LOW-COMPLEXITY RATE-DISTORTION OPTIMIZED VIDEO STREAMING

Jacob Chakareski[*†]*, John Apostolopoulos*[†]*, and Bernd Girod*[*]

[†]Streaming Media Systems Group
Hewlett-Packard Labs, Palo Alto, CA 94304

[*]Information Systems Laboratory
Stanford University, Stanford, CA 94305

ABSTRACT

This paper proposes two techniques for low-complexity rate-distortion (R-D) optimized streaming of packetized video. These techniques enable computing packet transmission schedules which satisfy a constraint on the average transmission rate while at the same time minimizing the average end-to-end distortion. Optimized packet schedules are computed with considerably lower complexity as compared to conventional algorithms for R-D optimized streaming, which makes these techniques suitable for on-line optimized streaming. Simulation experiments examine the performance of the proposed techniques using JVT/H.264 encoded video sequences and previous frame error concealment. The two techniques demonstrate substantial performance gains of 2-8 dB over a conventional streaming system that is not R-D optimized, which corresponds to a significant fraction of the gain achieved by current (high-complexity) R-D optimized schemes. Furthermore, this performance improvement is achieved with a complexity comparable to that of the conventional, non-R-D optimized, system.

1. INTRODUCTION

The basic goal of video streaming over the Internet is to maximize the reconstructed quality at the receiver, while overcoming the challenges of time-varying throughput, packet loss, and delays. A recent advance in streaming technology is the emergence of Rate-Distortion Optimized (RaDiO) streaming techniques [1–3] that take into account packet importance and knowledge about the channel in a Lagrangian rate-distortion cost function $J = D+\lambda R$. In this approach, schedules of packet transmissions are computed such that a constraint on the average transmission rate is met while minimizing at the same time the average end-to-end distortion. The performance improvements of the RaDiO techniques reported to date relative to non-Lagrangian heuristics are very encouraging.

A framework for RaDiO sender-driven streaming of packetized media has been proposed in [3]. The flexibility of the framework has allowed its application to a number of streaming scenarios such as [4]. Still, there were some important limitations of the initial framework that were overcome by an advanced framework for RaDiO video streaming proposed in [5], which was subsequently extended to cover streaming over multiple network paths, distributed streaming from multiple servers, streaming from an intermediate network proxy, and streaming with rich acknowledgements [6]. In general, however, the performance improvements due to the RaDiO streaming come at the price of increased computational complexity due to the optimization framework employed for computing the optimal schedules. This effect is exacerbated by the fact that optimal packet schedules need to be recomputed at every new transmission instance of video packets. Therefore, conventional RaDiO techniques may be too complex for current video streaming systems. To address this issue this paper proposes two techniques for low-complexity RaDiO streaming. The techniques again compute optimal packet schedules in a Lagrangian framework, however with a dramatically reduced complexity as compared to the conventional RaDiO techniques.

This paper continues by presenting our model for the communication process in Section 2, which is used by some of the systems when computing packet schedules. Section 3 presents the two proposed techniques for low-complexity RaDiO streaming. The complexity of the various techniques is briefly described in Section 4, and Section 5 examines and compares the performance of the various streaming techniques.

2. CHANNEL CHARACTERIZATION

The forward and the backward channel on a network path between a server and a client are modeled as independent time-invariant packet erasure channels with random delay. Hence, they are completely specified with the probabilities of packet loss ϵ_F and ϵ_B, and the probability densities of the transmission delay p_F and p_B, respectively. This means that if the media server sends a packet on the forward channel at time t, then the packet is lost with probability ϵ_F. However, if the packet is not lost, then it arrives at the client at time t', where the forward trip time $FTT = t' - t$ is randomly drawn according to the probability density p_F. Therefore, we let $P\{FTT > \tau\} = \epsilon_F + (1 - \epsilon_F) \int_\tau^\infty p_F(t)dt$ denote the probability that a packet transmitted by the server at time t does not arrive at the client application by time $t + \tau$, whether it is lost in the network or simply delayed by more than τ. Then similarly, $P\{BTT > \tau\} = \epsilon_B + (1 - \epsilon_B) \int_\tau^\infty p_B(t)dt$ denotes the probability that an acknowledgment transmitted by the client at time t does not arrive at the server by time $t + \tau$, whether it is lost in the network or simply delayed by more than τ. Finally, these induce a probability $\epsilon_R = 1-(1-\epsilon_F)(1-\epsilon_B)$ of losing a packet either on the forward or backward channel, and a round trip time distribution $P\{RTT > \tau\} = \epsilon_R+(1-\epsilon_R) \int_\tau^\infty p_R(t)dt$, where $p_R = p_F * p_B$ is the convolution of p_F and p_B. Note that $P\{RTT > \tau\}$ is the probability that the server does not receive an acknowledgement packet by time $t + \tau$ for a data packet transmitted at time t.

