Providing VCR Functionality in a Constant Quality Video-On-Demand Transportation Service

Wu-chi Feng, Farnam Jahanian, Stuart Sechrest
EECS Department, University of Michigan
1301 Beal Ave., Ann Arbor, MI 48109-2122
{wuchi, farnam, sechrest}@eecs.umich.edu

Abstract

The use of bandwidth smoothing techniques for the delivery of prerecorded compressed video has been shown to be an effective technique in reducing the network bandwidth requirements needed to play back video. Any fixed allocation scheme can affect the ability to allow for VCR functions. In this paper, we examine the impacts that bandwidth smoothing techniques have on the delivery of stored video. We introduce the notion of VCR-window, which allows users to have full VCR functionality in a limited window while not requiring increases in the bandwidth reservations. In addition, we introduce a resource reservation scheme that can be used with the VCR-window for the delivery of stored video. To test the applicability of the VCR-window in video-on-demand systems, we have digitized 15 full-length movies for experimental data. Our results indicate that we can provide time-limited VCR functionality while retaining a fairly high network utilization.

Key words: Video-on-demand, VCR functionality, buffering and smoothing, bandwidth reservation, MPEG

1 Introduction

The delivery of constant quality compressed video requires that the network adapt to large fluctuations in bandwidth. Bandwidth smoothing techniques have been shown to be effective in removing burstiness, making network resource scheduling simpler, but at the expense of added delay and requiring additional buffering [8,12,14]. For live video applications, the benefits of smoothing are constrained by the requirement that latency remain low between video capture and video playback. Stored video applications, on the other hand, have two major advantages in network scheduling over live video. First, *a priori* knowledge can be used to smooth out the stream to the extent that the buffer will allow, trading buffer bandwidth for network bandwidth. Second, the scheduling of the network bandwidth can be done well in advance of the playback of video. The use of advance reservations in such schemes, however, has implications for the ability to provide users with familiar video cassette recorder (VCR) functionality such as stop, rewind, and fast forward.

For constant quality video delivery, the video bandwidth requirements can be smoothed by prefetching data into a buffer, shifting bursts of large frames forward in time. Depending on the amount of buffering available, a frame may sit in the client's smoothing buffer for a shorter or longer time before it is played back. Long buffer residency times are often required to reduce the high peak bandwidth requirements. Despite these long buffer residency times, the rate of transmission and the rate of consumption remain coupled. Alterations in the consumption rate that occur with VCR functions will require alteration in the video delivery plan, lest the buffer overflow or underflow. A video-on-demand system must effectively handle the contradictory goals of smoothing versus responsiveness

For video-on-demand systems with little or no buffering, the clients and servers must be tightly coupled. Any change in consumption of video data from the client must be immediately and continuously handled by the server. With buffering, changes in consumption will still require an adjustment by the server. These adjustments, however, need not be made instantaneously. Many operations can be performed without requiring the delivery of any new data from the server. Larger buffers allow greater latitude in handling these stops, starts, and rewinds. With excess buffering specifically used for handling variations in consumption rate, it is possible to further decrease the required disruptions of the server by combining the changes in consumption into a few requests. The amount of disruptions the servers must handle will be proportional to the buffer size used. If a majority of the rate changes can be handled by the client machine, then the network and servers can devote their resources to the handling the special cases that may arise instead of handling cases which can be taken care of with appropriate buffering.

In this paper, we present a framework for providing VCR functionality and advance reservations of bandwidth within a bandwidth-smoothing, stored-video environment. We introduce the notion of *VCR-window,* the set of buffered frames within which, full-function VCR capabilities are available without requiring changes to the bandwidth reservations made. The size of the VCR-window is determined by the size of the client buffer. We expect that for reasonable buffer sizes a large proportion of VCR operations can be handled from the VCR-window, with the remainder requiring more involved client and server interactions. The VCR-window affects the way in-advance reservations can be handled. We will discuss the impact of providing VCR functionality on *a priori* bandwidth reservations and present a reservation system for interactive video-on-demand systems. To study the VCR-window and the effect providing VCR capabilities has on the in-advance reservation system, we have digitized 15 full-length movies, which

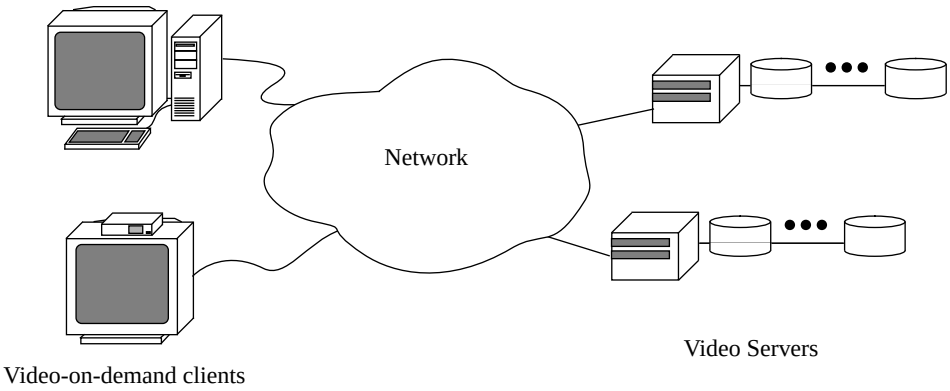

Figure 1: A Basic Video-On-Demand Architecture. This figure shows a basic video-on-demand server consisting of video servers, a network, and video-on-demand clients. The clients can be either a computer or set-top-box that contains hardware to interact with the network and a small disk for smoothing bandwidth requirements of the network.

account for over 32 GBytes of video data. Our results show that the VCR-window can be implemented with a small amount of additional buffering with little modifications necessary to the bandwidth reservations.

In the next section, we will present some background material, along with further motivation for investigating the problem of providing VCR functionality in bandwidth smoothing environments. In Section 3, we describe the VCR-window and a resource reservation scheme for stored video applications. In Section 4, we present our experimental results based on the 15 digitized movies along with one one-hour seminar that was presented in our department. Finally, we finish with some conclusions and directions for future work.

2 Background and Motivation

In this section, we discuss some of the assumptions about the video-on-demand system that we make for this paper. In addition, we present some background material on smoothing techniques that have been introduced in the literature and discuss the trade-off between smoothing and delay.

2.1 A Basic Video-On-Demand Architecture

Our basic video-on-demand system architecture consists of three main components: video servers, a network, and the clients. The servers will typically consist of large fast disks that deliver video to the clients [1,11,13,17]. These servers may also be part of a hierarchical video distribution system where less frequently requested video data is served from a tertiary archival server consisting of low cost storage devices such as tape or optical jukeboxes [7]. The network provides the pathway between the video servers and

their clients. The only assumption we make about the network is that it can provide network resource guarantees based either on some rate-based or real-time channel approach[2,24]. We also assume that the network provides some mechanism for in-advance reservations of bandwidth [5,10,23]. The clients consist of either desktop computers with support for digital video or a set-top-box devoted solely for viewing compressed video. We assume that clients have buffering available (either disk or RAM) for smoothing of network bandwidth requirements. In addition, we assume that the clients contain enough intelligence to create bandwidth allocation plans and to interact with the network and servers.

2.2 Bandwidth Smoothing Techniques

Bandwidth smoothing techniques fall into two broad categories: window based and non-window based bandwidth smoothing. Window-based smoothing techniques result in smoothing that occurs over some fixed interval [14,12]. Because the bandwidth smoothing is constrained to a fixed interval, this technique is useful for systems where the delay between the transmission and playback of a frame must adhere to some maximum delay requirement. Thus, window-based smoothing is particularly suitable for use in live video conferencing applications because they result in a maximum delay equal to the window size. For prerecorded video, however, one can exploit the *a priori* knowledge available so that smoothing can take full advantage of the buffer that is available[8].

