System Design and Network Requirements for Interactive Multimedia

Bing Zheng and Mohammed Atiquzzaman

***Abstract*—In recent years, there has been a strong interest in transmitting compressed video over packet switched networks, such as asynchronous transfer mode (ATM). Previous work has dealt with transmitting MPEG over constant bit rate (CBR) and variable bit rate (VBR) services of ATM. The available bit rate (ABR) service of ATM is expected to be much more cost effective than CBR or VBR. However, there hasn't been much work done on running interactive client/server applications (for example, video on demand) over ABR.**

In this paper, we have developed a framework to design interactive video systems transmitting MPEG video over the ATM ABR service. We have developed models to determine the network connection parameters required to run interactive client/server multimedia applications over an ATM network using the ABR service. We solve our model using real-time dynamic equilation (RTDE) analysis. We conclude that by proper dimensioning of the buffers at the client and the server, it is possible to run interactive video over the ATM ABR service.

***Index Terms*—Buffer dimensioning, MPEG-2 video, multimedia over ATM systems, video system design.**

I. Introduction

With the availability of high-speed networks, there has been a strong interest in multimedia applications over high speed networks. Multimedia applications, such as video on demand and video conferencing, are envisioned to be the most widely used applications in the future high speed networks [1]. The asynchronous transfer mode (ATM) is a multiservice network which is capable of carrying voice, video and data over the same network. Consequently, there has been a lot of interest in optimizing the transmission of multimedia over ATM networks. The ATM Forum has standardized four types of ATM bearer services: constant bit rate (CBR), variable bit rate (VBR), available bit rate (ABR) and unspecified bit rate (UBR).

A number of authors have studied transmission of video over ATM based networks; most of the studies have used the CBR and VBR service categories. In [2]–[7], the authors have discussed traffic shaping, congestion control, bandwidth allocation and rate control for VBR video over ATM, while in [8]–[11], the authors investigated the bandwidth requirement for VBR video. The authors in [12] proposed a statistical model for Markov modulated bursty traffic and the application in ATM bandwidth allocation. In [13], the authors discussed the error resilient protocol to overcome the cell loss of MPEG-2 video transmission over ATM networks, the authors in [14] and [15] discussed the algorithm and techniques for quality of service (QoS) guarantee with VBR service. On other hand, since the Internet is the most widely deployed and used network, the authors in [16] discussed the call admission and resource reservation for QoS guarantee in video over internet.

The digital storage media command and control (DSM-CC) protocol is an open protocol to enable widespread use of integrated video based multimedia services. A number of ATM network architectures using DSM-CC for delivery of multimedia have been described in [17]. Mao [18] has discussed the system architecture and protocol stack for delivery of interactive MPEG-coded video over ATM using hybrid fiber coax (HFC) and fiber to the curb (FTTC) as the delivery media. Design of video server to support many users has been studied by a number of authors [19], [20]. The buffering architecture and requirements for stored video on demand have been discussed in [21], [22]. The authors in [23] discussed the video-on-demand (VoD) server with unrestricted VCR function in multicast VoD system.

The ABR service has the highest utilization of network resources, and offers an acceptable quality of service at a low cost. The ABR service was initially intended for data applications which did not require stringent bandwidth guarantees. However, recent studies [24]–[27] have shown that the feedback based ABR service can be successfully used to transmit video over an ATM network. In [25], the authors proposed a scheme for successfully transmitting variable bit rate compressed video over ATM using the explicit-rate congestion control mechanism of ABR. A smoothing and rate adaptation algorithm to be used by the compressed video source in conjunction with the explicit rate based control scheme of ABR has been proposed in [26]. The authors in [27] use two levels of video quality in order to cope with the variable bandwidth nature of ABR connections. A lower quality image is transmitted when the network is congested. The feedback control mechanism and service architecture for MPEG video systems were studied in [28], [29]. Quality control of variable bit rate video over the ABR service was presented in [30]. The above studies were concerned with the transmission of *noninteractive video*. The *objective* of this paper is to develop a framework for the design of interactive video systems for *interactive multimedia* over the ABR service of ATM. We use VoD as an example of interactive multimedia application.

A VoD system running over a network (see Fig. 1) typically consists of a server and a client. The server sends precoded multimedia which is stored in a storage system such as redundant array of inexpensive disks (RAID). Depending on the type of connection, the server may use different modes of transmission. The server has a buffer to prevent cell losses during periods of network congestion. The client has a decoder and a buffer which is used to smooth out any fluctuation in the data arrival rate (jitter) [31], [32] from the network to the client, and to prevent cell loss at the client. The size of the buffer at the client,

Manuscript received May 3, 2003; revised December 30, 2003. This paper was recommended by Associate Editor S. Chen.

B. Zheng is with the New Focus Inc., San Jose, CA 95134 USA (e-mail: zhengbin@ieee.org).

M. Atiquzzaman is with the School of Computer Science, University of Oklahoma, Norman, OK 73019 USA (e-mail: atiq@ou.edu).

Digital Object Identifier 10.1109/TCSVT.2004.839982

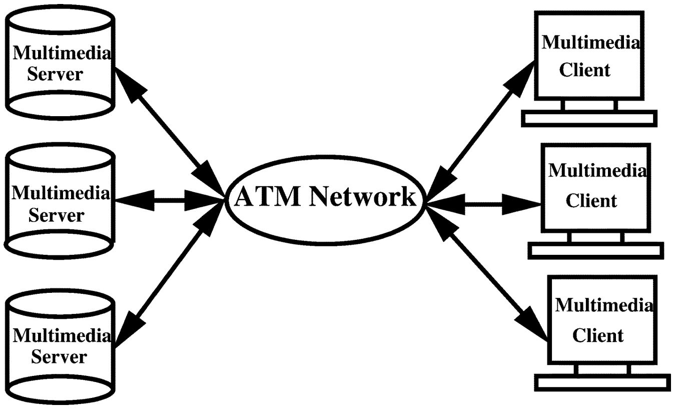

Fig. 1. Networked client/server video application.

and the mode used by the server to transmit the multimedia determines the cost of the system and the QoS at the client. QoS at the client includes issues such as cell loss due to buffer overflow at the client, lack of data due to buffer underflow at the client, delay between the server and the client (due to network round trip time, buffering delays at the server and client, and rate mismatch between the multimedia rate and the available bandwidth from the network), etc.

To make such an interactive client server based video system commercially viable, it is essential to reduce the operational cost, and the cost of client and server by optimizing the size of these expensive buffers. To maintain QoS at the client, the buffers should be dimensioned to avoid overflow or underflow. Such design decisions require the development of performance models to enable optimal design of interactive client-server systems operating over an ABR service.

