Optimal PET Protection for Streaming Scalably Compressed Video Streams With Limited Retransmission Based on Incomplete Feedback

Ruiqin Xiong, *Member, IEEE*, David Taubman, *Senior Member, IEEE*, and Vijay Sivaraman, *Member, IEEE*

***Abstract*—For streaming scalably compressed video streams over unreliable networks, Limited-Retransmission Priority Encoding Transmission (LR-PET) outperforms PET remarkably since the opportunity to retransmit is fully exploited by hypothesizing the possible future retransmission behavior before the retransmission really occurs. For the retransmission to be efficient in such a scheme, it is critical to get adequate acknowledgment from a previous transmission before deciding what data to retransmit. However, in many scenarios, the presence of a stochastic packet delay process results in frequent late acknowledgments, while imperfect feedback channels can impare the server's knowledge of what the client has received. This paper proposes an extended LR-PET scheme, which optimizes PET-protection of transmissed bitstreams, recognizing that the received feedback information is likely to be incomplete. Similar to the original LR-PET, the behavior of future retransmissions is hypothesized in the optimization objective of each transmission opportunity. As the key contribution, we develop a method to efficiently derive the effective recovery probability versus redundancy rate characteristic for the extended LR-PET communication process. This significantly simplifies the ultimate protection assignment procedure. This paper also demonstrates the advantage of the proposed strategy over several alternative strategies.**

***Index Terms*—Error protection, feedback, hybrid-ARQ, priority encoding transmission (PET), retransmission, scalable video.**

I. Introduction

THIS paper addresses robust transmission of scalable video streams through unreliable communication channels. Traditionally, the possibility of packet losses has been addressed either by *forward error correction* (FEC) or *automatic repeat request* (ARQ) retransmission. FEC approaches employ channel codes to protect streams against possible packet losses. They are often advocated for the transmission of real-time compressed data. A well known FEC framework for unequal error protection is the *priority encoding transmission* (PET) scheme of Albanese *et al.* [1]. Much recent research into the protection of scalably compressed imagery over packet-based networks [2]–[7] has adopted this perspective. For FEC approaches to utilize bandwidth efficiently, an accurate forecast on the channel behavior of ongoing transmissions is required. However, the network environment is always subject to unpredictable changes. Therefore, it may frequently occur that part of transmission bandwidth is wasted in providing unnecessarily strong protection, or the receiver is unable to recover the original uncoded source stream due to insufficient protection.

Retransmission of lost data may indeed not be possible in some cases due to stringent delivery time constraints. In many other cases, however, a limited number of retransmissions is actually possible. This perspective has been adopted by a considerable body of literature [8]–[13] which optimize the delivery of streaming media with limited retransmission. The combination of both limited retransmission and FEC has also been considered in a variety of settings [14]–[20]. More recently, the advantage of the PET framework in this context has also been studied, initially by Gan and Ma [20], and then by Taubman and Thie [21], [22]. The opportunity to retransmit at a future time actually relaxes the requirement to obtain an accurate forecast on the channel behavior of ongoing transmissions.

In *limited retransmission PET* (LR-PET) [21], [22], each frame is assigned a limited number of transmission opportunities, in each of which the stream elements with decreasing importance are protected using progressively weaker channel codes. Symbols in lost packets are allowed to be retransmitted, which will be protected again. To determine the optimal protection, the behavior of possible future retransmission of lost data is hypothesized and formulated in optimization objectives. This effectively constructs channel codes which extend into future transmission slots. Of course, longer channel codes offer better protection efficiency. More importantly, such extended channel codes are initially "hypothetical" and then determined progressively, as the transmissions are implemented one-by-one when feedback from earlier transmissions become ready.

By hypothesizing retransmission behaviour, a streaming server tends to be conservative in initial transmission opportunities, moving part of the protection to future retransmissions. In doing so, it expects that the uncertainty in channel behavior can be reduced with the help of feedback available at the time

Manuscript received June 18, 2009; revised December 02, 2009; accepted March 09, 2010. First published April 08, 2010; current version published August 18, 2010. This work was supported in part by the Australian Research Council (ARC) under Discovery Grant DP0774447, and in part by the National Basic Research Program of China (973 Program) under Grant 2009CB320904. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Amy Reibman.

R. Xiong was with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia. He is now with the School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: rqxiong@gmail.com).

D. Taubman and V. Sivaraman are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: d.taubman@unsw.edu.au; vijay.sivaraman@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2047908

of retransmission, so that the bandwidth wasted in providing excessive redundancy can be reduced.

A key assumption in the original LR-PET scheme [21], [22] is that feedback information is always available and complete, i.e., the outcome (receipt or loss) of every packet in a previous transmission is known by the sender, at the time when retransmission must be scheduled. This requires the round-trip transmission delay to be consistently smaller than the interval $\tau_{\rm d}$ between any two consecutive transmissions of a frame; it also requires the delivery of acknowledgment (ACK) messages to be reliable. In practice, of course, these conditions cannot always be satisfied. The possibility of ACK losses has been investigated in [23].

In addition to packet losses, this paper considers the influence of transmission delays. Due to the presence of stochastic transmission delays, the larger the time $\tau_{\rm d}$ that a sender waits before scheduling retransmission, the more likely it is to receive all acknowledgments, but this also increases the risk that retransmitted data will arrive too late at the receiver and become useless. Therefore, a sender may restrict the interval $\tau_{\rm d}$ to a small value to make a second or even more retransmissions possible, the benefits of which have been investigated in [24]. In these cases, feedback information is likely to be incomplete at the time of retransmission.

This paper proposes an extended LR-PET scheme to handle incomplete feedback, which is often caused by transmission delays; we call it random-delay LR-PET. To determine what data to retransmit and how much protection to use, not only the possibly incomplete feedback but also additional statistics from earlier transmission history are exploited to estimate the outcome of the previous transmission. Similar to the original LR-PET, we include into the optimization objective a set of hypotheses concerning the possible channel feedback and the effect of future retransmissions.

One important contribution of this paper is the development of a method to derive an effective recovery probability versus redundancy rate $(P - R)$ characteristic for the extended LR-PET communication process, taking all such future hypotheses into account, so that the ultimate protection assignment can be solved remarkably quickly. Our approach in this regard is similar to that which we have adopted in [25], to which we occasionally refer for thorough proofs of some key results. However, the two papers are concerned with different transmission schemes. In [25], we efficiently extend the original LR-PET scheme to the case of multiple retransmissions, while the present paper is concerned with a realistic (in fact completely general) retransmission scheme, which is subject to errors and delays. As a result, the present work involves a different retransmission mechanism and a more ellaborate collection of hypotheses.

The rest of this paper is organized as follows. Section II briefly reviews the PET framework for protecting scalable streams. Section III describes the LR-PET framework, including the transmission scenario and the overall optimization process. Section IV describes two strategies that might be used for the encoding and optimization in retransmission. Section V investigates the protection optimization for the primary transmission opportunity and develops a method to derive the effective $P - R$ characteristic of the whole LR-PET transmission, accounting for hypothetical retransmission. Section VI considers the redundancy constraints and the complexity issue. Experimental results demonstrating the efficacy of the approach are then reported in Section VII.

II. PET Framework

We briefly review the PET framework for protecting scalably compressed streams against packet erasures. Although this material may be found in [21] and [22], it helps to make the paper self-contained, and also to convey the notations and assumptions we employ for the rest of the paper.

A. Source Model

For the purpose of this paper, it is convenient to model any scalable video stream as a sequence of independently compressed "source frames" $\mathcal{F}[n]$, $n = 1, 2, 3, \ldots$, each consists of a collection of embedded elements $\mathcal{E}_q[n]$ of uncoded lengths $L_q[n]$ and utilities $U_q[n]$, $q = 1, 2, 3, \ldots, Q$, which progressively augment the quality of the reconstructed video. In the simplest case, each source frame corresponds to one video frame compressed by a scalable image coder such as EZW [26], SPIHT [27] or JPEG2000 [28]. In more general cases, each so-called source frame may actually represent a group of video frames jointly compressed by a scalable video coder such as [29]–[33]. Elements in each source frame $\mathcal{F}[n]$ exhibit a sequential dependency, $\mathcal{E}_1[n] \prec \mathcal{E}_2[n] \prec \cdots \prec \mathcal{E}_Q[n]$, which means that an element cannot be decoded or interpreted without obtaining all the previous elements in the same frame. We also assume a convex source utility-length characteristic, i.e.,

$$\frac{U_1[n]}{L_1[n]} \geqslant \frac{U_2[n]}{L_2[n]} \geqslant \cdots \geqslant \frac{U_Q[n]}{L_Q[n]}. \tag{1}$$

Since each source frame can be processed independently, we may omit the index n in some contexts for notation simplicity when this brings no confusion.

