Delivery Techniques

Developing hybrid bandwidth smoothing techniques that are aimed for both VCR interactivity as well as high-utilization of network channels are required. This involves both the interaction between various bandwidth smoothing plans interacting in harmony to increase bandwidth utilization and providing ways for dynamically changing the bandwidth allocations when the client's buffer allocation varies over time, such as on a workstation viewing video.

Interactive Video-on-Demand Servers

Much work has been focused on providing interactivity through the video-on-demand server. In the worst case, the interactive server techniques described in the literature will be important. In addition, providing a mechanism between using bandwidth smoothing plans and relying on the interactive server to provide VCR services (in the chaotic cases) is required.

Efficient Excess Capacity Allocation

For interactive services, efficient methods for allocating *contingency channels* are required. This involves the investigation of techniques for guaranteeing responsiveness in interactive video-on-demand systems. That is, guaranteeing a user a worst case delay between VCR interactions and the continuation of the playback of the video.

5. Conclusion

The work on resource allocation for video-on-demand systems has been fruitful, however, much work is still left to be done. Buffering techniques are attractive for video-on-demand systems that provide guarantees of service because these systems typically make guarantees based on the peak bandwidth requirement. The main challenge that I see in the near future is providing a way for buffering and interactive services to work in harmony.

6. References

[1] A. Dan, D. Sitaram, P.Shahabuddin, "Scheduling Policies for On-Demand Video Servers with Batching", In *Proc. ACM Multimedia 1994*, Oct. 1994, pp. 15-23.

[2] W. Feng, S. Sechrest, "Smoothing and Buffering for the Delivery of Stored Video", In *Proc. IS&T/SPIE Symposium on Multimedia Computing and Networking 1995*, San Jose, CA, Feb. 1995, pp. 234-242.

[3] W. Feng, S. Sechrest, "Critical Bandwidth Allocation for Delivery of Compressed Video", *Computer Communications*, Vol. 18, No. 10, Oct. 1995, pp. 709-717.

[4] W. Feng, "Rate-Constrained Bandwidth Smoothing for the Delivery of Stored Video", In *Proc. IS\&T/SPIE Symposium on Multimedia Computing and Networking*, San Jose, CA, Feb. 1997.

[5] W. Feng, F. Jahanian, S. Sechrest, "Optimal Buffering for the Delivery of Compressed Prerecorded Video" , In *Proc. IASTED/ISMM International Conference on Networks*, Jan. 1996. To appear in *Springer-Verlag Multimedia Systems Journal.*

[6] W. Feng, F. Jahanian, S. Sechrest, "Providing VCR Functionality in a Constant Quality Video-On-Demand Transportation Service", In *Proc. of the International Conf. on Multimedia Computing and Sy stems*, pp. 127-135, July 1996.

[7] W. Feng, J. Rexford, "A Comparison of Bandwidth Smoothing Techniques for the Transmission of Prerecorded Compressed Video", To appear in *IEEE INFOCOM 1997*, April 1997.

[8] J. M. McManus, K. W. Ross, "Video on Demand Over ATM: Constant-Rate Transmission and Transpor t," In *Proc. INFOCOM*, pp.1357-1362, Mar. 1996.

[9] J. Salehi, Z. Zhang, J. Kurose, D. Towsl ey, "Supporting Stored Video: Reducing Rate Variability and End-to-End Resource Requirements Through Optimal Smoothing", In *Proc. SIGMETRICS*, May 1996, pp.222--231.

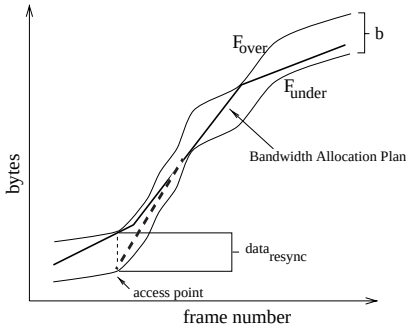

Figure 3: Supporting VCR Functionality: This figure shows the result of a scan to a random "access point" such as a long fast-forward. With no overlap of data in the buffer, the distance between the bandwidth smoothing plan and F_{under} at the access point must be made up in order to continue along the original bandwidth plan. The heavy dotted line shows a sample plan for resynchronizing to the original plan.