As a non-window based smoothing technique, the Critical Bandwidth Allocation (CBA) algorithm was introduced to smooth bandwidth requirements based on the *a priori* knowledge available about videos in stored video applications [8]. The CBA algorithm constructs a bandwidth allocation plan that consists of runs, each with a constant bandwidth allocation requirement. Each run within the CBA generated plan has a bandwidth equal to the minimum constant bandwidth requirement required to play back the run without overflowing or underflowing the buffer. While this technique results in bandwidth plans with the minimum number of bandwidth increases, it does not take full advantage of the buffer to minimize the number of bandwidth changes required. The Optimal Bandwidth Allocation (OBA) algorithm creates bandwidth plans for the delivery of stored video that have the fewest number of bandwidth changes possible given the client-side buffer size[9]. For our discussions, we will use the digitized movie *Speed* and a video we will call *Seminar.* The *Seminar* video was recorded at a talk presented at our department. The video consists of the speaker standing next to an overhead projector presenting their work. As shown in Figure 2, the bandwidth allocation plans generated by the OBA result in very few required bandwidth changes for the play-

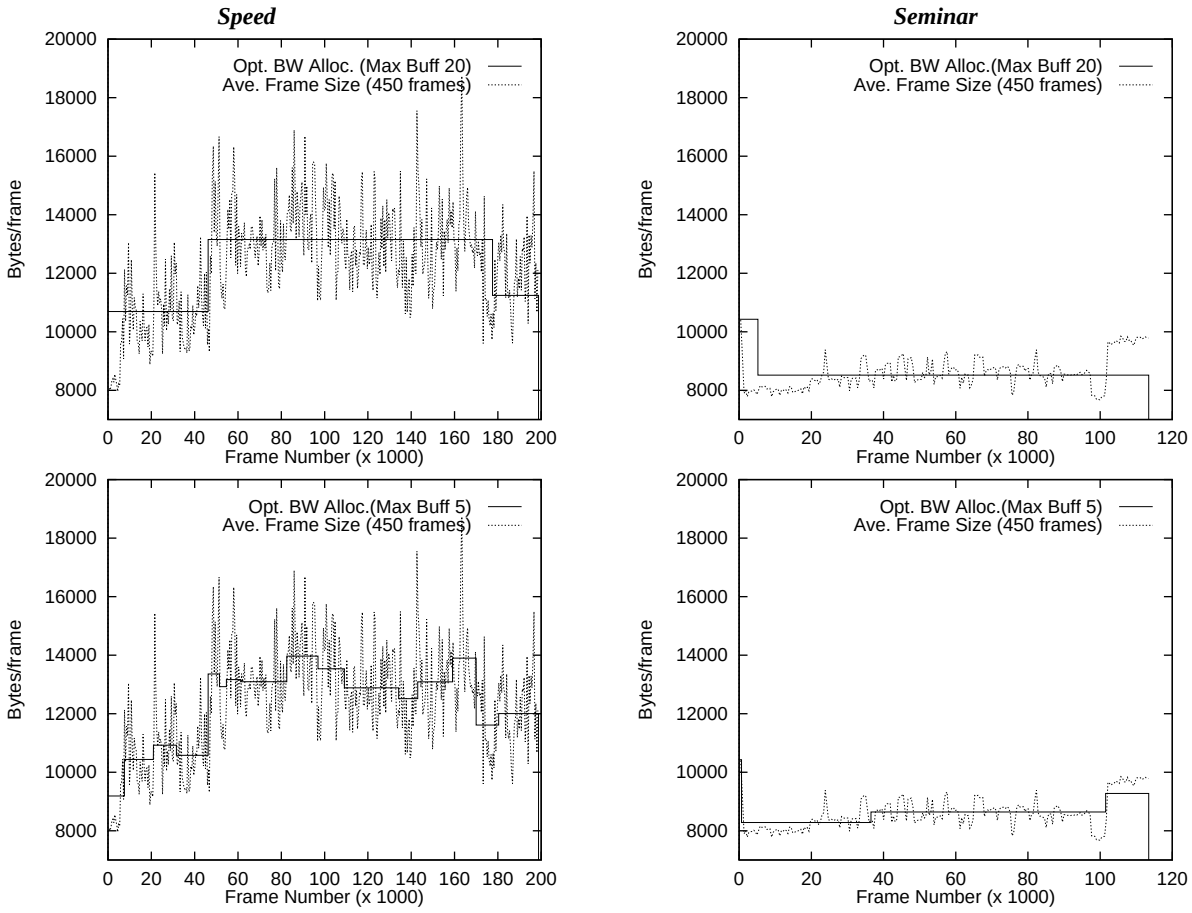

Figure 2: This figure shows the plan created using the optimal bandwidth allocation algorithm with a 20 MByte and 5 MByte buffer for the Motion-JPEG encoded videos *Speed* and *Seminar.* The dashed lines represent the average frame sizes for 15 second (450 frame) groups within the videos.

back of the videos. Note, the bandwidth plans generated by the CBA and OBA algorithms do not require any prefetching before the playback of the videos begin, hence, a high initial bandwidth may be required (as in the *Seminar* video). If the video request is made in advance, this initially high bandwidth requirement can be removed by prefetching data before the start of playback of the video. It is important to note that the *Seminar* and *Speed* videos are Motion-JPEG encoded, thus, they do not take advantage of temporal similarities between frames. This results in buffering and bandwidth estimates that are conservative for the results shown in this paper. In general, using MPEG encoded video streams instead of Motion-JPEG encoded video streams will result in one of two situations: 1) The actual buffer requirements will be smaller than presented (assuming the same buffer residency times) or 2) The buffer residency times will be much higher than the numbers presented (assuming the same buffer sizes). With the use of MPEG's B and

P frame types, the amount of buffering required to achieve the same amount of smoothing can be expected to be 4 to 10 times smaller.

Finally, burstiness within compressed video streams occurs at two levels. Short-term burstiness (or pattern burstiness) is the burstiness introduced by the video compression technique taking advantage of temporal knowledge between frames. This type of burstiness is exhibited in patterns of frame types such as those found in the MPEG I, P, and B frames [15]. Typically, an MPEG video is encoded with a regular repeating pattern of frames such as:

$$IBBPBBPBBIBBPBBI...$$

Thus, techniques which smooth based on some multiple of the pattern size are effective at removing the short-term burstiness. Long-term burstiness occurs from differences in scene variation. While window-based techniques are effective at removing short term burstiness, they are not as effective at removing long-term burstiness. Non-window based smoothing techniques are useful in removing both short-term and long-term burstiness because the smoothing is based on the optimal use of available buffering.

2.3 Buffering Versus Delay

Using the CBA or OBA bandwidth smoothing techniques results in a trade-off between buffering and delay. To smooth large frame-size peaks such as those found at the end of the *Seminar* video (see Figure 2), the data in the burst must be prefetched before the peak is played back. As a result, the buffer residency times (the time that a frame sits in the buffer) can be fairly substantial. The buffer residency times for the videos *Speed* and *Seminar* using the optimal bandwidth allocation plans from Figure 2 are shown in Figure 3. As shown by Figure 3, the buffer residency times are correlated to the amount of buffering used for smoothing. That is, the larger the buffer used, the higher the buffer residency times tend to be. In addition, the variance of buffer residency times will also depend on the long term burstiness of the video. As described in Figure 3, on average, the amount of time a frame spends in the buffer can be on the order of half a minute to a minute for a 20 MB buffer. For the *Speed* video, there was a larger variation in frame sizes throughout the movie, resulting in buffer residency times that also varied. For the *Seminar* video, the stream consisted of roughly the same size frames except at the end where a larger bursts of frames occurred. As a result, large buffer residency times were required to smooth out the large frame sizes at the end of the video. Incidentally, the end of the *Seminar* video consisted of the lights turning on, a panning of the speaker toward the center of the room, and a short question and answer session that included the first

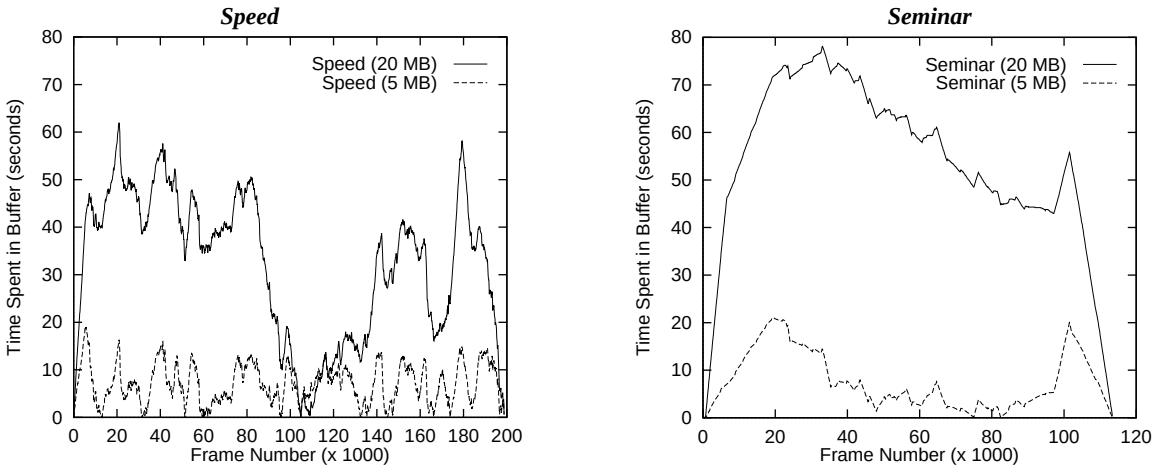

Figure 3: Buffer Residency Times. This figure shows the buffer residency times for the frames in the Motion-JPEG videos *Speed* and *Seminar* using a 5 and 20 MB buffer. For these videos and a 20 MB buffer, the average buffer residency time was 32.1 and 54.0 seconds for the *Speed* and *Seminar* videos, respectively. Note that for equivalent size buffers the residency times for MPEG encoded videos would be considerably larger.

row of listeners. We again reiterate that the buffer residency times are conservative due to the Motion-JPEG encoding. For MPEG encodings, the buffer residency times are expected to be much larger because of the tighter encoding, or similarly, the buffer residency times are expected to be the same for a much smaller smoothing buffer.