B. PET Encoding

In the PET framework, the delivery of media streams is arranged in a series of transmission slots $\mathcal{T}[n]$, $n = 1, 2, 3, \ldots$. Each source frame $\mathcal{F}[n]$ is assigned to the transmission slot with the same index, $\mathcal{T}[n]$. Source elements are protected by a family of (N, k) codes, all of which have the same codeword length N, but may have different source lengths k $(1 \leqslant k \leqslant N)$. We only consider *maximum distance separable* (MDS) codes,[1] which have the property that the receipt of any k symbols in a (N, k) codeword is sufficient to recover the source symbols. We use r to indicate the amount of redundancy. The value $r = 0$ is reserved for the special case that an element is not transmitted. For other r values $1 \leqslant r \leqslant N$, the (N, k) code used to encode source elements has $k = N + 1 - r$. Therefore

$$k_{\min}(r) = \begin{cases} N + 1 - r, & r > 0 \\ +\infty, & r = 0 \end{cases} \tag{2}$$

is the minimum number of packets that a receiver must receive to recover the elements protected using redundancy index r. In transmission slot $\mathcal{T}[n]$, once the redundancy index r_q of each

[1] In practice, we use Reed–Solomon codes.

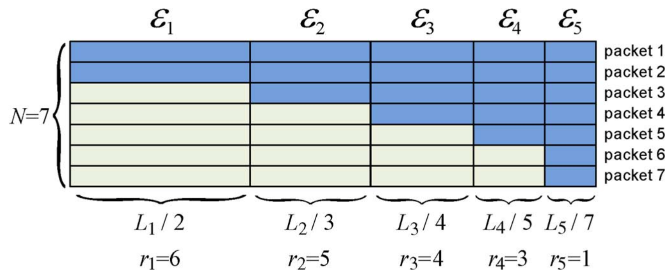


Fig. 1. Encoding of source elements in a PET frame. In this example, the PET frame consists of $N = 7$ packets. The dark shaded boxes stand for source symbols while the light shaded boxes stand for parity symbols.

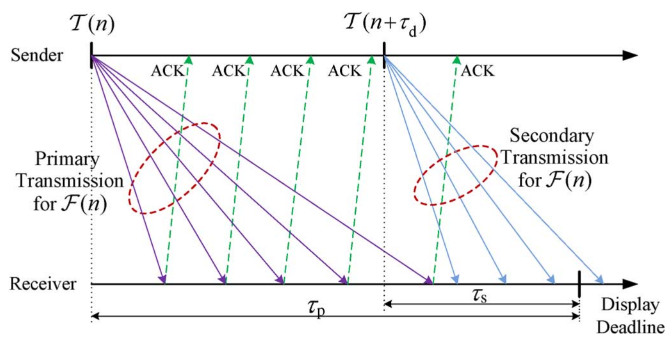


Fig. 2. LR-PET transmission of scalable media stream over random delay packet erasure channels.

element $\mathcal{E}_q$ in $\mathcal{F}[n]$ is determined, the PET scheme packages the encoded elements into N network packets, which we call a "PET frame," as illustrated in Fig. 1. $L_q/k_{\min}(r_q)$ source symbols are placed in each of the first $k_{\min}(r_q)$ packets, while each of the remaining $N - k_{\min}(r_q)$ packets contains $L_q/k_{\min}(r_q)$ parity symbols.

To guarantee that elements $\mathcal{E}_1$ through $\mathcal{E}_q$ are all recovered whenever $\mathcal{E}_{q+1}$ is recovered, the PET optimization procedure considers only those assignments satisfying $r_1 \geqslant r_2 \geqslant \ldots \geqslant r_Q$. It is not hard to show that strategies which do not satisfy this constraint are inherently suboptimal [21], [22].

C. PET Optimization

The essential step in PET optimization is to obtain the $P-R$ (recovery probability versus redundancy rate) characteristic of the communication process under current channel conditions. We introduce notation $\boldsymbol{\rho} = (\rho_0, \rho_1, \ldots, \rho_N)$, where ρ_k denotes the probability that exactly k out of the N packets in a PET frame are received. In plain PET, each element has only one opportunity for transmission. Therefore, according to the property of MDS codes, the recovery probability for elements protected with redundancy r is

$$P_{\boldsymbol{\rho}}(r) = \sum_{k\in[k_{\min}(r),N]} \rho_k. \tag{3}$$

The redundancy rate for elements protected in this case is

$$R(r) = \frac{N}{k_{\min}(r)} = \begin{cases} \frac{N}{N+1-r}, & r > 0 \\ 0, & r = 0 \end{cases}. \tag{4}$$

For a specific transmission slot $\mathcal{T}[n]$, the goal of PET optimization is to find a set of indices $\{r_q\}$ satisfying $r_1 \geqslant r_2 \geqslant \ldots \geqslant r_Q$, which maximizes $U(\{r_q\}) = \sum_q U_q[n]P_{\boldsymbol{\rho}}(r_q)$ subject to the transmission length constraint of this slot, i.e., $L(\{r_q\}) = \sum_q L_q[n]R(r_q) \leqslant L_{\max}[n]$. The maximum transmission length $L_{\max}[n]$ may vary from one slot to another. This can be converted to a family of unconstrained optimization problems parameterized by a quantity $\lambda > 0$. We find a set of indices $\{r_q(\lambda)\}$ which maximizes

$$J(\lambda) = \sum_q \{U_q[n]P_{\boldsymbol{\rho}}(r_q(\lambda)) - \lambda \cdot L_q[n]R(r_q(\lambda))\} \tag{5}$$

subject to $r_1(\lambda) \geqslant r_2(\lambda) \geqslant \cdots \geqslant r_Q(\lambda)$. Under the assumption of convex source utility-length characteristic, the optimal $\{r_q\}$ can be determined by maximizing

$$J_q(\lambda_q) = P_{\boldsymbol{\rho}}(r_q(\lambda)) - \lambda_q \cdot R(r_q(\lambda)) \tag{6}$$

for each q separately [21], [22], without violating the redundancy constraint $r_q(\lambda) \geqslant r_{q+1}(\lambda)$. Here

$$\lambda_q = \lambda \cdot L_q[n]/U_q[n] \tag{7}$$

is element-specific Lagrangian parameter. The solution to (6), $r_q = \arg\max_r[P_{\boldsymbol{\rho}}(r) - \lambda_q \cdot R(r)]$, can be quickly found by searching the upper convex hull of the $P_{\boldsymbol{\rho}}(r)$ versus $R(r)$ characteristic for the point whose slope is closest to but no less than λ_q. The ultimate solution is obtained by solving (6) and adjusting the parameter λ with a bisection search until the transmission length constraint of $\mathcal{T}[n]$ is satisfied as tightly as possible.

III. Limited Retransmission PET Framework

A. Transmission Scenario

We now consider streaming scalably compressed video using PET protection and limited retransmissions. The channel is a "*random-delay packet erasure channel*," in which each packet either arrives intact with some delays, or is entirely lost due to excessive delays or transmission errors. We assume that a backward channel exists so that a receiver can notify the sender about the receipt of each PET packet.

In this paper, we limit our discussion to the case that each source frame has only two transmission opportunities.[2] Fig. 2 illustrates the communication process. For a source frame $\mathcal{F}[n]$, its elements are first protected and transmitted using the PET frame of slot $\mathcal{T}[n]$. This is referred to as the primary transmission (PT) of $\mathcal{F}[n]$. Additional data of $\mathcal{F}[n]$ are allowed to be encoded and transmitted using the PET frame of slot $\mathcal{T}[n+\tau_{\rm d}]$, based upon the feedback of $\mathcal{T}[n]$ available at that time. This is referred to as the secondary transmission (ST) of $\mathcal{F}[n]$. For these transmitted data to be useful, it must arrive at the receiver before the display deadline of $\mathcal{F}[n]$. Suppose the constraint on total delivery time for each frame is τ. The delivery times allowed for PT and ST are $\tau_{\rm p} = \tau$ and $\tau_{\rm s} = \tau - \tau_{\rm d}$ respectively.