Interactive Playback

Interactive playback for buffered, interactive VOD systems is perhaps the hardest resource allocation facing VOD designers. As an illustration, consider the bandwidth smoothing plan in Figure 3. When a random access is made from VCR functions such as a long fast-forward, excess channel capacity may have to be allocated in order to resynchronize the plan with the original peak bandwidth requirement. As another illustration, consider a scan into an area that has a large number of large frame sizes. Under bandwidth smoothing these frames would have been prefetched in order to reduce the bandwidth requirement. However, a random access to these frames will require that (1) excess channel capacity be allocated, (2) reducing the quality of video until the plans are resynchronized, or (3) making the user wait until the buffer is filled. Due to the undeterministic nature of the interactions, providing guaranteed VCR interactivity can be difficult while maintaining a high network utilization. To aid in VCR functionality, ideas such as *contingency channels* can be useful[1], where excess channel capacity is allocated for temporary allocation to VCR functionality. In addition, using a bandwidth algorithm that minimizes both the rate-constraint and buffer residency times can be useful by minimizing the amount of data required on a resynchronization[4]. It may also be useful to use less of the client-side buffer for smoothing and more of the buffer to store data around the point of play for the client, allowing a moderate amount of VCR functions to be handled directly from the client-side buffer. This technique also allows the server greater latitude in responding to the VCR interactions because the VCR interactions can be initially handled by the buffer.

4. Challenges for Interactive Video-on-Demand Systems

For interactive VOD systems, many challenges face video-on-demand researchers. In this section, we briefly mention some of the issues that are of concern.

Smoothing Buffer and VCR Buffer Trade-off

We are currently investigating techniques for balancing between reducing the peak bandwidth requirement and providing VCR functionality. By not using the entire client-side buffer as a smoothing buffer, the minimum bandwidth requirement may not be achieved. However, as has been shown, much of the burstiness in variable-bit-rate video sources can be removed with small amount of buffering, typically on the order of a few megabytes for MPEG encodings and 10-15 megabytes for Motion-JPEG encodings. To aid in this effort, we are looking at ways of evaluating the trade-off between smoothing and VCR buffering, which involves the weighing of costs such as user VCR guarantees, the cost of network bandwidth, and the cost of the smoothing buffer itself.

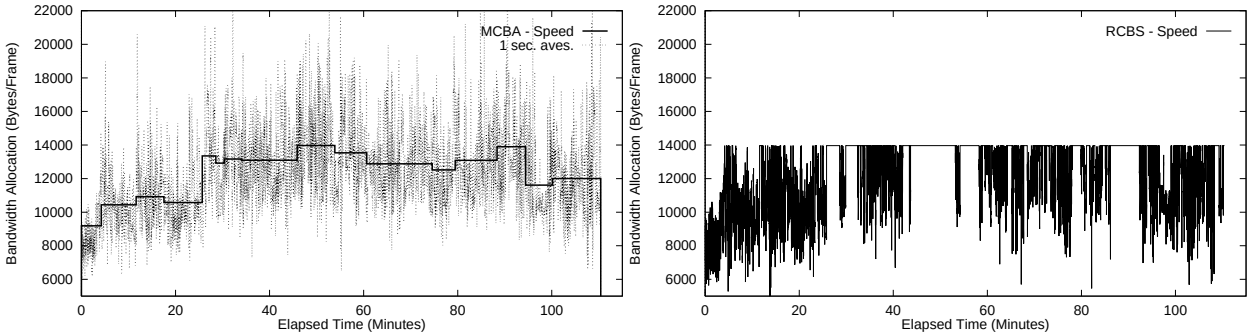

Figure 2: Bandwidth Smoothing Plans: This figure shows example bandwidth smoothing plans for the *minimum changes algorithm* and the *rate-constrained bandwidth smoothing algorithm* for the Motion-JPEG compressed movie *Speed* using a 5 megabyte buffer.

may also minimize the number of rate increases[2, 3], minimize the total number of rate changes[5], minimize the variability of rate changes[9], or minimize the buffer residency times[4] while providing for the continuous uninterrupted playback of video. A more in-depth discussion of the pre-1997 algorithms can be found in an upcoming comparison paper[7]. A sample minimum changes plan is shown in Figure 2. One important trait of these algorithms, however, is that with small amounts of buffering significant reductions in the burstiness of the video delivery are possible. Note that in Figure 2 the one second frame averages represents smoothing across 30 frames. Thus, the actual frame sizes are more bursty than are shown in the figure.

