Interactive video on demand over high speed networks

Bing Zheng [a] and Mohammed Atiquzzaman [b,*]

[a] *Department of Electrical and Computer Engineering, The University of Dayton, Dayton, OH 45469-0226, USA*
E-mail: zhengbin@flyernet.udayton.edu
[b] *School of Computer Science, University of Oklahoma, Norman, OK 73019-6151, USA*
Fax: +1 520 962 8422; E-mail: atiq@ou.edu

Abstract. Interactive Video on Demand with Instantaneous Access (IVoD-i) can offer users with VCR like functions by sending requests to video servers. Earlier work has shown that the Available Bit Rate Service of ATM networks can be effectively used for interactive video on demand. However, non optimal values of the connection setup parameters can affect the Quality of Service (QoS) at the user. This is because the ABR service only provides minimum bandwidth guarantee for a connection. The *objective* of this paper is *to provide QoS to IVoD-i over the ABR service of ATM*. To achieve this objective, we have developed an analytical framework to determine the ABR connection parameters to guarantee QoS to IVoD-i users. Our proposed framework has been tested with real life MPEG video traces. Results show that our framework can provide QoS to IVoD-i users. Moreover, our framework outperforms previous schemes in terms of startup delay, user buffer requirement, and jitter.

1. Introduction

Large bandwidth and low transmission error of ATM networks allow transmission of multimedia in real time [1]. It is widely believed that Video on Demand (VoD) will be one of the most widely used application in next generation networks. VoD offers users the capability to access video library on demand. From the access point of view, VoD can be classified into Interactive VoD (IVoD) where users can perform VCR like interactive functions (such as play, pause, fast forward, and fast backward), and Passive VoD (PVoD) where users can only receive and playback video sent from sources passively. From the point of view of service time, VoD can be classified into instantaneous service (VoD-i) where the server responds instantly (in the time of network latency), and delayed service (VoD-d) in which all user requests are serviced with a time delay (in the order of minutes) set by the system configuration. Therefore, IVoD consists of IVoD-i and IVoD-d, of which IVoD-i is the most appropriate for applications such as remote education, entertainment, etc. [2], where VCR like interactive function is desired.

Video is characterized by large bandwidth needed for transmission. Therefore, to reduce bandwidth requirement, video is compressed before it is transmitted over networks. Compressed video stream is bursty in nature; its data rate varies from frame to frame, which results in the required bandwidth varying from time to time. Therefore, to efficiently utilize network bandwidth, dynamic bandwidth negotiation/allocation mechanism is needed. Dynamic bandwidth negotiation has two purposes. First, it matches the allocated bandwidth with the video rate during normal transmission. Second, it allocates high bandwidth to video sources to compensate for high data consumption rate of interactive users.

The ABR service of ATM offers dynamic bandwidth allocation during the entire connection by using Resource Management (RM) cells. RM cells are sent by sources at regular intervals. Backward RM cells carry feedback information about network congestion and allowable data rate, which are used to adjust the sending rate of sources. The sources have dedicated rates during each interval. Although ABR has the advantage of dynamic bandwidth

[*] Corresponding author.

0926-6801/00/$8.00 © 2000 – IOS Press. All rights reserved

negotiation during the entire connection, the allocation of the Allowable Cell Rate (ACR) depends on the level of network congestion. Different from CBR and VBR, which have high statistical guarantee of bandwidth for the entire connection period, ABR only offers minimum guarantee of bandwidth with Minimum Cell Rate (MCR). Because video transmission normally lasts for a long period (i.e., typically a movie has 120 minutes), during connection setup, it is impossible to predict how network congestion will change with time. Therefore, to use ABR for interactive video transmission with acceptable QoS, the following three problems must be solved. The first problem is setting up a feasible ABR connection to provide acceptable QoS to an interactive user. The second is the choice of values of the ABR parameters and the requirement for interactive users. The third problem is to determine the QoS that can be obtained by a user.

1.1. Related work

The first model concerning connection setup parameters for VoD over ATM CBR service was proposed in [3]. However, their model only supports passive VoD (PVoD) with two shortcomings: long start delay (37 second) and large user buffer (23 MBytes). Authors in [4] proposed an algorithm to exploit the temporal structure of MPEG video to reduce required bandwidth for CBR transmission. However, the open loop characteristic of CBR lacks the capability of responding to user's interactive operation by dynamic bandwidth allocation within channel. To solve the bandwidth renegotiation problem, renegotiated CBR (RCBR) which is an open loop renegotiation mechanism [5–7] was proposed. Unlike ABR where the in band RM cells can return network congestion status and available bandwidth to sources, there is no mechanism in RCBR to inform the sources of the available bandwidth since there is no closed loop feedback. Therefore, the most difficulty of RCBR is to predict the next available bandwidth [8,9].

Although CBR service is simple for connection setup, its disadvantage is low bandwidth utilization. To improve the bandwidth utilization, video transmission over ATM VBR service is widely studied. The issues include traffic shaping and rate control [10,11], bandwidth allocation/management [12,13], congestion control [14], source model and behavior [15–18]. In [19], improving quality of transmitted video by using Usage Parameter Control (UPC) was studied. However, VBR does not provide feedback mechanism for bandwidth renegotiations after connection setup. In [20–22], authors proposed renegotiations VBR (RVBR). However, since renegotiations are source initiated, there is no way for the network to acknowledge sources of congestion status or of newly available bandwidth. Therefore, the source can not utilize the newly available bandwidth efficiently [8].

The closed loop feedback mechanism of ABR makes it suitable for interactive video transmission. In [23], authors studied cell loss performance for rate based flow control for ABR video transmission. Authors in [24] discussed a scheme with closed-loop feedback for congestion control in video transmission using ABR. A non-zero MCR was set to obtain the guaranteed service. Although authors in [25] studied the interactive performance between the user and source, they mainly investigated the user request probability. The quality of service at the user was not studied. In [26], authors studied delay performance for video transmission over the ATM ABR service. In [8], authors discussed transporting compressed video over the ABR service. The feedback from the network is used to control the encoder rate to adapt to available network bandwidth. Two criteria are used for adaptation of the encoder rate. One is demand prediction by using video source model to predict future possible video stream rate; the other is to adjust the encoder rate to match the allocated rate from backward RM cells. Real time video conference was discussed. Interactive performance described in the paper was between the network and the sender. Users did not have VCR like function. Although the authors mentioned the necessity to derive requirements for ABR connection parameters, they did not provide any explicit answer.