The large buffer residency times introduced by non-window based smoothing techniques have a direct impact on the ability to provide users with VCR capabilities. If random access to any point is to be allowed (while keeping the video quality constant), the network may have to contend with a potentially large required burst in bandwidth above the originally allocated bandwidth. This large burst of extra bandwidth may be required to make up for the absence of buffering (and delay) to help reduce the bandwidth requirements. Thus, providing VCR capabilities in a bandwidth smoothing environment can be a difficult task.

3 A Constant Quality Video-on-Demand Service

Any video-on-demand system must bring together several interrelated issues such as disk scheduling at the server, reserving underlying network bandwidth, and charging users for network usage. For systems that deliver constant quality video and use bandwidth smoothing to reduce peak bandwidths, the bandwidth allocations (especially if made in advance) are somewhat rigid to change because of the buffer residency requirement to smooth bandwidth requirements. In this section, we describe a video-on-demand service that has several key features: constant quality video delivery, VCR functionality (the *VCR-window*), and reservations in advance. Before describing the VCR-window, we first describe the type of inter-

activity that we expect to see in future video-on-demand systems. We then present the VCR-window and describe a reservation in advance system that can be used in conjunction with the VCR-window.

3.1 VCR Interactivity

In a bandwidth-smoothing video-on-demand system, providing unconstrained full-function VCR capabilities can cause major problems with the ability to deliver the required video data due to lack of network resources. By looking at the expected interactions during the playback of video, the video-on-demand system may be able to somewhat constrain the interactions, but still support a large majority of the video interactions that users will require. For video-on-demand services, it is our belief that video-on-demand users will typically change the access pattern during the playback of a video that fall into one of the four categories below:

- *Pause/Stop* - the user stops the movies for a short time to answer a phone call, go to the kitchen, etc.
- *Rewind* - the user rewinds the video to play back part of the video that they did not understand
- *Examine* - the user stops the VCR to examine more closely a portion of the video. As an example, a user may be watching a football game and wants to see a certain play a couple of times in slow motion to see why it did or didn't work.
- *Fast forward scan* - the user scans past parts of the video such as commercials in the program.

In the future, we believe that users may also require all of these functions from a video-on-demand system, although the actual distribution of access patterns within these categories may change. As an example, consider the operation fast-forward scan, which typically gets used to fast-forward through commercials. Currently, it is unclear how commercials will play a role in future video-on-demand systems. Clearly, if all users are going to fast-forward scan by commercials, then it would not make sense for companies to pay for slots within the playback of the video. We would expect that in such an environment that commercial messages may become a service, whereby, the video providers allow users to access commercials by companies on demand. For the rest of this paper, we will assume that fast-forward scans past commercials will not be required, however, we will not rule out the possibility of fast-forward scans in our discussion. As a result of this assumption, we expect that many of the accesses will be in a localized area within the video, thus, providing a limited window of full function VCR capabilities may suffice for most applications. In addition, by limiting the window size, the network bandwidth reservation levels may not need to be altered, and the required interactions with servers and networks may be minimized.

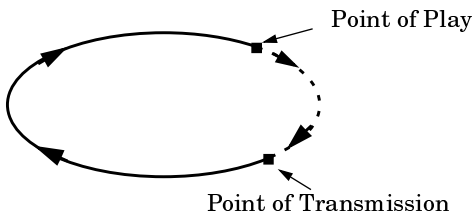

Figure 4: This figure shows the conceptual model of buffering for video data. The point of play (POP) and the point of transmission (POT) move clockwise. The solid line represents the rewind area, while the dashed line represents the amount of buffering that is in use for prefetching (smoothing)

3.2 The VCR window

To allow for VCR functionality, we propose a different model of video delivery which allows users to have full function VCR controls in a limited window called the *VCR window.* In our model of video transfer, we allow all VCR functions to occur at anytime within the course of playback but limit the range of accessible data without having to renegotiate the reserved bandwidth. We define the notion of the point of play (POP) to be the furthest frame in the video that has been viewed by the user and the point of transmission (POT) as the furthest frame in the movie that resides in the client buffer. Our model then consists of viewing the buffer as a circular buffer, in which, the POP and POT traverse the circumference in a clockwise manner (see Figure 4). During the delivery of video to the client, the POT will always be ahead of the POP. In addition, the distance the POT is ahead of the POP will be the amount of buffer space used for prefetching. The remaining part of the circumference is the amount of data that has been played back and is still in the buffer. Thus, when the buffer is nearly full, the POT will be just behind the POP, and when the buffer is nearly empty, the POP will be just behind the POT. Note, if the POT ever passes the POP or the POP passes the POT, we have buffer overflow and buffer underflow, respectively.

Using these definitions of the POT and POP, we make the observation that we can allow the user to have full function VCR capabilities in the area that the POP leads the POT without changing the bandwidth reservation level. One major drawback of this method is that when the buffer is nearly full the POP will not lead the POT by any significant amount. To ensure that the rewind area has some minimum amount of data, we define the *rewind buffer* to be the closest distance that the POT can approach the POP. For clarity we will refer to the distance that the POP leads the POT as the *rewind area* (or the VCR-window) for the rest of this paper. Figure 5 shows the resulting two cases when the buffer is full and when the buffer is empty.

Figure 5: Buffer Limit Conditions. Figures (a) and (b) show the cases that occur during playback when the buffer is empty and full, respectively. The solid line represents data that has been played back but not removed from the buffer, while the dashed line represents data that has been transmitted but not played back.

Formally, we can define the amount of available data in the rewind area on the *i*th frame as

$$RewBuffSize(i) = MaxBuff - \left(\left(\sum_{j=0}^{i} BwAlloc(j)\right) - \left(\sum_{j=0}^{i} FrameSize(j)\right)\right)$$

where,

- *MaxBuff* be the maximum buffer size including the *Rewind Buffer.*
- *BwAlloc(k)* be the bandwidth allocation on frame *k*. Note, we assume that the bandwidth allocation plan is allocated in bytes/frame.
- *FrameSize(k)* be the frame size of the *k*th frame.

This equation takes the difference between the total bandwidth received and the total bandwidth played back (i.e. the amount of data in the buffer) and subtracts it from the amount of buffering available. This equation does not, however, calculate the amount of video that is actually available but calculates its aggregate size. We can calculate the additional amount of rewind buffer needed to have *T* frames available in the buffer on the *i*th frame as

$$AddBuffReq(i, T) = \left(\sum_{k=max(0,\, i-T)}^{i} FrameSize(k)\right) - RewBuffSize(i)$$

This equation is essentially the size of the *T* frames needed in the rewind area with the amount of data already in the rewind area subtracted. If the user requires that 100% of the time the buffer has *T* frames in it, then the additional amount of buffering needed is simply the maximum AddBuffReq() over all frames within the movie.

To allow for VCR capabilities, when a user starts moving in the rewind area, the data flow from the server is stopped. The flow is then restarted only when the playback point reaches the POP. Thus, only two interactions to the server and network are required to support the VCR-window, one to stop the data flow, and one to start the data flow again. Because the delivery of bandwidth starts at exactly the same point in

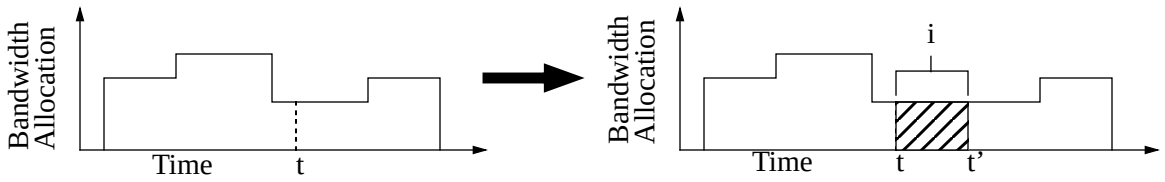

Figure 6: VCR-Window Example. This figure shows the adjustment in the bandwidth allocation plan that is made when a user uses the VCR controls for *i* time units (include the time to get pack to the POP that it was at). The remaining portion of the bandwidth allocation plan is then shifted by the amount of time spent in the rewind area. In this case, it is shifted *i* time units.

which the data was stopped, no changes in the bandwidth reservation level will be necessary while allowing for a small window of full function VCR capabilities. For long term bandwidth reservations, the only modification necessary will be the extension of the bandwidth requirement by the amount of time that was spent in the rewind area. Using this model for VCR functionality, the operations stop/pause, rewind, and examine can be provided to users. As an example, consider the bandwidth allocation plan shown in Figure 6. At time *t*, the user decides to stop the playback and examine the video that was just played. Suppose that the total time it takes for the user to examine the video and get back to time *t* is *i* time units. We then simply move all bandwidth allocations after the time *t* in the original bandwidth allocation plan to start at time *t'*, the time at which playback started again. Note, by shifting the bandwidth allocation plan after time *t* by *i* time units, the resultant bandwidth reservation has been modified. We will discuss how the reservation scheme can be modified to handle this change in the next section.

