Open-loop Video Distribution with Support of VCR Functionality ⋆,⋆⋆

Ernst W. Biersack

Institut Eurécom, BP 193, 06904 Sophia-Antipolis, France

Alain Jean-Marie

LIRMM, Université de Montpellier II, 161 Rue Ada, 34392 Montpellier Cedex 5, France

Philippe Nain

INRIA Sophia-Antipolis, 2004 Route des Lucioles, BP 93, 06902 Sophia-Antipolis Cedex, France

Abstract

Scalable video distribution schemes have been studied for quite some time. For very popular videos, open-loop broadcast schemes have been devised that partition each video into segments and periodically broadcast each segment on a different channel. Open-loop schemes provide excellent scalability as the number of channels required is independent of the number of clients. However, open-loop schemes typically do not support VCR functions. We will show for open-loop video distribution how, by adjusting the rate at which the segments are transmitted, one can provide VCR functionality. We consider deterministic and probabilistic support of VCR functions: depending on the segment rates chosen, the VCR functions are supported either 100% of the time or with very high probability. For the case of probabilistic support of PLAY and Fast Forward only we model the reception process as a semi-Markov accumulation process. We are able to calculate a lower bound on the probability of successfully executing Fast Forward actions.

Key words: Video on demand, interactivity, semi-Markov accumulation process, stochastic bounds.

⋆ This work is part of the DisCont project, supported by a COLOR'2002 grant of INRIA Sophia-Antipolis, France.

⋆⋆Respective e-mails of the authors: `erbi@eurecom.fr`, `ajm@lirmm.fr`, and `Philippe.Nain@sophia.inria.fr`.

1 Introduction

1.1 Classification

VoD systems can be classified in *open–loop systems* [12] and *closed–loop systems* [10,16]. In general, open loop VoD systems partition each video into smaller pieces called *segments* and transmit each segment on a separate channel at its assigned transmission rate. Those channels may be logical, implemented with an adequate multiplexing. All segments are transmitted periodically and indefinitely. The first segment is transmitted more frequently than later segments because it is needed first in the playback. In open–loop systems there is no feedback from the client to the server, and transmission is completely one–way. In closed-loop systems, on the other hand, there is a feedback between the client and the server. Closed–loop systems generally open a new unicast/multicast stream each time a client or a group of clients issues a request. To make better use of the server and network resources, client requests are batched and served together with the same multicast stream.

Open-loop systems use segmentation in order to reduce the network bandwidth requirements, which makes them highly scalable because they can provide Near Video on Demand (NVoD) services at a fixed cost *independent* of the number of users. In this paper, we will show how open-loop NVoD schemes can support VCR functions, which are defined as follows:

PLAY Play the video at the basic video consumption rate, b;
PAUSE Pause the playback of the video for some period of time;
SF/SB Slow forward/Slow backward: Playback the video at a rate equal to $Y_S \times b$ for some period of time. We have $Y_S < 1$;
FF/FB Fast forward/Fast backward: Playback the video at a rate equal to $X_F \times b$ for some period of time. We have $X_F > 1$.

1.2 Related Work

Most VoD systems do not support VCR functions. It is assumed that users are passive and keep playing the video from the beginning until the end without issuing any VCR function. However support of VCR functions makes a VoD service much more attractive. Most research on interactive VoD focuses on closed-loop schemes [1,6,15,13]. To support VCR functions such as Fast-Forward (FF), all these schemes serve the client who issues a FF command via *a dedicated unicast transmission*, referred to as contingency channel. When the client returns into PLAY state, (s)he joins again the multicast distribution. It is obvious that such a solution is not very scalable since it requires sepa-

rate contingency channels and also explicit interaction with the central server. Thus, open-loop schemes are particularly well suited when: a) the number of users grows large, or b) the communication medium has no feedback channel, which is the case in satellite or cable broadcast systems.

Very little work has been done to support VCR functions in open-loop VoD schemes [2,4,8,14]. Except for the paper by Fei *et al.* [8], all the other schemes only consider PAUSE or discrete jumps in the video. Fei *et al.* propose a scheme called "staggered broadcast" and show how it can be used together with what they call "active buffer" management to provide limited interactivity. In staggered broadcast, the whole video of duration L is periodically transmitted on N channels at the video consumption rate b. Transmission of the video on channel i starts $t_s = L/N$ time units later than channel $i-1$. Depending on the buffer content and the duration of the VCR action, the VCR action may be possible or not. In the case that the VCR action is not possible, it is approximated by a so-called discontinuous interactive function where the viewing jumps to the closest (with respect to the intended destination of the interaction) point of the video that allows the continuous playout after the VCR action has been executed.

The big difference between the related work and our scheme is that up to now, the support of VCR functions either required a major extension of the transmission scheme (e.g. contingency channels) or was very restricted (e.g. staggered broadcast). We will demonstrate the feasibility of deterministic support of VCR functions in open-loop VoD systems by *increasing the transmission rate* of the different segments. While this idea looks very straightforward, it has been, to the best of our knowledge, never proposed before.

The rest of the paper is organized as follows. We first describe the so-called tailored transmission scheme, then discuss how to adapt this scheme to support VCR functions. For the case of PLAY and FF user interactions we develop an analytical model that allows the computation of a lower bound on the probability that a user interaction can be successfully executed and then provide some quantitative results. The paper ends with a brief conclusion.

2 Open-loop NVoD

2.1 Introduction

Many different open-loop NVoD schemes have been proposed in the literature; for a survey see [12]. These schemes typically differ in the way a video is partitioned into segments and can be classified mainly in three categories:

- Schemes that partition the video in different length segments and transmit each segment at the basic video consumption rate [9,19];
- Schemes that partition the video in equal-length segments and decrease the transmission rate of each segment with increasing segment number [3];
- Hybrid schemes that combine the two above methods [14,20].

In the following, we will present in more detail the scheme called *tailored transmission* scheme that was proposed by Birk and Mondri [3] and is a generalization of many of the other open-loop NVoD schemes previously described.

2.2 Tailored Transmission Scheme

The base version of the tailored transmission scheme works as follows. A video is partitioned into N equal-length segments. Each segment is transmitted periodically and repeatedly on its own channel. A client who wants to receive a video starts by listening to one, more, or all channels and records these segments.

We shall need the following notation:

- s_i denote the time the client starts recording segment i;
- w_i denote the time the client has entirely received segment i;
- v_i denote the time the client starts viewing segment i;
- r_i denote the transmission rate of segment i [bits/sec];
- D denote the segment size [bits];
- b denote the video consumption rate [bits/sec].