The video buffer verifier (VBV) is a model *hypothetical* decoder buffer that will not overflow or underflow when fed with a conforming MPEG bit stream. The above definition of VBV is used to define a compliant MPEG stream, where the transmission medium is modeled as an *instantaneous transfer medium* from the encoder buffer to the decoder buffer. In the system being considered in this paper, the MPEG video stream that is fed to the decoder buffer arrives from the network, with highly variant bandwidth and jitter (delay variation). The MPEG video arriving at the input to our decoder, therefore, can not be assumed to be VBV compliant, and the definition of VBV does not apply to our scenario of video over packet switched networks.

When the ABR service is used to transmit multimedia from the server, the server has to open an ABR connection which involves negotiating a set of connection setup parameters with the network. The values of the connection parameters depend on issues such as the mode of transmission from the server, the amount of buffer at the client, the network congestion status, the network delay between the server and the client, etc. It is therefore crucial to determine the optimal connection parameters in order to reduce the operational cost and the QoS at the client.

Most of the previous efforts on transmitting video over ATM have focused on either source behavior or user performance, and are usually based on the CBR and VBR service categories. Moreover, they have not studied the interactive nature of the client and its impact on the design of the server and client. Since ABR provides a much more cost effective service for video than offered by CBR or VBR, it is crucial to study the system design, performance, and networking requirements of *interactive* video over the ATM ABR service. The authors are not aware of any *interactive video system over the ATM ABR service*. In this paper, we develop a systematic method to *design and analyze the performance of interactive Video on Demand Systems, and determine the required network connection parameters* when the system is running over an ATM ABR connection. The challenge we tackle in this paper is to design interactive video systems which will operate over the time-variable bandwidth channel of the ATM ABR service.

The contributions of this paper are as follows.

- Proposed *new models of the client and the server* which take into account the interactive nature of the client and the uncertainties in the available bandwidth from the network.
- Developed analytical models to determine the *minimum fill level* at the client buffer to allow continuous playout of video by avoiding buffer underflow.
- Developed models to determine the *optimal buffer size* at the client and the server.
- Proposed techniques to evaluate the values of the *ABR connection parameters* (i.e., the RM cell parameters).
- Evaluated the effect of *group of picture (GOP) size* and level of *user interactivity* on the buffering requirements at the client and server.
- Evaluated *delay and delay variation* between the server and client.

Since analytical techniques offer greater insight into the functionality of a system, and allow a fast method of fine tuning the system parameters, we carry out our study using Markov chains and real-time dynamic equilation (RTDE) techniques. The *novelty* of our proposed system model and analysis techniques is their applicability to interactive VoD system analysis and design. Our model can serve as a framework for designing interactive client server based video systems over bandwidth constrained channels.

The rest of the paper is organized as follows. In Section II, we give an overview of networked VoD systems over the ATM ABR service. In Section III, we propose a new client model and its interaction with the ABR service, analyze its operation, determine the minimum fill level of the buffer to avoid underflow and obtain the required minimum buffer size. In Section V, we propose a novel server model which is compatible with the ABR service category, obtain the minimum buffer size required at the server, and obtain the RM cell parameters required to transmit video over the ABR service. Numerical results are presented in Section VII, followed by conclusions in Section VIII.

II. VoD Over ATM ABR Service

A VoD system using the ATM ABR is shown in Fig. 2. In this study, we use an interactive VoD system where the client can perform interactive behavior (trick modes) such as fastforward and fastbackward. Corresponding to various interactive operations of the client, the server has different operating states. Both the server and the client have buffers to smooth the data

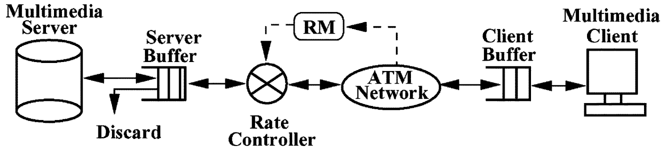

Fig. 2. VoD system over ATM ABR service.

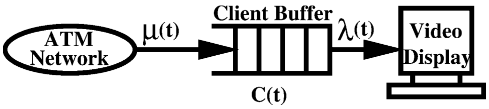

Fig. 3. Client model.

stream and reduce data loss. In our discussion, we assume the following.

- The client can perform VCR-like interactive function such as request a video, playback, fastforward (FFW), fastbackward (FBW), and stop.
- The server can respond to the client requests by sending video at different rates.
- There exist sufficient bandwidth in the transmission channel i.e., probability of network congestion is small and may last for a short time as compared to the duration of a GOP.
- The FFW/FBW operation lasts for a very short time as compared to the playback time.
- During a FFW/FBW operation, the video data consumption rate by the client is k time higher than its playback rate. The video server will try to send video data at k' times higher rate than the playback rate.

Based on the above assumptions, we develop the client behavior model in the next section.

III. Client Model and Operating Principle

In this section, we develop a model for the operation of the client. It will be used in later sections to derive expressions for buffer dimensioning at the client and the server. The client is modeled as consisting of a buffer and a video decoder/display as shown in Fig. 3. The buffer is required to smooth out fluctuations in the rate at which the client receives data from the network. We characterize the client behavior and operating modes as follows.

- The playback video stream coming to the buffer at time t has a frame rate $f(t)$ and each frame is of size $s(t)$. The bit rate $\mu(t)$ of the video stream is obtained by multiplying the frame rate with the frame size

$$\mu(t) = f(t)s(t). \tag{1}$$

- The client buffer, having a maximum size of C_u, has a fill level $C(t)$ at time t. We assume that the buffer is initially empty.
- The client can perform VCR-like functions as described in the previous section.

A. Client States

The four operating states of the client (see Fig. 4), are modeled by a Markov chain as follows.

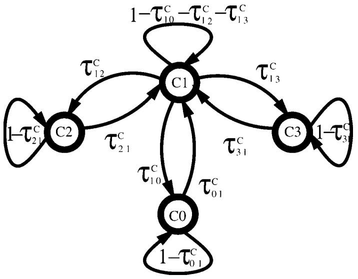

Fig. 4. State diagram of client.

1) *State C0:* The *stop* state where the client is not receiving or consuming video stream.

2) *State C1:* The *playback* state where the client receives data at the playback speed. In this state, the client consumes frames of sizes $s(t)$ at a rate $q(t)$ at time t. The data consumption rate, $\lambda_1(t)$, is given by

$$\lambda_1(t) = q(t)s(t). \tag{2}$$

3) *States C2 and C3:* In the *fastforward* (FFW) state (C2) or the *fastbackward* (FBW) state (C3), the client sends a FFW/FBW request to the server and consumes the current content of the client buffer at a speed which is k times faster than the rate at which it consumes data during playback. Therefore, the FFW and FBW rates given by $\lambda_2(t)$ and $\lambda_3(t)$ are given by

$$\lambda_2(t) = \lambda_3(t) = kq(t)s(t). \tag{3}$$

B. Steady State Distribution

The state transition probabilities between the client states are shown in Fig. 4, where $\tau^c_{i,j}$ represent the state transition probability from state i to state j of the client. The client operates as follows.