To decide what data to retransmit, the sender needs to know the number (identified by k) of PT packets which successfully arrive at the receiver. However, the sender only knows the

[2]Although it is interesting to also extend this to more transmissions, as we done in [24], special consideration is required to control the complexity in optimization. This is beyond the scope of this paper.

number (identified by h) of PT packets acknowledged by the time of retransmission. Obviously, h may be smaller than k due to the following reasons: 1) ACK messages may be lost; 2) the round-trip delay of a packet may be larger than the retransmission interval τ_{d} so that the ACK message is still on its way.

We use $\boldsymbol{\rho}^1 = \{\rho_k^1\}$ and $\boldsymbol{\rho}^2 = \{\rho_k^2\}$ to represent arrival probabilities for each number of packets k, in the PT and ST respectively. We use $\boldsymbol{\eta} = \{\eta_h\}$ to denote the probabilities of each h and $\boldsymbol{\rho}_{|h}^1 = \{\rho_{k|h}^1\}$ to represent the conditional probabilities of each k given h. Although the exact value of k for the current frame is unknown, it can be estimated based upon the value of h for the same frame, along with the statistic $\boldsymbol{\rho}_{|h}^1$ obtained from earlier transmission history. This can be done as follows: 1) the sender and the receiver record the h and k values of past transmissions, respectively; 2) the sender and the receiver exchange their recorded information periodically; 3) the sender or the receiver learns $\boldsymbol{\rho}_{|h}^1$ by observing the statistical relationship between h and k. One thing to note is that the mechanism to send back ACK messages may be implemented in many different ways. However, the approach developed in this paper applies equally well to any reasonable acknowledgment mechanisms.

B. LR-PET Optimization

Fig. 2 illustrates the communication process from the perspective of a specific source frame. We now consider a specific transmission slot. Slot $\mathcal{T}[n]$ transmits not only the PT data of $\mathcal{F}[n]$, but also the ST data of $\mathcal{F}[n-\tau_{\mathrm{d}}]$. The streaming server needs to select the primary redundancy indices $r_q[n]$ for each element in $\mathcal{F}[n]$ as well as the secondary redundancy indices $\mathbf{s}_q[n-\tau_{\mathrm{d}}]$ for each element in $\mathcal{F}[n-\tau_{\mathrm{d}}]$. Here $\mathbf{s}_q[n-\tau_{\mathrm{d}}]$ may be a scalar or a vector of redundancy index, depending upon how the secondary retransmission is implemented. The optimization objective of slot $\mathcal{T}[n]$ is to maximize the utility of both $\mathcal{F}[n]$ and $\mathcal{F}[n-\tau_{\mathrm{d}}]$ at the receiver, subject to the transmission length constraint of this slot. This is essentially to balance rate allocation between the primary transmission of $\mathcal{F}[n]$ and the secondary transmission of $\mathcal{F}[n-\tau_{\mathrm{d}}]$.

To appreciate the nature of this problem, note that the utility of $\mathcal{F}[n]$ depends not only upon $r_q[n]$, but also on $\mathbf{s}_q[n]$, which is unknown until the secondary transmission of $\mathcal{F}[n]$ really occurs in the future slot $\mathcal{T}[n+\tau_{\mathrm{d}}]$. Importantly, the secondary transmission can only be hypothesized at the time of primary transmission. In order to account for the complex interaction between primary transmission and secondary transmission, we adopt a global perspective [21]; a similar perspective is adopted in [12].

We introduce a collection of hypotheses concerning what might happen in the secondary transmission of $\mathcal{F}[n]$ as follows. For hypothetical value for h,[3] we write $\mathbf{s}_q^h[n]$ as the secondary redundancy indices the streaming server would be expected to adopt for elements in $\mathcal{F}[n]$. We define $P(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$ and $R(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$ as the expected recovery probability and the expected total redundancy rate, respectively, for the two transmissions (PT plus ST) of element $\mathcal{E}_q[n]$, if the element is protected using redundancy index $r_q[n]$ in its PT and using $\{\mathbf{s}_q^h[n]\}_h$ in its collection of hypothetical ST. The expected utility of $\mathcal{F}[n]$ at the receiver is $E[U[n]] = \sum_q U_q[n] P(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$ while the expected total transmission length of $\mathcal{F}[n]$ is $E[L[n]] = \sum_q L_q[n] R(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$. From a global perspective, the optimization objective for transmission slot window W is to maximize the total utility $E[\sum_{n\in W} U[n]]$ subject to a constraint on the total transmission length $E[\sum_{n\in W} U[n]]$ [21]. This can be converted to an unconstrained optimization objective parameterized by a quantity $\lambda > 0$, i.e., maximizing $E[\sum_{n\in W} U[n]] - \lambda E[\sum_{n\in W} L[n]]$. As source frames are independent, this is equivalent to maximizing $E[U[n]] - \lambda E[L[n]]$, i.e.,

$$\sum_q \Big\{ U_q[n] P\left(r_q[n], \{\mathbf{s}_q^h[n]\}_h\right) - \lambda \cdot L_q[n] R\left(r_q[n], \{\mathbf{s}_q^h[n]\}_h\right) \Big\} \quad (8)$$

for each n. Here λ can be interpreted as a type of quality parameter, which ultimately controls the selection of both the primary redundancy index $r_q[n]$ and the secondary redundancy indices $\{\mathbf{s}_q^h[n]\}_h$.

For the original LR-PET scheme [21], [22], it has been proved that for streams with convex source utility-length characteristics, we can optimize the protection of each element separately according to a specific quality parameter λ and the redundancy indices obtained in this way automatically satisfy the redundancy constraints which are necessary for the additive formulation of utility to be valid. [25] extends this conclusion to a more general LR-PET scheme with multiple retransmission opportunities. Following the same idea, for the extended LR-PET scheme in this paper, we adopt the strategy of optimizing each element separately by decomposing (8) into

$$P\left(r_q[n], \{\mathbf{s}_q^h[n]\}_h\right) - \lambda_q \cdot R\left(r_q[n], \{\mathbf{s}_q^h[n]\}_h\right) \quad (9)$$

(with $\lambda_q = \lambda \cdot L_q[n]/U_q[n]$). In Section VI, we will discuss the redundancy constraints and justify the previously mentioned strategy.

The formulation (8) and (9) emphasize the balance between the primary transmission and the secondary transmission of $\mathcal{F}[n]$. The key idea behind (8) and (9) is that the primary transmission of $\mathcal{F}[n]$ in slot $\mathcal{T}[n]$ and the secondary transmission of $\mathcal{F}[n]$ in slot $\mathcal{T}[n+\tau_{\mathrm{d}}]$ are assumed to be optimized in accordance with the same quality parameter λ [21]. In other words, $r_q[n]$ and $\{\mathbf{s}_q^h[n]\}_h$ are optimized jointly. What we are ultimately interested in is how $r_q[n]$ is regulated by the quality parameter λ. To figure out this, we need to evaluate the convex hull of the $P(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$ versus $R(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$ characteristic. For this purpose, we first solve $\mathbf{s}_q^h[n]$ for each case of hypothetical secondary transmission (depending upon $r_q[n]$ and h), which will be discussed in Section IV. Then we evaluate $P(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$ and $R(r_q[n], \{\mathbf{s}_q^h[n]\}_h)$ for each candidate value of $r_q[n]$, regulating all hypothetical retransmissions using the same parameter λ. This actually forms the P-R characteristic of the whole communication process, which will be discussed in Section V.

[3] Recall that h identifies the information the sender actually receives concerning the outcome of the primary transmission.