To assure the *continuous* playout of the video we require that each segment is fully received before its playout starts, i.e. $v_i \geq w_i$. Given a segment size D and $v_i - s_i \geq w_i - s_i = D/r_i$, the transmission rate r_i of segment i must satisfy the following condition to assure a continuous playout of the video:

$$r_i \geq \frac{D}{(v_i - s_i)} \ . \tag{1}$$

If the client starts recording *all* segments at the same time, i.e. $s_i = t_0$, Birk and Mondri have shown (Lemma 1 in [3]) that the transmission rate will be *minimal* and is given as

$$r_i^{min} = \frac{D}{(w_i - t_0)} \ . \tag{2}$$

Without loss of generality, we may assume that $t_0 = 0$, which implies $w_i =$

$i \times D/b$, where D/b is the *duration* of a segment. Then, $r_i^{min} = b/i$, and the total server transmission bandwidth is

$$R_t^{min} = b \sum_{i=1}^{N} \frac{1}{i} \sim b \times \ln(N) \ .$$

Figure 1 illustrates the tailored transmission scheme for the case of minimal transmission rates. The client starts receiving all segments at time t_0. The shaded areas for each segment contain exactly the content of that segment as received by the client who started recording at time t_0. A client is not expected to arrive at the starting point of a segment; instead a client begins recording at whatever point (s)he arrives at, and stores the data for later consumption. Therefore, the startup latency of is the scheme corresponds to the segment duration D/b.

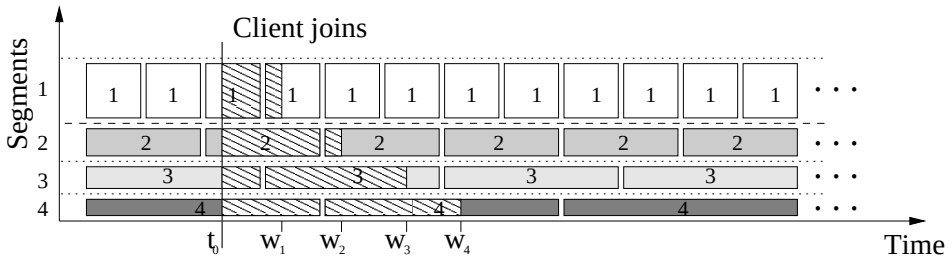

Fig. 1. An example of the tailored transmission scheme with minimal transmission rates.

3 How to Support VCR Functions

Given the base scheme of the tailored transmission with its minimal transmission rates, we will show how to adapt (increase) the segment transmission rates to support VCR functions. To convey the main idea, we will limit ourselves first to the case where the only VCR function possible is FF. In fact, the FF is the only VCR action that "accelerates" the consumption of the video, which possibly can lead to a situation where the consumption of the video gets ahead of the reception of the video. We present a solution that makes sure that any FF command issued can be successfully executed. The other user interactions such as SF, SB, FB or PAUSE can be accommodated by buffering at the client side. From now on, we therefore consider only two states: PLAY and FF.

We make the following two central assumptions:

- The client has enough disk storage to buffer the contents of a large portion of the video;

- The client has enough network access and disk I/O bandwidth to start receiving the N segments at the same time.

The trend for terminal equipments appears to be that more and more storage capacity is available. Actually, there already exist products that meet the above assumptions. An example is the digital video recorder by TiVo [18] that can store up to 60 hours of MPEG II video and, connected to a satellite feed, can receive transmissions at high data-rates.

However, for the case where the assumption on storage does not hold, we also know how to support VCR functions: the idea will remain the same, only the individual segment transmission rates required will be higher. The scheme we propose may be adapted to this situation. Note that the trade-off between the storage capability of the client and segment transmission rates for the case of NVoD has already been explored by Birk and Mondri [3].

3.1 Deterministic Support

Whenever a client issues a FF command, the video is viewed at a playout rate X_F faster than the normal rate, i.e. the consumption of the video occurs at a rate equal to $X_F \times b$ and each segment will be consumed after $\frac{D}{b \times X_F}$ units of time instead of D/b units of time in case of PLAY. As a consequence, the viewing times of all segments not yet viewed will be "advanced" in time. To obtain a *deterministic* guarantee that every FF command issued during the viewing of a video can be executed, we consider the *worst case scenario* where the client views the whole video in FF.
Let v_i^{FF} denote the time the client starts viewing segment i, given that (s)he has viewed segments $1, 2, \ldots, i-1$ in FF mode. We can compute the v_i^{FF} as follows [1]

$$v_i^{FF} = v_1 + \frac{i-1}{X_F} \times \frac{D}{b} \ .$$

If the client starts recording all segments at the same time, i.e. $s_i = t_0$, we can compute, similar to (2), the transmission rate r_i^{FF} that allows unrestricted FF interactions as

$$r_i^{FF} = \frac{D}{(v_i^{FF} - t_0)} = \frac{D}{(v_1 + \frac{i-1}{X_F} \times \frac{D}{b} - t_0)} \ .$$

[1] The playout and therefore VCR actions do not start before segment 1 has been entirely received; we therefore have $v_1^{FF} = v_1$.

If we assume that $t_0 = 0$, which implies $v_1 = D/b$, the expression simplifies to

$$r_i^{FF} = \frac{bX_F}{X_F + i - 1} \ . \tag{3}$$

3.2 Probabilistic Support for FF

In the previous subsection, we have computed the minimal transmission rates r_i^{FF} such that all the FF interactions issued can be realized. We have considered the worst case scenario where the client views the whole video in FF mode. While a client might do so, we think that it is much more likely that the viewing of a video will alternate between PLAY and FF modes (and possibly other VCR actions). We will in the following use a model for the viewing behavior where a user strictly alternates between PLAY and FF. We refer to this behavior as **S-FF** (for Simple FF).

Our goal is to support FF interactions with *high probability* while transmitting each segment at a rate lower than r_i^{FF}. To this purpose we define the rates r_i^I as follows:

- The server transmits the segments $i \in \{2, \ldots, N\}$ at a rate $r_i^I = A r_i^{min}$, where A is the **rate increase factor**, with $1 \leq A \leq X_F$, and r_i^{min} is computed in (2);
- The server transmits segment 1 at rate $r_1^I = r_1^{min}$, still because the playout does not start before segment 1 has been entirely received.

4 Analytical Model for the S-FF Model

In this section we will compute a closed-form lower-bound on the probability that a segment is *successfully consumed* by the client. Segment i is successfully consumed by the client if segment $i + 1$ is available to him/her before the consumption of segment i has been completed; otherwise we will say that the consumption of segment i has *failed*. A failure is resolved once the next segment is entirely available to the client. It is worth pointing out that failures may occur both in mode PLAY and in mode FF, as shown on Figure 2.