- The client always start from state C0 and can only go to state C1.
- The client can perform a FFW or FBW directly from state C1; after the completion of the FFW or FBW, it returns back to state C1.

According to the properties of Markov chain [33], a stationary state exists for a long-run behavior of the above client model. Let the steady state probability vector of the client states be represented by $V^c = (V^c_0, V^c_1, V^c_2, V^c_3)$ which satisfies

$$V^c = V^c P^c \tag{4}$$

where the transition matrix P^c is given by

$$P^c = \begin{bmatrix} 1-\tau^c_{0,1} & \tau^c_{0,1} & 0 & 0 \\ \tau^c_{1,0} & 1-\tau^c_{1,2}-\tau^c_{1,3}-\tau^c_{1,0} & \tau^c_{1,2} & \tau^c_{1,3} \\ 0 & \tau^c_{2,1} & 1-\tau^c_{2,1} & 0 \\ 0 & \tau^c_{3,1} & 0 & 1-\tau^c_{3,1} \end{bmatrix}. \tag{5}$$

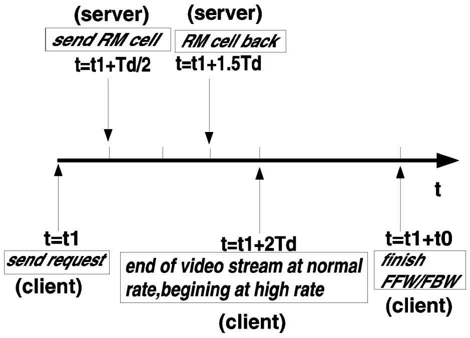

Fig. 5. Timing diagram to calculate $C_{\min}$.

We solve the above equations to obtain the steady state probability vector of the client states as follows:

$$\begin{aligned} V_0^c &= \frac{1}{1+\frac{\tau_{0,1}^c}{\tau_{1,0}^c}\left(1+\frac{\tau_{1,2}^c}{\tau_{2,1}^c}+\frac{\tau_{1,3}^c}{\tau_{3,1}^c}\right)} \\ V_1^c &= \frac{\tau_{0,1}^c}{\tau_{1,0}^c\left(1+\frac{\tau_{0,1}^c}{\tau_{1,0}^c}\left(1+\frac{\tau_{1,2}^c}{\tau_{2,1}^c}+\frac{\tau_{1,3}^c}{\tau_{3,1}^c}\right)\right)} \\ V_2^c &= \frac{\tau_{0,1}^c\tau_{1,2}^c}{\tau_{1,0}^c\tau_{2,1}^c\left(1+\frac{\tau_{0,1}^c}{\tau_{1,0}^c}\left(1+\frac{\tau_{1,2}^c}{\tau_{2,1}^c}+\frac{\tau_{1,3}^c}{\tau_{3,1}^c}\right)\right)} \\ V_3^c &= \frac{\tau_{0,1}^c\tau_{1,3}^c}{\tau_{1,0}^c\tau_{3,1}^c\left(1+\frac{\tau_{0,1}^c}{\tau_{1,0}^c}\left(1+\frac{\tau_{1,2}^c}{\tau_{2,1}^c}+\frac{\tau_{1,3}^c}{\tau_{3,1}^c}\right)\right)}. \end{aligned} \tag{6}$$

At time t, the expected data consumption rate by the client can be found by

$$E[\lambda(t)] = \sum_{i=1}^{3} V_i^c \lambda_i(t). \tag{7}$$

Since the fastforward and fastbackward rates $\lambda_2(t)$ and $\lambda_3(t)$ are k times the playback rate $\lambda_1(t)$, $E[\lambda(t)]$ can be written as

$$E[\lambda(t)] = [V_1^c + k(V_2^c + V_3^c)]\lambda_1(t). \tag{8}$$

In the next section, we use the above model for client behavior to develop the buffering requirements at the client.

IV. Minimum Client Buffer Size

In this section, we use the RTDE method to determine the minimum amount of buffer required at the client to avoid any client buffer underflow. We assume that the client buffer is empty at time $t=0$. The condition for no buffer overflow or underflow during the time period $[0,t]$, is given by

$$0 \le C(t) \le C_u, \qquad \text{for all } t \ge 0. \tag{9}$$

The client buffer accumulation $C(t)$, during the period $[0,t]$ is obtained from

$$C(t) = \int_0^t (\mu(t) - \lambda(t))dt. \tag{10}$$

Assume that at time $t = t_1$, the client switches from playback to FFW (or FBW) which has a duration of time t_0 as shown in Fig. 5. The server will react to this operation after a single trip delay time $T_d/2$, where T_d is the round trip delay in the ATM network. The server needs time T_d to obtain the required FFW/FBW bandwidth of $k'\mu(t)$, and after time $T_d/2$ the FFW/FBW video stream reaches the client. Therefore, from time t_1 to $t_1 + 2T_d$, the input data rate to the client buffer will still be $\mu(t)$. We denote the amount of data received by the client during this duration as Q_{in1}. From time $t_1 + 2T_d$ to $t_1 + t_0$, the input data rate to the client buffer will be $k'\mu(t)$. We denote the amount of data received by the client buffer during this duration as Q_{in2}. During the FFW/FBW operation, the client will consume an amount of data equal to Q_d. Therefore, Q_{in1}, Q_{in2}, and Q_d are given by

$$Q_d = \int_{t_1}^{t_1+t_0} k\lambda_1(t)d(t) \tag{11}$$

$$Q_{in1} = \int_{t_1}^{t_1+2T_d} \mu(t)d(t) \tag{12}$$

$$Q_{in2} = \int_{t_1+2T_d}^{t_1+t_0} k'\mu(t)d(t). \tag{13}$$

During the entire duration of the FFW/FBW operation, the following condition must be met to avoid any starvation at the client:

$$C(t_1) + Q_{in1} + Q_{in2} - Q_d \ge 0. \tag{14}$$

Substituting Q_d, Q_{in1}, Q_{in2} in the above equation, we get

$$\begin{aligned} C(t_1) + \int_{t_1}^{t_1+2T_d} \mu(t)d(t) + \int_{t_1+2T_d}^{t_1+t_0} k'\mu(t)d(t) \\ - \int_{t_1}^{t_1+t_0} k\lambda_1(t)d(t) \ge 0. \end{aligned} \tag{15}$$