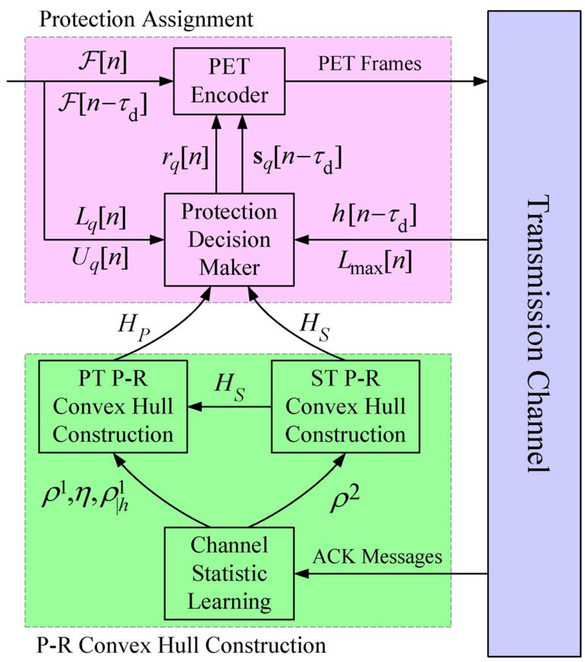


Fig. 3. Overall LR-PET optimization procedure.

Fig. 3 illustrates how the LR-PET optimization works. The streaming server appears on the left side of the channel while the client is on the right side (omitted). The whole procedure consists of two parts. The P-R convex-hull construction part (at lower-left of Fig. 3) learns statistics $\boldsymbol{\rho}^1$, $\boldsymbol{\rho}^2$, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}^1_{|h}$ based upon feedback information (e.g., ACK messages) from the channel and constructs the P-R convex-hulls $\mathcal{H}_{\mathrm{S}}$ for the secondary transmission alone and $\mathcal{H}_{\mathrm{P}}$ for the whole communication process. This activity has a relative high complexity. However, it is independent of the media stream elements to be delivered. In real applications, we may reduce the complexity by performing the convex-hull construction episodically — e.g., only when nontrivial changes in channel statistics are detected. The protection assignment part (at upper-left of Fig. 3) decides the redundancy indices $r_q[n]$ for $\mathcal{F}[n]$ and $\mathbf{s}_q[n-\tau_{\mathrm{d}}]$ for $\mathcal{F}[n-\tau_{\mathrm{d}}]$ by searching the P-R convex-hull $\mathcal{H}_{\mathrm{P}}$ and $\mathcal{H}_{\mathrm{S}}$ respectively, using element-specific quality parameters controled by λ. This process has a very low complexity but must be executed for every transmission slot. In actual LR-PET optimization, the quality parameter λ is typically determined slot-by-slot using bisection search until the total transmission length of current slot $\mathcal{T}[n]$ (depending upon $r_q[n]$ and $\mathbf{s}_q[n-\tau_{\mathrm{d}}]$) satisfies the transmission length constraint $L_{\max}$ as tightly as possible. The chosen value of λ may vary from one slot to another. However, to provide good overall transmission performance and pleasing experience of service, λ should be kept as constant as possible [12]. This can be achieved by employing transmission buffers. Details of the optimization procedure are discussed in the remainder of this paper.

C. Example Channel Model

We describe a simple channel model, which can be used to analyze the proposed LR-PET scheme. In this model, the loss and delay behavior of each packet is independent and identically distributed. The loss probability of each packet is p. The transmission delay d of each packet satisfies a *Gamma distribution*, i.e., d has a probability density function

$$f_{\alpha,\beta}(d) = d^{\alpha-1}\frac{e^{-d/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \tag{10}$$

with mean $\alpha\beta$ and variance $\alpha\beta^2$. We use p_F, α_F, β_F to denote the parameters of the forward channel. With the *upper incomplete gamma function* defined as

$$\Gamma(\alpha, x) = \int_x^\infty t^{\alpha-1}e^{-t}dt$$

the probability that a PET packet fails to get through the forward channel within time t after its transmission is

$$\tilde{p}_{PKT}(t) = p_F + (1-p_F)\frac{\Gamma(\alpha_F, t/\beta_F)}{\Gamma(\alpha_F)}. \tag{11}$$

Then the probabilities $\boldsymbol{\rho}^1$ and $\boldsymbol{\rho}^2$ can be calculated by

$$\rho^1_k = \binom{N}{k}\cdot(1-\tilde{p}_{PKT}(\tau_{\mathrm{p}}))^k\cdot(\tilde{p}_{PKT}(\tau_{\mathrm{p}}))^{N-k} \tag{12}$$

$$\rho^2_k = \binom{N}{k}\cdot(1-\tilde{p}_{PKT}(\tau_{\mathrm{s}}))^k\cdot(\tilde{p}_{PKT}(\tau_{\mathrm{s}}))^{N-k}. \tag{13}$$

Now we use p_B, α_B, β_B to denote the parameters of the backward channel. Suppose each packet is acknowledged individually. Then $p' = 1-(1-p_F)(1-p_B)$ represents the probability that either a PET packet or its ACK is lost. Therefore, the probability that a PET packet is still unacknowledged at time t after its transmission is

$$\tilde{p}_{ACK}(t) = p' + (1-p')\int_t^\infty \left(f_{\alpha_F,\beta_F} * f_{\alpha_B,\beta_B}\right)(x)dx \tag{14}$$

where $*$ denotes convolution. Then $\boldsymbol{\eta}$ can be calculated by

$$\eta_h = \binom{N}{h}\cdot(1-\tilde{p}_{ACK}(\tau_{\mathrm{d}}))^h\cdot(\tilde{p}_{ACK}(\tau_{\mathrm{d}}))^{N-h}. \tag{15}$$

Furthermore, the conditional packet arrival probability $\boldsymbol{\rho}^1_{|h}$ can be calculated by

$$\rho^1_{k|h} = \binom{N-h}{k-h}\cdot(1-\tilde{p}^*(\tau_{\mathrm{d}},\tau_{\mathrm{p}}))^{k-h}\cdot(\tilde{p}^*(\tau_{\mathrm{d}},\tau_{\mathrm{p}}))^{N-k}. \tag{16}$$

Here $\tilde{p}^*(t_1,t_2)$, $t_1 < t_2$ represents the probability that a PET packet fails to get through the forward channel within time t_2 after its transmission, under the condition that it is unacknowledged at time t_1 after its transmission. This conditional probability can be calculated by

$$\tilde{p}^*(t_1,t_2) = \frac{\tilde{p}_{PKT}(t_2)}{\tilde{p}_{ACK}(t_1)}. \tag{17}$$

Although we formulate $\boldsymbol{\rho}^1$, $\boldsymbol{\rho}^2$, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}^1_{|h}$ only for the previously mentioned example channel, the random-delay LR-PET scheme and the optimization procedure we develop in this paper

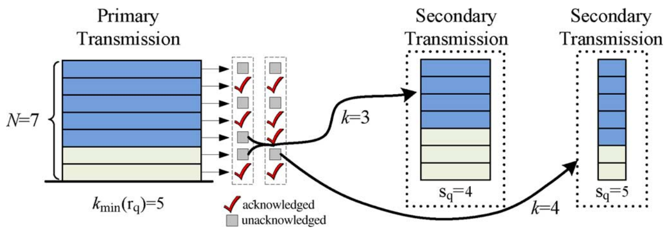


Fig. 4. PET encoding for the primary transmission and the secondary transmission (for two hypothetical cases) using the RACA strategy. During the retransmission of $\mathcal{E}_q$, symbols in $k_{\min}(r_q) - k$ of the $N - k$ lost packets are encoded and transmitted as a new single element.

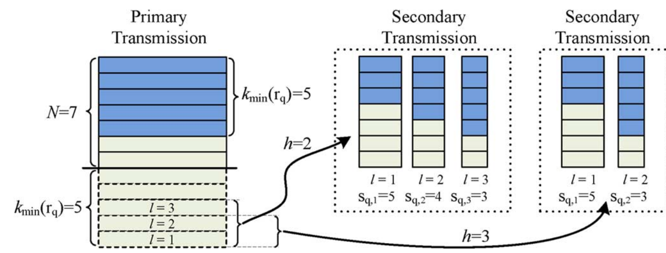


Fig. 5. PET encoding for the primary transmission and the secondary transmission (for two hypothetical cases) using the RICA strategy. During the retransmission of $\mathcal{E}_q$, symbols in $k_{\min}(r_q) - h$ of the packets represented by dashed boxes are encoded and transmitted, with symbols from each packet regarded as a new subelement.

can work with channels with arbitrary loss and delay characteristics. Actually, the probabilities required by the optimization procedure can all be learned by collecting statistics from the channel.

