Real-Time Video on the Web using Dynamic Rate Shaping

Stephen Jacobs and Alexandros Eleftheriadis
Department of Electrical Engineering
Columbia University
New York, NY 10027

Abstract

We propose an architecture which can support video services in networks which have no Quality of Services (QoS) guarantees. We demonstrate that there is a need for such services today and also in the future, when QoS will be built into most networks. We then describe our current Internet-based system, which combines both image processing and networking techniques in order to provide good quality video without harming the performance of the Internet. These techniques are shown to be superior to previous work for the environment of video services operating in the Internet. We also present experimental results gathered in both a controlled environment and in external wide area connections across the Internet.

1 Introduction

The underlying technologies of today's Internet are not sufficient to support Quality of Service (QoS) guarantees. The evolution of the base technologies of the Internet could result in any number of possibilities, including ATM backbones, IP switching, or fully deployed ATM networks. Regardless of the specifics, the general consensus is that the network infrastructure of the future will have QoS and that users will be able to request connections with or without QoS.

Certainly connections which reserve resources will demand a higher cost and users may not always want to pay this extra cost for a particular video service. If a user is watching a movie, perhaps it would be worthwhile. However, just browsing a video database may not warrant the extra cost since the user may browse many video streams before deciding on the one that is being sought. Thus, some video services may not warrant the extra cost of QoS.

Another motivation for being able to provide good quality video services without QoS is that some networks may never be able to provide QoS guarantees. Wireless networks fit into this category because there may be no way to guarantee a certain bandwidth when so many other variables exist that affect the throughput on the wireless link.

The technique for developing video services without QoS involves an explicit attempt at avoiding network congestion. Clearly, network congestion hurts the performance of all users of the network. The goal is to send only the data that can fit into the network at a particular time. This requires both a networking and an image processing approach. From the networking perspective, an estimate of the available bandwidth in the network must be found. Then, from the image processing perspective, a technique for shaping the compressed video into that available bandwidth is necessary. In the next sections we discuss previous work as well as our proposal for both the networking and image processing techniques.

2 Internet Video Architecture

Figure 1 shows a generic architecture for supporting adaptable media applications on networks without QoS. The server consists of three main parts. The first one is the part which can shape a specific media into a desired rate. In the case of video this might be frame dropping, as previously mentioned. The second section is the media pump which reads data from the buffer that contains the adaptable media and sends the data into the network using RTP/UDP/IP. The congestion window is the third part of the server. The media pump only sends out data when the congestion window deems it appropriate. The client and/or the network provide feedback which is used to determine the congestion window.

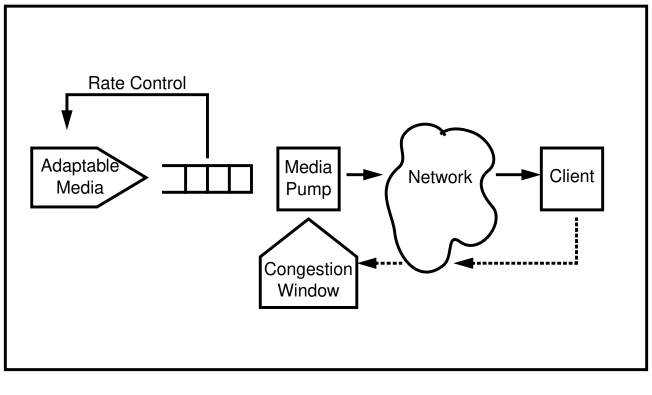

Figure 1: Internet video system architecture.

If the adaptable media is filling the buffer at a rate, R, and the congestion window is allowing the media pump to send data out only at an average rate $S < R$, then the buffer will begin to fill. This information about the buffer occupancy is fed back into the adaptable media object to decrease the rate. The rate control algorithm, which uses information about buffer occupancy to change the rate of the media, is an essential part of the system.

3 Bandwidth Adaptation

As real-time services have become more prevalent over the last few years, there has been a growing concern that the current Internet infrastructure may not be able to support them, leading perhaps, to a "congestion collapse." This is a valid concern in general since many real-time services send their bits through the network without concern for congestion control or avoidance.

Network bandwidth estimation has been explored extensively. Many algorithms provide a somewhat different view of the state of the network [2, 3, 10, 11]. Despite the different techniques for finding the available bandwidth, probably the most important factor in designing a congestion avoidance algorithm is that it does not degrade the quality of other traffic on the network. For this reason, it is important to ensure that the technique used provides fairness among the different traffic types, i.e. ftp, http, audio, video, etc.

