CONGESTION-DISTORTION OPTIMIZED SCHEDULING OF VIDEO OVER A BOTTLENECK LINK

Eric Setton and Bernd Girod

Information Systems Laboratory, Stanford University,
{esetton, bgirod}@stanford.edu

ABSTRACT

Recent research on multimedia scheduling has focused on minimizing the rate-distortion cost of transmission policies. This tradeoff only partially reflects the network congestion a media stream may generate on bandwidth-limited channels. We introduce the concept of congestion-distortion optimized streaming and propose a scheduler which attempts to minimize the Lagrangian cost of end-to-end delay and media distortion. Experiments for video on a simulated network illustrate the performance of the scheduler. Compared to a state-of-the-art scheduler, the proposed algorithm reduces end-to-end delay by approximately 50% while improving the video quality by up to 3 dB.

1. INTRODUCTION

The growth of multimedia streaming over wired and wireless networks has been a major incentive for finding standard-compatible optimized algorithms to enhance the performance of traditional communication systems. In a typical low latency scenario, packets of encoded media only have a short amount of time to reach the receiver before being displayed. Hence, particular care needs to be dedicated to packet scheduling in order to prioritize transmissions and enable potential retransmissions of important packets.

A framework for performing rate-distortion optimized scheduling of multimedia (RaDiO) is receiving increasing attention in the video streaming community. This scheduling method considers the unequal contribution of different portions to the overall distortion of a multimedia data stream. Its aim is to find an optimal schedule for the packets of a stream, which minimizes the expected Lagrangian cost of rate and distortion. In [1], the formalization of this scheduling process is presented and an optimization procedure which overcomes the exponential complexity of the problem is described. Substantial gains using a RaDiO scheduler are reported for video streaming, notably in [2] and in [3] where a more realistic video distortion model is proposed.

Streaming a real-time media stream might increase the congestion of bottleneck links. This may disrupt the transmissions of other users and even delay the delivery of the media stream itself. This effect is not adequately captured by a rate-distortion (R-D) tradeoff. A potential solution to this problem is to design a streaming system which abides to TCP friendliness at the transport layer [4]. We propose to incorporate congestion control in the scheduler itself and perform congestion-distortion optimized (CoDiO) scheduling of multimedia. The purpose of this work is to analyze the advantages of using a congestion-distortion tradeoff instead of a R-D tradeoff. As congestion is an increasing convex function of the transmission rate, the scheduler should achieve similar R-D performance to RaDiO. In addition, it is expected to shape the traffic optimally to reduce delay over bottleneck links.

In the following, we describe the network scenario we consider and explain the channel model used to estimate the delay distribution of packets sent over the network. In Section 3, we present the congestion-distortion optimized (CoDiO) scheduling algorithm for video and show how to replace rate with congestion in the framework proposed in [1]. In Section 4, we analyze experimental results obtained on a simulated network and compare the performance of CoDiO and RaDiO for different scenarios.

2. CHANNEL MODEL

We consider the route between a video server and a client as a succession of high bandwidth links shared by many users and ended by a bottleneck last hop. The delay over the first portion of the path may be modelled as a random variable following the gamma distribution [5], whereas the delay on the last hop is determined by the capacity of the link and the size of the queue. The resulting end-to-end delay between the server and the

This work was funded in part by Sony Electronics through the Stanford Networking Research Center and in part by NSF grant CCR-0325639.

client is then a gamma distribution parameterized by a time-varying shift reflecting the delay at the bottleneck. This shift reflects the variations of the size of the bottleneck queue, depicted in Fig. 1, which can be estimated given the arrival times of packets at the bottleneck, their sizes which we will denote throughout the paper by B_i and the capacity of the link denoted by C.

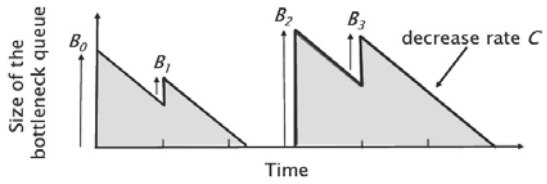

Fig. 1. Backlog at the bottleneck queue

The capacity of the bottleneck link may be estimated by transmitting back-to-back packets over the network as in [6]. For simplicity, we assume throughout the paper that this value is known at the server and that the last hop is not shared with any other streams. In addition we also assume that the server has an accurate estimate of the parameters of the gamma distribution characterizing the high bandwidth links as in [1, 2, 3]. Given this information and the history of past transmissions, the server may compute at each time an approximate delay distribution and use this distribution to estimate the probability that a packet has arrived at the client.