IV. Retransmission Strategies

In this section, we investigate two retransmission strategies that might be employed in LR-PET. The two strategies are based upon different knowledge of $\boldsymbol{\rho}^1_{|h}$.

A. Retransmission With Assumed Complete Acknowledgment (RACA)

The first considered transmission strategy is the original LR-PET scheme [21]. In this case, the sender considers received acknowledgement information to be complete, such that $k = h$. Equivalently, the sender assumes that $\boldsymbol{\rho}^1 = \boldsymbol{\eta}$ and $\rho^1_{k|h} = \delta_{kh} = [k = h]$ without actually attempting to estimate or use $\rho^1_{k|h}$. This transmission scheme is summarized as follows.

If the primary redundancy index r_q assigned to element $\mathcal{E}_q$ satisfies $k_{\min}(r_q) \leqslant k$, retransmission is not needed (by assumption). Otherwise, the receiver needs exactly $k_{\min}(r_q) - k$ of the $N - k$ lost packets to recover $\mathcal{E}_q$. The LR-PET sender regards the codeword symbols of $\mathcal{E}_q$ in these $k_{\min}(r_q) - k$ packets as a new element to be transmitted. The effective uncoded length of $\mathcal{E}_q$ is reduced from L_q to $L_q\theta(r_q, k)$, where $\theta(r, k)$ is defined by

$$\theta(r,k) = \begin{cases} 1 - \frac{k}{k_{\min}(r)}, & k < k_{\min}(r) \\ 0, & k \geqslant k_{\min}(r). \end{cases} \tag{18}$$

The data to retransmit for $\mathcal{E}_q$ is protected in the ST PET frame using a new redundancy index, s^k_q. To guarantee that $\mathcal{E}_1$ through $\mathcal{E}_q$ are all recovered whenever $\mathcal{E}_{q+1}$ is recovered, the optimization procedure only considers protection assignments satisfying $s^k_q \geqslant s^k_{q+1}, \forall q$. Fig. 4 illustrates the PET encoding in the PT and the ST.

When r_q and $h = k$ are known, the optimization objective (9) using this retransmission strategy is reduced to finding a redundancy index s^k_q which maximizes

$$J_{q,k}(\lambda_q) = P_{\boldsymbol{\rho}^2}\left(s^k_q\right) - \lambda_q\theta(r_q,k)R\left(s^k_q\right). \tag{19}$$

This leads to $s^k_q = \arg\max_r[P_{\boldsymbol{\rho}^2}(r) - \lambda_q\theta(r_q,k)R(r)]$. The solution can be easily found by searching the upper convex hull $\mathcal{H}_{\mathrm{S}}$ of the $P_{\boldsymbol{\rho}^2}(r) - R(r)$ characteristics for the point whose slope is closest to but no less than $\lambda_q\theta(r_q, k)$. The optimization of r_q for this strategy has been thoroughly analyzed in [21], [22], taking into account retransmission hypotheses.

B. Retransmission With InComplete Acknowledgment (RICA)

Our second considered transmission strategy is the one proposed in this paper, in which the sender is supposed to have acquired a nontrivial estimate of $\boldsymbol{\rho}^1_{|h}$. In particular, the sender is aware of the likelihood that the acknowledgment for some successfully received packets may be absent. For the receiver to get maximum benefits from the retransmitted data and those successfully received but unacknowledged packets simultaneously, the extra redundancy provided by a retransmission should not overlap with the redundancy contained in previous transmissions.

To do this, we employ the encoding strategy introduced in [23]. For element $\mathcal{E}_q$ assigned with a primary redundancy index r_q, $N + k_{\min}(r_q)$ codeword packets are constructed using a $(N + k_{\min}(r_q), k_{\min}(r_q))$ MDS code. Among these codeword packets, only the initial N packets are transmitted in the PT. Recall that h denotes the number of PT packets acknowledged by the time of retransmission. If $k_{\min}(r_q) \leqslant h$, retransmission is not necessary. Otherwise, $k_{\min}(r_q) - h$ out of the final $k_{\min}(r_q)$ symbol packets constructed during the PT are considered for retransmission. The codeword symbols of $\mathcal{E}_q$ in each of these packets are regarded as new source elements of uncoded length $L_q/k_{\min}(r_q)$. We call them *sub-elements* of $\mathcal{E}_q$, denoted by $\mathcal{E}_{q,l}$, $l = 1, 2, \ldots, k_{\min}(r_q) - h$. Subelement $\mathcal{E}_{q,l}$ is protected in the ST PET frame using redundancy index $s^h_{q,l}$. Fig. 5 illustrates PET encoding in both the PT and the ST.

We arrange the subelements of $\mathcal{E}_q$ in decreasing order of their protection strength, i.e.,

$$s^h_{q,1} \geqslant s^h_{q,2} \geqslant \cdots \geqslant s^h_{q,k_{\min}(r_q)-h}, \quad \forall q. \tag{20}$$

As a result, subelements $\mathcal{E}_{q,1}$ through $\mathcal{E}_{q,l-1}$ are always recovered whenever $\mathcal{E}_{q,l}$ is recovered. We introduce an inverse series $\{\tilde{s}_{q,k^h}\}_{h \leqslant k < k_{\min}(r_q)}$ defined by $\tilde{s}^h_{q,k} = s^h_{q,k_{\min}(r_q)-k}$. Recall that k denotes the number of PET packets from the PT which actually arrived at the receiver. Obviously, retransmission is relevant only for cases $k \in [h, k_{\min}(r_q) - 1]$. To recover $\mathcal{E}_q$, the receiver requires successful delivery of the first $l = k_{\min}(r_q) - k$ subelements, or equivalently, the lth subelement. The probability of achieving this is $P_{\boldsymbol{\rho}^2}(s^h_{q,l}) = P_{\boldsymbol{\rho}^2}(\tilde{s}^h_{q,k})$. Therefore, the optimization objective (9) using this retransmission strategy is reduced to finding a set of indices $\{\tilde{s}^h_{q,k}\}$ which maximize

$$J_q(\lambda_q)=\sum_{k=h}^{k_{\min}(r_q)-1}\left(\rho^1_{k|h}P_{\boldsymbol{\rho}^2}\left(\tilde{s}^h_{q,k}\right)-\lambda_q\cdot\frac{1}{k_{\min}(r_q)}R\left(\tilde{s}^h_{q,k}\right)\right). \quad (21)$$

Unfortunately, indices $\{\tilde{s}^h_{q,k}\}$ produced by optimizing each subelement separately do not necessarily satisfy the constraints (20), without which the expression (21) is invalid. Consider, however, the cumulative conditional probabilities $\phi^k_h=\sum_{i=h}^k \rho^1_{i|h}$ and $\Phi^M_h=\{\phi^{h-1}_h,\phi^h_h,\ldots,\phi^{M-1}_h\}$. It turns out that without any loss of optimality, we can restrict our attention to the *lower* convex hull of the ϕ^k_h characteristic in Φ^M_h, $M=k_{\min}(r_q)$. Specifically, Appendix A proves that, if $k_0<k_1<\cdots<k_m$ is an enumeration of superscripts of points on this convex hull, the optimal solution must satisfy $\tilde{s}^h_{q,k}=\tilde{s}^h_{q,k_j}$ for all $k\in(k_{j-1},k_j]$. This allows us to reformulate (21) as

$$J_q(\lambda_q)=\sum_{j=1}^{m}\left(\left(\phi^{k_j}_h-\phi^{k_{j-1}}_h\right)P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,k^h_j}\right)-\lambda_q\cdot\frac{k_j-k_{j-1}}{k_{\min}(r_q)}R\left(\tilde{s}^h_{q,k_j}\right)\right) \quad (22)$$

and maximize (22) separately for each j, without violating the constraint (20) by virtue of convexity.

By defining a modified version of $\{\rho^1_{k|h}\}$ according to slopes on this convex hull as

$$\breve{\rho}_{k|h}=\frac{\phi^{k_{j+1}}_h-\phi^{k_j}_h}{k_{j+1}-k_j},\quad \forall k\in(k_j,k_{j+1}] \quad (23)$$

the optimal indices $\tilde{s}^h_{q,k}$ can be found from the convex hull of the $P_{\boldsymbol{\rho}^2}(r)-R(r)$ characteristic using slope threshold $\lambda_q/(k_{\min}(r_q)\breve{\rho}_{k|h})$.