Authors in [9] proposed dynamic bandwidth allocation for compressed MPEG video transmission over ATM. They developed a model to determine the required bandwidth and user buffer size; trace driven simulation was carried out. However, their model did not take user interactive performance and network latency into consideration. Moreover, their studies did not concern transmission delay and jitter which are important QoS criterion for video transmission.

1.2. Objective of this paper

As described above, although considerable work has been done in video transmission over ATM, previous work did not answer the question of setting up an ATM ABR connection to support interactive user with acceptable QoS. Before interactive VoD over ATM with acceptable QoS can be run, the following problems must be solved. First, for a given video, what are the ABR connection parameters that are needed to provide acceptable QoS? Second, are these connection parameters feasible? And third, what kind of QoS will be achieved under these connection parameters? We attempt to answer these questions in this paper. The *objective* of this paper is *to provide acceptable QoS to IVoD-i transmission over the ATM ABR service.* The *goals* of this paper are three folds. The first is *to develop an analytical technique to determine if a feasible ABR connection can be setup for a given interactive user and a given video*. The second is *to select ABR connection parameters, such as MCR, PCR, ICR and user buffer size, to guarantee acceptable QoS to IVoD-i users*. The third is *to determine the QoS that will be obtained by users for a given set of connection parameters*. Our *contributions* are two fold. The first is that *our analysis technique takes the network latency and user interactivity into consideration*. The second is that *we developed an analytical method to calculate the ABR connection parameters which will provide acceptable QoS to the user*. Our technique has the *advantages* of *short start delay, small user buffer requirement and bounded end to end delay*.

The rest of the paper is organized as follows. In Section 2, the system model and assumptions for IVoD-i over ATM ABR are presented. In Section 3, we present a theory to determine if a feasible ABR connection can be setup, and to qunatify the ABR connection parameters and user buffer requirements. In Section 4, we present our simulation configuration, simulation results, and discussion. Section 5 concludes the findings of this paper.

2. System model for interactive VoD with instantaneous access

2.1. Model for interactivity

Our IVoD-i system is modeled in Fig. 1. Video source is connected to ATM network through a source side switch; an interactive user is attached to ATM network through a user side switch.

Video is compressed and stored in MPEG video. MPEG frames are grouped into a frame structure called Group of Pictures (GoP). A GoP starts with an I frame followed by B and P frames. According to the work in [27], the GoP size can be described by a Markov chain. Therefore, video sources can predict their rate. In MPEG video, I frame is the most important frame for restoring a picture. In our model, when the video source is in the interactive mode (fast forward or fast backward), only I frames are sent. An interactive user is modelled by two operation modes: playback mode and interactive mode. The user stays in the playback and interactive modes with exponential distribution [28] times with parameters α and β respectively, as shown in Fig. 2. This model was also used in [29] where the authors were mainly concerned with the number of users supported by a given system.

When a user switches to interactive mode, it issues a request to the video source. At the same time, it starts consuming buffered video at a rate with a speed factor of K times the playback rate. Since the most important features for an interactive user are the interactive operations such as fast forward and fast backward, to simplify the analysis, we only consider the playback and interactive modes for users.

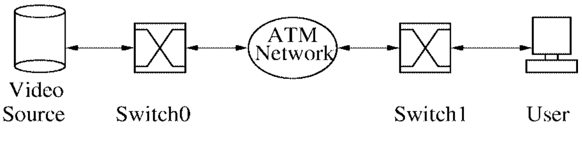

Fig. 1. IVoD-i system model.

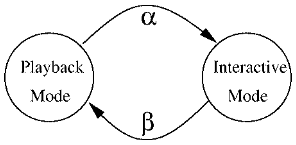

Fig. 2. Interactive user.

2.2. Notations

To facilitate further discussion, we define the following notations:

- F: Frame rate of video in frames/second.
- B: User buffer size in ATM cells.
- N: Total frame number for a movie.
- x_i: Number of cells for the ith frame. In this paper, MPEG compressed video, which is one of the most popular compressed formats, is used as video source. As shown in cited reference, it is well known that x_i is predictable.
- n_T: Number of frames transmitted before the first RM cell comes back.
- n_L: Number of frames the network will transmit at current rate before responding to user's new request, this reflects the network latency.
- I_{CR}: Initial Cell Rate (ICR) in cells/second.
- $A_{CR}(i)$: Allowable Cell Rate (ACR) for ith frame in cells/second.
- M_{CR}: Minimum Cell Rate (MCR) in cells/second.
- P_{CR}: Peak Cell Rate (PCR) in cells/second.
- K: Speed factor for interactive fast forward or fast backward operation at user.
- α: Time parameter for user at playback mode.
- β: Time parameter for user at interactive mode.

2.3. Assumptions

In the following analysis, we assume:

- When the network is under heavy congestion, the available bandwidth for the video source is MCR;
- When network is without any congestion, the video source will get PCR if it requires it;
- In optimal case, available cell rate will match the video rate.

3. Determination of ABR connection parameters for a feasible connection

ABR differs from other service types in two aspects. First, it provides a closed loop feedback mechanism, which makes it possible that sources can respond to network congestion dynamically by negotiating its future required bandwidth with the network. Second, Transit Buffer Exposure (TBE) of ABR makes it possible that sources can build up user buffers quickly (i.e., in round trip time). However, since an ABR connection only guarantees the MCR rate, in determining the bandwidth allocation during the entire connection, it is very important to negotiate proper connection parameters when connection is setup. If MCR is too low, it can not match the consumption rate at the user. User buffer will have risk of underflow resulting in users waiting for video, and degraded QoS. On the other hand, if PCR is too high, after users finish interactive access and return to playback, because of network latency, video will continue arriving at the PCR rate for some time. This may result in overflow at the user buffer and degraded QoS. Therefore, to guarantee QoS for the entire connection period, it is important to properly set MCR, PCR and ICR.

3.1. Determination of Minimum Cell Rate M_{CR}

Lemma 1. *For bandwidth constrained ABR connections, for all $n_T \in [1, N]$, the minimum cell rate M_{CR} to guarantee no underflow for an interactive user is*:

$$M_{CR} = \max\left(\frac{F}{n - n_T}\left(\frac{\beta + \alpha K}{\alpha + \beta}\sum_{i=1}^{n} x_i - \frac{I_{CR} n_T}{F}\right)\right), \tag{1}$$

where $n \in [n_T + 1, N]$.