3.3 Access Outside the VCR window

Scans to points outside the VCR window will require renegotiations with the network and server. For long fast-forwards, the consumption rate originally anticipated will now be compressed in time, resulting in the need for more bandwidth than was originally planned for. We expect that these interactions may not occur very frequently, nonetheless, they should not be disallowed. For the renegotiation of bandwidth reservations in these cases, we expect that ideas such as the notion of *contingency channels* proposed by researchers at IBM will be useful [4]. In addition, the application of the CBA techniques, which allow for low-latency start of playback of video, will be useful in the efficient allocation of bandwidth for the contingency channel [8]. Because the VCR-window filters many of the interactions that will be required by the use of buffering, the contingency channels can be more efficiently allocated to handling the special cases that may arise during the playback of video.

3.4 A Video-on-Demand Resource Reservation Scheme

Resource reservations are an important part of network management for both in-advance and on the fly reservations because the network can then accurately estimate the bandwidth requirements that the clients need. For stored video-on-demand services, the ability to provide reservations of bandwidth in advance can make the job of resource allocation easier[16]. The work on resource reservation schemes have identified two key components that are necessary for resource reservations: the bandwidth requirement (level) and the duration that the bandwidth requirement is needed [10,5,23]. Without providing these, the authors argue that resource reservations in-advance then becomes a difficult task. In addition, Ferrari, Gupta, and Ventre point out that scheduling of bandwidth based on some fixed interval (period) will reduce the fragmentation that the reservation scheme will have to contend with[10]. Finally, it is commonly agreed upon that advance reservations will consist of two distinct phases, an admission control phase where the reservation is admitted and an enforcement phase where the bandwidth allocation is enforced.

While having an in-advance reservation is attractive for the ability to determine network load in-advance, stored video-on-demand applications have an additional benefit available. If the client machine (set-top box) is always connected to the network, then the video-on-demand system can begin downloading data to the client before the reservation actually begins. Thus, the video-on-demand system can take advantage of idle cycles by processing in-advance reservations that are to begin in the future. In the rest of this section, we will describe our in-advance reservation model and then the extensions necessary for the VCR-window.

Our in-advance reservation model is a periodic-based reservation scheme with a minimum bandwidth allocation period of 30 seconds. The user machine/set-top-box creates a bandwidth allocation plan based on the period boundaries and then passes this plan to the server and network for admission control. The network managers then compare the bandwidth requirements of the new channel and compare it to the available bandwidth allocation plan offered by the user. The example in Figure 7 shows sample requests that may be sent to the network manager. By using 30 second bandwidth allocation periods, the network manager only needs to evaluate 180 slots for a 90 minute video, reducing the complexity of the admission control algorithm. The network manager then allocates the available resources to the new channel if available. If the available bandwidth does not exist, then the network manager can either 1) offer a new starting time which can satisfy the bandwidth allocation plan or 2) create a different network path to the server which can satisfy the request. Handling of these conditions is beyond the scope of this paper.

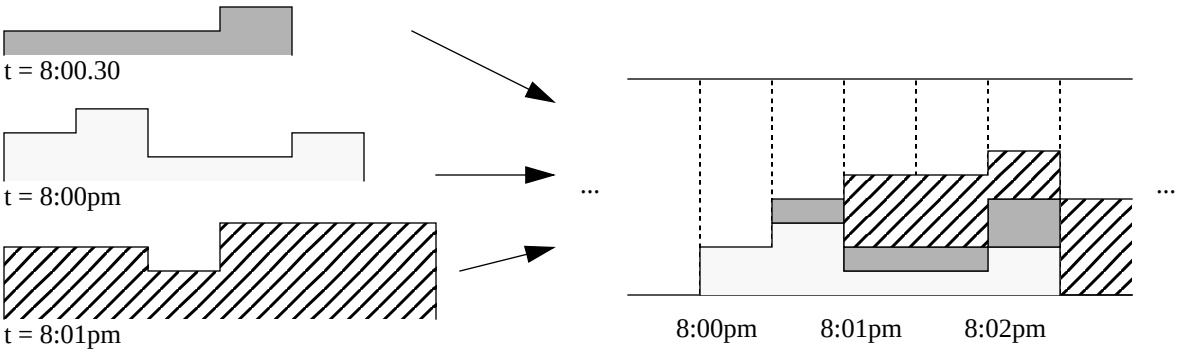

Figure 7: Resource Reservation Scheme. This figure shows the reservation of bandwidth for three sample streams. Bandwidth is reserved in 30 second intervals to reduce fragmentation of bandwidth.

Passing bandwidth plans as shown in Figure 7 to the network manager creates bandwidth plans that are rigid in nature. That is, bandwidth is allocated based on the bandwidth allocation plan that was passed in. In order to allow VCR-window functionality, the resource reservation system must reserve bandwidth based on the expected delay to be introduced by the user. The total time that the video can be delayed must be declared at admission control. The actual amount of time delay will depend on the guarantees that the user expects and the quality of service expected if the delay bounds are exceeded. This delay may be also determined by economic factors (i.e. how much a user is willing to pay for bandwidth that they may not use). For now, we assume the worst case for this delay, in that, all of the delay can occur at any interval. Therefore, in the calculation of the bandwidth allocation plan used for admission control, we create a bandwidth allocation plan that reserves the data such that at each point within the movie the video can be stopped for the maximum delay. Let T be the delay (in frames) for VCR functionality that is required from the user. Then the new bandwidth allocation plan (in bytes/frame) can be defined as

$$NewBwPlan(i) = max(BwPlan(i), BwPlan(max(0, i - T)))$$

Figure 8 shows a sample bandwidth allocation plan calculation that has the expected VCR-induced delay built into the bandwidth allocation plan.

4 Experimentation

The success of the VCR-window concept will depend on how much data will be available in the rewind area at given time as well as the amount of buffering required for use as the rewind buffer. As the amount of buffering devoted to the rewind buffer increases, the amount of smoothing available diminishes. The success of the reservation system will depend on the effectiveness of the video-on-demand system to uti-

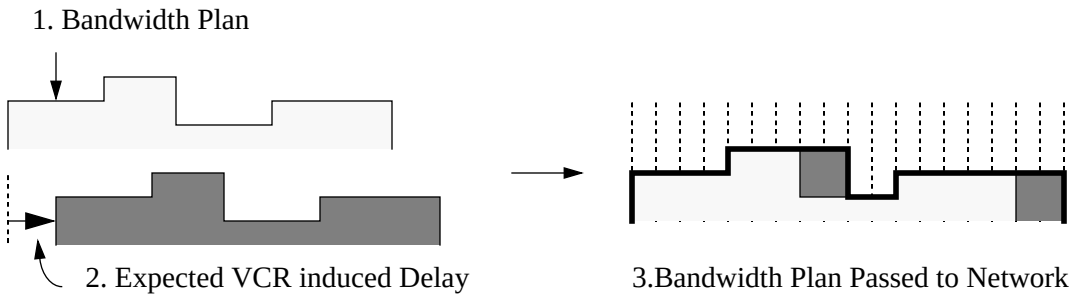

Figure 8: Bandwidth Reservation Calculation. 1) Client machine creates bandwidth allocation plan. 2) Client machine creates second plan that is delayed by the expected delay. 3) Bandwidth plan that combines the maximum bandwidth requirements of both plans is created and passed to the network and server as part of admission control. This plan is denoted by the heavy solid line.

lize its bandwidth. In order to fully understand the impact of buffering on the VCR-window and the associated in-advance reservation system, we have digitized 15 full-length movies along with 1 seminar that was presented at the University of Michigan, the video we have called *Seminar*. In this section, we will first describe the video data that was captured and then discuss our experimental results. For the experiments, we used the OBA algorithm with prefetching on the initial run.

4.1 A Trip to the Movies

To aid in the capture of digital video, we used a Miro VideoDC1tv capture board and a Pentium P90 processor based system. The Miro Video capture board is a Motion-JPEG compression board that captures full screen video in real-time. We do not have the equipment to perform rapid MPEG encodings. Because the basic routine for encoding I-frames within an MPEG video are derived from the JPEG compression standard, the frame sizes for our experimental video data are roughly equivalent to all I-frame encoded MPEG video movies. The CBA and OBA algorithms are most sensitive to scene content changes and not pattern burstiness, however, the size of the resulting streams strongly affects the buffer requirements. MPEG encoded video could achieve the same performance with smaller buffers or allow an expanded VCR-window with similar buffers. We, therefore, expect that the results presented here are somewhat conservative compared to a system using MPEG as its compression standard. The Miro Video board digitized the movies at 640x480 and then subsampled them to 320x240 with guaranteed VHS picture quality.