The previous discussion focuses on optimizing subelement redundancy indices for a specific element $\mathcal{E}_q$. Actually, for different q, the r_q value and the convex hull associated with Φ^M_h, $M=k_{\min}(r_q)$ may vary. Generally, we need to consider Φ^M_h for all $M>h$. Fig. 6 clearly illustrates how the convex hull associated with Φ^M_h varies with the value of M. Actually, the convex hull for Φ^{M+1}_h can be iteratively calculated easily based upon the result for Φ^M_h because $\Phi^{M+1}_h=\Phi^M_h\cup\{\phi^M_h\}$. To indicate the dependence upon M, we revise the notations $\{k_j\}$ and $\{\breve{\rho}_{k|h}\}$ to $\{k^M_j\}$ and $\{\breve{\rho}^M_{k|h}\}$, respectively. It is not hard to observe from Fig. 6 that

$$\breve{\rho}^M_{k|h}\geqslant\breve{\rho}^{M+1}_{k|h},\quad \forall h\leqslant k\leqslant M-1. \quad (24)$$

Once we obtain convex hulls of Φ^M_h for all $M>h$, the optimal retransmission indices can be generally formulated as

$$\tilde{s}^h_{q,k}=\operatorname{argmax}_r\left[P_{\boldsymbol{\rho}^2}(r)-\lambda_q\theta_h(r_q,k)R(r)\right] \quad (25)$$

with the function $\theta_h(r_q,k)$ defined by

$$\theta_h(r,k)=\frac{1}{k_{\min}(r)\breve{\rho}^{k_{\min}(r)}_{k|h}}. \quad (26)$$

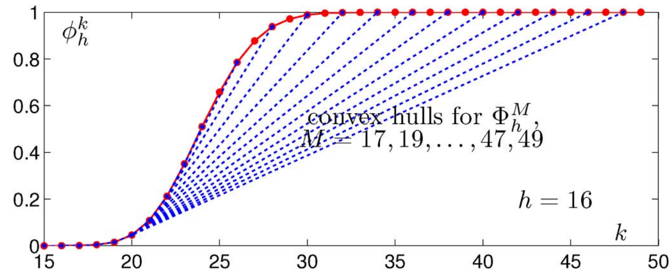


Fig. 6. *Lower* convex hulls formed from the data series $\Phi^M_h=\{\phi^{h-1}_h,\phi^h_h,\ldots,\phi^{M-1}_h\}$ with different M values.

In the special case where feedback is surely complete at the time of retransmission (i.e., $\rho^1_{k|h}=\delta_{kh}$), the convex hull of Φ^M_h contains only ϕ^{h-1}_h and ϕ^{M-1}_h. Thus, the $\{\breve{\rho}^M_{k|h}\}$ constructed from this convex hull is

$$\breve{\rho}^M_{k|h}=\frac{1}{M-h},\quad \forall k\in[h,M-1]. \quad (27)$$

In this case, $\theta_h(r,k)=((k_{\min}(r)-h)/k_{\min}(r))=\theta(r,h)$, $\forall k\in[h,k_{\min}(r)-1]$. This means all subelements of $\mathcal{E}_q$ are protected using the same redundancy index as the one used in the RACA strategy.

V. Optimization for Primary Transmission

Now we consider protection assignment in the primary transmission. Recall that, at the time to assign r_q, the channel behavior and the future retransmission can only be hypothesized. The key task is to evaluate the expected recovery probability and redundancy rate of the LR-PET communication process, considering all the possible hypothetical retransmission cases. Since LR-PET using the RACA strategy has already been thoroughly investigated in [21], [25], we only consider the RICA strategy in this section.

We first introduce notation for the convenience of further discussion. We use $C=\{T_i=(P_i,R_i,\mathbf{V}_i)\}_i$, listed in the increasing order of R_i, to describe any protected communication process. Each triplet T_i corresponds to one possible protection assignment, with $\mathbf{V}_i$ representing the protection parameters, P_i and R_i representing the recovery probability and redundancy rate associated with it, respectively. $\mathbf{V}_i$ may include only a primary redundancy index, in the simplest case of a single transmission, or also the set of hypothetical retransmission redundancy indices, if the secondary transmission exists. But we are ultimately interested only in the optimal choice of primary redundancy index. In all cases, a practical implementation of our proposed algorithm only needs to explicitly store values r (i.e., r_q) of $\mathbf{V}_i$ for each T_i.

Among all the protection assignment possibilities, we are only interested in the cases which achieve an optimal trade off between P_i and R_i. We define $\mathcal{H}_C=\{T_{i_1},T_{i_2},\ldots\}$, where the subscript indices $\{i_1,i_2,\ldots\}$ are chosen so that the coordinate point set $\mathbb{P}_{\mathcal{H}_C}=\{(R_i,P_i)|T_i\in\mathcal{H}_C\}$ constitutes the upper convex hull of $\mathbb{P}_C=\{(R_i,P_i)|T_i\in C\}$. We call $\mathcal{H}_C$ the convex hull of C. Fig. 7 illustrates the relationship between $\mathcal{H}_C$

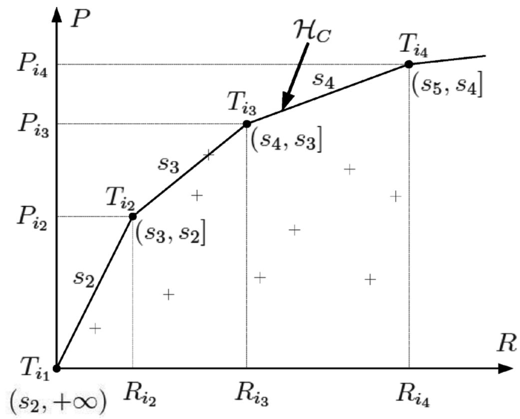


Fig. 7. Convex hull $\mathcal{H}_C$ of an effective channel code C. The points marked with "$+$" indicate the elements in C but not in $\mathcal{H}_C$. The interval labeled under each T_{i_j} is the λ range in which V_{i_j} is the optimal protection assignment solution.

and C. Obviously, the $P-R$ slope values on the convex hull $\mathcal{H}_C$, defined by

$$s_j = \begin{cases} \frac{P_{i_j} - P_{i_{j-1}}}{R_{i_j} - R_{i_{j-1}}}, & j > 1 \\ +\infty, & j = 1 \end{cases} \tag{28}$$

satisfy $s_1 > s_2 > \cdots > 0$. Since we are mainly interested in $\mathcal{H}_C$ instead of C, we notationally define $P_{\mathcal{H}_C}(j) = P_{i_j}$, $R_{\mathcal{H}_C}(j) = R_{i_j}$, $r_{\mathcal{H}_C}(j) = r_{i_j}$, $S_{\mathcal{H}_C}(j) = s_j$ and $\mathbb{S}^1(\mathcal{H}_C) = \bigcup_i \{S_{\mathcal{H}_C}(i)\}$. The subscript $\mathcal{H}_C$ here is to specify the relevant convex hull. We further define the following operator:

$$\mathfrak{S}(\mathcal{H}_C, \lambda) = \max\{i | S_{\mathcal{H}_C}(i) \geq \lambda\} \tag{29}$$

to search $\mathcal{H}_C$ for the protection assignment which is optimal for a given λ value. We then have:

Property 1: For all $\mathcal{H}_C$, $\mathfrak{S}(\mathcal{H}_C, \lambda)$ is a nonincreasing piecewise-constant function of λ. In particular, $\mathfrak{S}(\mathcal{H}_C, \lambda) = i$ for all $\lambda \in (S_{\mathcal{H}_C}(i+1), S_{\mathcal{H}_C}(i)]$. These intervals are illustrated in Fig. 7.