Proof. Corresponding to the nth frame, playback time is n/F. Received video data S_{in} by the user is:

$$S_{in} = \sum_{i=n_T+1}^{n} \frac{A_{CR}(i)}{F} + \frac{I_{CR} n_T}{F}. \tag{2}$$

The first term in Eq. (2) is the number of cells sent at ACR after the first RM cell comes back. The second term represents the number of cells sent before the first RM cell comes back. Since n_T is the time in terms of the number of frames transmitted before the first RM cell comes back, and F is the frame rate in frames/second. Therefore, according to the ABR service rules, the frames transmitted in $A_{CR}(i)$ rate is from $n_T + 1$ to n. The number of cells transmitted in I_{CR} rate is therefore $I_{CR} n_T / F$. Note that, if $A_{CR}(i)$ is equal to the constant A_{CR}, we will have $A_{CR}(n - n_T)/F$ cells transmitted in A_{CR} rate, which has the same unit as the term $I_{CR} n_T / F$. For interactive users, it will run at playback mode with probability $\beta/(\alpha + \beta)$. During the period $[1/F, n/F]$, it will consume $(\beta/(\alpha+\beta)) \sum_{i=1}^{n} x_i$ cells from its buffer. It also will have the probability of $\alpha/(\alpha+\beta)$ in running at the interactive mode; the effective cells consumed is $(\alpha K/(\alpha+\beta)) \sum_{i=1}^{n} x_i$. Therefore, the number of consumed cells at the user is S_{out}:

$$S_{out} = \frac{\beta + \alpha K}{\alpha + \beta}\sum_{i=1}^{n} x_i. \tag{3}$$

The difference $S(\alpha, \beta, n)$ between received and consumed cells can be obtained by subtracting Eq. (3) from Eq. (2):

$$S(\alpha, \beta, n) = \sum_{i=n_T+1}^{n} \frac{A_{CR}(i)}{F} + \frac{I_{CR} n_T}{F} - \frac{\beta + \alpha K}{\alpha + \beta}\sum_{i=1}^{n} x_i. \tag{4}$$

In an ABR connection, ICR is normally set to be high. Therefore, during $n \in [1, n_T]$, $S(\alpha, \beta, n) \geqslant 0$ always holds. Only $n \in [n_T + 1, N]$ need to be considered. For ABR connections, if the network is in heavy congestion, ACR will shift toward MCR. In the worst case, it will run at MCR. To guarantee no underflow at user buffer in the worst case, it must have:

$$\frac{M_{CR}(n - n_T)}{F} + \frac{I_{CR} n_T}{F} - \frac{\beta + \alpha K}{\alpha + \beta}\sum_{i=1}^{n} x_i \geqslant 0. \tag{5}$$

Therefore, the Minimum Cell Rate M_{CR} is given by:

$$M_{CR} = \max\left(\frac{F}{n - n_T}\left(\frac{\beta + \alpha K}{\alpha + \beta}\sum_{i=1}^{n} x_i - \frac{I_{CR} n_T}{F}\right)\right), \tag{6}$$

for all $n \in [n_T + 1, N]$. □

3.2. Determination of Peak Cell Rate P_{CR}

If the network is not in congestion, sources can be allocated their requested high bandwidth. When a user finishes interactive operation, data will keep coming at the rate P_{CR} for some time due to network latency. At time n/F, overflow will happen if $S(\alpha,\beta,n)\geqslant B$, where B is the user buffer size.

Lemma 2. *For a given user buffer size B and ICR, the Peak Cell Rate P_{CR} must satisfy the following conditions for no overflow at the user during the entire connection period*:

$$P_{CR} = \min\left(\frac{F}{n-n_T}\left(\frac{\beta+\alpha}{\alpha}B + \frac{\beta+\alpha K}{\alpha} \times \sum_{i=1}^{n} x_i - \frac{\beta+\alpha}{\alpha}\frac{I_{CR}n_T}{F} - \frac{\beta}{\alpha}\sum_{i=n_T+1}^{n} x_i\right)\right), \tag{7}$$

where $n \in [n_T+1, N]$.

Proof. In this case, a source will work in an interactive mode with probability $\alpha/(\alpha+\beta)$, during which it will send video at the rate P_{CR}. It will be in playback mode with probability $\beta/(\alpha+\beta)$, during which it will send video at the rate $A_{CR}(i)$ which matches the video rate. So, user received data S_{in} is:

$$S_{in} = \frac{I_{CR}n_T}{F} + \sum_{i=n_T+1}^{n} \frac{\beta}{\alpha+\beta}\frac{A_{CR}(i)}{F} + \frac{\alpha}{\alpha+\beta}\frac{P_{CR}(n-n_T)}{F}. \tag{8}$$

The first term is the received data sent at I_{CR} rate, the second term is the received data sent at the rate $A_{CR}(i)$ during normal playback, the third term is received data sent at rate P_{CR} during fastforward/fastbackward. The user consumed data is still in the form as described in Eq. (3).

The difference function $S(\alpha,\beta,n)$ is:

$$S(\alpha,\beta,n) = \sum_{i=n_T+1}^{n} \frac{\beta}{\alpha+\beta}\frac{A_{CR}(i)}{F} + \frac{\alpha}{\alpha+\beta} \times \frac{P_{CR}(n-n_T)}{F} + \frac{I_{CR}n_T}{F} - \frac{\beta+\alpha K}{\alpha+\beta}\sum_{i=1}^{n} x_i. \tag{9}$$

For no overflow for a given user buffer B, it requires:

$$S(\alpha,\beta,n) - B \leqslant 0. \tag{10}$$

By substituting Eq. (9) into Eq. (10), we have:

$$\frac{\alpha}{\alpha+\beta}\frac{P_{CR}(n-n_T)}{F} \leqslant B + \frac{\beta+\alpha K}{\alpha+\beta}\sum_{i=1}^{n} x_i - \frac{I_{CR}n_T}{F} - \frac{\beta}{\alpha+\beta}\sum_{i=n_T+1}^{n}\frac{A_{CR}(i)}{F}. \tag{11}$$

Since the ACR matches the rate of video, we can replace $\sum_{i=n_T+1}^{n}(A_{CR}(i)/F)$ by $\sum_{i=n_T+1}^{n} x_i$. The above equation proves Lemma 2. From the above two Lemmas, we can choose M_{CR} and P_{CR} when setting up an ABR connection to guarantee QoS at user. However, both M_{CR} and P_{CR} depend on the Initial Cell Rate I_{CR}. □

3.3. Determination of Initial Cell Rate I_{CR}

Lemma 3. *For a connection with latency n_L, to guarantee QoS for an interactive user, I_{CR} must satisfy*:

$$I_{CR} \geqslant \frac{F}{n_T}\left(K\sum_{i=1}^{n_T} x_i + (K-1)\sum_{i=n_T+1}^{n_T+n_L} x_i\right). \tag{12}$$