We will assume that the client alternates between both modes of consumption. More precisely, we introduce two independent renewal sequences of rvs $\{S_P(n)\}_n$ and the $\{S_{FF}(n)\}_n$, where $S_P(n)$ and $S_{FF}(n)$ will represent the duration of the n-th PLAY and FF periods, respectively.

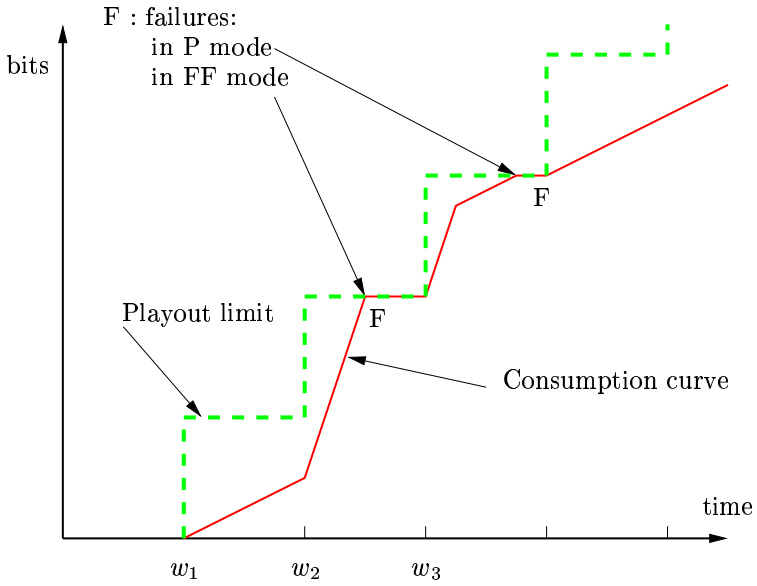

Fig. 2. Failures occurring in PLAY and FF modes.

For modeling purposes, and also because we believe this assumption corresponds to a reasonable behavior of the client, we will assume that the remaining duration of a PLAY or FF period when a failure occurs is *resumed* when the next segment is available to the client. This corresponds, for instance, to the situation where the client wants to reach a particular point in the video or avoid a particular scene, regardless of the failures that (s)he may encounter while viewing the video.

In order to ensure a probabilistic support for FF (cf. Section 3.2) recall that segment i is transmitted at rate $r_i^I = Ar_i^{\min} = Ab/i$ [bits/sec]. Therefore, the i-th segment will be entirely available to the client at time $w_i = iD/Ab$. The continuous playout of segment i requires that at the viewing time v_i, this segment has been entirely received, that is $v_i \geq w_i$. Segment $i = 2, 3, \ldots, N$ will fail if this inequality does not hold. The continuous playout of the video requires that all segments be on time, namely,

$$v_i \geq w_i \quad i = 2, 3, \ldots, N.$$

Recall that $v_1 = w_1$ since the client cannot start viewing the first segment before it has been entirely received.

The number L of segments on time is given by

$$L = \sum_{i=1}^{N} \mathbf{1}_{\{v_i \geq w_i\}}$$

where $\mathbf{1}_A$ stands for the indicator function of the event A, from which we

deduce the mean number of segments on time

$$\mathbb{E}[L] = \sum_{i=1}^{N} \mathbb{P}(v_i \geq w_i). \tag{4}$$

Denote by $R(t)$ the number of bits of the video which have been consumed by the client in $[v_1, v_1 + t)$. Clearly,

$$v_i = \inf\{t > 0 \,:\, R(t) \geq (i-1)D\} \quad i = 2, 3, \ldots, N. \tag{5}$$

Computing $\mathbb{P}(v_i \geq w_i)$ in closed-form for all i is not an easy task. Indeed, it is related to computing the distribution of the length of a busy period in a fluid queue fed by a Markov-Modulated Rate Process. In the present paper, we will content ourselves with the derivation of an elementary lower bound.

To derive this lower bound, we consider the *semi-Markov accumulation process* $\{Q(t), t > 0\}$ which is constructed as follows: during a PLAY period $Q(t)$ *continuously* increases with the rate b and during a FF period it *continuously* increases with the rate bX_F. More precisely, for $t > 0$,

$$\begin{aligned} Q(t) = &\sum_{n\geq 0} \mathbf{1}_{\{T_n < t \leq T_n + S_P(n+1)\}} \left[b(t - Z_{FF}(n)) + bX_F Z_{FF}(n)\right] \\ &+ \sum_{n\geq 0} \mathbf{1}_{\{T_n + S_P(n+1) < t \leq T_{n+1}\}} \left[bZ_P(n+1) + bX_F(t - Z_P(n+1))\right] \end{aligned}$$

with $T_n := \sum_{i=1}^{n}(S_P(i)+S_{FF}(i))$, $Z_P(n) := \sum_{i=1}^{n} S_P(i)$, $Z_{FF}(n) := \sum_{i=1}^{n} S_{FF}(i)$. By convention $T_0 = Z_P(0) = Z_{FF}(0) = 0$.

By construction of $Q(t)$ and $R(t)$ it is obvious that (see Figure 3)

$$R(t) \;\leq\; Q(t) \quad t \;\geq\; w_1. \tag{6}$$

Observe that both processes $\{R(t), t \geq w_1\}$ and $\{Q(t), t \geq w_1\}$ would be identical in the absence of failures. We see from (6) and the definition (5) that $v_i \geq w_i$ will hold if $Q(w_i) < (i-1)D$, which implies that

$$\mathbb{P}(v_i \geq w_i) \geq \mathbb{P}(Q(w_i) < (i-1)D).$$

Hence, cf. (4),

$$\mathbb{E}[L] \geq \sum_{i=1}^{N} \mathbb{P}(Q(w_i) < (i-1)D)). \tag{7}$$

For the transmission scheme we described in Section 3.2, the segment arrival times are given by $w_i = iD/Ab$ for $i \geq 2$, but the analysis above actually holds for any *reception schedule* of segments given by a sequence $\{w_i; i \geq 1\}$.