Our architecture uses the TCP Congestion Control (TCP-CC) algorithm as a congestion to form the congestion window. TCP is not used for transport and retransmissions are performed selectively. An algorithm which always retransmits increases the delay, which in general, is unacceptable for real-time applications and certainly for video.

The reason for using TCP-CC is that TCP streams work well together and today's Internet is proof of that. They operate according to a greedy but "socially-minded" and cooperative algorithm which attempts to get as much bandwidth as possible, but backs off substantially during congestion. Using TCP-CC can make real-time traffic look as harmless as a file transfer to the network and still maintain relatively low delay [7].

The Internet is only one type of network which has no QoS guarantees. ATM-ABR provides a guaranteed minimum QoS, but informs the server when more bandwidth is available. In this case, the bandwidth estimate is explicit in that the network furnishes the exact bandwidth which is available. Also, as mentioned earlier, wireless may never have QoS guarantees.

4 Adaptable Media

Just as the Internet is one type of network which has no QoS guarantees, video is only one type of media which can be adapted on the fly. For environments where minimal delay is more important than perfect quality, the rates of both audio and still images can be adapted. Thus, although this paper focuses on video and the Internet, some of the concepts also apply to any combination of non-QoS network and adaptable media, i.e. wireless and audio.

Techniques for adapting compressed video on the fly have been limited in the past. One technique is to adjust the quantization parameters at the encoder based on the state of the network [9]. Although this works quite well for live video, it cannot work for precompressed streams. Another commonly used technique for changing the bit rate of video is to drop frames [4]. However, frame dropping alone is a crude technique which provides only a coarse approximation to the available bandwidth since the smallest unit of data which can be removed is an entire frame. Although subjective tests have not been completed, it seems intuitive that very low frame rate video is perceived as less valuable for many applications.

To solve this problem, we developed a technique called Dynamic Rate Shaping (DRS) [5, 6]. DRS provides the ability to dynamically change the bit rate of a precompressed stream. In its simplest form, DRS selectively drops coefficients from the MPEG-1 or MPEG-2 bit stream which are least important in terms of image quality.

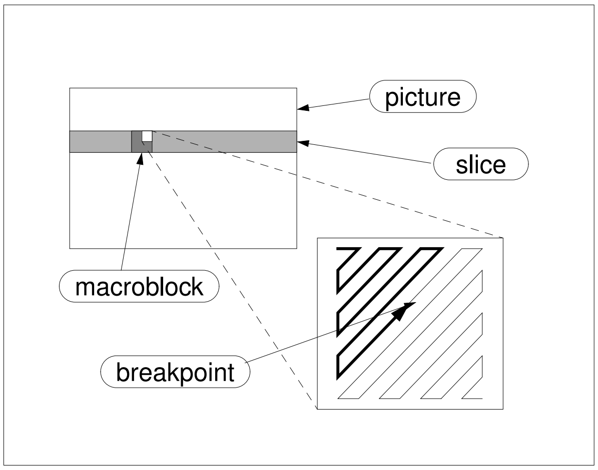

Figure 2: DRS operation.

Figure 2 shows how DRS works. In MPEG, each picture consists of one or more slices and each slice consists of one or more macroblocks. A macroblock consists of four blocks which have been DCT compressed and run-length coded. The low frequency coefficients, which are more important for subjective image quality, are in the upper left of the block, while the high frequency coefficients are in the lower right. The DCT coefficients are encoded in a zigzag fashion to improve compression efficiency.

DRS operates on the compressed video bit stream, eliminating DCT coefficients run-lengths. The coefficients to be eliminated are determined using Lagrangian optimization, resulting from an operational rate-distortion formulation. The problem is complicated by the fact that MPEG coding utilizes predictive and interpolative modes for motion compensation, and thus any modification of the bit stream will result in error propagation. We have experimentally shown in [5] that, if decisions within each frame are optimal, then ignoring the accumulated error does not impact the quality in any way. Thus the algorithm can operate in a memoryless mode which has significantly less complexity, while achieving essentially optimal (within 0.3 dB) performance. Also, DRS is much less complex than a complete decoder, making it feasible to run on a video server.

Of course, there is overhead in MPEG so that a lower bound exists on the rate of the shaped bit stream. The lower bound is reached once every coefficient except the DC coefficients are dropped from every block. At this point, further reduction can be achieved only by frame dropping, barring any recoding operations.

DRS can meet any reasonable bandwidth estimate exactly. This means that the bandwidth estimate is more fully utilized. It also maintains the original frame rate of the video. Lastly, DRS decouples the adaptable media from the encoder so that it can be used for both live and stored video streams. A combination of DRS and frame dropping will probably yield better perceptual results than either technique working alone, especially for large rate reductions.