Proof. We consider the situation that the user issues interactive operation right after connection is setup. The network will require time n_L/F to respond to the user's request. The number of received cells can be expressed as:

$$S_{in} = \frac{I_{CR} n_T}{F} + \sum_{i=n_T+1}^{n_T+n_L} \frac{A_{CR}(i)}{F}. \tag{13}$$

The number of cells consumed by a user is given by:

$$S_{out} = K \sum_{i=1}^{n_T+n_L} x_i. \tag{14}$$

In the optimal case, during the latency period, $A_{CR}(i)$ will match the video rate, and video will arrive at the user at the rate of P_{CR} after a delay of n_L/F. To avoid underflow, the following condition must be satisfied:

$$\frac{I_{CR} n_T}{F} + \sum_{i=n_T+1}^{n_T+n_L} \frac{A_{CR}(i)}{F} - K \sum_{i=1}^{n_T+n_L} x_i \geqslant 0. \tag{15}$$

To obtain the minimum required value for I_{CR}, replacing $\sum_{i=n_T+1}^{n_T+n_L}(A_{CR}(i)/F)$ with $\sum_{i=n_T+1}^{n_T+n_L} x_i$ in optimal case, we get:

$$I_{CR} \geqslant \frac{F}{n_T}\left(K \sum_{i=1}^{n_T} x_i + (K-1) \sum_{i=n_T+1}^{n_T+n_L} x_i\right). \tag{16}$$

□

From the ABR service rule, MCR, PCR and ICR are negotiated when connection is setup. MCR must not be greater than PCR. Lemmas 1, 2, and 3 answer the questions regarding the choice of connection setup parameters. We immediately have:

Theorem 1. *For given user buffer size B and Initial Cell Rate I_{CR}, there exists a feasible ABR connection to guarantee QoS at the user for a given interactive level if and only if*:

$$\max\left(\frac{F}{n-n_T}\left(\frac{\beta+\alpha k}{\alpha+\beta}\sum_{i=1}^{n} x_i - \frac{I_{CR} n_T}{F}\right)\right) \leqslant \min\left(\frac{F}{n-n_T}\left(\frac{\beta+\alpha}{\alpha}B + \frac{\beta+\alpha k}{\alpha}\sum_{i=1}^{n} x_i - \frac{\beta+\alpha}{\alpha}\frac{I_{CR} n_T}{F} - \frac{\beta}{\alpha}\sum_{i=n_T+1}^{n} x_i\right)\right), \tag{17}$$

for all $n \in [n_T, N]$.

Theorem 2. *For given network latency n_L and video fastforward/fastbackward speed factor K during the interactive mode, Initial Cell Rate I_{CR} can be determined by*:

$$I_{CR} \geqslant \frac{F}{n_T}\left(K \sum_{i=1}^{n_T} x_i + (K-1) \sum_{i=n_T+1}^{n_T+n_L} x_i\right). \tag{18}$$

Theorem 3. *The user will not suffer from long start up delay. Start up delay is of the same order as the round trip time.*

From Lemma 2, user buffer B can be expressed as:

$$B \geqslant \frac{\alpha}{\alpha+\beta}\frac{P_{CR}(n-n_T)}{F} + \frac{I_{CR}n_T}{F} + \frac{\beta}{\alpha+\beta}\sum_{i=n_T+1}^{n} x_i - \frac{\beta+\alpha K}{\alpha+\beta}\sum_{i=1}^{n} x_i. \tag{19}$$

Therefore, for given level of user interactivity and Initial Cell Rate, there exists a user buffer size B corresponding to a Peak Cell Rate.

4. Simulation configuration and results

The simulations were carried out using OPNET 5.1. Interactive VoD sources that can respond to requests issued by interactive VoD clients were implemented in OPNET 5.1. In our simulation, we used real MPEG video trace to drive the VoD sources. *Starwar and Soccer* [30] trace files were used. For a typical interactive multimedia system, previous results show that the level of interactivity $\alpha/(\alpha+\beta)$ is less than 20% for education applications where the fastforward/fastbackward is frequently used [31]. For interactive users watching movies delivered by video server, the observed interactive level $\alpha/(\alpha+\beta)$ is less than 2% and the speed factor K is equal to three [29]. In our simulation, we considered an interactivity level which is between the above values.

From the *Lemmas and Theorems* in Section 3, for a given user buffer size $B = 2$ Mbits (or 4717 ATM cells) and user interactivity level $\alpha/(\alpha+\beta) = 5\%$, the ABR connection parameters for the above MPEG video are shown in Table 1, where $K = 3$ as in [29].

4.1. Network configuration for simulation

The simulation configuration is shown in Fig. 3. System configuration parameters for ABR connection are shown in Tables 1 and 2, where *Fixed Delay* is the propagation delay from a source to its corresponding client.

Four sources are connected to the source side of switch *switch0*. All connections are OC-3. Four users are connected to the user side of switch *switch1* via OC-3. The OC-1 connection between *switch0* and *switch1* acts as the bottleneck link. *source1* and *source2* act as dynamic load to the bottleneck link and two switches with CBR and VBR traffic respectively. The load of CBR and VBR services on the OC-1 bottleneck link is 50% of its total capacity. ERICA scheme is used for *switch0* and *switch1*. The simulation time was 1000 seconds (about 17 minutes). Frame rate was set to 25 frames/second. In our simulation, two video sources started transmission at the same time. Since different video has different frame sizes (I, P, B) and/or GoP (Group of Picture), therefore, even in the case that two video sources started sending at the same time, the synchronization between the I frames will soon be lost.

4.2. Simulation results

In Figs 4 to 11 and Figs 14 to 15, x-axis represents the simulation time in minutes. In Figs 12 to 13, x-axis represents the delay suffered by the video.

Table 1

Calculated ABR connection parameters

Parameter	Starwar	Soccer
ICR (Mbps)	1.4	8.9
MCR (Mbps)	0.3	0.8
PCR (Mbps)	10.1	10.0

4.2.1. Playback trace and user buffer requirement

Figures 4 and 5 show the playback MPEG video frame sequences at the user. It is seen that *Starwar* has a very small frame size (about 25 kbits for I frame, 1.4 kbits for B frame, and 3 kbits for P frame) at the beginning. Therefore, *Starwar* needs a small ICR. On the contrary, *Soccer* has a very large frame size (about 145 kbits for I frame, 18 kbits for B frame, and 44 kbits for P frame) at the beginning. Therefore, *Soccer* needs a very large ICR. This is also shown in Table 1.