3. Delivering Video in Video-on-Demand Systems

For supporting the playback of stored video streams, I feel that the VOD service must provide for at least three levels of guaranteed service: Strict Playback, Quasi-Interactive Playback, and Interactive Playback. To understand why, we will discuss the necessary support for these three methods

Strict Playback

In the strict playback mode, the user is forced to watch the video from the beginning to end without any change in the consumption rate. Creating bandwidth plans using one of the unlimited buffer algorithms makes resource allocation within the VOD system the simplest due to the fewer number of interactions (and for the CBA algorithm, the monotonicity of bandwidth allocations). In addition, for limited client-side buffers, using one of the minimal peak bandwidth algorithms is attractive for reducing the amount of resources allocated for guaranteed service. As an example, the minimum changes bandwidth algorithm allows the server to allocate resources on the order of tens of minutes with approximately 10 megabytes of client-side buffer and 2 to 3 Mbps streams. Because the bandwidth plans are fixed, the server is free to schedule bandwidth as tightly as possible, using peaks in some plans to fill valleys in others.

Quasi-Interactive Playback

In the quasi-interactive playback mode, the user is allowed to have limited VCR interactions. In this mode, a method called the *VCR window* can be used to allow users full access to their videos[6]. This is based upon the observation that the common VCR functions rewind, examine, stop, and pause can be handled by the client-side buffer *without* requiring additional bandwidth from the network and server to service. The *VCR window* can be appended with additional *VCR buffering* in order to provide a larger VCR window for the user. Because the user can change its consumption rate, the streams cannot be scheduled as tightly as with in the strict playback case. However, in a system where the users abide by the *VCR window* the network and server resources can be allocated based on the peak bandwidth requirement *without* worrying about the user issuing a bandwidth request above and beyond what has been reserved.

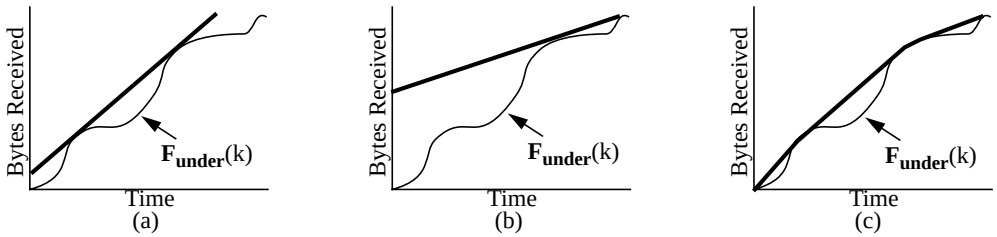

Figure 1: Buffer Unlimited plans. Figure (a) shows a constant bandwidth allocation transmission plan where bandwidth is plentiful. Figrue(b) shows a constant bandwidth allocation where the bandwidth may be limited as in the Internet. Figure (c) shows the critical bandwidth allocation algorithm transmission plan.

succinct, we first discuss what types of bandwidth smoothing algorithm have been presented and then discuss how they relate to resource allocation within video-on-demand (VOD) systems.

2. Creating Bandwidth Allocation Plans

For the discussion of bandwidth smoothing algorithms, two types of client-side buffers are worth discussing. For systems that have large client side buffers, the main concern with the delivery of stored video is to not underflow the client-side buffer. For a video that consists of n frames, where frame i requires f_i bytes of storage, the server must always transmit quickly enough to avoid buffer underflow, where