By separating the fourth term into integral intervals $[t_1, t_1+2T_d]$ and $[t_1+2T_d, t_1+t_0]$, we rewrite the above expression as

$$\begin{aligned} C(t) \ge \int_{t_1}^{t_1+2T_d} (k\lambda_1(t) - \mu(t))d(t) \\ + \int_{t_1+2T_d}^{t_1+t_0} (k\lambda_1(t) - k'\mu(t))d(t). \end{aligned} \tag{16}$$

Since the FFW/FBW operation lasts for a short period of time, using expected values for video stream rates and rewriting the above equation in average form during the integral interval, we get the expression for client buffer fill level as

$$\begin{aligned} C(t) \ge 2T_d(kE[\lambda(t)] - E[\mu(t)]) \\ +(t_0 - 2T_d)(kE[\lambda(t)] - k'E[\mu(t)]). \end{aligned} \tag{17}$$

$E[\lambda(t)]$ and $E[\mu(t)]$ are the expected values of the video rate. For no buffer underflow (starvation), we can approximate the above expression by taking the larger of the two which gives us

$$\begin{aligned} C(t) \ge 2(k-1)T_d \max(E[\lambda(t)], E[\mu(t)]) \\ +(t_0 - 2T_d)(k - k')\max(E[\lambda(t)], E[\mu(t)]). \end{aligned} \tag{18}$$

Normally, $t_0 \gg T_d$. Under optimal condition where the input data rate to the client buffer is the same as the rate of data consumption by the client, $(k - k') = 0$ in (18). This gives us the

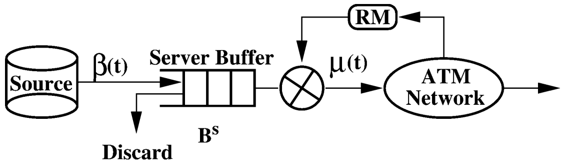

Fig. 6. Server model.

minimum buffer fill level ($C^c_{\min}$) in the optimal case (which also sets a limit on the minimum buffer size) required at the client to prevent underflow as follows:

$$C^c_{\min} \geq 2(k-1)T_d \max(E[\lambda(t)], E[\mu(t)]). \qquad (19)$$

In the worst case, for example when the network is congested, there will be a mismatch between the rates at which the client receives and consumes data, i.e., the $(k-k') \neq 0$ in (18). In such a case, the maximum client buffer fill level $C^c_{\max}$ to prevent starvation is given by

$$\begin{aligned} C^c_{\max} \geq\ & 2(k-1)T_d \max(E[\lambda(t)], E[\mu(t)]) \\ & + (t_0 - 2T_d)(k-k') \max(E[\lambda(t)], E[\mu(t)]). \qquad (20) \end{aligned}$$

From (19) and (20), we observe that the *required client buffer fill level for no starvation* is decided by the network size (or FRTT), the level of network congestion, the level of client interactivity, and the video rate (reflected in $E[\lambda(t)]$). In the next section, we develop a server model, and develop the buffering requirements.

V. Server Model and Operating Principle

The server consists of a video source (see Fig. 6) and a shaping buffer to smooth the video traffic before injecting it into the ATM ABR connection. The backward RM cell from the ATM network determines the bit rate at which the server is allowed to inject data into the ABR connection. During connection setup, the server negotiates ABR traffic parameters [34], such as the peak cell rate (PCR), initial cell rate (ICR), minimum cell rate (MCR), etc.

The server uses precoded video (MPEG) as a video source. MPEG video consists of I-, P-, and B-frames which are grouped in a special structure called GOP. Each GOP includes one I-frame followed by a number of B- and P- frames. A GoP is denoted by MmNn which represents a total of n frames in the GOP with $(m-1)$ number of B-frames between successive anchor frames. For example, the sequence of frames in an M3N9 GOP are given by IBBPBBPBB.

The server operates as follows.

- At the start of transmission, the server sets its *ICR* equal to *PCR* and the *MCR* equal to the mean bit rate of B-frames which has the lowest bit rate in an MPEG stream.
- During transmission, the *ACR* is approximately equal to the mean bit rate of the video.
- At the beginning of each GOP, it asks for a bandwidth equal to *PCR* which is less than the bit rate of the I-frames; after sending the I-frame, it slow down to the *ACR* rate.
- When the server receives a FFW or FBW request from the client, it continues sending at the current rate, and then sends an RM cell requesting a higher bandwidth. After receiving the backward RM cell, the server sends video data at the authorized higher rate.

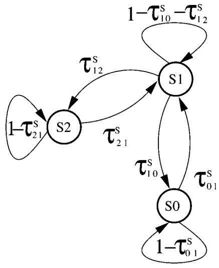

Fig. 7. State diagram of server.

TABLE I
RM Cell Field Parameters

Parameter	Value
PCR	decided by Eq. (22), (23)
MCR	mean bit rate for B frame in a GOP
ICR	PCR
ACR	decided by Eq. (22), (23)
TBE	$C(t)_{\min}$

A. Server States and Steady-State Probabilities

The server sends data at different rates depending on the state (playback, FFW, FBW, etc.) of the client. The following server states can therefore be identified (Fig. 7).

- *State S0:* Represents the stop state, no data is sent.
- *State S1:* Represents the state where the server is sending data for playback at the client.
- *State S2:* Represents the state where the server is sending data at a high speed, corresponding to the FFW or FBW state of the client.

B. Server Buffer Size and ABR Connection Setup Parameters

In this section, we determine the amount of buffering required at the server. Because it depends on the ABR connection parameters negotiated during connection setup, we also determine the connection parameters (shown in Table I) in this section. The Transient Buffer Exposure (*TBE*) parameter in the RM cell is the number of cells that the server can send during the connection setup period before the return of the first RM cell [34]. To ensure that the client has the required minimum buffer fill level, we set *TBE* equal to the minimum client buffer size. Let

- T_f = time duration of a frame;
- $\beta_I, \beta_P, \beta_B$ = mean bit rates of I-, P-, and B-frames, respectively;
- $E[\beta]$ = ABR of an MmNn GOP transmitted by the server in the playback state;
- B^s = minimum buffer size for server;
- $C(t)_{\min}$ = minimum fill level of client buffer at time t.

Compressed video from the hard disk is sent by the server to the server buffer which is then sent to the network depending on the bandwidth available from the network. If the available bandwidth is less than the rate at which the server sends video to the buffer, data will accumulate at the server buffer. To determine the optimal server buffer size, the dynamic variation of the buffer accumulation at the server should be zero for each GOP

$$\delta(\text{buffer accumulation}) = 0. \tag{21}$$

Then, the *PCR* and *ACR* must satisfy the following relationship:

$$(\beta_I - PCR) \le ACR(n-1) - \beta_P\left(\frac{n}{m} - 1\right) - \beta_B\frac{n(m-1)}{m} \tag{22}$$

where, the left hand side is the data accumulated from an I-frame due to β_I being higher than the PCR. The right hand side represents the amount of data sent to the network during the remaining $(n-1)$ frames of the GOP at the ACR rate. The amount of data sent during this time consists of the B- and P-frames of the GOP and the excess data from the I-frame.