3. CONGESTION-DISTORTION OPTIMIZED SCHEDULING

In this section, we describe how to determine an optimal transmission schedule for the packets of a video stream. This schedule indicates when the packets of the stream will be sent from the server to the client, if at all. To limit the exponential number of possible schedules, discrete transmission times are used and the time horizon covered by the schedule is limited. Furthermore, rather than optimizing jointly the schedule for all the packets of the stream, only a small number of packets are selected and the optimization is performed iteratively for each packet.

The aim of CoDiO is to determine a schedule minimizing the expected Lagrangian cost $D + \lambda\Delta$, where D is the distortion of the received video stream and Δ is the end-to-end delay which serves as the congestion metric. To minimize this objective function, CoDiO selects the most important packets in terms of video quality, and transmits them in an order which minimizes the average backlog of the bottleneck queue. For example, the I frames of a video stream will be transmitted in priority whereas B frames might be dropped. In addition, CoDiO will avoid transmitting packets in large bursts as this has the worse effect on the queuing delay. In the following, we describe how to evaluate the expected end-to-end delay and distortion corresponding to a given transmission schedule of a set of l packets. This elementary step is repeated several times to evaluate the performance of different schedules and choose the schedule which performs best.

3.1. Determining the end-to-end delay

Unlike rate in the RaDiO framework, end-to-end delay is not additive and the contribution of each packet cannot be derived separately. For a given transmission schedule, the rate output by the server may be used to derive the size of the bottleneck queue as a function of time. This in turn leads to the average value of the end-to-end delay over the time horizon considered. In this section we present in more detail each of these steps.

We denote by t_j successive transmission times and by $\pi_i(t_j)$ binary variables representing whether or not the transmission of packet i of size B_i is scheduled at time t_j. The transmitted rate at time t_j is:

$$R(t_j) = \sum_{i=1}^{l} \pi_i(t_j) B_i \tag{1}$$

For clarity, Eq. (1) does not explicitly consider retransmissions. In particular, a retransmitted packet may be acknowledged before the end of the time horizon and removed from the transmission queue regardless of the transmission schedule. To keep the rate estimation accurate, we only consider the schedule of a packet until its most probable acknowledgement time.

The additional information needed to derive the size of the bottleneck queue is the time each packet reaches the bottleneck. In this derivation, we assume the delay introduced by the high bandwidth links is constant and equal to the mean of the actual delay distribution. This approximation is valid when the variance is limited. Knowing the capacity of the bottleneck link, the transmitted rate and this constant delay, the size of the queue may easily be computed. A typical illustration of the size of the queue as a function of time is shown in Fig. 1.

The corresponding average end-to-end delay is simply the sum of the average value of this function taken over the time horizon considered and of a constant term which reflects the delay over the high bandwidth links.

3.2. Determining the video distortion

The expected value of the distortion for the video stream decoded by the client is computed as in [3]. Namely, if copy error concealment is used, an undecodable frame is replaced with the nearest correctly decoded frame for display. Hence, to capture the effect of packet loss on the video quality, only a limited number of display outcomes need to be identified and associated with different distortions. Let $D(s, f)$ denote the distortion resulting from substituting frame s to frame f, the expected distortion when displaying frame f is:

$$D(f) = \sum_{s} D(s, f) Pr\{s\} \qquad (2)$$

In Eq. (2), $Pr\{s\}$ represents the probability that frame s is displayed instead of f. This probability may be computed, as described in [3], by combining the probabilities that different packets do not reach the client by their playout deadline.