$$F_{under}(k) = \sum_{i=0}^{k} f_i$$

indicates the amount of data consumed by the client by frame k, where $k=0,1,...,n-1$. Using this function, several plans for the delivery of stored video are possible as shown in Figure 1. For a constant bandwidth allocation, where bandwidth is plentiful, choosing a slope (and hence, bandwidth allocation) such that the line is as close to F_{under} as possible minimizes the amount of buffer for a single constant bandwidth allocation plan. For networks, where the bandwidth may be limited, choosing a slope that is equivalent to the maximum available bandwidth by setting the allocation line tangent to F_{under} provides for a plan of delivery that is continuous with the rate constraint. Note the point where the plan crosses the y-axis is the amount of prefetch required for continuous delivery. Finally, the critical bandwidth allocation (CBA) algorithm creates a convex hull around F_{under} resulting in a plan that has monotonically decreasing bandwidth requirements[2]. Using a CBA algorithm has several advantages. First, the buffer requirement is smaller than or equal to any single constant bandwidth allocation plan. Second, the plan bandwidth requirements are monotonically decreasing, making admission control simpler. The admission control simply involves determining whether there exists enough bandwidth to start the flow of data.

For buffer limited clients, a client should receive no more than

$$F_{over}(k) = b + \sum_{i=0}^{k} f_i$$

by frame k to prevent buffer overflow of the playback buffer (of size b). Consequently, any valid server plan should stay within the *river* defined by F_{under} and F_{over}. That is,

$$F_{under}(k) \le \sum_{i=0}^{k} c_i \le F_{over}(k)$$

where c_i is the transmission rate during frame slot i of the smoothed video stream.

Based on this framework, several algorithms have been developed in order to traverse down the river defined by F_{under} and F_{over}. Most algorithms minimize the peak bandwidth requirement given the fixed client-side buffer[2, 3, 4, 5, 9]. While minimizing the peak bandwidth requirement, a smoothing algorithm

Resource Allocation in Stored Video-On-Demand Systems

Wu-chi Feng
Department of Computer and Information Science
The Ohio State University
Columbus, OH 43210

Abstract

The delivery of prerecorded compressed video streams that are constant-quality encoded has been the focus of much research. Because stored video streams can be analyzed *before* the flow of data begins, a client-side buffer can be used to smooth the bandwidth request for the server as well as network resources. This paper highlights the methods that have been proposed for utilizing a client-side buffer for the delivery of stored video-on-demand systems. In addition, this paper explores the feasibility of providing interactive service within buffered video-on-demand systems.

1. Introduction

The delivery of compressed prerecorded video has received a fair amount of multimedia and networking attention in the recent past. Because the data for a particular movie can be analyzed *a priori* to the actual flow of data, the network and server resource requirements greatly differ for stored and live video applications[2]. Using the *a priori* information, a bandwidth plan for the delivery of the stored video can be created that takes advantage of a client-side buffer. This buffer can be used to prefetch large bursts of frames before they occur, resulting in a *smoothed* video stream. The bandwidth plan that is created, however must limit the amount of prefetching to prevent buffer overflow and must provide enough prefetching to prevent buffer underflow.

Given a fixed size buffer, several *bandwidth smoothing* algorithms have been introduced in the literature that are provably optimal under certain constraints. In particular, plans with the minimum peak bandwidth requirements for the duration of playback are possible, using the fixed client-side buffer. In addition, these algorithms can also minimize the number of rate increases[2, 3], minimize the total number of rate changes[5], minimize the variability of rate changes[9], or minimize the buffer residency times[4] while providing for the continuous uninterrupted playback of video.

In this paper, we highlight some of the major results that have been developed for the delivery of stored video in video-on-demand systems. In addition, we will present some of the challenges left for buffering in video-on-demand systems and some of the first steps in solving some of these problems. In particular, we argue for several key points for resource allocation in buffered, interactive video-on-demand systems. First, video-on-demand systems must provide levels of guaranteed service, allowing the user to choose the amount of interactivity (and hence, the guarantees) that they require. Second, the use of buffering provides a powerful mechanism for the guaranteed delivery of stored video-on-demand services. Finally, while the use of buffering may make providing interactive control such as fast-forward and rewind operations harder, the benefits of buffering are too important to go unused.

In the following section, we discuss some of the work that has appeared in the literature and how it fits in to the delivery mechanism for the delivery of stored video across networks. We will then take a look at the problem of providing interactive control in these environments. To make the discussion a little more