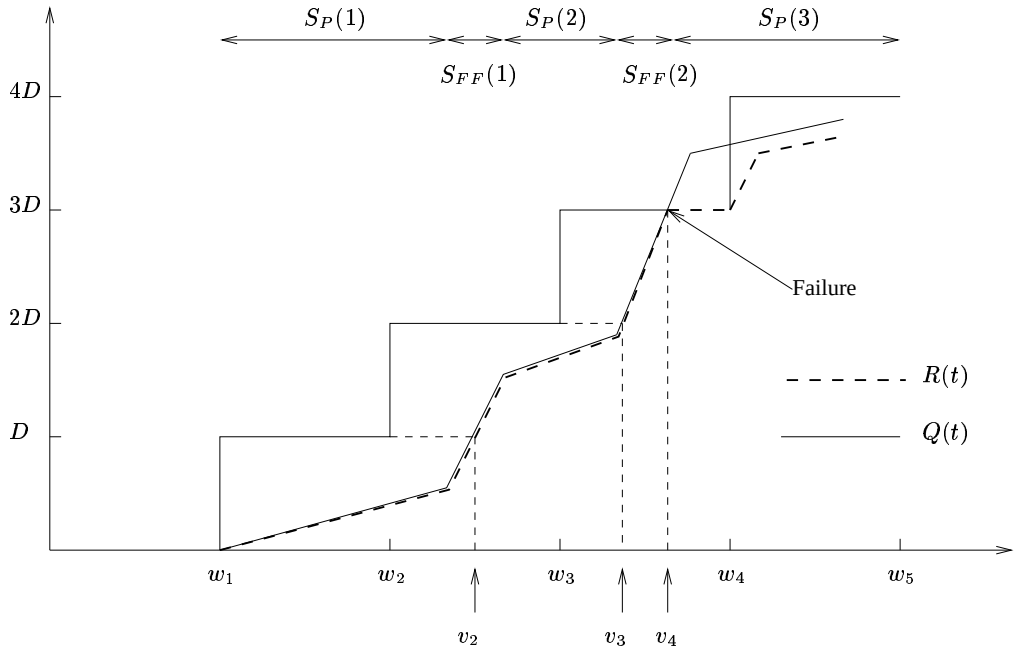

Fig. 3. Comparison of $Q(t)$ and $R(t)$.

In Section 5, we present results for determining $\mathbb{P}(Q(T) < x)$, for any T and x. These results are actually obtained for any semi-Markov accumulation process with a finite state-space (see Section 5.2). When $S_P(n)$ and $S_{FF}(n)$ are exponentially distributed random variables (rvs) with respective means $1/\tau_P$ and $1/\tau_{FF}$, we can apply the formulas in Section 5.3. First, use (16) with $r_1 = b$ and $r_2 = bX_F$. Then, use the formulas for $q_{ij}(x)$ (the density of the distribution of Q, conditionally on the start/end states) with $\alpha = \tau_P$ and $\beta = \tau_{FF}$. The probabilities $\mathbb{P}(Q(T) < x)$ are then obtained by numerical integration.

5 Semi-Markov Accumulation Process

In this section, we develop a framework for evaluating the workload distribution generated in a given time-interval by a semi-Markov accumulation process with an arbitrary (but finite) state-space.

After defining the process (Section 5.1), we show that the Laplace transforms of the sought distributions satisfy the linear system of equations (15). Finally, we apply the formula to the case of a two-state continuous-time Markov process (Section 5.3), where the Laplace transform can be inverted to obtain the density of the distribution.

We first construct formally the accumulation process from a semi-Markov process. Let $\mathcal{E} = \{1, 2, \ldots, K\}$ be a finite state-space. Let

- $\{S_i(n)\}_n$ be a sequence of i.i.d. rvs, for each $i \in \mathcal{E}$;
- $\{Z(n)\}_n$ be a homogeneous discrete-time Markov chain on the state-space $\mathcal{E}$, with one-step transition matrix $\mathbf{P} = (p_{ij})_{i,j\in\mathcal{E}}$.

The semi-Markov process $\{X(t), t \geq 0\}$ is defined jointly with a sequence $\{T_n\}_n$ of jump times as

$$\begin{aligned} T_{n+1} &= T_n \ + \ S_{Z(n+1)}(n) \\ X(t) &= Z(n+1) \qquad T_n \leq t < T_{n+1} \\ X(t) &= Z(0) \qquad 0 \leq t < T_0 \end{aligned}$$

with T_0 some nonnegative rv.

The accumulation process $Q(t)$ is such that while the process $X(t)$ is in state i, $Q(t)$ accumulates at a constant rate r_i. Formally, $\{Q(t), t \geq 0\}$ is constructed as follows: set $Q(0) = 0$ and let

$$\begin{aligned} Q(t) &= Q(T_n) + r_{Z(n+1)}\ (t - T_n) \qquad && T_n \leq t < T_{n+1} \\ Q(t) &= r_{Z(0)}\ t && 0 \leq t < T_0\ . \end{aligned} \tag{8}$$

This construction is illustrated in Figure 4. The upper part shows the evolution of the discrete-time Markov chain $Z(n)$, and of $X(t)$. The lower part displays $Q(t)$ as a function of the jump times T_n. The accumulation rates are such that $0 = r_3 < r_1 < r_2$.

5.2 Distribution of $Q(t)$

Let $Q^{i;r}(T)$ denote the quantity accumulated in $[0, T)$ given that $X(0) = i$ and $T(0) = r$. In other words, the process X starts in state i with a residual sojourn time r in this state. Similarly, denote $Q^{i;S_i}(T)$ the same quantity, but given that $X(0) = i$ with a residual time $T(0)$ distributed according to the total sojourn time distribution (i.e., as if a transition to state i had occurred at time 0).

Depending on the problem to be solved, one may be interested in the distribution of $Q^{i;S_i}(T)$ or that of $Q^{i;\widetilde{S}_i}(T)$, where $\widetilde{S}_i$ is the forward recurrence time of S_i. The latter corresponds to the case where the semi-Markov process $\{X(t)\}$

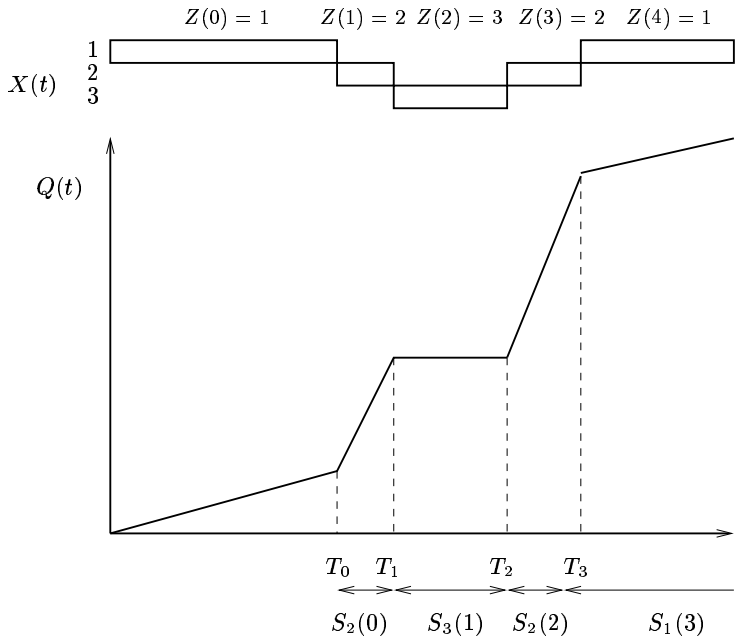

Fig. 4. Construction of the accumulation process

is stationary. The common procedure for computing these distributions is to compute that of $Q^{i,r}(T)$ for an arbitrary r, and then integrate with respect to the proper distribution.