Now we consider the protection assignment problem. Viewed from the primary transmission, where h — the number of acknowledged packets from the PT — is not yet known, the expected recovery probability for the LR-PET procedure of two transmissions is

$$P^2\left(r_q, \{\tilde{s}^h_{q,k}\}\right) = P_{\boldsymbol{\rho}^1}(r_q) + \sum_{h \in \mathbb{K}(r_q)} \left(\eta_h \sum_{k=h}^{k_{\min}(r_q)-1} \breve{\rho}^{k_{\min}(r_q)}_{k|h} P_{\boldsymbol{\rho}^2}\left(\tilde{s}^h_{q,k}\right) \right) \tag{30}$$

and the expected overall redundancy rate is

$$R^2\left(r_q, \{\tilde{s}^h_{q,k}\}\right) = R(r_q) + \sum_{h \in \mathbb{K}(r_q)} \left(\eta_h \sum_{k=h}^{k_{\min}(r_q)-1} \frac{1}{k_{\min}(r_q)} R\left(\tilde{s}^h_{q,k}\right) \right). \tag{31}$$

Here, $\mathbb{K}(r) = [0, k_{\min}(r) - 1]$ is the set of hypotheses concerning h, for which retransmission of an element protected with redundancy r may be beneficial. The objective for optimizing primary redundancy index r_q is to maximize

$$J^2_q(\lambda_q) = P^2\left(r_q, \{\tilde{s}^h_{q,k}\}\right) - \lambda_q \cdot R^2\left(r_q, \{\tilde{s}^h_{q,k}\}\right) \tag{32}$$

where for each r_q value and each hypothesis h, the hypothetical retransmission indices $\{\tilde{s}^h_{q,k}\}$ are determined as described in Section IV-B.

Suppose $\mathcal{H}_{\mathrm{S}}$ is the P-R convex hull for the ST alone, formed from the set $\{(P_{\boldsymbol{\rho}^2}(r), R_{\boldsymbol{\rho}^2}(r), r)\}$ specified by (3) and (4) over all possible r. The optimal choice of retransmission indices is, thus

$$\tilde{s}^h_{q,k} = r_{\mathcal{H}_{\mathrm{S}}}\left(\mathfrak{S}\left(\mathcal{H}_{\mathrm{S}}, \lambda_q \cdot \theta_h(r_q, k)\right)\right). \tag{33}$$

We introduce functions

$$P^2(r, \tilde{\lambda}) = P_{\boldsymbol{\rho}^1}(r) + \sum_{h \in \mathbb{K}(r)} \left(\eta_h \sum_{k=h}^{k_{\min}(r)-1} \breve{\rho}^{k_{\min}(r)}_{k|h} \cdot P_{\tilde{\mathcal{H}}}\left(\mathfrak{S}\left(\tilde{\mathcal{H}}, \tilde{\lambda}\theta_h(r,k)\right)\right) \right) \tag{34}$$

$$R^2(r, \tilde{\lambda}) = R(r) + \sum_{h \in \mathbb{K}(r)} \left(\eta_h \sum_{k=h}^{k_{\min}(r)-1} \frac{1}{k_{\min}(r)} \cdot R_{\tilde{\mathcal{H}}}\left(\mathfrak{S}\left(\tilde{\mathcal{H}}, \tilde{\lambda}\theta_h(r,k)\right)\right) \right). \tag{35}$$

(with $\tilde{\mathcal{H}} \triangleq \mathcal{H}_{\mathrm{S}}$ for notational convenience), so that according to (33) the objective (32) can be reformulated as maximizing

$$J^2_q(\lambda_q) = P^2(r_q, \lambda_q) - \lambda_q \cdot R^2(r_q, \lambda_q) \tag{36}$$

over all possible r_q. Although the original optimization problem involves the choice $\tilde{\lambda} = \lambda_q$ as indicated by (36), it turns out that we can free up the parameter $\tilde{\lambda}$, forming an extended optimization procedure in which the objective is to maximize

$$J^2_q(\lambda_q) = P^2(r_q, \tilde{\lambda}) - \lambda_q \cdot R^2(r_q, \tilde{\lambda}) \tag{37}$$

over all combinations of r_q and $\tilde{\lambda}$. As we will show, this extended optimization problem turns out to be equivalent with the original one, in the sense that 1) they both yield the same optimal value for r_q; 2) the optimal value for $\tilde{\lambda}$ is essentially identical to λ_q.

First of all, from (34), (35) and Property 1, we easily have the following property:

Property 2: For given r, $P^2(r, \tilde{\lambda})$ and $R^2(r, \tilde{\lambda})$ are nonincreasing piecewise-constant functions of $\tilde{\lambda}$. If we enumerate the slopes in $\mathbb{S}^2(\tilde{\mathcal{H}}, r) = \bigcup_{h \in \mathbb{K}(r)} \bigcup_{k=h}^{k_{\min}(r)-1} (\mathbb{S}^1(\tilde{\mathcal{H}})/\theta_h(r,k))$ in decreasing order as $s'_1 > s'_2 > \ldots > s'_{m'}$, $P^2(r, \tilde{\lambda})$ and $R^2(r, \tilde{\lambda})$ remain constant over the interval $\tilde{\lambda} \in (s'_{i+1}, s'_i]$.

Therefore, for any given r, we can restrict the parameter $\tilde{\lambda}$ in (34) and (35) to the discrete set $\mathbb{S}^2(\tilde{\mathcal{H}}, r)$. There are only $m' = |\mathbb{S}^2(\tilde{\mathcal{H}}, r)|$ discrete points on the $P^2(r, \tilde{\lambda})$ versus $R^2(r, \tilde{\lambda})$

curve, with the ith point generated by setting $\tilde{\lambda} = s'_i$. It is not hard to prove the following property:[4]

Property 3: The $P-R$ slopes on the $P^2(r,\tilde{\lambda})$ versus $R^2(r,\tilde{\lambda})$ curve, defined by

$$\tilde{S}_r^2(i) = \begin{cases} \frac{P^2(r,s'_i)-P^2(r,s'_{i-1})}{R^2(r,s'_i)-R^2(r,s'_{i-1})}, & 1 < i \leqslant \left|\mathbb{S}^2(\tilde{\mathcal{H}},r)\right| \\ +\infty, & i=1 \end{cases}$$

satisfies $\tilde{S}_r^2(i) = s'_i$.

Now we can prove the following key observation in this paper:

Corollary 1: The optimization objectives (36) and (37) yield the same optimal choice for r_q.

Proof: For any candidate value of r_q, the $P^2(r_q,\tilde{\lambda})$ versus $R^2(r_q,\tilde{\lambda})$ characteristic is convex, since $\{s'_i\}$ strictly decreases with i. Therefore, a potentially optimal $\tilde{\lambda}$ which maximizes (37) must satisfy

$$\tilde{\lambda} = s'_v, \quad v = \max\left\{i | s'_i \geqslant \lambda_q,\ s'_i \in \mathbb{S}^2(\tilde{\mathcal{H}},r_q)\right\}. \tag{38}$$

This means $\lambda_q \in (s'_{v+1}, s'_v = \tilde{\lambda}]$. For any λ_q in this interval, we have $P^2(r_q,\lambda_q) = P^2(r_q,\tilde{\lambda})$ and $R^2(r_q,\lambda_q) = R^2(r_q,\tilde{\lambda})$ — c.f. Property 2. In this case, the objectives (36) and (37) become equal. They must yield the same optimal choice for r_q. ■

We conclude that the solution to the original optimization objective (32) can be found by searching the convex hull $\mathcal{H}_{\mathrm{P}}$ formed from the triplet set $\{(P^2(r,\tilde{\lambda}), R^2(r,\tilde{\lambda}), r)\}$ generated by (34) and (35) with all combinations of r and $\tilde{\lambda}$

$$r_q = r_{\mathcal{H}_{\mathrm{P}}}\left(\mathfrak{S}(\mathcal{H}_{\mathrm{P}},\lambda_q)\right). \tag{39}$$

VI. Redundancy Constraints and Complexity

The optimization procedure described in previous sections assumes that the protection for each element can be optimized independently. However, since elements in each source frame exhibit sequential dependencies, to ensure validity of the additive utility formulation in previous optimization objectives, the following constraints must be satisfied:

$$r_q \geqslant r_{q+1}, \quad \forall q \tag{40}$$
$$\tilde{s}^h_{q,k} \geqslant \tilde{s}^h_{q+1,k}, \quad \forall q,h,k. \tag{41}$$

A. Constraints on Retransmission Redundancy

Based upon (33) and Property 1, the constraint (41) is surely satisfied if we have $\lambda_q \cdot \theta_h(r_q,k) \leqslant \lambda_{q+1} \cdot \theta_h(r_{q+1},k), \forall q$, i.e.,

$$\frac{\lambda_q}{k_{\min}(r_q)} \frac{1}{\breve{\rho}_{k|h}^{k_{\min}(r_q)}} \leqslant \frac{\lambda_{q+1}}{k_{\min}(r_{q+1})} \frac{1}{\breve{\rho}_{k|h}^{k_{\min}(r_{q+1})}}, \quad \forall q. \tag{42}$$

Since $r_q \geqslant r_{q+1}$, we have $k_{\min}(r_q) \leqslant k_{\min}(r_{q+1})$ so that the inequality for the second term on both sides of (42) is satisfied according to (24). Therefore, we only need to consider the validity of inequality $(\lambda_q/k_{\min}(r_q)) \leqslant (\lambda_{q+1}/k_{\min}(r_{q+1}))$.