When the server is performing a FFW/FBW operation, it will send data at a rate $kE[\beta]$ which is the maximum transmission rate of the server. Hence the *minimum value of PCR* should be set to $kE[\beta]$

$$\text{PCR} \ge kE[\beta]. \tag{23}$$

If the PCR is set to $kE[\beta]$, by substituting it into (22), we have

$$\beta_I - kE[\beta] \le \text{ACR}(n-1) - \beta_P\left(\frac{n}{m} - 1\right) - \beta_B\frac{n(m-1)}{m} \tag{24}$$

$$\text{or,} \quad \text{ACR}(n-1) \ge \beta_I + \beta_P\left(\frac{n}{m} - 1\right) + \beta_B\frac{n(m-1)}{m} - kE[\beta] \tag{25}$$

from which we obtain the *requirement for ACR* during playback as

$$\text{ACR} \ge \frac{n-k}{n-1}E[\beta]. \tag{26}$$

The I-frame usually has the highest data rate among all the frames. The server uses *PCR* to send data during the I-frame, and since $PCR < \beta_I$, the server buffer will need to buffer the excess data from the I-frame. Therefore, the *minimum server buffer size*, B_s, is decided by the accumulation of the excess data in time T_f, and is given by:

$$B^s_{\min} = (E[\beta_I] - \text{PCR})T_f. \tag{27}$$

As an example, lets take a typical M3N9 GOP of an MPEG-2 video with $\beta_I = 8.25$ Mb/s, $\beta_P = 2.25$ Mb/s, $\beta_B = 0.6$ Mb/s, $E[\beta] = 1.817$ Mb/s, and $T_f = 0.033$ sec [24]. Equation (27) [with PCR being determined by (23)] gives a minimum server buffer size of 11.67 kbytes at PCR = 5.468 Mb/s and ACR = 1.363 Mb/s.

VI. Delay and Delay Variation

Interactive video application is sensitive to delay and delay variation. In this section, we will study the delay and delay variation of sending the video from the server to the client. The client and sever buffers will introduce queuing delay and the delay variation (jitter) due to the variable bit rate nature of compressed video traffic in a packet switched network. We define delay variation as the difference between the maximum and minimum delay between the server and client for a GOP. Let $D_{\min}$ and $D_{\max}$ denote the minimum and maximum values of the delay during a GOP. The delay variation (Δ) is therefore expressed as

$$\Delta = D_{\max} - D_{\min}. \tag{28}$$

The delay in a client/server type system is the sum of the delay in the server, the network and the client. In this paper, we only consider the delay and delay variation at server and client buffers. Since ABR service guarantees only the MCR, in the case of heavy network congestion, the data rate from the server will be equal to the MCR given by $E[\beta_B]$. We now calculate the maximum accumulation in the server buffer during one GOP. Because the minimum outgoing rate is $E[\beta_B]$, the maximum server buffer accumulation during a GOP is given by the rate difference between the I- and P-frame rate and the outgoing rate as follows:

$$B^s_{\max} = (E[\beta_I] - E[\beta_B])T_f + (E[\beta_P] - E[\beta_B]) \times \left(\frac{n}{m} - 1\right)T_f. \tag{29}$$

The first term is due to the accumulation from an I-frame, whereas the second part is due to the accumulation from $((n/m) - 1)$ number of P-frames in a GOP. The B-frame does not contribute to the accumulation because the *MCR* is equal to $E[\beta_B]$. From Little's formula, $D^s_{\max}$ and $D^s_{\min}$ at server are obtained by dividing the maximum and minimum server buffer fill level which are given by (29) and (27) with the average rate as follows:

$$D^s_{\max} = \frac{B^s_{\max}}{E[\mu(t)]} \tag{30}$$

$$D^s_{\min} = \frac{B^s_{\min}}{E[\mu(t)]}. \tag{31}$$

When client performs interactive operation, it requires a bandwidth which equal to the fastfowrd/fastbackword rate of the multimedia stream. From (19) and (20) for the minimum and maximum client buffer occupancy $C^c_{\min}$ and $C^c_{\max}$, the maximum and minimum delay at the client $D^c_{\max}$ and $D^c_{\min}$ are expressed as

$$D^c_{\max} = \frac{2(k-1)T_d \max(E[\lambda(t)], E[\mu(t)])}{E[\lambda(t)]} + \frac{(t_0 - 2T_d)\delta \max(E[\lambda(t)], E[\mu(t)])}{E[\lambda(t)]} \tag{32}$$

$$D^c_{\min} = \frac{2(k-1)T_d \max(E[\lambda(t)], E[\mu(t)])}{E[\lambda(t)]} \tag{33}$$

where $\delta = k - k'$ is the rate mismatch factor between the rate required by the client to perform fastforward/fastbackword operation and the rate allocated by the network. In the best case of the network having no congestion, it will assign the required bandwidth to the server, i.e., $k' = k$. In worst case, the network assigns the guaranteed (MCR) rate only, i.e $k' < k$. In

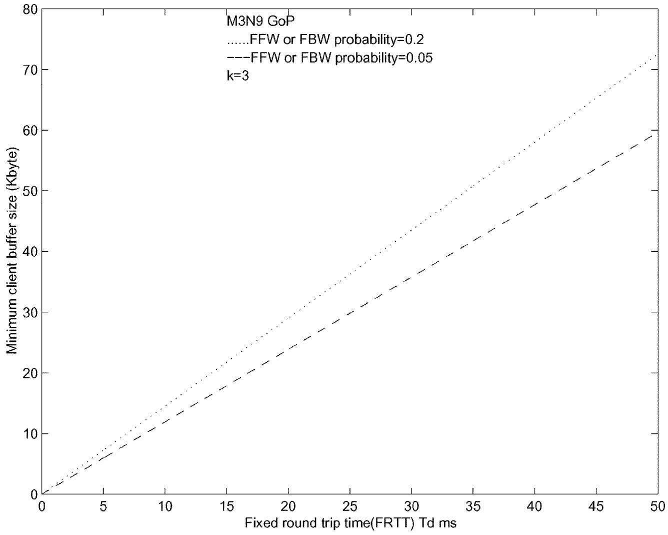

Fig. 8. Minimum client buffer size versus FRTT for M3N9 GOP.

our case, because $MCR = E[\beta_B]$, the minimum value of k' is $E[\beta_B]/E[\beta]$. The maximum value of δ is given by

$$\delta_{\max} = k - \frac{E[\beta_B]}{E[\beta]}. \tag{34}$$

For $k = 3$ and using the parameters given in Section V for a typical MPEG video, $\delta_{\max} = 2.67$.