We are therefore interested in the distribution of $Q^{i;r}(T)$, jointly with that of $X(T)$, namely $\mathbb{P}(Q^{i;r}(T) \leq x, X(T) = j)$. We shall actually compute the Laplace-Stieltjes Transform (LST)

$$\hat{Q}_T^{i,j;r}(\nu) \;=\; \mathbb{E}\Big[e^{-\nu Q^{i;r}(T)}\mathbf{1}_{\{X(T)=j\}}\Big] \;=\; \int_0^\infty e^{-\nu x}\mathrm{d}\mathbb{P}(Q^{i;r}(T) \leq x, X(T) = j)\ . \tag{9}$$

The computation below may be seen as a generalization of the analysis developed by Cox and Miller in [5, §9.3] for alternating renewal processes (i.e. $K = 2$).

First, if $r \geq T$, then no jump occurs before time T, and since $X(0) = i$, then $X(T) = i$ and $Q(T) = r_i\ T$. In that case,

$$\hat{Q}_T^{i,j;r}(\nu) \;=\; \mathbb{E}\Big[e^{-\nu r_i T}\mathbf{1}_{\{i=j\}}\Big] \;=\; e^{-Tr_i\nu}\ \delta_{i,j} \tag{10}$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

On the other hand, if $r < T$, then at least one jump occurs in the time-interval $[0, T)$, and conditioning on the state reached after the first jump (i.e. $Z(1)$) then using the stationarity and independence of the underlying sequences, we

have

$$\hat{Q}_T^{i,j;r}(\nu) \;=\; \sum_{k\in\mathcal{E}} p_{ik} e^{-r\ r_i\nu}\ \mathbb{E}\left[e^{-\nu Q^{k;S_k}(T-r)}\mathbf{1}_{\{X(T-r)=j\}}\right]\ . \tag{11}$$

We now compute the Laplace transform of $\hat{Q}_T^{i,j;r}(\nu)$ with respect to T. With the help of (10)-(11), we obtain

$$\begin{aligned}
&\int_0^\infty e^{-\mu T}\hat{Q}_T^{i,j;r}(\nu)\ \mathrm{d}T \\
&= \int_0^r e^{-\mu T}e^{-Tr_i\nu}\ \delta_{i,j}\ \mathrm{d}T \\
&\quad + \int_r^\infty e^{-\mu T}\sum_{k\in\mathcal{E}} p_{ik}e^{-rr_i\nu}\ \mathbb{E}\left[e^{-\nu Q^{k;S_k}(T-r)}\mathbf{1}_{\{X(T-r)=j\}}\right]\ \mathrm{d}T \\
&= \frac{1-e^{-(\mu+r_i\nu)r}}{\mu+r_i\nu}\ \delta_{i,j} \;+\; \sum_{k\in\mathcal{E}} p_{ik}\int_0^\infty e^{-\mu(r+T)}e^{-rr_i\nu}\ \mathbb{E}\left[e^{-\nu Q^{k;S_k}(T)}\mathbf{1}_{\{X(T)=j\}}\right]\ \mathrm{d}T \\
&= \frac{1-e^{-(\mu+r_i\nu)r}}{\mu+r_i\nu}\ \delta_{i,j} \;+\; e^{-(\mu+r_i\nu)r}\sum_{k\in\mathcal{E}} p_{ik}\int_0^\infty e^{-\mu T}\ \mathbb{E}\left[e^{-\nu Q^{k;S_k}(T)}\mathbf{1}_{\{X(T)=j\}}\right]\ \mathrm{d}T\ .
\end{aligned} \tag{12}$$

A relation involving only the rvs $Q^{i;S_i}(T)$ is obtained from (12) by integrating both sides with respect to r, considered to be distributed as S_i. Let $S_i(r)$ denote the distribution function of S_i and let $S_i^*(s) = \mathbb{E}\left[e^{-sS_i}\right]$ be its LST. Introduce also the notation

$$\begin{aligned}
K_{i,j}(\mu,\nu) &= \int_0^\infty e^{-\mu T}\ \mathbb{E}\left[e^{-\nu Q^{i;S_i}(T)}\mathbf{1}_{\{X(T)=j\}}\right]\ \mathrm{d}T \\
&= \int_0^\infty e^{-\mu T}\ \int_0^\infty e^{-\nu q}\ \mathrm{d}\mathbb{P}(Q^{i;S_i}(T)\le q, X(T)=j)\ \mathrm{d}T\ .
\end{aligned} \tag{13}$$

Then, we have

$$\begin{aligned}
&K_{i,j}(\mu,\nu) \\
&= \int_0^\infty\int_0^\infty e^{-\mu T}\ \hat{Q}_T^{i,j;r}(\nu)\ \mathrm{d}T\ \mathrm{d}S_i(r) \\
&= \int_0^\infty \frac{1-e^{-(\mu+r_i\nu)r}}{\mu+r_i\nu}\ \delta_{i,j}\ \mathrm{d}S_i(r) \\
&\quad + \int_0^\infty e^{-(\mu+r_i\nu)r}\ \mathrm{d}S_i(r)\ \sum_{k\in\mathcal{E}} p_{ik}\int_0^\infty e^{-\mu T}\ \mathbb{E}\left[e^{-\nu Q^{k;S_k}(T)}\mathbf{1}_{\{X(T)=j\}}\right]\ \mathrm{d}T \\
&= \frac{1-S_i^*(\mu+r_i\nu)}{\mu+r_i\nu}\ \delta_{i,j} \;+\; S_i^*(\mu+r_i\nu)\ \sum_{k\in\mathcal{E}} p_{ik}K_{k,j}(\mu,\nu)\ .
\end{aligned} \tag{14}$$