3. TWO TECHNIQUES FOR LOW-COMPLEXITY RADIO STREAMING

We first introduce some necessary notation. Let there be L frames in the video sequence. In the following, we use the terms frames

Jacob Chakareski was a summer researcher at HP Labs, Palo Alto. The authors would like to thank Wai-tian (Dan) Tan and Susie Wee of HP Labs, Palo Alto, for the useful discussions on the present work.

and packets interchangeably, as it is assumed that every video frame comprises a single transmission packet. Note, however, that our analysis also applies to the more general case when video frames are packetized into multiple packets. Now, let $D(k)$ denote the total MSE distortion that afflicts a video sequence associated with the single isolated loss of frame k. Figure 1 illustrates the distortion $D(k)$ caused by losing frame k.

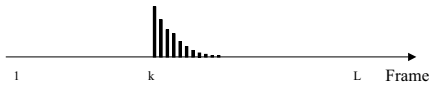

Fig. 1. Loss of single frame k induces distortion in later frames. $D(k)$ is the total distortion summed over all affected frames.

The distortion $D(\boldsymbol{k})$ associated with a packet loss pattern $\boldsymbol{k} = (k_1, k_2, \ldots, k_n)$ of length n is simply modeled as

$$D(\boldsymbol{k}) = \sum_{i=1}^{n} D(k_i).$$

Note that the above model assumes additivity of the distortions associated with individual packet losses, ignoring interdependencies between the effects of lost packets, which does not necessarily hold true when individual packet losses are not spaced sufficiently far apart with respect to the intra-refresh period, as recognized in [7]. Still, due to its simplicity and convenience for mathematical manipulations the additive model has found a number of applications in streaming and modelling of packetized media, such as [8–10]. Therefore, we also employ this model in our two techniques for low-complexity (LC) RaDiO streaming, as discussed next.

3.1. Technique 1: LC RaDiO 1

Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_L)$ be the vector of transmission schedules or policies, one for every frame in the video sequence. Then, using our additive model from above, we define the expected distortion $\mathcal{D}(\boldsymbol{\pi})$ for the video presentation to be as follows

$$\mathcal{D}(\boldsymbol{\pi}) = \sum_{l=1}^{L} D(l)\epsilon(\pi_l),$$

where $\epsilon(\pi_l)$ is the expected error, or the probability that frame l is not delivered to the client on time given the transmission policy π_l. Similarly, we define $R(\boldsymbol{\pi}) = \sum_{l=1}^{L} B_l \rho(\pi_l)$ to be the transmission rate induced by the policy vector $\boldsymbol{\pi}$, where B_l is the size of frame l in bytes and $\rho(\pi_l)$ is the expected cost, or the expected number of transmitted bytes per source byte (under policy π_l).

Formally, we are interested in finding the policy vector $\boldsymbol{\pi}$ that minimizes $\mathcal{D}(\boldsymbol{\pi})$ subject to a constraint on $R(\boldsymbol{\pi})$. As in conventional RaDiO techniques, we achieve this by minimizing the Lagrangian $J(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}) + \lambda R(\boldsymbol{\pi})$ for some Lagrange multiplier $\lambda > 0$, thus achieving a point on the lower convex hull of the set of all achievable distortion-rate pairs. However, note that due to the employed additive model our expression for the expected distortion $\mathcal{D}(\boldsymbol{\pi})$ is much simpler than those typically found in conventional RaDiO techniques such as [3, 5]. In particular, we replace the individual distortion reductions Δd_l associated with every frame l in the conventional RaDiO techniques with the total distortions $D(l)$ given above. This allows us to implicitly account for decoding dependencies and error concealment without an increase in computational complexity. On the other hand, conventional RaDiO techniques have to explicitly account for these two in their model for $\mathcal{D}(\boldsymbol{\pi})$, leading to significant computational overhead since conventional RaDiO techniques have to take an expectation over a much larger number of prospective events associated with receiving or not receiving on time various video packets.