From above discussion, the maximum queuing delays at the server and client buffers $(D_{\max})$ can be obtained by adding the maximum delay at client with the maximum delay at server. The minimum queuing delay $D_{\min}$ is the maximum value of $D^c_{\min}$ and $D^s_{\min}$

$$D_{\max} = D^c_{\max} + D^s_{\max} \tag{35}$$

$$D_{\min} = D^c_{\min} + D^s_{\min}. \tag{36}$$

From (28), the delay variation can therefore be expressed as

$$\Delta = D^c_{\max} + D^s_{\max} - (D^c_{\min} + D^s_{\min}). \tag{37}$$

We observe from (37) that to minimize the delay variation, the rate mismatch between the allocated rate and the requested rate should be minimized. The relative numerical results are given in the following section.

VII. Numerical Results

The relationship between the minimum client buffer size versus FRTT is shown in Figs. 8 and 9 corresponding to two levels of client interactivity for GOPs of M3N9 and M3N15. As expected from (19), it is seen that the the minimum client buffer size increases linearly with an increase in the FRTT. For a particular value of FRTT, if the probability of the client being in the FFW/FBW states increases, the required minimum buffer size increases. For small values of FRTT, the required minimum client buffer size does not depend on the FFW/FBW probability. On the other hand, for large values of FRTT, the required client buffer size varies significantly with the FFW/FBW probability.

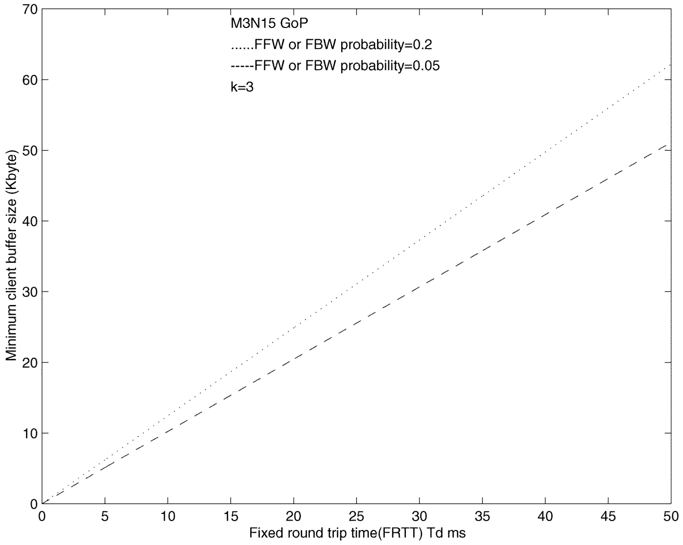

Fig. 9. Minimum client buffer size versus FRTT for M3N15 GOP.

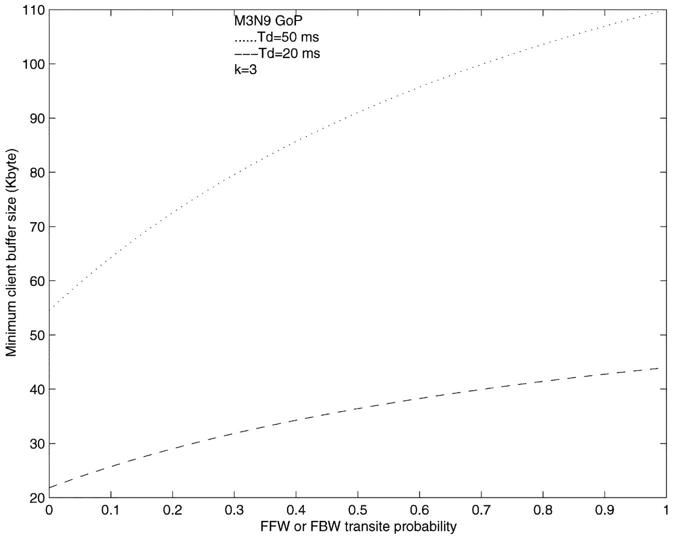

Fig. 10. Minimum client buffer size versus FFW/FBW probability for M3N9 GOP.

For both small and large values of FRTT, M3N15 GOP needs a smaller client buffer size than that required for M3N9 GOP. This is because M3N15 GOP has a relatively smaller burst compared to that of M3N9 GOP.

Figs. 10 and 11 show the minimum client buffer size as a function of the FFW/FBW probability for T_d =50 and 20 ms and for GOPs of M3N9 and M3N15, respectively. The dynamic fill level of the server buffer for different GOPs is shown in Fig. 12. It is seen that the server buffer fill level initially increases due to the high data burst coming from the I-frame. This is followed by B-frames having data rates lower than the rate at which the buffer sends data to the network. This results in a reduction of the buffer fill level. Since P-frames contain more data than B-frames, the buffer fill level goes up momentarily whenever the P-frames enter the server buffer. At the end of a GOP, the server buffer level fills to zero as required by (21). This further proves that the RM cell parameters set by (22) and (26) are correct.

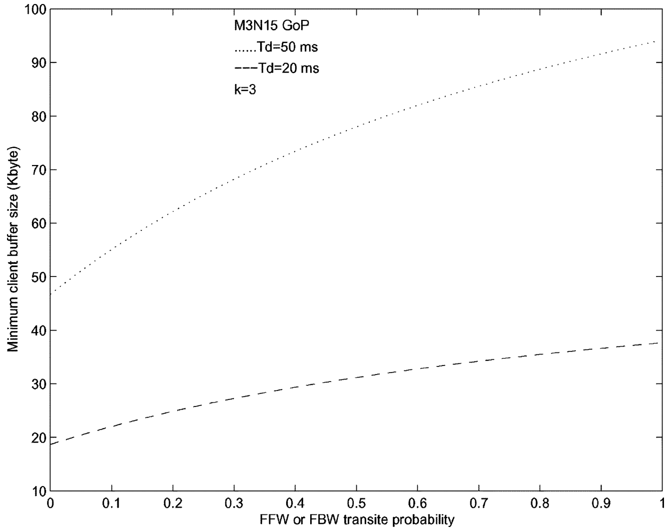

Fig. 11. Minimum client buffer size versus FFW/FBW probability for M3N15 GOP.