This is a system of linear equations from which the required Laplace transforms can be computed. To see this better, define the matrices

$$\mathbf{K} \;=\; (K_{i,j}(\mu,\nu))_{(i,j)\in\mathcal{E}\times\mathcal{E}} \qquad \mathbf{S} \;=\; \mathrm{diag}\,(S_i^*(\mu+r_i\nu))_{i\in\mathcal{E}}$$
$$\mathbf{L} = \mathrm{diag}\left(\frac{1}{\mu+r_i\nu}\right)_{i\in\mathcal{E}}$$

where $\mathrm{diag}(a_i)_{1\le i\le m}$ denotes the $m\times m$ diagonal matrix with elements $a_1,\ldots,a_m$. Then, (14) rewrites as

$$\begin{aligned}\mathbf{K} &= \mathbf{L}\;(\mathbf{I}-\mathbf{S}) \;+\; \mathbf{SPK}\\ \mathbf{K} &= (\mathbf{I}-\mathbf{SP})^{-1}\;\mathbf{L}\;(\mathbf{I}-\mathbf{S})\;. \end{aligned} \tag{15}$$

The matrix $\mathbf{I}-\mathbf{SP}$ is invertible because the spectral radius of $\mathbf{SP}$ is less than 1 when $\Re(\mu+r_i\nu)>0$ for all i. This follows by application of a standard bound on the spectral radius ([11, Cor. 6.1.5]): $\rho(\mathbf{SP}) \le \max_i \sum_j |(\mathbf{SP})_{ij}| = \max_i |S_i^*(\mu+r_i\nu)|$. This is less than one in the specified domain, from well known properties of Laplace transforms.

Once the matrix $\mathbf{K}$ is computed, other initial conditions of the process $\{X(t),\ t\ge 0\}$ may be investigated.

- For instance, if the residual sojourn time in state i is r, then the distribution is obtained using (12), that is

$$\int_0^\infty e^{-\mu T}\hat{Q}_T^{i,j;r}(\nu)\;\mathrm{d}T \;=\; \frac{1-e^{-(\mu+r_i\nu)r}}{\mu+r_i\nu}\;\delta_{i,j} \;+\; e^{-(\mu+r_i\nu)r}\sum_{k\in\mathcal{E}} p_{ik}K_{k,j}(\mu,\nu)\;;$$

- If the residual sojourn time in state i is given by $\widetilde{S}_i$, the forward recurrence time of S_i (in other words, if $\{X(t), t\ge 0\}$ is stationary), then integrating (12) gives, with obvious notation

$$\widetilde{K}_{i,j}(\mu,\nu) = \int_0^\infty e^{-\mu T}\;\mathbb{E}\left[e^{-\nu Q^{i;\widetilde{S}_i}(T)}\mathbf{1}_{\{X(T)=j\}}\right]\;\mathrm{d}T$$
$$\widetilde{\mathbf{K}} = \mathbf{L}\;(\mathbf{I}-\widetilde{\mathbf{S}}) \;+\; \widetilde{\mathbf{S}}\mathbf{PK}\;.$$

Remark: A simple extension of this derivation shows that the accumulation process may be generalized by replacing the constant-rate process by any stationary process with independent increments. The formulas above hold with the term "$r_i\nu$" replaced by some $\phi_i(\nu)$ characteristic of the process (see [7, Eq. (7.3') p. 419]). For instance, for the Poisson process with rate r, $\phi(\nu) = r(1-e^{-\nu})$, whereas for a diffusion process with drift r and variance σ^2, $\phi(\nu) = r\nu+\frac{1}{2}\sigma^2\nu^2$.

In this section, we address the case of a two-state, continuous-time Markov process, with infinitesimal generator

$$\mathbf{Q} \;=\; \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \;.$$

Let $Q_{r_1,r_2}(T)$ denote the quantity accumulated during the interval $[0,T)$ when accumulation rates in states 1 and 2 are r_1 and r_2, respectively. In distribution, we have

$$Q_{r_1,r_2}(T) \;=\; r_1\, Q_{1,0}(T) \;+\; r_2\, (T - Q_{1,0}(T)) \;=\; r_2\, T \;+\; (r_1 - r_2)\, Q_{1,0}(T)\;. \tag{16}$$

Computing the distribution of $Q_{r_1,r_2}(T)$ is therefore reduced to computing the distribution of $Q_{1,0}(T)$, which is the visit time in state 1 during the interval $[0,T)$. We therefore take $r_1 = 1$ and $r_2 = 0$ and apply formulas of Section 5.3. We assume that the residual time in the initial state has the same distribution as the total sojourn time. Observe that due to the memoryless property of the exponential distribution, S_i and $\widetilde{S}_i$ have the same distribution. The relevant matrices are:

$$\mathbf{L}^{-1} \;=\; \begin{pmatrix} \mu+\nu & 0 \\ 0 & \mu \end{pmatrix}\;, \qquad \mathbf{P} \;=\; \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\;, \qquad \mathbf{S} \;=\; \begin{pmatrix} \dfrac{\alpha}{\alpha+\mu+\nu} & 0 \\ 0 & \dfrac{\beta}{\beta+\mu} \end{pmatrix}\;.$$

Using (15), we obtain

$$\mathbf{K} \;=\; \frac{1}{(\mu+\beta)(\mu+\nu+\alpha) - \alpha\beta} \begin{pmatrix} \mu+\beta & \alpha \\ \beta & \mu+\nu+\alpha \end{pmatrix}\;.$$

The last step is to invert the Laplace transform $K_{ij}(\mu,\nu)$ with respect to ν and μ. From the definition (13), this will give the density of the distribution of $Q(T)$.