Figs 6 and 7 show the arrival and consumption rates of video at the client. It can be seen that the consumption curve at the user always follows the arrival curve. Figures 8 and 9 show the dynamic occupancy of user buffer during the transmission period. Buffer occupancy varies with the video traffic. Referring to Figs 4 and 5, the higher the video traffic burst, the larger the changes of user buffer occupancy. It is seen that for a given video, a connection

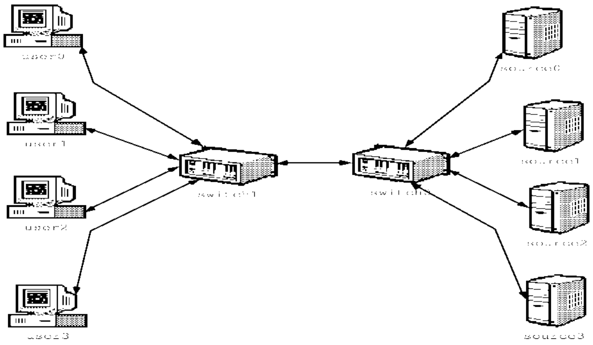

Fig. 3. System configuration for simulation.

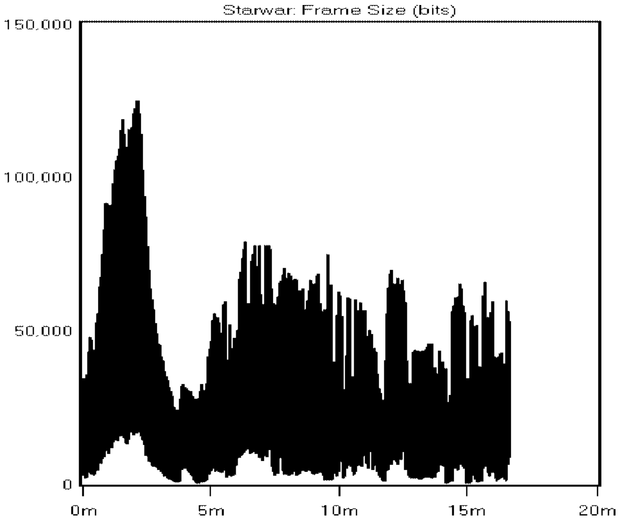

Fig. 4. Playback MPEG video *Starwar* frame trace at user versus simulation time.

Table 2
System configuration for ABR connections

Source	Dest	Service	Traffic	Fixed Delay
source0	user0	ABR	MPEG	50 ms
source3	user3	ABR	MPEG	52 ms

with the ABR parameters calculated with our framework will guarantee no underflow or overflow of the user buffer. Moreover, the user buffer occupancy will not exceed the user buffer size. From Fig. 8, for *Starwar*, the largest user buffer occupancy is around 1.2 Mbits. Even for *Soccer*, which has the largest burst frame, the largest value for user buffer occupancy does not exceed 1.5 Mbits, which is less than the given user buffer of size 2 Mbits as shown in Fig. 9. Therefore, our model is suitable for different types of video.

4.2.2. End to end delay

Figures 10 and 11 show end to end delay at the user. It is seen that start up delay for *Starwar* and *Soccer* are about 63 ms and 115 ms respectively. Recall that our network configuration has a round trip time (RTT) of about 100 ms. Therefore, we conclude that the start up delay is of the same order as the RTT value. The difference in start up delay between *Starwar* and *Soccer* comes from the traffic characteristic of these video. From Figs 4 and 5, since *Soccer* has much larger frame size (145 kbits for I frame, 18 kbits for B frame, and 44 kbits for P frame)

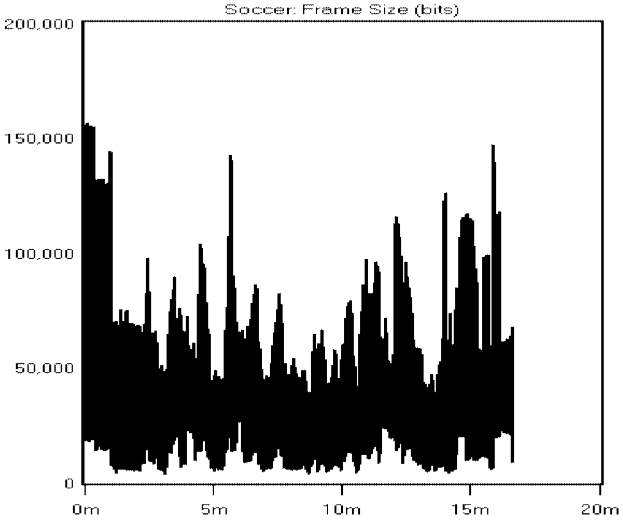

Fig. 5. Playback MPEG video *Soccer* frame trace at user versus simulation time.

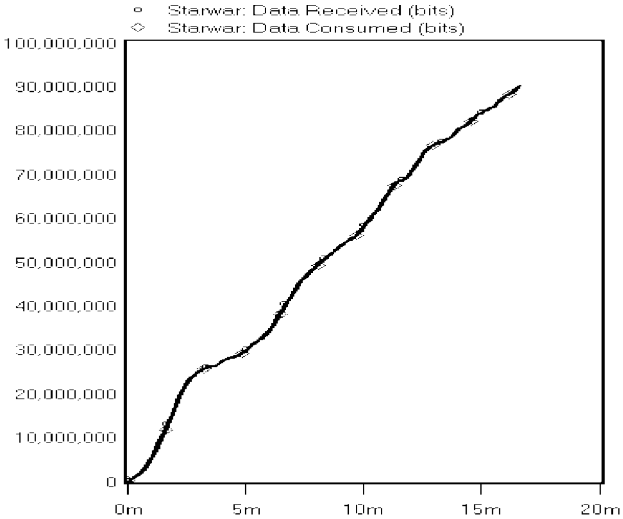

Fig. 6. Video *Starwar* arrival and consumption at user versus simulation time.

than *Starwar* (25 kbits for I frame, 1.4 kbits for B frame, and 3 kbits for P frame), it needs more time to transmit its frame to the user. Therefore, *Soccer* has larger start up delay. Although start up delay varies between videos, it is seen that the start up delay is of the same order as RTT. Therefore, users will not suffer from long start up delay, as pointed out in Theorem 3 in Section 3.