In addition, due to the additive model we can also compute the optimal transmission schedules with a much smaller complexity. Specifically, it can be shown that due to the independence between different video frames imposed by the additive model the optimal policy vector $\boldsymbol{\pi}^*$ that minimizes $J(\boldsymbol{\pi})$ is given by

$$\pi_l^* = \arg\min_{\pi} \epsilon(\pi) + \lambda' \rho(\pi), \qquad l = 1, \ldots, L \tag{1}$$

where $\lambda' = \lambda B_l / D(l)$. Note that in the expression above we need to cycle through every frame only once when computing the optimal transmission policies. On the other hand, to solve for $\boldsymbol{\pi}^*$ conventional RaDiO techniques employ a gradient descent algorithm which iteratively cycles through all frames till convergence. We compute the optimal individual policies in (1) by enumerating all possible policies π, plotting the error-cost performances $\{(\rho(\pi), \epsilon(\pi))\}$ in the error-cost plane, and producing an operational error-cost function for our scenario.

In the following, we provide expressions for the expected error-cost for a policy π in the case of sender-driven streaming

$$\begin{aligned} \epsilon(\pi) &= \prod_{j:a_j=1} P\{FTT > t_{DTS} - t_j\}, \\ \rho(\pi) &= \sum_{j:a_j=1} \prod_{k<j:a_k=1} P\{RTT > t_j - t_k\}, \end{aligned}$$

where $t_0, t_1, \ldots, t_{N-1}$ are N discrete transmission opportunities at which a video packet can be transmitted prior to its delivery deadline t_{DTS}. Furthermore, a_i are transmission actions according to which the sender sends ($a_i = 1$) or does not send ($a_i = 0$) the video packet at every transmission opportunity t_i. In essence, the transmission actions a_i comprise the transmission policy π, i.e., $\pi = (a_0, a_1, \ldots, a_{N-1})$. Finally, for more details on computing the optimal individual policies in RaDiO streaming, we refer the reader to prior works such as [3].

3.2. Technique 2: LC RaDiO 2

This streaming technique is even less computationally complex than the prior one. In the following, we explain the operations of Technique 2 in detail. Let $\mathcal{W}$ be a window of packets considered for transmission at a given time instant. We would like to find the subset of packets from $\mathcal{W}$ that should be transmitted such the total distortion associated with the video packets from $\mathcal{W}$ is minimized, while at the same time a transmission rate constraint is met.

Alternatively, the problem can be phrased as follows. What is the subset of packets $\boldsymbol{k} \in \mathcal{W}$ that should be dropped, i.e., not transmitted, such the total distortion is minimized, while meeting at the same time a rate constraint for the packets from $\mathcal{W}$ that will be transmitted. Using the additive model for the total distortion from Section 3 and the method of Lagrange multipliers this alternative formulation of the optimization problem under consideration can be written as

$$\boldsymbol{k}^* = \arg\min_{\boldsymbol{k} \in \mathcal{W}} D(\boldsymbol{k}) + \lambda R(\mathcal{W} \setminus \boldsymbol{k}) \tag{2}$$

where $\lambda > 0$ is the Lagrange multiplier and "\" denotes the operator "set difference".

Due to the additive model employed in (2), the solution $\boldsymbol{k}^*$ can be easily found by associating a utility in terms of distortion per bit for every packet $j \in \mathcal{W}$ defined as $\lambda_j = D(j)/R(j)$ [10]. Then, we simply decide to drop all packets $j \in \mathcal{W}$ for which the following is true: $\lambda_j < \lambda$. In other words, we drop all the packets from $\mathcal{W}$ that have utilities smaller than the Lagrange multiplier λ.