Using our testbed, we captured 15 full-length movies at a range of 0.85 to 1.61 bits per pixel. The statistics for these videos are shown in Table 1. In digitizing the video data, we attempted to capture a variety of different movies in order to examine the effects each had on the VCR-window and the reservation system. The *Beauty and the Beast* video is an animated Walt Disney movie, resulting in scenes with a lot of high

Movie	Tot.Size (GB)	Length	Ave. Bit Rate (Mbps)	Ave. Bits per pixel	Ave. Frame Size (bytes)	Frame Size std. dev.
Beauty and Beast	1.816	1 hour 20 min	3.039	1.24	12661.41	3580
Big	2.262	1 hour 42 min	2.963	1.21	12345.90	2366
Croc. Dundee	1.816	1 hour 34 min	2.586	1.06	10772.97	2336
E.T.	1.780	1 hour 50 min	2.165	0.88	9021.89	2574
1993 NCAA Final Four	1.206	41 min	3.949	1.61	16455.80	4138
Home Alone 2	2.352	1 hour 55 min	2.732	1.12	11382.89	2480
Honey, I Blew Up the Kid	2.122	1 hour 25 min	3.321	1.36	13835.90	3183
Hot Shots 2	1.920	1 hour 24 min	3.064	1.25	12765.98	3240
Jurassic Park	2.501	2 hours 03 min	2.727	1.11	11362.97	3252
Junior	2.705	1 hour 47 min	3.363	1.37	14013.38	3188
Rookie of the Year	2.221	1 hour 39 min	2.984	1.22	12434.66	2731
Seminar	0.976	1 hour 03 min	2.065	0.85	8604.06	592
Sister Act	2.063	1 hour 36 min	2.856	1.17	11901.60	2608
Sleepless in Seattle	1.720	1 hour 41 min	2.275	0.93	9477.39	2459
Speed	2.459	1 hour 55 min	2.970	1.21	12374.23	2707
Total Recall	2.343	1 hour 49 min	2.875	1.17	11977.92	2692

Table 1: Digitized Video Statistics. This table shows the statistics gathered for the data used in the experimentation of this paper. The fifteen movies and one seminar represent 32.3 Gigabytes of data and approximately 25.5 hours of video data.

frequency components as well as scenes that had large areas of constant color. The *1993 NCAA Final Four* video is a documentary describing the NCAA Final Four basketball tournament, resulting in many of the scenes with lots of detail. As a result, the *1993 NCAA Final Four* video had the highest average bit rate. The rest of the movies are a mix of conventional entertainment containing a wide range of scene content, including digital effects and animations. The *Seminar* video, as previously mentioned, contains a single scene and, thus, contains the smallest variation in frame sizes.

4.2 VCR-window Experimentation

The success of the VCR-window will depend on the amount of rewind buffer required for the guarantees expected from the user. Using the CBA or OBA algorithms with a reasonably large buffer, the number of times when the buffer will be full (and therefore limiting the rewind area) will be small. Thus, very little additional buffering for use as a rewind buffer may be required.

Figure 9 shows the histogram of the amount of video data that is available in the rewind area for a 25 and 50 Mbyte buffer with the rewind buffer size set to 0. For the *Speed* video, using a 25 MByte buffer resulted in having over half a minute of video in the rewind area 53% of the time while using a 50 MByte buffer resulted in having over half a minute of video in the rewind area 75% of the time. In addition, 15

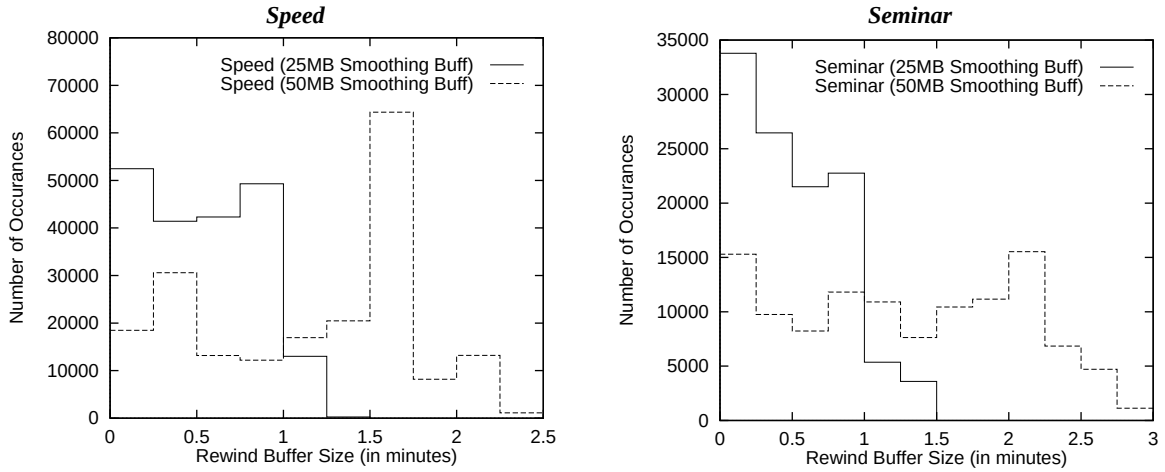

Figure 9: Histogram of Buffer Rewind Times. These graphs show the distribution of buffer rewind times for *Speed* and *Seminar* (rewind buffer size = 0). For 25 and 50 Mbyte buffers, this resulted in having 30 seconds of video available 52.8% and 75.3% of the time, respectively, for the video *Speed*. Similarly, the *Seminar* video had 30 seconds of video available 46.9% and 77.9% of the time, respectively.

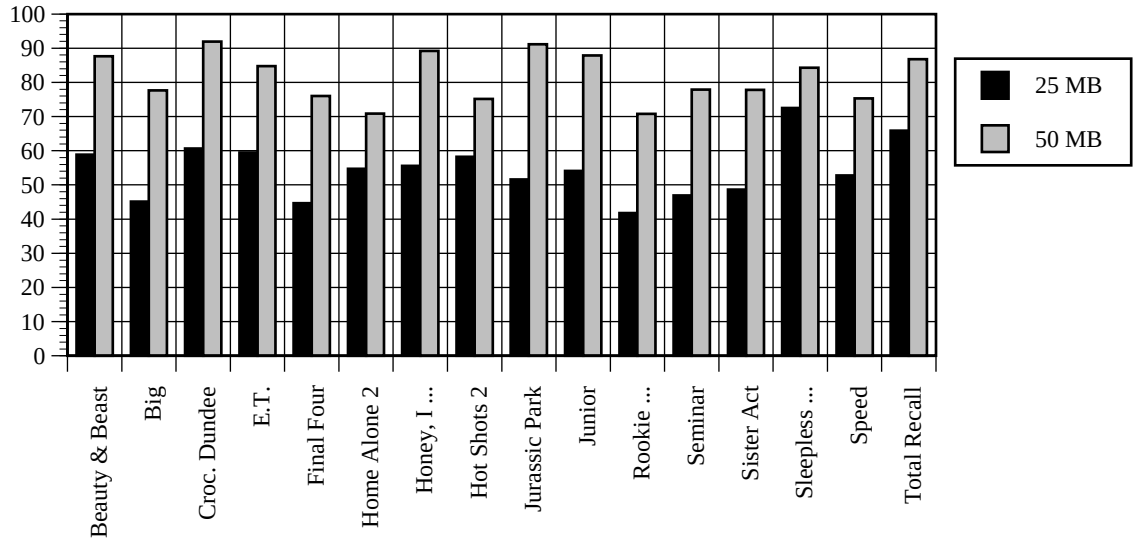

Figure 10: Buffer Rewind Times for all Movies. This figure shows the percentage of time that rewind area contains more than 30 seconds of video for the 25 and 50 MByte smoothing buffer with the rewind buffer size set to 0.

seconds of video was available 74% and 91% of the time for the 25 and 50 Mbyte buffers. The *Seminar* video exhibited similar numbers to the *Speed* video. As shown in Figure 10, the percentage of time that the rewind area contained more than 30 seconds of video for the rest of the video data exhibited similar numbers also. Typically, the 25 and 50 MB buffers resulted in the rewind area with 30 seconds of video 45-60% and 75-90% of the time, respectively.