4. EXPERIMENTS AND RESULTS

The performance of both RaDiO and CoDiO was evaluated by simulating the transmission of a video stream in the network simulator ns-2. The results presented are for a network path with two hops. The first hop is a high-bandwidth 47.5Mbps T3 link which is filled with a 22Mbps flow of exponential cross-traffic. The second is a low bandwidth 50 kbps link which only carries video traffic. In the experiments, both schedulers perform predictions of the delay distribution obtained by dynamically estimating the time varying backlog of the bottleneck, as described in Section 2. The difference between these two schedulers resides only in the cost function they seek to minimize. For RaDiO this is the Lagrangian cost of rate and distortion $D + \lambda R$, for CoDiO of end-to-end delay and distortion $D + \lambda\Delta$.

Encoded video streams are generated by encoding the sequences *Foreman* and *Mother and Daughter* with the H.263+ video codec using a 2-layer encoding at 10 frames per second. For these sequences the group of picture (GOP) length is 20 and the GOP structure is IPPP... Enhancement layer frames are predicted by both the base layer frame and the previous enhancement layer frame. The rate for both the base layer and the enhancement the layer is 32kbps for the *Foreman* sequence and 32kbps and 37kbps respectively for the *Mother and Daughter* sequence[1]. The playback delay is fixed at 600 ms and video quality is measured in terms of PSNR of the decoded luminance component. Average congestion is measured by computing the average end-to-end delay of small probe packets sent every 10 ms over the channel.

[1]Note that if no scheduling was used the total bit-rate would overwhelm the bottleneck link.

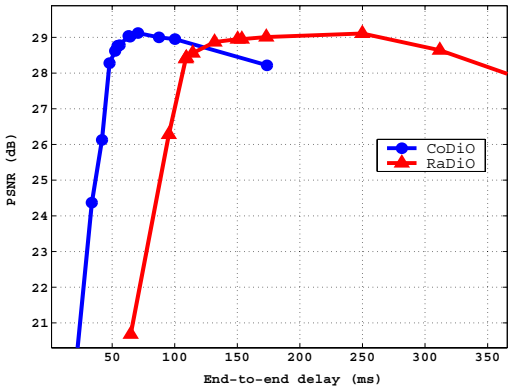

Fig. 2. $\Delta - D$ performance comparison

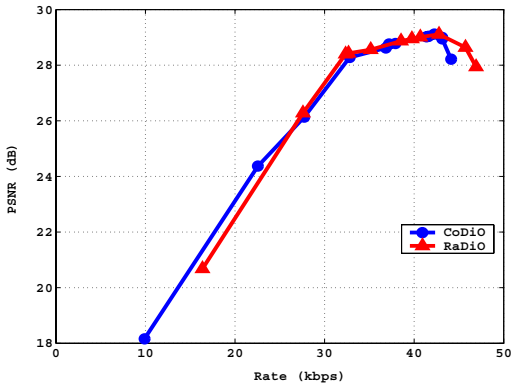

Fig. 3. $R - D$ performance comparison

The graphs in Fig. 2 and Fig. 3 show the congestion-distortion and the rate-distortion performance of CoDiO and RaDiO. The different points are obtained by giving different values to the Lagrange multiplier. When λ is large, more importance is given to self generated congestion (for CoDiO) or to rate (for RaDiO); only a small number of packets are transmitted and video quality is limited. When λ is small, hardly no control is imposed on the output rate of the schedulers. This creates congestion at the bottleneck link and affects performance by generating late loss. Optimal performance is achieved, for both schedulers for intermediate values of λ. The R-D performance for both schedulers is almost identical. In addition, for a given video quality, CoDiO reduces the end-to-end delay by approximately half, as illustrated in Fig. 2. The reason for this performance improvement resides in the optimal rate shaping performed by CoDiO and shown

in Fig. 4. The transmissions for RaDiO often occur in bursts which are reflected by the large delay spikes seen in Fig. 4. For CoDiO the only delay spikes are caused by the transmission of large I frames. The rest of the time, the rate is smooth and the queue backlog is low. As a consequence, a user sharing the bottleneck bandwidth with a video stream transmitted using CoDiO would experience much less delay over this link. Figure 5 shows comparable improvements for CoDiO for the case when there is a 5% packet erasure rate on the bottleneck link.