5 Rate Control

The rate control algorithm provides updated rate information to the adaptable media source based on the buffer occupancy between the adaptable media and the media pump. Other work has been done in this area, but primarily in the context of feedback to an encoder [1, 8]. The goals for a rate control algorithm in this environment are to:

1. prevent buffer overflow, which will cause delays,
2. prevent buffer underflow, unless the adaptable media source is at the maximum rate,
3. keep the buffer at a desired occupancy, B_d,
4. converge quickly to a new output rate, and
5. minimize the size of the oscillations around the new output rate.

Since MPEG video consists of several different frame types whose sizes vary greatly, estimates of the buffer occupancy must be taken as averages over no less than a one second interval to avoid momentary fluctuations in the buffer occupancy due to variations in the input rate. In our case, we are using a five and sometimes ten second interval to get further smoothing due to variations in the output rate. The goal of the smoothing is to obtain a trend in the buffer occupancy, not an instantaneous occupancy. The rate control algorithm is invoked at the end of each interval. Another impetus for choosing a large interval is that rapidly changing image quality is perceptually disagreeable.

6 Performance Results

To verify the effectiveness of our system, we ran two sets of experiments. The first set consisted of a client and server with a controlled bottleneck in between. The bottleneck was set up with a certain bottleneck rate and a maximum buffer size. It reads in packets destined for the client and adds them to the queue. At the same time, it sends out packets onto the network as if the network were operating at the bottleneck rate. It does this by delaying subsequent packets until the time that it would take to send a packet at the bottleneck rate. If the incoming rate is faster than the bottleneck rate, the queue will start to build. If the buffer size is then exceeded, packets are dropped.

6.1 Controlled Bottleneck

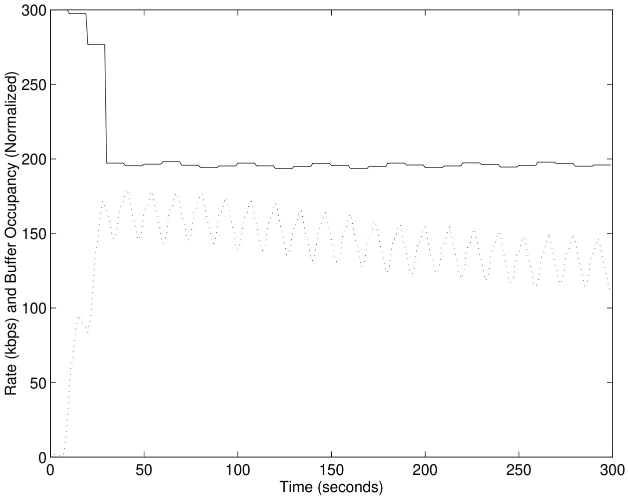

Figure 3: The rate (solid) and buffer occupancy (dashed) for a stream originally encoded at 300 kbps going through a 200 kbps bottleneck. The buffer occupancy has been normalized to fit on the same graph as the rate.

Figure 3 shows the effects of an MPEG-2 stream originally encoded at 300 kbps going through a 200 kbps simulated bottleneck with a buffer depth of 10 packets.[1] The duration of the connection is five minutes. The convergence of the rate is quite fast; part of the delay is the time to attain the desired buffer occupancy, which is half of the total buffer size, located at 150 on this figure. Note also, that the oscillations in the rate are very small, only 5 kbps peak to peak.

The oscillations seen in the buffer occupancy of Figure 3 are due to the fact that the rate only changes every 10 second in this case. But they are not important because it is the rate that must stay stable, not the buffer occupancy.

6.2 Wide Area Network

The second set of experiments was performed using the Internet itself. The stream was sent from a computer in our laboratory in New York to a computer located 13 hops away in New Jersey, where the packets were read in and sent back immediately to a client in our laboratory again, traversing another 13 hops on the way back. This experiment is shown in Figure 4. Again the original stream is 300 kbps MPEG-2.

[1]The original and rate shaped streams can be retrieved from ftp://wakko.ctr.columbia.edu/share. The files names are benhur.300.mpg and benhur.200.mpg.