The addition of a rewind buffer will shift the histograms of buffer rewind times to the right, thus increasing the amount of time that is available in the rewind area. Figure 11 shows, the amount of buffering

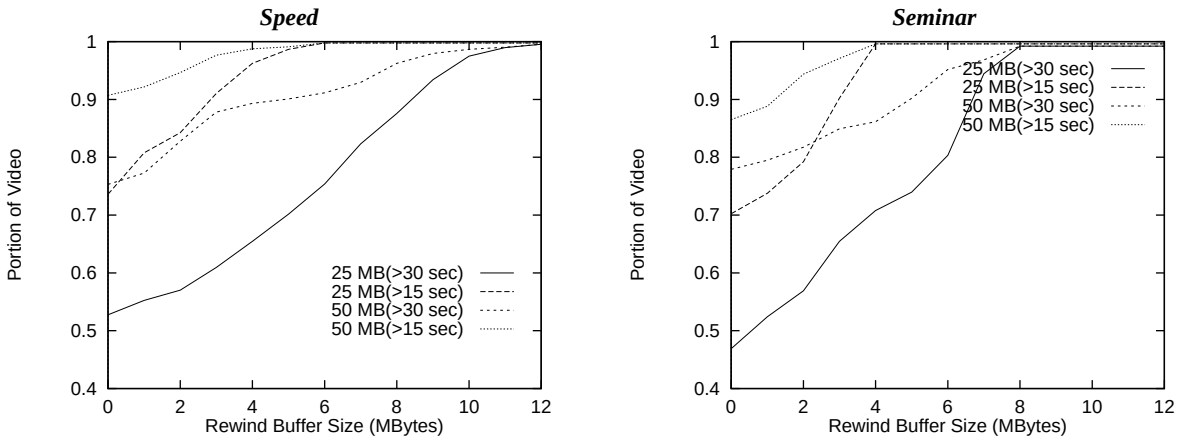

Figure 11: Rewind buffer size required for percentage of video above required limit for the movie *Speed*. As an example, with a 25 MByte smoothing buffer, in order to achieve 15 seconds of buffering 90% of the time, roughly 3MBytes of rewind buffer is required above the 25 MByte smoothing buffer.

needed for the rewind buffer size in order to have the rewind area contain a certain percentage of video in the rewind buffer greater than 15 and 30 seconds. Note, these buffer rewind sizes are in addition to the 25 and 50 MByte buffers used for the smoothing of bandwidth requirements. As expected the lines for the same time (15 and 30 seconds) approached the same required rewind buffer size because this buffer size is determined by the same point (area) within the video. In addition, the amount of required rewind buffer space decreased as the size of the smoothing buffer increased. This was mainly due to the larger buffer sizes having more rewind area on average. In order to achieve at least 15 seconds of video in the rewind area 95% of the time for the movie *Speed*, only 4 and 2 MBytes of rewind buffer were required for the 25 and 50 MByte smoothing buffers, respectively. The *Seminar* video approached the 100% line faster than the *Speed* video. This is due to the smaller (and more constant) average bit rate of the *Seminar* versus the *Speed* video. If we take the average frame sizes for the videos and multiply it by the number of frames in 30 seconds of video (900 frames), the *Speed* video results in 11.1 MB while the *Seminar* video results in 7.7 MB. It is interesting to note that these videos approached 100% near these values.

Figure 12 shows the percentage of time that 30 seconds of movie is available when using an 8 MByte rewind buffer. The highest bit rate video *Final Four* resulted in the smallest percentage of rewind times greater than 30 seconds, while the 3 smallest bit rate videos (*E.T, Crocodile Dundee,* and *Sleepless in Seattle*) resulted in the highest percentage of rewind times. This suggests that the rewind size is roughly correlated to the average bit rate that the encoded movie has. Thus, we expect that the use of tighter encoding schemes such as MPEG with B and P frames will reduce the overall requirement of the rewind buffer size.

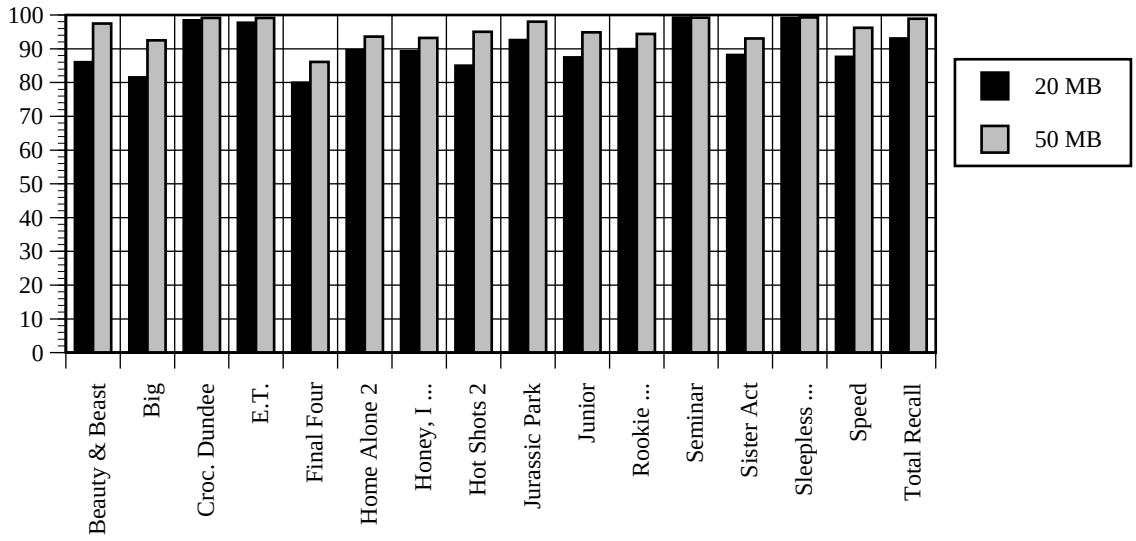

Figure 12: Buffer Rewind Size Measurements. This figure shows the percentage of time that 30 seconds of video is available using an 8 MByte rewind buffer in addition to the 25 and 50 MByte smoothing buffer.

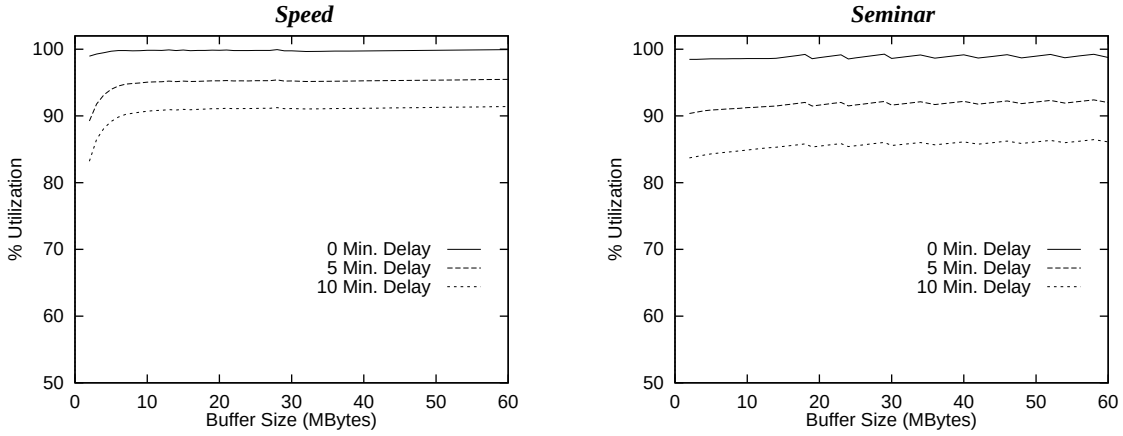

Figure 13: Reservation Utilization. These graphs show the reservation utilization for the video *Speed* and *Seminar* with reservations made on a 30 second period and reserved with the maximm additional delay expected during playback.

4.3 Bandwidth Reservations

For bandwidth reservations, one of the main concerns is the actual network utilization versus the amount of bandwidth that was allocated. For example, if the amount of bandwidth reserved is around 95% but only 50% of the bandwidth is actually used, then the reservation scheme may need to be modified to be more effective in utilizing the network. As shown in Figure 13, the reservations based on a 30 second period with no extra delay built into the reservation plan for VCR functionality yield reservation utilizations between 99% and 100% of what was reserved. Thus, the OBA allocation (with prefetching on the first run) yields bandwidth allocation plans that utilize nearly all the bandwidth reserved based on 30 sec-

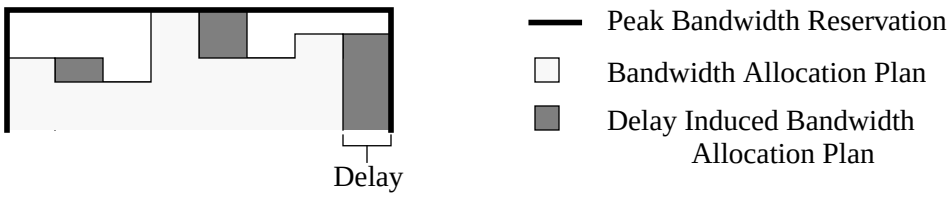

Figure 14: Peak Bandwidth Reservation. The heavy solid line shows the creation of a peak bandwidth allocation plan. This bandwidth allocation plan is used for the advanced reservations made in Figure 15.

ond periods and suffer very little internal fragmentation of bandwidth allocations. Furthermore, the utilizations can be increased by aligning bandwidth change boundaries with the periodic boundaries. Three trends are worth noting in Figure 13. First, for very small buffer sizes (< 5 MB), the utilization is hurt by two things, more bandwidth changes that are not aligned with periodic boundaries and more bandwidth is reserved for the delay constraint. Second, because the utilization for the OBA algorithm is quite high, the utilization for the streams reach their limits fairly quickly. Finally, the *Seminar* video has lower utilization for the 5 and 10 minute delays because the video is shorter in length, thus, the 5 and 10 minute delay reservations make up a larger portion of their reservations. With tighter encoding mechanisms, the utilization can be expected to be higher with no other modifications to the buffer size or delay for VCR functionality.