Figures 12 and 13 show the relative distribution of delay for *Starwar* and *Soccer*. For *Starwar*, the minimum end to end delay is around 60 ms with the maximum value of 105 ms for the entire video. The distribution of end to end delay is close to the lower end value. Similar results can be seen for *Soccer*; its minimum end to end delay is around 70 ms with a maximum value of 125 ms. Although the end to end delay varies between videos, we can conclude two important results. First, the minimum end to end delay is only a little higher than the fixed propagation delay. Second, the largest end to end delay is about two times of the minimum delay. The delay distribution satisfies the requirement described in ATM Forum Traffic Management 4.0.

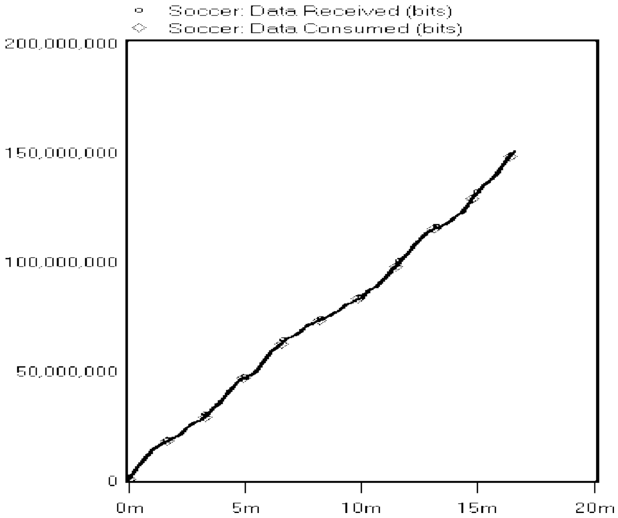

Fig. 7. Video *Soccer* arrival and consumption at user versus simulation time.

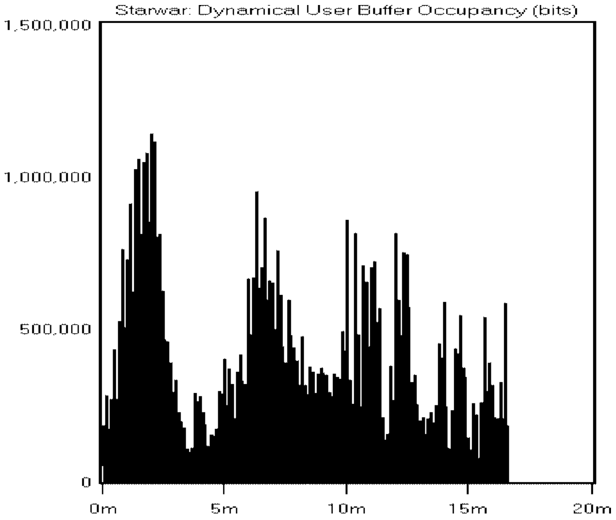

Fig. 8. Dynamic user buffer occupancy for video *Starwar* versus simulation time.

4.2.3. Jitter

For real time MPEG video transmission, delay variance or jitter between the adjacent transmitted frames is another important QoS criterion. Since MPEG frame decode is based on the GoP, i.e., the decode of B frame depends on the P frame which will arrive at user after the B frame. Too large delay variance between adjacent frame will not only affect the picture quality, but also waste the network bandwidth resource due to out of date. Therefore, one of the goal for quality acceptable MPEG video transmission is to limit the delay variance in a small range as possible. Figures 14 and 15 show delay variation at users. It is seen that delay variation is very small. The largest value is about 3.2 ms for video *Soccer*, and 1.3 ms for video *Starwar*. It is also seen that delay variance varies with the bursts of video traffic. The higher the burst, the larger the delay variance.

From the results given in Section 4.2.1, 4.2.2 and 4.2.3, we compare the achieved QoS and buffer requirement at user of our technique with the results from [3,9] in Table 3. Note that jitter performance was not presented in [3] and [9].

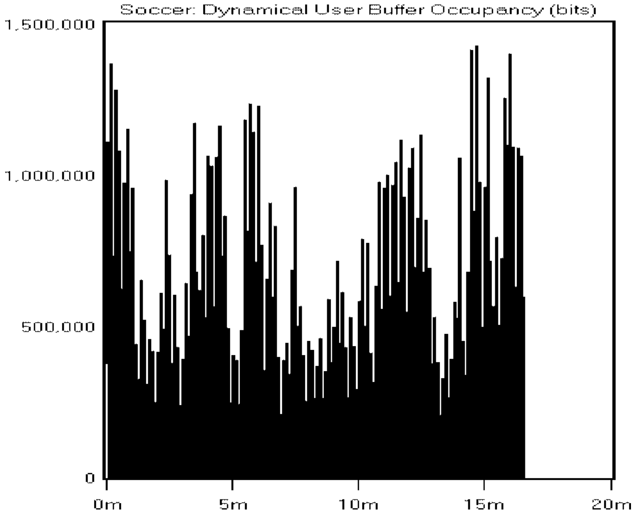

Fig. 9. Dynamic user buffer occupancy for video *Soccer* versus simulation time.

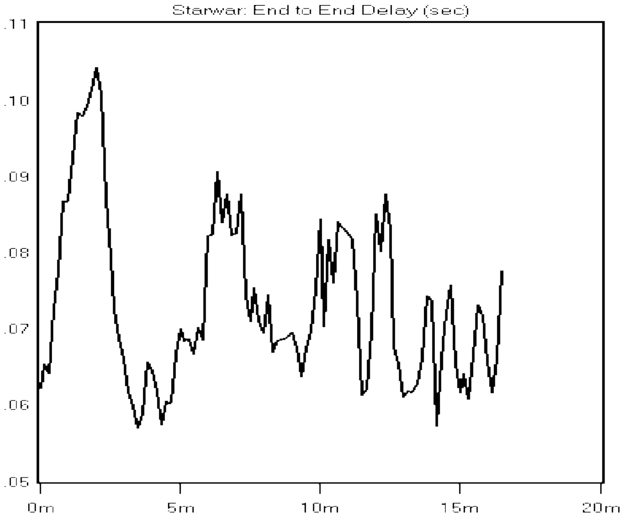

Fig. 10. End to end delay at user for video *Starwar* versus simulation time.

5. Conclusion

In this paper, we developed an analytical framework for interactive video on demand with instantaneous access capacity (IVoD-i) over ATM network with ABR service. The video source is pre-stored MPEG video, which can respond to requests issued by users. Users can perform VCR like interactive function. The *techniques for determining ABR parameters for a feasible connection are developed*. For a given video and level of user interactivities, we can determine the parameters ICR and MCR for an ABR connection. We have also developed a relationship between user buffer size and PCR. A connection which has been setup with these parameters will guarantee an acceptable QoS at user.