4. COMPLEXITY

This section briefly describes the computational requirements of the various RaDiO techniques for steady-state operation (ignoring the transients at the beginning and end of each session), and provides upper bounds on the number of operations per video packet. The complexity of LC RaDiO 1 is on the order of $|\mathcal{W}|2^N$, where N is the number of transmission opportunities over which a transmission policy is computed and $|\mathcal{W}|$ is the size of the transmission window $\mathcal{W}$. This is because a policy is computed for a video packet j as long as that packet is in the transmission window. However, once packet j is acknowledged, there is no further computing cost associated with it even though j might still be in the transmission window. Furthermore, LC RaDiO 2 requires at most $|\mathcal{W}|$ computing operations (on average $\frac{1}{2}|\mathcal{W}|$) to find the appropriate location for a video packet j (based on its utility per bit λ_j) in the sorted list of packets that are already in the transmission window $\mathcal{W}$. Note that this operation is performed once per packet, when it first enters the transmission window. Finally, the complexity of conventional RaDiO streaming is on the order of $N_i|\mathcal{W}|(M+2^N)$, where N_i is the number of iterations of the gradient descent algorithm (typically 2-3), and M is the number of concealment events that are considered for a video packet (typically about 30).

5. EXPERIMENTAL RESULTS

This section investigates the end-to-end distortion-rate performance for streaming packetized video content using different algorithms. The video sequences are coded using JM 2.1 of the JVT/H.264 video compression standard. Three standard test video sequences in QCIF format are used, Foreman, Mother and Daughter (MthrDhtr), and Carphone. Each sequence is coded at 10 fps, resulting in 130 coded frames, with a constant quantization level for an average Y-PSNR of about 36 dB, and a Group of Pictures (GOP) size of 20 frames, where each GOP consists of an I frame followed by 19 consecutive P frames. Four closed-loop streaming systems are employed in the experiments, out of which three are RaDiO. *Conv. RaDiO* is a streaming system that employs a conventional RaDiO technique for packet scheduling such as the one from [5]. *LC RaDiO 1* and *LC RaDiO 2* are streaming systems that employ respectively the two techniques for low-complexity RaDiO packet scheduling presented in this paper. Finally, the streaming system labelled *Oblivious* is a conventional streaming system which does not take into account the importance of individual packets in terms of reconstruction distortion. In particular, when making transmission decisions, *Oblivious* does not distinguish between two packets that contain two different P frames, except for the size of the packets. Therefore, *Oblivious* randomly chooses between two P-frame packets of the same size, for example, when it needs to reduce the number of transmitted packets. Similarly, transmissions of new packets and retransmissions of old lost packets are also performed in a random order by this system.

In all four systems, packets are considered for transmission in overlapping windows of variable size, similar to, e.g., [3]. At each transmission opportunity *LC RaDiO 2* and *Oblivious* consider for retransmission only those packets from the transmission window whose last transmission has not been acknowledged within $\mu_R + 3\sigma_R$ seconds from the current transmission opportunity, where μ_R and σ_R are the mean and standard deviation, respectively, of the round-trip time. The Lagrange multiplier λ is fixed for the entire presentation for all three RaDiO systems. Performance is measured in terms of the luminance peak signal-to-noise ratio (Y-PSNR) in dB of the end-to-end distortion, averaged over the duration of the video clip, as a function of the average transmission rate (Kbps) on the forward channel. In the experiments we use 600 ms for the playback delay.

