table 1 summarizes the transmission performances evaluated by the loss of information, expressed by the average number of undisplayed macroblocks per frame \(U_{av}\) and the total number of undisplayed macroblocks \(U_t\), for various traffic scenarios (3 encoding methods \(EM\) and 3 rate control methods \(RCM\), at different values of the cell loss rate \(CLR\) and for a target \(s_t=4.5\). The table contains also the average transmission rate \(TR\) (Mb/s) obtained in the experiments.

Table 2 presents a comparison of the average bit rate obtained for a minimum quality level \(s_{min}\). The table indicates that for the advertisement sequence, the average rate for VBR (VRG and ARG) is about one-half of the CBR/CGR rate, but the difference diminish at higher \(s\). For the VC sequence, the data rates are nearly the same for all the three methods at a given target quality values. The new proposed algorithm offers a slide improvement for the VC sequence, due to the narrow amount of necessary bandwidth.

The last type of test on the image quality reflects the performances of the different encoding algorithms when transmitting an image video sequence at 10 pictures/second at different average rates.

![Fig. 2. PSNR for different encoding methods](image)
Fig. 2 shows the average peak signal-to-noise ratio PSNR obtained at an identical buffer size of the encoder and of the decoder of value r·6400. For that reason, in the diagram the channel rate $R_c$ appears as normalized, $R_c=1$ representing 640 Kbits/second. The average quantizer step was modified between 2 and 30, in step of two.

The following traffic performances were tested by simulation:

- **The scalability with the number of destinations of the cell rate**

To study this performance, the number of destinations in the multicast connection is varied, while holding all other parameters fixed. The collected statistics are: the average overall cell rate of the source and the average high priority cell rate of the source. The tests show that the feedback mechanism’s average high priority cell rate decreases rapidly and is dominated by MCR when the number of destinations is greater than 50, quite it ensures an average overall cell rate relatively constant as the number of destinations is increased. The transmission rate is about 5 times greater then a CBR connection with the same amount of bandwidth.

- **The scalability with the number of destinations of the quality of video**

The quality of video received by destinations is compared using as indication the average $P_{SNR}$, by measuring the quality of video received by two types of destinations participating to the feedback mechanism: congested and uncongested. The tests shown that uncongested destinations receive video of high quality, because the feedback mechanism allows the video source to increase its overall cell rate and consequently use unutilized bandwidth.

V . CONCLUSIONS

The paper proposes a new feedback mechanism for controlling the rate of a video stream with low and high priority cells on a multicast ATM point-to-multipoint connection. The performances of the proposed feedback mechanism, associated with the encoding algorithm, were investigated on a series of simulations. In these simulations, two sets of constraints were introduced: constraints due to the end-to-end delay needed to maintain real-time video and constraints due to the proposed algorithms for congestion prevention and for encoding. Results from these simulations showed that the video feedback mechanism is capable of providing better quality than CBR service at the same amount of available bandwidth and is comparable with a VBR type algorithm. The degree of reduction of the video quality when the propagation delay increases is not significant even for a wide area environment. The feedback mechanism can adapt to varying degrees of congestion in the network in order to improve video quality when possible.

Further work intends to suppress the negative effect of data partitioning by using SNR scalability and to test the feedback mechanism on a multipoint-to-multipoint connection.

REFERENCES