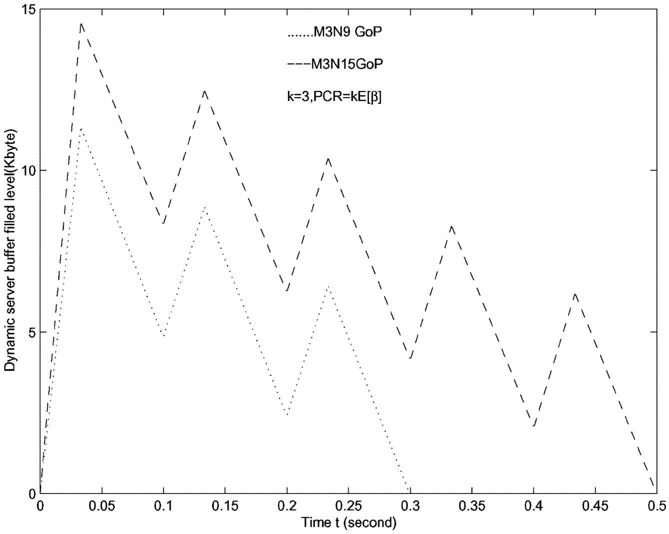

Fig. 12. Dynamic server buffer fill level.

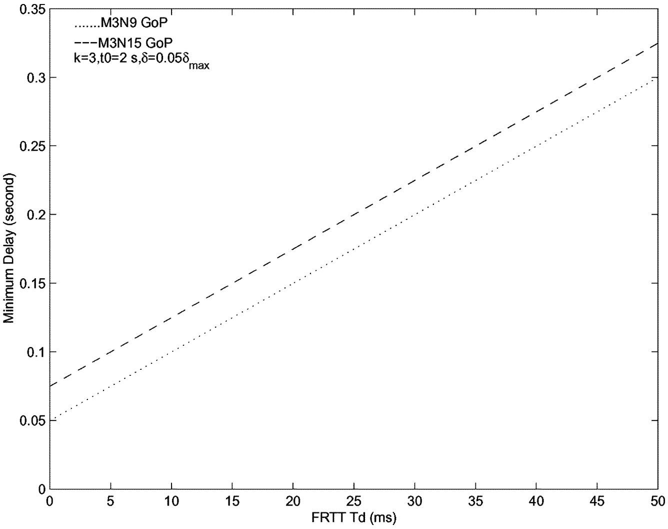

Fig. 13. Minimum delay versus FRTT.

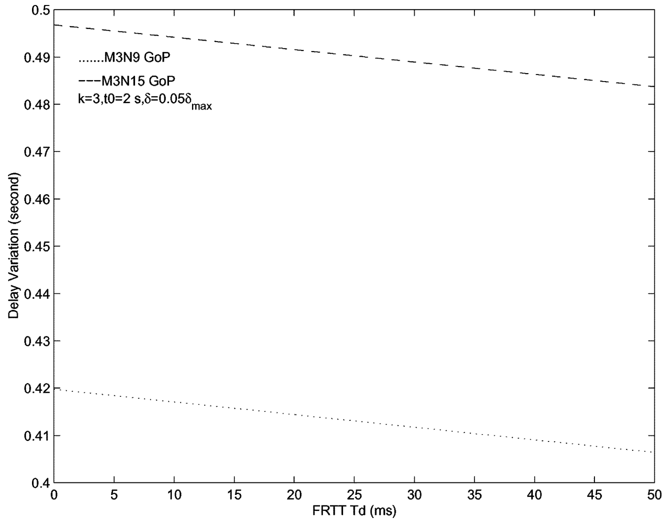

Fig. 14. Delay variation versus FRTT.

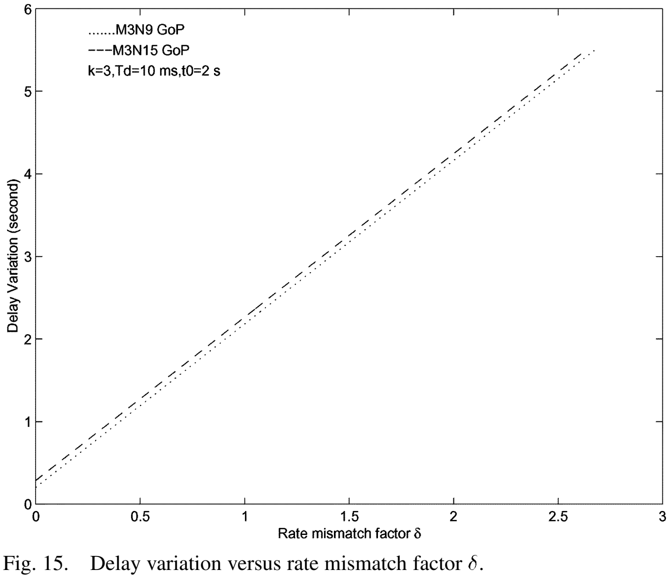

Fig. 15. Delay variation versus rate mismatch factor δ.

The minimum delay versus FRTT is shown in Fig. 13. As shown in (31), (33), and (36), the minimum delay is linearly proportional to the FRTT; the contribution from the delay in the server buffer gives a starting value of minimum delay between the server and the client. Regardless of the network size, the minimum delay increases linearly with FRTT. We also find that the value of GOP has a small effect on the minimum delay.

The delay variation versus the FRTT is shown in Fig. 14 for two values of GOPs. As shown in (30)–(34), and (37), the delay variation is decided by the difference between the duration of interaction (t_0) and the FRTT (T_d). Therefore, as T_d increases, the delay variation will decrease linearly. As shown in (30) and (31), the video data rate contributes to the delay variation from the server. Therefore, the GOP has a large effect on delay variation as seen in Fig. 14.

The delay variation versus the rate mismatch factor is given in Fig. 15. It is concluded that the delay variation increases linearly with the rate mismatch factor for a given FRTT and a FFW/FBW duration. To have a bounded delay variation in order to offer an acceptable QoS for multimedia transmission, the rate mismatch factor must be minimized and bounded.

VIII. Conclusion

We have established the design criteria and networking requirements for an interactive video on demand system carrying MPEG-2 video over the ATM ABR service. First, we proposed the client and server models for the above system. Second, by using the RTDE analysis, we have determined the values of the required connection parameters of the ABR service. Third, we have formulated analytical expressions to determine the minimum buffer size at the server and the client to prevent underflow at the client. We have also looked at the interactions between the client and the server involving various trick modes of the client operation.

We conclude that the client buffer size depends directly on the FFW/FBW probability; the higher the probability, the larger is the required size of the buffer. Our second conclusion is that the larger the GOP of an MPEG-2 stream, the lower is the required size of the client buffer. This is because the larger the value of the GOP, the less bursty is the video stream which in turn requires less buffering at the client.

The delay variation is sensitive to the rate mismatch factor. If the rate mismatch can be minimized and bounded, for a given network size and interactive operation, the delay variation will be bounded. GOP structure has no effect on delay variation and minimum delay.

We also found that there is a tradeoff between the required buffer size, the desired QoS, and the value of the GOP of an MPEG-2 video. If the aim of the system designer is to reduce the burstiness of the data and to decrease the required client buffer size, a high value of GOP should be employed. On the other hand, to improve the QoS of the video for a given network specification, a small value of GOP should be used to obtain better picture quality. Since our study is based on bandwidth constrained channels, our model and analysis techniques can be used as a framework to study interactive client server based multimedia applications over the Internet.