The QoS at users is evaluated with real life MPEG video traces. Results obtained have shown that an ABR connection with parameters calculated by our model offers acceptable QoS to interactive users. Users will not

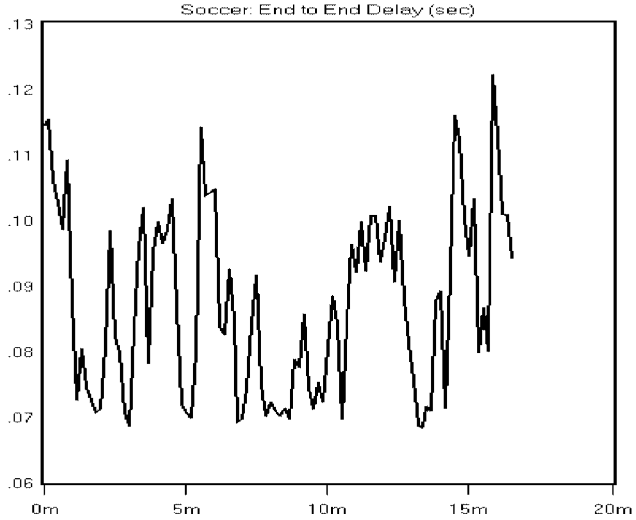

Fig. 11. End to end delay at user for video *Soccer* versus simulation time.

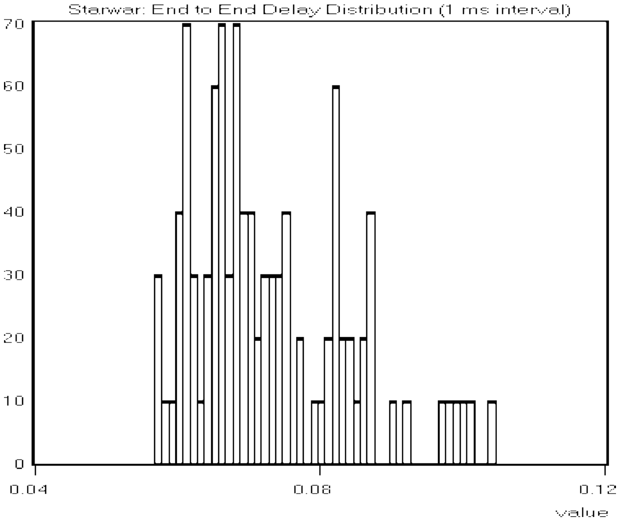

Fig. 12. Distribution of end to end delay for video *Starwar* at user.

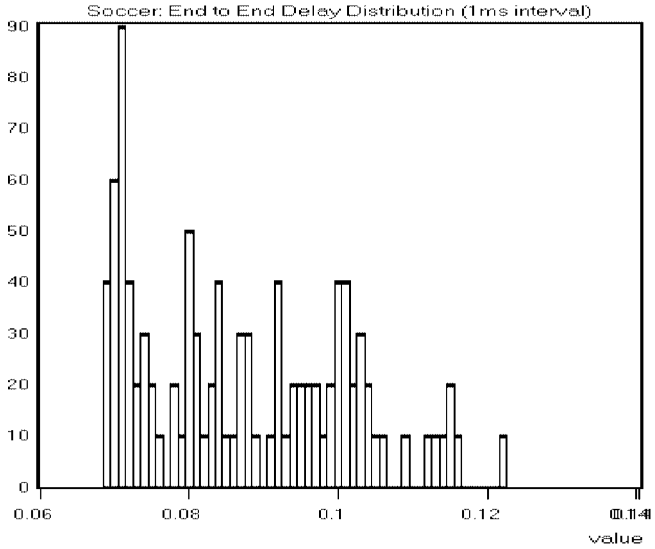

Fig. 13. Distribution of end to end delay for video *Soccer* at user.

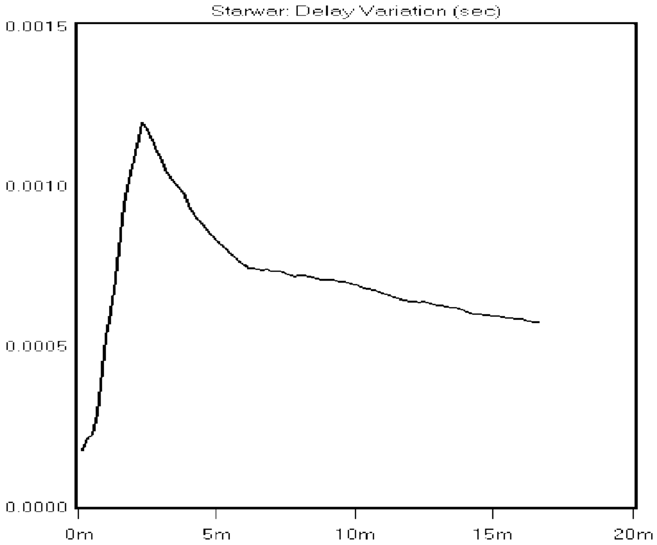

Fig. 14. Delay variation for video *Starwar* at user versus simulation time.

Table 3

Comparison of achieved QoS and buffer requirement at user

	User buffer	Start Delay	Jitter	Interactive
CRTT [3]	23 Mbytes	37 second	N/A	No
DBA [9]	0.7 Mbytes	0.48 second	N/A	No
Our Model	0.25 Mbytes	around RTT	4 ms	Yes

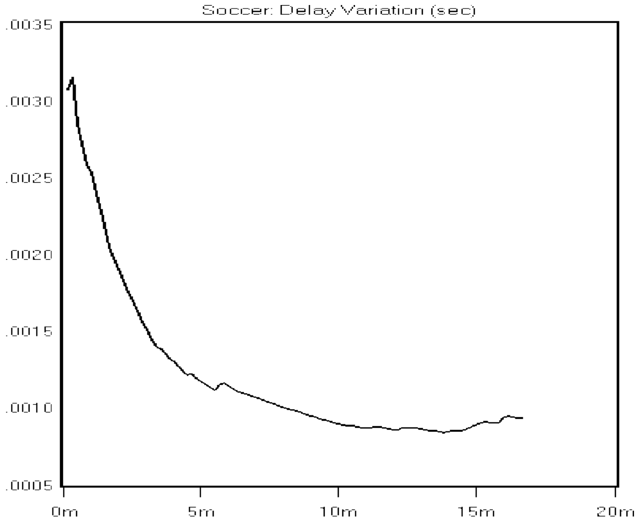

Fig. 15. Delay variation for video *Soccer* at user versus simulation time.

suffer from long start up delay. The start up delay is of the same order as the round trip time. The end to end delay is bound within an acceptable range. Also, the amount of video consumed by the user is well matched with the rate at which video is received by the user. This result in small user buffer requirement at the client.