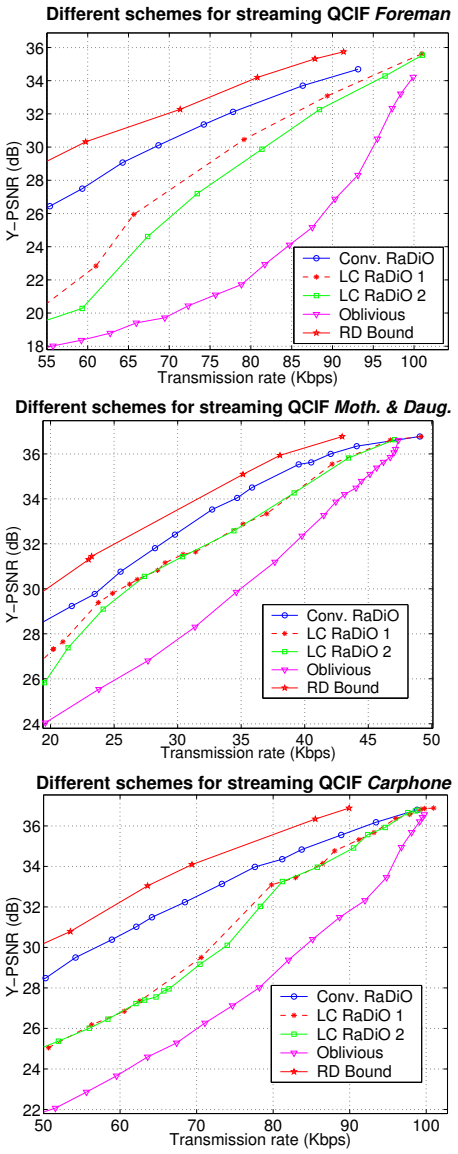

Fig. 2. R-D performance for streaming Foreman (top), Mother and Daughter (middle) and Carphone (bottom).

The forward and the backward channel on the network path

between the server and the client are modeled as follows. Packets transmitted on these channels are dropped at random, with a drop rate $\epsilon_F = \epsilon_B = \epsilon = 10\ \%$. Those packets that are not dropped receive a random delay, where for the forward and the backward delay densities p_F and p_B we use identical shifted Gamma distributions with parameters (n, α) and right shift κ, where $n = 2$ nodes, $1/\alpha = 25$ ms, and $\kappa = 50$ ms for a mean delay of $\kappa + n/\alpha = 100$ ms and standard deviation $\sqrt{n}/\alpha \approx 35$ ms.

For comparison purposes, we also plot the performance of an "ideal" R-D optimal system denoted as "RD bound". Specifically, the performance of "RD bound" is computed using the R-D characteristics of the video sequence and the characteristics of the channel in the following manner. The communication channel between the sender and the receiver acts as a packet erasure channel with a drop probability of ϵ_F. Then, if the sender transmits at a data rate R_s, the data rate observed at the receiver is $R_r = (1 - \epsilon_F)R_s$ (assuming independence between packet losses and packet size). Then, for every data rate R_s at which a sender can transmit the distortion performance of "RD bound" is computed as the smallest possible distortion for R_r using an optimal pruning algorithm and the video's R-D characteristics.

The performance of the four systems are examined in Figure 2. For streaming Foreman, *Conv. RaDiO* outperforms both *LC RaDiO 1* and *LC RaDiO 2* over the whole range of transmission rates under consideration, with a margin of roughly 1-2 dB at the high end of transmission rates and increasing as the transmission rate decreases. Similarly, *LC RaDiO 1* and *LC RaDiO 2* outperform *Oblivious* with a significant margin over the whole range of transmission rates, with a gain of at least 6 dB for rates of 65-90 Kbps, and a maximum gain of about 8 dB at 80 Kbps. Furthermore, *LC RaDiO 1* provides about 0.5 dB gain over *LC RaDiO 2* at high rates, increasing to almost 2 dB at lower rates. Finally, the performance loss of *Conv. RaDiO* with respect to "RD bound" is on the order of 1-2 dB and is due to the late loss of media (packets arriving at the receiver after their delivery deadline) and unnecessary retransmissions due to late or lost acknowledgements.

For streaming the Mother and Daughter sequence, once again *Conv. RaDiO* outperforms both *LC RaDiO 1* and *LC RaDiO 2* over the whole range of transmission rates under consideration, while *LC RaDiO 1/2* do the same with respect to *Oblivious*. However, the differences in performance is not as pronounced as for Foreman. For example, the performance difference between *Conv. RaDiO* and *LC RaDiO 1/2* is within about 1 dB. Similarly, *LC RaDiO 1/2* provide 2-4 dB gain over *Oblivious* for most transmission rates. Finally, *LC RaDiO 1* and *LC RaDiO 2* provide very similar performance over the whole range of transmission rates. The smaller range of performance variation across the four systems for Mother & Daughter is due to its comparably less motion than Foreman, leading to a smaller reduction in quality incurred for a lost or late packet since error concealment can be more effective.