In this paper, we have considered ATM ABR as the transport network. This work could be extended to IP networks where the real-time streaming protocol (RTSP) (a client-server multimedia presentation control protocol, designed to address the needs for efficient delivery of streamed multimedia over IP networks) could be considered for streaming MPEG video stream over IP networks.

References

[1] D. Deloddere, W. Verbiest, and H. Verhille, "Interactive video on demand," *IEEE Commun. Mag.*, vol. 32, no. 5, pp. 82–88, May 1994.

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

[3] P. P. Mishra, "Fair bandwidth sharing for feedback controlled VBR video traffic," in *Proc. IEEE GLOBECOM*, Singapore, Nov. 1995, pp. 1102–1108.

[4] H. Kanakia, P. P. Mishra, and A. R. Reibman, "An adaptive congestion control scheme for real time packet video transport," *IEEE/ACM Trans. Netw.*, vol. 3, no. 6, pp. 671–682, Dec. 1995.

[5] H. Kanakia, P. P. Mishra, and A. Reibman, "An adaptive congestion controlscheme for real-time packet video transport," in *Proc. ACM SIGCOMM*, 1993, pp. 20–31.

[6] K. Karahara, Y. O. M. Murata, and H. Miyahara, "Performance analysis of reactive congestion control for ATM network," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 4, pp. 651–661, May 1995.

[7] M. Hamdi, J. W. Roberts, and P. Rolin, "Rate control for VBR video coders in broad-band networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 1040–1051, Aug. 1997.

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

[9] C. J. Beckman, "Dynamic bandwidth allocation for interactive video application over corporate network," in *Proc. IEEE COMPCON*, 1996, pp. 219–225.

[10] X. Li, S. Paul, and M. Ammar, "Layered video multicast with retransmission evaluation of hierarchical rate control," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar. 1998, pp. 1062–1072.

[11] F. F. N. Pereira and J. M. S. Nogueira, "Transmission of MPEG video over ATM-based networks utilizing dynamic bandwidth negotiation," presented at the IEEE MMNS, Versailles, France, Nov. 1998.

[12] Q. Ren and H. Kobayashi, "Diffusion approximation modeling for Markov modulated bursty traffic and its applications to bandwidth allocation in ATM networks," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 679–691, Jun. 1998.

[13] P. Cuenca, A. Garrido, F. Quiles, and L. Orozco-Barbosa, "Some proposals to improve error resilience in the MPEG-2 video transmission over ATM networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar. 1998, pp. 668–675.

[14] R. Coelho and S. Tohme, "A generic smoothing algorithm for real-time variable bit rate video traffic," *Comput. Netw. ISDN Syst.*, vol. 29, pp. 2053–2066, 1998.

[15] J. Zhang and J. Hui, "Applying traffic smoothing techniques for quality of service control in VBR video transmissions," *Comput. Commun.*, vol. 21, no. 4, pp. 375–389, Apr. 1998.

[16] S. Verma, R. K. Pankaj, and A. Leon-Garcia, "Call admission and resource reservation for guaranteed quality of service (GQoS) services in internet," *Comput. Commun.*, vol. 21, no. 4, pp. 362–374, Apr. 1998.

[17] V. Balabanian, L. Casey, and N. Greene, "Digital storage of media-comand and control protocol applied to ATM," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 5, pp. 1162–1172, Aug. 1996.

[18] W. Mao, "Broadband network delivery of interactive digital video using ATM," *J. VLSI Signal Process.*, vol. 17, no. 2/3, pp. 255–268, Nov. 1997.

[19] T. S. Chua, J. Li, B. C. Ooi, and K. L. Tan, "Disk striping strategies for large video-on-demand servers," in *Proc. 4th ACM Int. Multimedia Conf.*, Boston, MA, Nov. 18–20, 1996, pp. 297–306.

[20] D. Jadev, C. Srinilta, and A. Choudhary, "Batching and dynamic allocation techniques for increasing the stream capacity of an on-demand server," *Parallel Comput.*, vol. 23, no. 12, pp. 1727–1734, Dec. 1997.

[21] A. D. Gelman, S. Halfin, and W. Willinger, "On buffer requirement for store-and-forward video on demand," in *Proc. IEEE GLOBECOM*, 1991, pp. 976–980.

[22] S. Sengodan and V. O. K. Li, "A shared buffer architecture for interactive VOD servers," in *Proc. IEEE INFOCOM*, Kobe, Japan, Apr. 1997, pp. 1343–1350.

[23] E. L. Abram-Profeta and K. G. Shin, "Providing unrestricted vcr functions in multicast video-on demand servers," in *Proc. IEEE ICMCS*, 1998, pp. 66–75.

[24] L. G. Roberts, "Can ABR service replace VBR service in ATM network," in *Proc. COMPCON Conf.*, Piscatway, NJ, 1995, pp. 346–348.

[25] T. V. Lakshman, P. P. Mishra, and K. K. Ramakrishram, "Transporting compressed video over ATM networks with explicit rate feedback control," in *Proc. IEEE INFOCOM*, Kobe, Japan, Apr. 1997, pp. 38–47.

[26] N. G. Duffield, K. K. Ramakrishnan, and A. R. Reibman, "An algorithm for smoothed adaptive video over explicit rate network," *IEEE/ACM Trans. Netw.*, vol. 6, no. 6, pp. 717–728, Dec. 1998.

[27] K. B. Younes and K. Begain, "Scalable video on demand on ABR in ATM networks," in *Proc. IEEE Int. Conf. ATM*, Colmar, France, Jun. 1998, pp. 51–58.

[28] B. J. Vickers, M. Lee, and T. Suda, "Feedback control mechanism for real-time multipoint video services," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 3, pp. 512–530, Apr. 1997.

[29] B. J. Vickers and T. Suda, "An ATM service architecture for the transport of adaptively encoded live video," in *Proc. ICCCN*, Washington, D.C., Oct. 1996, pp. 179–186.

[30] W. Luo and M. El Zarki, "Quality control for VBR video over ATM networks," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 1029–1039, Aug. 1997.

[31] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to MPEG-2*. London, U.K.: Chapman & Hall, 1997.

[32] P. Pancha and M. El Zarki, "MPEG coding for variable bit rate video transmission," *IEEE Commun. Mag.*, vol. 32, no. 5, pp. 54–66, May 1994.

[33] K. S. Trivedi, *Probability and Statistics with Reliability, Queuing and Computer Science Application*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[34] "ATM Forum Traffic Management Specifications. Ver. 4.0, Tech. Rep.," ATM Forum, 1996.