The expected overall utilization of the network is not captured by the graphs in Figure 13 because they do not capture the peak reservations which may affect other bandwidth allocation plans. To establish a "lower bound" on the expected network utilization, we modified the bandwidth allocation plans to have both the peak bandwidth reservation for the *entire* video and the expected VCR induced delay. A sample graph allocation is shown in Figure 14. Thus, the peak bandwidth allocation makes the reservation for the *entire* movie as one constant bandwidth reservation. We then graphed the expected bandwidth utilization based on these peak bandwidth allocations instead of the bandwidth reservations described in Section 3.4. As shown by Figure 15, we see that the bandwidth utilization has dropped from those shown in Figure 13. The bandwidth utilizations, however, are still reasonable. For bandwidth plans that use a 25 MByte smoothing buffer, no rewind buffer, and have an extra 5 minutes of delay in the bandwidth reservations, the movie *Speed* had a utilization of 90.7% while the *Seminar* video had a utilization of 92.7%. Thus, even for peak bandwidth reservations with 5 extra minutes reserved, we expect that the bandwidth utilization can be held fairly high. The peak bandwidth reservations did not affect the *Seminar* video as much as the *Speed* video because it had less variation between frame sizes resulting in smaller peaks when then occurred.

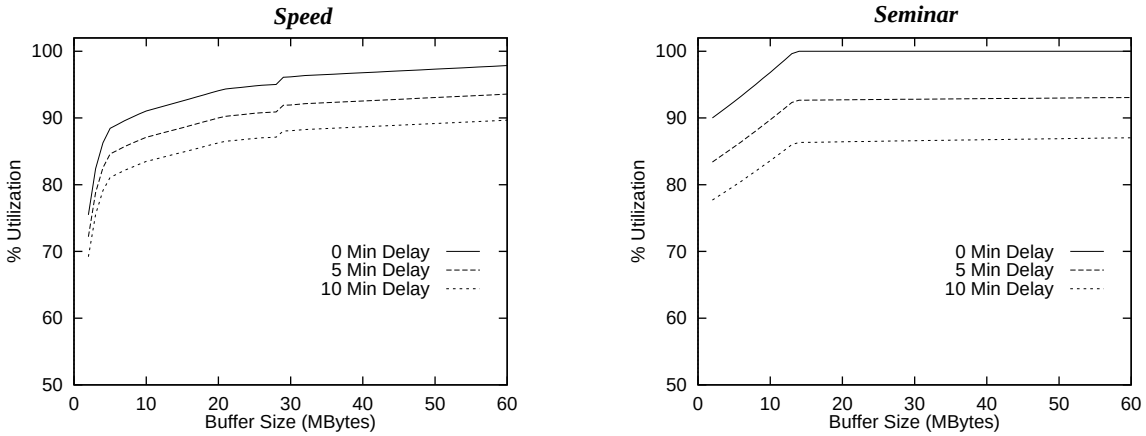

Figure 15: Peak Reservation Utilization. These graphs show the peak reservation utilization for the video *Speed* and *Seminar* with reservations made at the peak bandwidth allocation and with the maximm additional delay expected during playback.

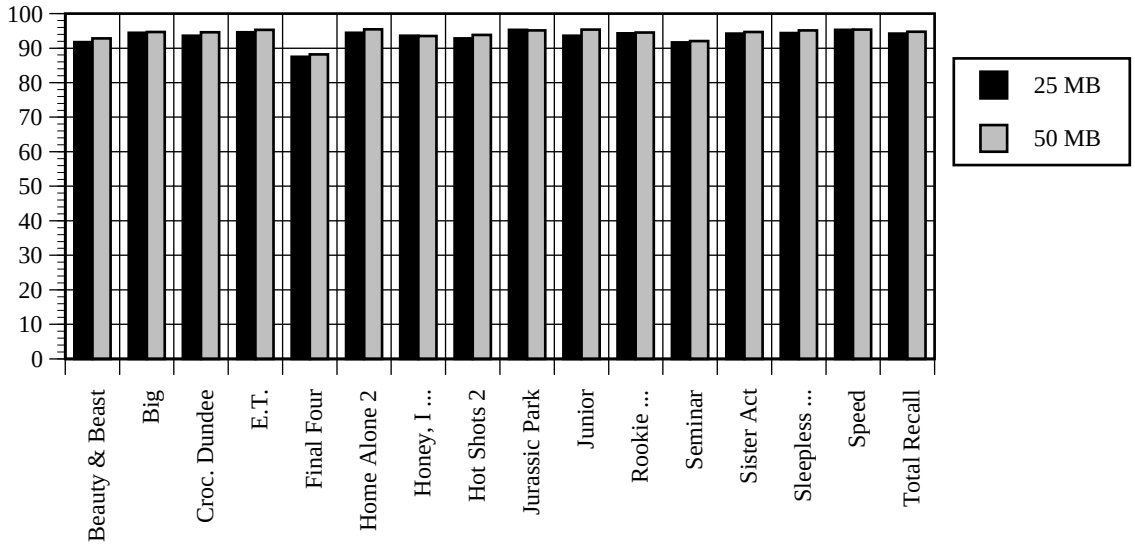

Figure 16: Reservation Utilization for Other Video Data. This figure shows the reservation utilization for bandwidth plans that are allocated in 30 second periods and have 5 minutes of delay added to the reservations. All utilizations were in the 90% to 95% range with the exception of the *Final Four* video.

Finally, Figure 16 and Figure 17 show the normal reservation utilizations and peak reservations utilizations in the same exact way that Figure 13 and Figure 15 were made, respectively. In Figure 16, the normal reservation scheme with an additional 5 minute delay for both 25 and 50 MByte buffers are shown for all movies. They resulted in utilization ranging from 90% to 95% with the exception of the *Final Four* video. This result is expected as the *Final Four* video is only 41 minutes in length, thus, the extra 5 minutes accounts for 11% of the video. As expected the peak utilizations are generally less than the normal reservation method. In addition, the 25MB buffers are affected more because they cannot remove the peak burstiness as much as with a 50 MB buffer. Nonetheless, they exhibited fairly good utilizations. The video

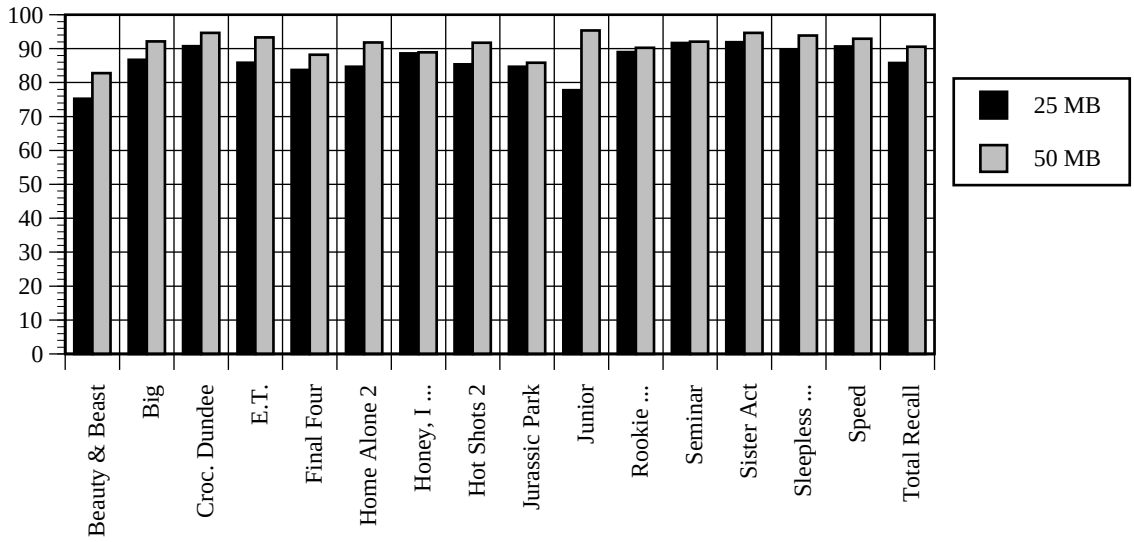

Figure 17: Peak Reservation Utilization for Other Video Data. This figure shows the peak reservation utilization for bandwidth plans that are allocated in 30 second periods and have 5 minutes of delay added to the reservations.