Finally, the relative performance results for streaming Carphone are similar to those observed for streaming Foreman, therefore we do not discuss these results in great detail. We just note that for transmission rates greater than 80 Kbps the performance of *Conv. RaDiO* and *LC RaDiO 1/2* are within about 1 dB, and the gains of *LC RaDiO 1/2* over *Oblivious* are about 3 dB for most of the range of transmission rates.

Several important observations follow from these experiments. *Conv. RaDiO* outperforms the other streaming systems with a margin that is usually substantial. This is expected as *Conv. RaDiO* employs an optimization framework for computing its transmission schedules that is far more sophisticated and accurate than the techniques employed by the other systems. At the same time, this comes at the price of a much higher computational complexity. Therefore, it is encouraging to see that *LC RaDiO 1/2* provide a significant fraction of the performance provided by *Conv. RaDiO* while requiring significantly less complexity. Moreover, the appeal of *LC RaDiO 1/2* becomes even stronger when we note the substantial performance gains, reaching up to 8 dB, that they offer over systems such as *Oblivious*, which can be thought of as a representative example of streaming systems used in practice today. In particular, *LC RaDiO 2* provides this significant performance gain with a complexity that is of the same order as that of *Oblivious*.

Furthermore, as the playout delay increases, we expect that the difference in performance between the RD bound and *Conv. RaDiO* and *LC RaDiO 1/2* will decrease. This is because the number of possible retransmissions per media packet will increase and moreover the time interval between retransmissions can be increased thus allowing any prospective acknowledgements to arrive before the next retransmission takes place. Hence, all of these R-D optimized schemes will provide significantly better performance than *Oblivious*, and more importantly we expect that *LC RaDiO 1/2* will provide approximately the same performance as *Conv. RaDiO* while requiring significantly lower complexity.

6. CONCLUSIONS

This paper proposes two techniques for low-complexity R-D optimized packet scheduling. The techniques enable computing optimal schedules for packet transmissions with a considerably lower complexity than that of conventional algorithms for R-D optimized streaming. Our experimental results demonstrate that the proposed techniques provide substantial performance improvements over conventional non-R-D optimized streaming systems, and a significant fraction of the benefits provided by the high-complexity R-D optimized streaming systems. This is very promising as our techniques have low computational complexity and therefore are quite suitable for on-line optimized streaming.

7. REFERENCES

[1] M. Podolsky, S. McCanne, and M. Vetterli, "Soft ARQ for layered streaming media," Tech. Rep. UCB/CSD-98-1024, University of California, Computer Science Division, Berkeley, CA, Nov. 1998.

[2] Z. Miao and A. Ortega, "Optimal scheduling for streaming of scalable media," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2000.

[3] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Trans. Multimedia*, 2001, submitted.

[4] P. A. Chou and A. Sehgal, "Rate-distortion optimized receiver-driven streaming over best-effort networks," in *Proc. Int'l Packet Video Workshop*, Apr. 2002.

[5] J. Chakareski and B. Girod, "Rate-distortion optimized packet scheduling and routing for media streaming with path diversity," in *Proc. Data Compression Conference*, Mar. 2003.

[6] J. Chakareski and B. Girod, "Rate-distortion optimized video streaming with rich acknowledgments," in *Proc. Visual Communications and Image Processing*, Jan. 2004.

[7] Y.J. Liang, J. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter," in *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Apr. 2003.

[8] E. Masala and J.C. de Martin, "Analysis-by-synthesis distortion computation for rate-distortion optimized multimedia streaming," in *Proc. IEEE Int'l Conf. Multimedia and Exhibition*, July 2003.

[9] J. Chakareski, J. Apostolopoulos, W.-T. Tan, S. Wee, and B. Girod, "Distortion chains for predicting the video distortion for general packet loss patterns," in *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, May 2004.

[10] J. Chakareski, J. Apostolopoulos, W.-T. Tan, S. Wee, and B. Girod, "R-D hint tracks for low-complexity R-D optimized video streaming," in *Proc. IEEE Int'l Conf. Multimedia and Exhibition*, June 2004.