Although this work is based on ATM ABR service, its basic principle and analysis techniques will be valuable for scenarios where the network can provide closed loop feedback and support bandwidth negotiation. The main focus of this work is to provide interactive VoD. Therefore, reducing the start up delay and avoiding client buffer over/under flow was of strong interest. Future work will concentrate on cell loss under various connection parameters.

References

[1] C. Tryfonas and A. Varma, MPEG-2 transport over ATM networks, *IEEE Communications Surveys* **2**(4) (1999), 24–33.

[2] D. Deloddere, W. Verbiest and H. Verhille, Interactive video on demand, *IEEE Communication Magazine* **32**(5) (1994), 82–88.

[3] J.M. McManus and K.W. Ross, Video on demand over ATM: Constant rate transmission and transport, *IEEE Journal on Selected Areas in Communications* **14**(6) (1996), 1087–1098.

[4] M. Krunz and S.K. Tripath, Exploiting the temporal structure of MPEG-2 video for the reduction of bandwidth requirement, in: *IEEE INFOCOM '97*, Kobe, Japan, 1997, pp. 143–150.

[5] M. Grossglauser, S. Keshav and D. Tse, RCBR: A simple and efficient service for multiple time scale traffic, in: *Proceedings of SINGCOMM '95*, Boston, 1995, pp. 219–230.

[6] N.G. Duffield, K.K. Ramakrishnan and A.R. Reibman, An algorithm for smoothed adaptive video over explicit rate network, *IEEE/ACM Transaction on Networking* **6**(6) (1998), 717–728.

[7] A.M. Adas, Using adaptive linear prediction to support real time VBR video under RCBR network service model, *IEEE/ACM Transaction on Networking* **6**(5) (1998), 635–644.

[8] T.V. Lakshman, P.P. Mishra and K.K. Ramakrishnan, Transporting compressed video over ATM networks with explicit rate feedback control, *IEEE Transaction on Networking* **7**(5) (1999), 710–723.

[9] L. Zhang and H. Fu, A novel scheme of transporting pre-stored MPEG video to support video on demand service, *Computer Communications* **23** (2000), 133–148.

[10] M. Graf, VBR video over ATM: Reducing network resource requirement through endsystem traffic shaping, in: *IEEE INFOCOM '97*, Kobe, Japan, 1997, pp. 48–57.

[11] M. Hamdi, J.W. Roberts and P. Rolin, Rate control for VBR video coders in broad-band networks, *IEEE Journal on Selected Areas in Communications* **15**(6) (1997), 1040–1051.

[12] C.J. Beckman, Dynamic bandwidth allocation for interactive video application over corporate network, in: *IEEE COMPCON '96*, 1996, pp. 219–225.

[13] Y. Yang and S.S. Hemami, Separate source and channel rate selection for video over ATM, in: *IEEE Data Compression Conference Proceedings*, Snowbird, UT, USA, 2000, p. 581.

[14] H. Kanakia, P.P. Mishra and A.R. Reibman, An adaptive congestion control scheme for real time packet video transport, *IEEE/ACM Transaction on Networking* **3**(6) (1995), 671–682.

[15] R. Grunenfelder and J.P. Cosmas, Characterisation of video codes as autogressive moving average processes and related queueing system performance, *IEEE J.Selected Areas in Communication* **9**(April) (1991), 284–293.

[16] D.P. Heyman, A. Tabatabai and T.V. Lakshman, Statistical analysis and simulation study of video teleconference traffic in ATM network, *IEEE Trans. Circuits and Systems for Video Technolgoy* **2**(1) (1992), 49–59.

[17] G. Ramamuthy and B. Sengupta, Modeling and analysis of a variable bit video multiplexer, in: *IEEE INFOCOM '92*, 1992, pp. 817–827.

[18] K. Chandra and A.R. Reibman, Modeling one and two layer variable bit rate video, *IEEE/ACM Transactions on Networking* **7**(3) (1999), 398–413.

[19] F.J. Kuo, J.S. Wu and D.L. Keng, Usage parameter control schemes for improving MPEG video quality over ATM networks, *Computer Communications* **22** (1999), 1585–1591.

[20] H. Zhang and E. Knightly, Red-vbr: A new approach to support vbr video in packet switching network, in: *Proceedings of NOSSDAV '95*, Durham, 1995, pp. 307–310.

[21] D.J. Reininger, D. Raychaudhuri and J.Y. Hui, Bandwidth renegotiation for VBR video over ATM networks, *IEEE Journal on Selected Areas in Communication* **14**(6) (1996), 1076–1086.

[22] H. Zhang and E. Knightly, Renegotiation-based approach to support delay-sensitive VBR video, *ACM Journal of Multimedia System* **5**(3) (1997), 164–176.

[23] A. Dagiuklas and M. Ghanbari, Rate-based flow control of video services in ATM networks, in: *IEEE GLOBECOM '96*, London, 1996, pp. 284–288.

[24] B. Ahn, K.-H. Cho, H. Song and J. Park, Design of rate-based congestion control scheme for MPEG video transmission in ATM network, in: *IEEE GLOBECOM '97*, Phoenix, 1997, pp. 1690–1694.

[25] G. Bianchi and R. Melen, The role of local storage in supporting video retrieval services on ATM networks, *IEEE/ACM Transactions on Networking* **5**(6) (1997), 882–892.

[26] J.P. Zhang and S. Gara, Improvement in transferring compressed video over ATM networks with enhanced ABR flow control, in: *IEEE GLOBECOM98*, Sydney, Australia, 1998, pp. 3128–3133.

[27] O. Rose, Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM system, research report, Institute of Computer Science, University of Wurzburg, February, 1995.

[28] V.O.K. Li, W. Liao, X. Qiu and E.W.M. Wong, Performance model of interactive video on demand system, *IEEE Journal on Selected Areas in Communications* **14**(6) (1996), 1099–1109.

[29] J.K. Dey-Sircar, J.D. Salehi, J.F. Kurose and D. Towsley, Providing VCR capabilities in large scale video servers, in: *ACM International Conference on Multimedia*, San Francisco, 1994, pp. 25–32.

[30] O. Rose, http://nero.informatik.uni-wuerzburg.de/mpeg/traces, March, 1995.

[31] P. Branch, G. Egan and B. Tonkin, Modeling interactive behaviour of a video based multimedia system, in: *IEEE ICC99*, Vancouver, BC, Canada, 1999.