The inversion can be performed using general rules and tables for Laplace transforms (see e.g. [17]). Inverting with respect to ν is straightforward, because we have a rational function of degree 1 in ν. We obtain:

$$\int_0^\infty e^{-\mu T}\; \frac{\mathrm{d}}{\mathrm{d}q}\mathbb{P}(Q^{i;S_i}(T) \le q, X(T) = j)\; \mathrm{d}T = e^{-\alpha q}\; e^{-\mu q}\; \exp\left(\frac{\alpha\beta q}{\beta+\mu}\right)\;.$$

For the inversion with respect to μ, we use in particular the following properties:

$$\begin{aligned}
\mathcal{L}^{-1}(e^{-\mu x} f(\mu), \mu; T) &= \mathcal{L}^{-1}(f(\mu), \mu; T - x)\ \mathbf{1}_{\{T \geq x\}} \\
\mathcal{L}^{-1}(f(\mu + \beta), \mu; T) &= \mathcal{L}^{-1}(f(\mu), \mu; T)\ e^{-\beta T} \\
\mathcal{L}^{-1}(e^{a\mu}, \mu; T) &= \left(\frac{a}{T}\right)^{1/2}\ I_1\left(2\sqrt{aT}\right)\ +\ \delta_0(T) \\
\mathcal{L}^{-1}(e^{a\mu}/\mu, \mu; T) &= I_0\left(2\sqrt{aT}\right) \\
\mathcal{L}^{-1}(e^{a\mu}/\mu^2, \mu; T) &= \left(\frac{T}{a}\right)^{1/2}\ I_1\left(2\sqrt{aT}\right)
\end{aligned}$$

where $\mathcal{L}^{-1}(g(s), s; t) = G(t)$ if $g(s) = \int_0^\infty \exp(-st) G(t) \mathrm{d}t$ (i.e. inverse of the Laplace transform $g(s)$ at point t), $I_n(\cdot)$ is the modified Bessel function of the first kind and order n (see e.g. [17, p. 7]) and $\delta_a(t)$ is the Dirac function at point a.

Define

$$q_{ij}(x)\ =\ \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(Q(T) \leq x, X(T) = j | X(0) = i)$$

for $x \geq 0$. We finally find, with $f(x) = 2\sqrt{\alpha\beta x(T-x)}$:

$$\begin{aligned}
q_{11}(x) &= e^{-\alpha T}\delta_T(x)\ +\ e^{-\alpha x}\ e^{-\beta(T-x)}\ I_1(f(x))\ \sqrt{\frac{\alpha\beta x}{T-x}}\ \mathbf{1}_{\{T \geq x\}} \\
q_{12}(x) &= \alpha e^{-\alpha x}\ e^{-\beta(T-x)}\ I_0(f(x)) \\
q_{21}(x) &= \beta e^{-\alpha x}\ e^{-\beta(T-x)}\ I_0(f(x)) \\
q_{22}(x) &= e^{-\beta T}\delta_o(x)\ +\ e^{-\alpha x}\ e^{-\beta(T-x)}\ I_1(f(x))\ \sqrt{\frac{\alpha\beta(T-x)}{x}}\ \mathbf{1}_{\{T \geq x\}}\ .
\end{aligned}$$

In order to obtain the distribution functions $\mathbb{P}(Q(T) \leq x, X(T) = j | X(0) = i)$, the Laplace transforms $K_{ij}(\mu, \nu)/\nu$ should be inverted. This leads to more involved series which shall not be reproduced here.

6 Numerical Results

We have applied the bound in (7) to a video of length 2 hours = 7200 sec. We have varied the segment size from 200 sec to 800 sec. The number N of segments varies inversely from 36 to 9. The playout factor for FF is $X_F = 3$. This is a standard value for VCRs, also used in other papers. We consider two

different duration ratios $\tau_{FF}/\tau_P = 2$ and 5 (that is: PLAY periods last 2 times, resp. 5 times longer than FF periods). The parameters chosen are detailed in Table 1. We have displayed in this table the average "natural" consumption rate of the video, given by

$$b_N \;=\; b \times \frac{\frac{1}{\tau_P} \;+\; \frac{X_F}{\tau_{FF}}}{\frac{1}{\tau_P} \;+\; \frac{1}{\tau_{FF}}} \;=\; b \times \frac{\tau_{FF} \;+\; \tau_P\, X_F}{\tau_{FF} \;+\; \tau_P} \;.$$

Table 1
Parameters of the numerical experiments.

$1/\tau_P$	$1/\tau_{FF}$	A	b_N/b
45	9	1.4	1.33
45	9	1.3	1.33
60	30	1.9	1.67
60	30	1.8	1.67
60	30	1.7	1.67

In order to compare the performance of our scheme for videos of different lengths, we have measured the *probability of success*:

$$\pi_s \;=\; \frac{\mathbb{E}[L]}{N} \;.$$

The results should depend on how the natural rate b_N compares to the rate increase factor A. If $b_N/b < A$, then the law of large numbers will force the "natural" consumption curve $Q(t)$ (and therefore $R(t)$) to lie below the playout limit with large probability. Note that this effect may be long to appear if b_N/b is close to A. If $b_N/b > A$, then the converse effect appears. In that case, it also turns out that the actual curve $R(t)$ records a large number of failures.

Another effect may kick in: the probability that a failure occurs within segment i may depend on i. First, the time between w_1 and w_2 ($=D \times (2/A - 1)/b$) is smaller than the typical inter-arrival time between segments $w_{i+1} - w_i = D/Ab$. This may give a significant advance of data, and with few (large) segments, may result in a large success probability. On the other hand, when $b_N/b < A$, the first segments tend to be vulnerable to fluctuations in the consumption rate and have a smaller success probability. But if $b_N/b > A$, the first segments are more likely to be played out without failures than later ones.

The results are reported in Figure 5. The curve for $1/\tau_P = 45$, $1/\tau_{FF} = 9$ and $A = 1.3$ exhibits the poorest performance. This was expected, since $b_N/b > A$ in this case. Note however that the accuracy of the bound is not good for small

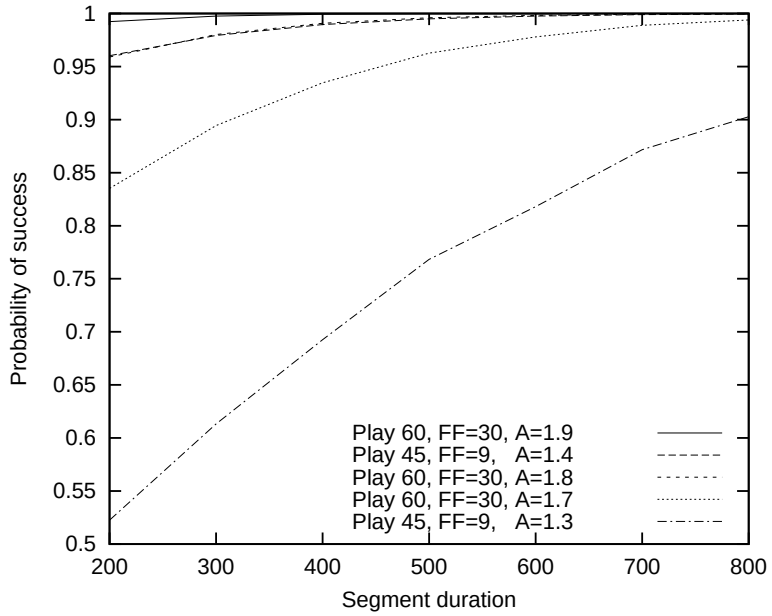

Fig. 5. Lower bounds on the probability of success $\mathbb{E}[L]/N$ for segments of the whole video.

values of the segment length (see Table 2), and that the probability of success is actually larger than 80%.