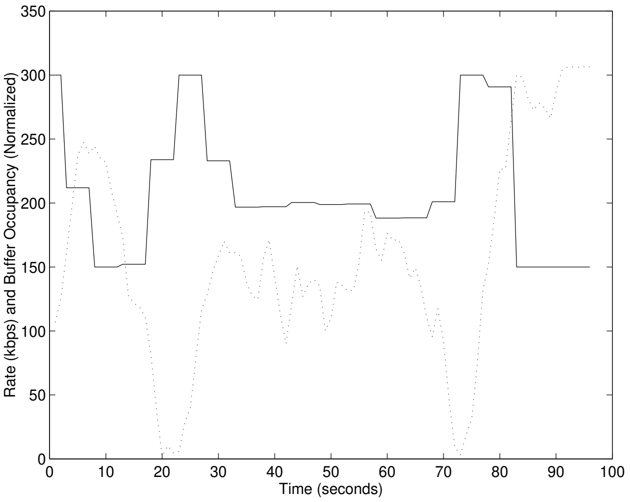

Figure 4: The rate (solid) and buffer occupancy (dashed) for a stream originally encoded at 300 kbps going through the Internet, traversing 26 hops. The buffer occupancy has been normalized to fit on the same graph as the rate.

The experiment was run over a 20 minute period, of which this is a 100 second excerpt. The rate and the buffer occupancy vary substantially, as one would expect given the rate fluctuations in the Internet.

It is important to note from this figure that the reactions of the rate control algorithm are consistent with the criterion described in Section 5. There is no buffer underflow except when the rate is at the maximum, which is acceptable. There is however buffer overflow which can be seen between 90 and 100 seconds. At this point, although the rate has gone down to the lower limit of 150 kbps, the buffer continues to fill. When the occupancy reaches the maximum, DRS must wait for the buffer to empty before proceeding, which introduces visible delays at the client. This experiment was done using only DRS and not frame dropping. If frame dropping were used too, the rate could be reduced even further.

At 10 seconds, the buffer begins to empty. Since the rate control algorithm tries to maintain an occupancy of half the maximum, it increase the rate at 20 seconds. But then, just before 30 seconds, the buffer begins to fill again and the rate must cut back. During the next 30 seconds, the algorithm has found a rate which maintains the desired occupancy, since the available bandwidth in the network has momentarily stablized.

7 Concluding Remarks

We have presented a novel architecture for supporting video services in a non-QoS network, namely the Internet. Several experiments have been performed to verify performance in both a controlled environment and in external wide area connections in the Internet. Our results show that our proposed architecture provides a stable system with good quality video which does not degrade the quality of other Internet traffic.

References

[1] R. Bollow, *Video Transmission Using the Available Bit Rate Service*, Master's Thesis, Berlin University of Technology.

[2] L. S. Brakmo and S. W. O'Malley, "TCP Vegas: New techniques for congestion detection and avoidance," in *SIGCOMM Symposium on Communications Architectures and Protocols*, London, United Kingdom, Aug. 1994, pp. 34–35.

[3] I. Busse, B. Deffner, and H. Schulzrinne, "Dynamic QoS control of multimedia applications based on RTP," in *First International Workshop on High Speed Networks and Open Distributed Platforms*, St. Petersburg, Russia, June 1995.

[4] Z. Chen, S. M. Tan, R. H. Campbell and Y. Li, "Real Time Video and Audio in the World Wide Web," in *World Wide Web Journal*, January 1996, Volume 1.

[5] A. Eleftheriadis and D. Anastassiou, "Constrained and General Dynamic Rate Shaping of Compressed Digital Video," in *Proceedings, 2nd IEEE International Conference on Image Processing*, Washington, DC, October 1995, pp. III.396–399.

[6] A. Eleftheriadis and D. Anastassiou, Meeting Arbitrary QoS Constraints Using Dynamic Rate Shaping of Coded Digital Video, in *Proceedings, 5th International Workshop on Network and Operating System Support for Digital Audio and Video*, Durham, New Hampshire, April 1995, pp. 95–106.

[7] S. Jacobs and A. Eleftheriadis, "Providing Video Services over Networks without Quality of Service Guarantees," in *World Wide Web Consortium Workshop on Real-Time Multimedia and the Web*, Sophia-Antipolis, France, October 24-25, 1996.

[8] H. Kanakia, P. Mishra, and A. Reibman, *An adaptive congestion control scheme for real-time packet video transport*, in SIGCOMM Symposium on Communications Architectures and Protocols, San Francisco, California, Sept. 1993, pp. 20-31.

[9] A. Ortega and M Khansari, "Rate Control for Video Coding over Variable Bit Rate Channels with Applications to Wireless Transmission," in *Proceedings of the 2nd IEEE International Conference on Image Processing (ICIP'95)*, Washington, DC, Oct 1995.

[10] K. K. Ramakrishnan and R. Jain, "A binary feedback scheme for congestion avoidance in computer networks," in *ACM Transactions on Computer Systems*, vol. 8, May 1990, pp. 158–181.

[11] T. Sakatani, "Congestion avoidance for video over IP networks," *Multimedia Transport and Teleservices Proceedings*, Nov. 13-15, 1994. pp. 256–273.