Beauty and the Beast video was affected the most by using a peak bandwidth reservation. The reason this occurred was that a peak of very large frame sizes could not be overcome with smoothing in a small area within the movie. We expect that these singular peaks will easily fit into the valleys of other bandwidth allocation reservations. Some movies exhibited no change in utilization from the reservation utilization to the peak reservation utilization. In these cases, the amount of buffering in the reservation utilization was enough to remove almost all of the burstiness and thus does not get affected as much by using the peak bandwidth requirement. In general, however, the overall expected bandwidth utilization of the network can be expected to be fairly high.

5 Conclusion

In this paper, we have introduced the notion of *VCR-window* which allows a user to have full function VCR capabilities within a constrained region that does not change the bandwidth allocation requirements. We have also shown that providing 30 seconds of video available for 90% of the time can be implemented with a small amount of additional buffering, even for loosely encoded Motion-JPEG video. We expect that the majority of interactions that occur during the playback of video can be accounted for by using this technique. For user who want a guarantee of some amount of video *always* available in the rewind area, the required rewind buffer size is determined by a few frames within the movie. For users who are willing to settle for lesser guarantees during the examine or scan phases, the VCR functionality can always be pro-

vided by the server which can fit the required video into the reserved channel capacity. Work on supporting scan operations from the server can be found in [3,6,22], while modifying compressed video to fit within a specified channel capacity can be found in [18,19].

We have also presented a periodic, in-advance resource reservation scheme to be used in conjunction with the VCR-window. The optimal bandwidth allocation algorithm results in very high network bandwidth utilization even under periodic scheduling boundaries. This is mainly due to the optimal bandwidth allocation algorithm minimizing the number of bandwidth changes as well as the peak required bandwidth. Nonetheless, the total amount of smoothing available depends on the long-term burstiness of the data itself. Using the advance reservation scheme in conjunction with the optimal bandwidth allocation algorithm, allowing users 5 to 10 minutes of "VCR-time" can be provided without degenerating the utilization. We can expect that the lower bound for network utilization will be at least 80 percent. The 5 to 10 minutes of extra reserved "VCR-time" can be allocated for users to browse commercials or previews of other movies, assuming that they fit into the bandwidth reservation or are viewed at a slightly lower quality.

In the event more "random" access patterns are required such as jumps or scans of more than a couple of minutes in the video are required, renegotiation of bandwidth will most likely be required or the reservation of bandwidth with that is a lot higher than actually used. For random accesses, it is probably more beneficial for the underlying network services to use approaches found in live-video applications or reserve part of the bandwidth for contingency channels that are used in the "difficult" cases. The size and magnitude of these contingency channels will depend on the percentage of times that the users in the video-on-demand system stray from the VCR-window. While we expect that the frequency of these occurrences will be quite small, the video-on-demand system should provide this flexibility. Finally, for random accesses, the use of indexing schemes to allow access at distinct points within a video may allow the bandwidth requirements to be handled in a more efficient manner for accesses outside the VCR window.

6 References

[1] D. Anderson, Y. Osawa, R. Govindan, "A File System for Continuous Media", *ACM Transactions on Computer Systems*, Nov, 1992, pp 311-337.

[2] C.M. Aras, J.F. Kurose, D.S. Reeves, H. Schulzrinne, "Real-time Communication in Packet Switched Networks", *Proceedings of the IEEE*, Vol. 82, No. 1, pp. 122-139, Jan. 1994.

[3] M.S. Chen, D.D. Kandlur, P.S. Yu, "Support for Fully Interactive Playout in a Disk-Array-Based Video Server", In Proceedings of *ACM Multimedia 1994*, San Francisco, CA, Oct. 1994, pp 391-398.

[4] A. Dan, P. Shahabuddin, D. Sitaram, D. Towsley,‘‘Channel Allocation Under Batching and VCR Control in Movie-On-Demand Servers”, *IBM Research Report RC19588*, Yorktown Heights, NY, 1994.

[5] M. Degermark, T. Kohler, S. Pink and O. Schelen, “Advance Reservations for Predictive Service”, In Proceedings of *5th Intl. Workshop on Network and Operating System Support for Digital Audio and Video*, Durham, New Hampshire, April 18-21, 1995, pp. 3-14.

[6] J. Dey-Sircar, J. Salehi, J. Kurose, D. Towsley,“Providing VCR Capabilities in Large-Scale Video Servers”, In Proceedings of *ACM Multimedia 1994*, San Francisco, CA, Oct. 1994, pp 25-32.

[7] C. Federighi, L. Rowe,‘‘A Distributed Hierarchical Storage Manager for a Video-on-Demand System”, In Proceedings of *1994 IS&T/SPIE Symposium on Electronic Imaging: Science and Technology*, San Jose, CA Feb. 1994.

[8] W. Feng, S. Sechrest,‘‘Critical Bandwidth Allocation for the Delivery of Compressed Prerecorded Video”, *Computer Communications*, Oct. 1995.

[9] W. Feng, F. Jahanian, S. Sechrest,‘‘An Optimal Bandwidth Allocation Strategy for the Delivery of Compressed Prerecorded Video”, CSE-Technical Report 260-95, Sept. 1995.

[10] D. Ferrari, A. Gupta and G. Ventre, “Distributed Advance Reservation of Real-Time Connections”, In Proceedings of *5th Intl. Workshop on Network and Operating System Support for Digital Audio and Video*, Durham, New Hampshire, April 18-21, 1995, pp. 15-26.

[11] D.J. Gemmell, H.M. Vin, D. Kandlur, P.V. Rangan, L. Rowe,‘‘Multimedia Storage Servers: A Tutorial”, *IEEE Computer*, May 1995, Vol. 28, No. 5, pp.40-49.

[12] Pawan Goyal, Harrick M. Vin,“Network Algorithms and Protocol for Multimedia Servers”, In *Proceedings of INFOCOM 1996* (to appear).

[13] D. Kandlur, M. Chen, Z.Y. Shae,“Design of a Multimedia Storage Server” , In *IS&T/SPIE Symposium on Electronic Imaging Science and Technology*, San Jose, CA, Feb. 1994.

[14] S. Lam, S. Chow, D. Yau, “An Algorithm for Lossless Smoothing of MPEG Video”, In *ACM SIGCOMM Conference Proceedings*, 1994.

[15] D.J. LeGall, ‘‘A Video Compression Standard for Multimedia Applications,’’ *Communications of the ACM*, Vol. 34, No. 4, (Apr. 1991), pp. 46-58.

[16] T.D.C. Little, D. Venkatesh,“Prospects for Interactive Video-On-Demand”, *IEEE Multimedia*, Vol. 1, No. 3, Fall 1994, pp. 14-24.

[17] P. Lougher, D. Shepherd,“The Design of a Storage Sever for Continous Media”, *The Computer Journal*, Vol. 36, No. 1, Feb. 1993, pp 32-42.

[18] P. Pancha, M. El Zarki,“MPEG Coding for Variable Bit-Rate Video Transmission”, *IEEE Communications Magazine*, Vol. 32, No.5, May 1994, pp 54-66.

[19] P. Pancha, M. El Zarki,“Prioritized Transmission of Variable Bit Rate MPEG Video”, In *IEEE GLOBECOM 1992*, Dec. 1992, pp 1135-1139.

[20] P. Pancha, M. El Zarki, “Bandwidth Allocation Schemes for Variable Bit Rate MPEG Sources in ATM Networks”, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 3, No. 3, June 1993, pp. 190-198.

[21] D. Reininger, D. Raychaudhuri, et. al, “Statistical Multiplexing of VBR MPEG Compressed Video on ATM Networks”, In *IEEE INFOCOM 1993*, March 1993, pp 919-926.

[22] P. J. Shenoy, H. M. Vin, “Efficient Support for Scan Operations in Video Servers”, In Proceedings of *3rd ACM Conference on Multimedia*, October, 1995.

[23] L. Wolf, L. Delgrossi, R. Steinmetz, S. Schaller and H. Wittig, “Issues of Reserving Resources in Advance”, In Proceedings of *5th Intl. Workshop on Network and Operating System Support for Digital Audio and Video*, Durham, New Hampshire, April 18-21, 1995, pp. 27-37.

[24] H. Zhang, S. Keshav,‘‘Comparison of Rate-Based Service Disciplines”, In Proceedings of *ACM SIGCOMM ‘91*, Sept. 1991, pp. 113-121.