The other curves exhibit a probability of success larger than 85% for $1/\tau_P = 60$, $1/\tau_{FF} = 30$ and $A = 1.7$ (which is just slightly larger than b_N/b), and larger than 95% for the three other sets of parameters. The curves with $A = 1.4$ and $A = 1.8$ almost coincide. The experiments show that choosing a parameter A only slightly larger than the expected consumption rate of the user, coupled with sufficiently large segment sizes, achieves a very reasonable success probability.

The accuracy of the bound (7) is not good in relative terms, as demonstrated in Table 2. In this table, the bound is compared with values obtained by simulating a million replications of a playout of the entire video. The relative accuracy improves when D increases; this is explained by the fact that the law of large numbers has more effects when segments are longer.

The accuracy is however sufficient to assess the efficiency of the rate increase technique, and may be used to optimize the parameter A, in a compromise between the probability of success and the bandwidth requirements. Such an optimization is outside the scope of this paper.

Table 2
Comparison of the lower bound on the success probability (B) with simulations (S); $1/\tau_P = 45$, $1/\tau_{FF} = 9$.

$A = 1.4$			$A = 1.3$		
D	B	S	D	B	S
200	0.9602	0.9837	200	0.5226	0.8099
300	0.9794	0.9898	300	0.6132	0.8110
400	0.9896	0.9941	400	0.6926	0.8259
500	0.9949	0.9970	500	0.7683	0.8512
600	0.9976	0.9985	600	0.8180	0.8727
700	0.9989	0.9993	700	0.8716	0.9021
800	0.9995	0.9997	800	0.9027	0.9212

7 Conclusions

We have shown how by increasing the segment transmission rates for the tailored transmission scheme one can provide either deterministic or probabilistic support of user interactions. Since the FF action is the most "challenging" one to support, we restricted our analysis to a viewing behavior where only PLAY and FF are allowed. We first derived deterministic guarantees for satisfying all possible FF actions. Since the deterministic guarantees were based on the pessimistic assumption that the user watches the whole video from start to end in FF mode, we then defined a model for the viewing behavior (S-FF model) that consists of the user alternating between the PLAY and the FF modes.

For the S-FF model, we derived an analytic expression for a lower bound on the success probability. The reception of the segments is modeled as a semi-Markov accumulation process that allows the computation of the amount of video data received. While supporting VCR functions (and in particular FF) requires an increase in the segment transmission rates, our results indicate that this increase remains "moderate". The analytical results obtained for the S-FF are still pessimistic ones in the sense that a user who executes not only PLAY and FF but also actions such as PAUSE of SF will reduce the rate at which the video is consumed compared to the case of the S-FF model. In future extensions of this research, we shall exploit the theoretical formalism for accumulation processes that we have developed in this paper in order to handle various user's behavior and other VCR functions.

References

[1] E.L. Abram-Profeta and K.G. Shin. Providing unrestricted VCR functions in multicast video-on-demand servers. In *Proceedings of Multimedia Computing and Systems*, pages 66–75, June 1998.

[2] K.C. Almeroth and M.H. Ammar. The role of multicast communication in the provision of scalable and interactive video-on-demand service. *IEEE Journal on Selected Areas in Communications*, 14(6):1110–1122, August 1996.

[3] Y. Birk and R. Mondri. Tailored transmissions for efficient near-video-on-demand service. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, pages 226–231, June 1999.

[4] S.W. Carter, D.D.E. Long, and J.-F. Pâris. An efficient implementation of interactive video-on-demand. In *Proceedings of Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 172–179, 2000.

[5] D.R. Cox and H.D. Miller. *The Theory of Stochastic Processes.* Methuen & Co LTd, London, 1965.

[6] A. Dan, P. Shahabuddin, D. Sitaram, and D. Towsley. Channel allocation under batching and VCR control in video-on-demand systems. *Journal of Parallel and Distributed Computing*, 30(2):168–179, 1995.

[7] J.L. Doob. *Stochastic Processes.* Wiley, 1953.

[8] Z. Fei, M.H. Ammar, I. Kamel, and S. Mukherjee. Providing interactive functions through active client buffer management in partitioned video broadcast. In *Proc. NGC 1999*, volume 1736 of *LNCS*, pages 152–169. Springer Verlag, November 1999.

[9] L. Gao, J. Kurose, and D. Towsley. Efficient schemes for broadcasting popular videos. In *Proceedings od the International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, July 1998.

[10] L. Gao and D. Towsley. Supplying instantaneous video-on-demand services using controlled multicast. In *Proceedings of IEEE Multimedia Computing Systems*, pages 117–121, June 1999.

[11] R.A. Horn and C.R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[12] A. Hu. Video-on-Demand broadcasting protocols: A comprehensive study. In *Proceedings of IEEE INFOCOM*, volume 1, pages 508–517, Anchorage, Alaska, USA, April 2001.

[13] W. Liao and V.O. Li. The split and merge protocol for interactive video on demand. *IEEE Transactions on Multimedia*, 4(4):51–62, October 1997.

[14] J.-F. Pâris. An interactive protocol for video-on-demand. In *IEEE International Performance, Computing and Communication Systems*, April 2001.

[15] H.K. Park and H.B. Ryou. Multicast delivery for interactive video-on-demand service. In *Proceedings of the IEEE International Conference on Information Networking (ICOIN)*, pages 46–50, Tokyo, Japan, January 1998.

[16] S. Ramesh, I. Rhee, and K. Guo. Multicast with cache (Mcache): An adaptive zero-delay Video-on-Demand service. In *Proceedings of IEEE INFOCOM*, April 2001.

[17] M.R. Spiegel. *Schaum's Outline of Theory and Problems of Laplace Transforms.* McGraw-Hill, New York, 1956.

[18] TiVo. What is TiVo: Technical specifications. From TiVo's site at URL: http://www.tivo.com/discover/tech_specs.asp, 2001.

[19] S. Viswanathan and T. Imielinski. Pyramid broadcasting for video on demand service. In *Proceedings Multimedia Conference*, San Jose, CA, February 1995.

[20] P.-F. You and J.-P. Pâris. A better dynamic broadcasting protocol for video on demand. In *Proceedings of the IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pages 84–89, Phoenix, AZ, April 2001.