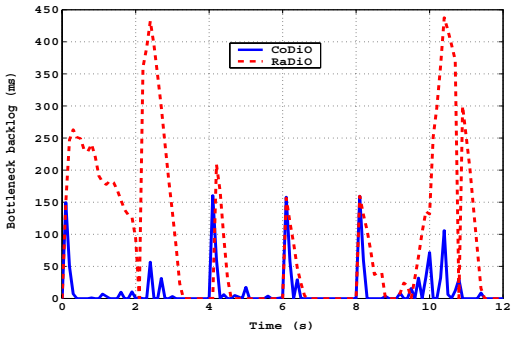

Fig. 4. Size of the bottleneck queue

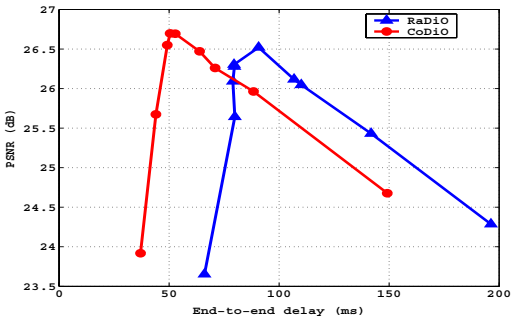

Fig. 5. 5% packet erasure rate on the last hop

Uncontrolled self-generated congestion may also have negative effects on the application itself. Figure 6 illustrates the congestion-distortion of CoDiO and RaDiO for the *Mother and Daughter* sequence. For this video clip, the instantaneous bit rate is such that by optimally shaping the rate, the scheduler yields an increase in the maximal achievable quality of 3 dB while also significantly reducing end-to-end delay.

5. CONCLUSIONS

We present a congestion-distortion optimized scheduling algorithm for video streaming. In lieu of bit-rate, the scheduler strives to minimize end-to-end packet delay while also minimizing distortion. Experiments over

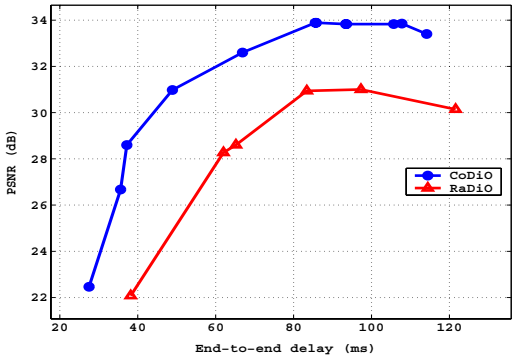

Fig. 6. $\Delta - D$ performance comparison for the *Mother and Daughter* sequence

a simulated network illustrate improved performance over the state-of the art scheduling algorithm RaDiO. At a constant rate, the proposed scheduler reduces the average end-to-end delay by approximately 50%. By reducing self-generated congestion it also improves video quality by up to 3 dB.

6. ACKNOWLEDGMENTS

The authors would like to thank Mark Kalman for providing the code for his implementation of RaDiO.

7. REFERENCES

[1] Philip A. Chou and Zhourong Miao, "Rate-distortion optimized streaming of packetized media," *Microsoft Research Technical Report MSR-TR-2001-35*, Feb. 2001.

[2] Jacob Chakareski and Bernd Girod, "Rate-distortion optimized packet scheduling and routing for media streaming with path diversity," *Proc. IEEE Data Compression Conference, Snowbird, UT*, Apr. 2003.

[3] Mark Kalman, Prashant Ramanathan, and Bernd Girod, "Rate-distortion optimized streaming with multiple deadlines," *Proc. International Conference on Image Processing, Barcelona, Spain*, Sept. 2003.

[4] I. V. Bajic, O. Tickoo, A. Balan, S. Kalyanaraman, and J. W. Woods, "Integrated end-to-end buffer management and congestion control for scalable video communications," *Proc. IEEE ICIP 2003, Barcelona, Spain*, vol. 3, pp. 257–260, Sept. 2003.

[5] A. Mukerjee, "On the dynamics and significance of low frequency components of internet load," *Internetworking: Res. and Experience*, vol. 5, pp. 163–205, Dec. 1994.

[6] Vern Paxson, *Measurement and analysis of end-to-end Internet dynamics*, Ph.D. dissertation UC Berkeley, 1997.