We consider two consecutive elements, $\mathcal{E}_q$ and $\mathcal{E}_{q+1}$. When $\mathcal{E}_q$ has a sufficiently larger utility-length ratio than $\mathcal{E}_{q+1}$ (i.e., λ_q sufficiently smaller than λ_{q+1}), $\mathcal{E}_q$ may be assigned a stronger protection (i.e., $k_{\min}(r_q) < k_{\min}(r_{q+1})$). However, it turns out in most scenarios that $k_{\min}(r_q)$ varies with λ_q but varies much slower than λ_q itself.

[4]Proof for this property can be found in [25].

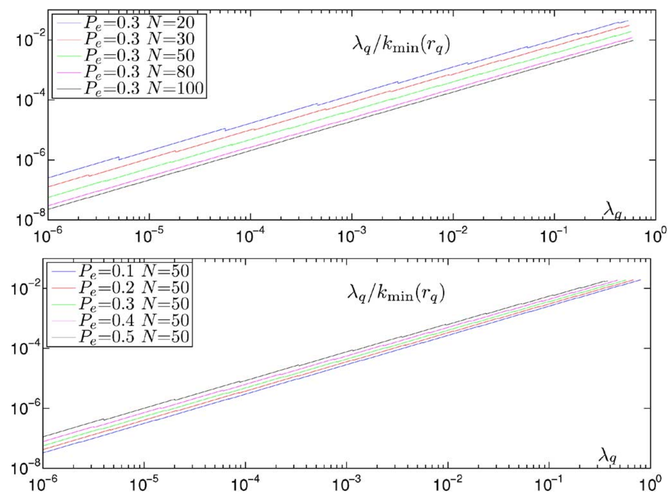


Fig. 8. Relationship between λ_q and $\lambda_q/k_{\min}(r_q)$ for plain PET over IID channels. (a) Effect of different channel code lengths. (b) Effect of different packet loss ratios.

Fig. 8 illustrates the relationship between $\lambda_q/k_{\min}(r_q)$ and λ_q for PET over IID channels. It is evident that $\lambda_q/k_{\min}(r_q)$ basically increases with λ_q, but jumps down slightly at some critical values of λ_q. This discontinuity comes from the discreteness of channel codes. Generally, we can find a threshold $\zeta > 1$ so that $(\lambda_q/k_{\min}(r_q)) \leqslant (\lambda_{q+1}/k_{\min}(r_{q+1}))$ is always true so long as $(\lambda_{q+1}/\lambda_q) > \zeta, \forall q$. This can be satisfied by merging neighboring elements which have close utility-length ratios.

Although the observations in Fig. 8 are based upon a single PET transmission over IID channels, the LR-PET optimization procedure for two or more transmissions over real-world channels produces similar results. For practical PET implementation (e.g., $N=50$) over typical channels (e.g., $P_e \leqslant 0.3$), the threshold ζ is very close to 1. Indeed, violation of the constraint (41) rarely occurs even if we do not perform element merging. In our implementation, we ignore the constraint (41) when constructing $\mathcal{H}_{\mathrm{P}}$, because the influence of enforcing (41) for the *hypothetical secondary transmissions* is trivial on $\mathcal{H}_{\mathrm{P}}$. However, for the *actual secondary transmission*, we enforce this constraint (41).

B. Constraints on Primary Redundancy

Now we consider the constraint (40). This involves the ***redundancy embedding property***, which is valid for $\mathcal{H}_C$ if $r_{\mathcal{H}_C}(i) \geqslant r_{\mathcal{H}_C}(j)$ for all $i > j$. If $\mathcal{H}_{\mathrm{P}}$ has this property, (40) is satisfied automatically according to (39) — recall Property 1 and the convexity of the source utility-length characteristic which implies $\lambda_q \leqslant \lambda_{q+1}, \forall q$.

We proved in [25] that, for LR-PET based upon complete feedback, the effective channel codes formed by any number of transmissions always satisfy the redundancy embedding property. The proof in [25] is based upon two facts: (I) for any r,

$\tilde{J}(\lambda, r) = (1/\lambda) \cdot P^n(r, \lambda) - R^n(r, \lambda)$ is a continuous function of λ; and (II) if $r_1 > r_2$, $P^n(r_1, \lambda) \geqslant P^n(r_2, \lambda)$ for all λ. Here $P^n(r, \lambda)$ and $R^n(r, \lambda)$ are generalized versions of $P^2(r, \lambda)$ and $R^2(r, \lambda)$, representing the expected recovery probability and redundancy rate, respectively, for n transmissions in which the primary one is protected by r while subsequent ones are regulated using λ as their Lagrangian parameter.

For LR-PET based upon incomplete feedback, (I) is still valid;[5] however, the conclusion for (II) is slightly complicated. The statement (II) means that, if subsequential transmissions are all regulated by the same parameter λ, a stronger protection in the initial transmission always leads to a higher recovery probability. For the extended LR-PET, the validity of (II) depends upon the usefulness of feedback information, which in turn depends upon the time at which retransmission is scheduled.

One extreme case is $\tau_{\mathrm{d}} \approx 0$, i.e., the secondary transmission is scheduled immediately after the primary transmission. In this case, no feedback is available to guide retransmission; moreover, both transmissions are equally efficient. Thus, the overall recovery probability is determined only by total redundancy rate, regardless of how it is partitioned between the two transmissions. All the $P^2(r, \lambda)$ versus $R^2(r, \lambda)$ curves (with different r values) approximately overlap with the $P - R$ curve of the primary transmission alone, except that they provide finer granularity. In such situation, violation of the redundancy embedding property occasionally occurs, due to the discreteness of channel codes we use.

However, in a real streaming system, setting retransmission interval to $\tau_{\mathrm{d}} \approx 0$ is obviously not an option the streaming server is likely to choose. Instead, the interval τ_{d} is likely chosen such that a nontrivial amount of feedback can be received by the time of retransmission. In such situation, a stronger protection in the initial transmission always means a smaller amount of data to be retransmitted in any case of channel behaviour, which in turn increases the effective utility-length ratio of the retransmission data and, thus, the redundancy assigned to them regarding a specific λ parameter. Therefore, we can expect that $P^2(r_1, \lambda) \geqslant P^2(r_2, \lambda)$ is true for all $r_1 > r_2$, when τ_{d} is not close to zero. Indeed, this has been confirmed by our experiments.

Generally, in any case of τ_{d}, we can enforce (II) by truncating the $P^2(r, \lambda)$ versus $R^2(r, \lambda)$ curves. For each r value, we compare $P^2(r, \lambda)$ to $P^2(r+1, \lambda)$, with λ varying from $+\infty$ to 0. When $\lambda = +\infty$, $P^2(r, \lambda)$ is obviously smaller than $P^2(r+1, \lambda)$ because $P^2(r, +\infty) = P_{\boldsymbol{\rho}^1}(r), \forall r$. As λ decreases, such difference may be reduced. If $P^2(r, \lambda)$ becomes larger than $P^2(r+1, \lambda)$ at $\lambda = \lambda_{\mathrm{end}}$, we remove later points on the $P^2(r, \lambda)$ versus $R^2(r, \lambda)$ curve, which are generated from (34) and (35) using $\lambda \leqslant \lambda_{\mathrm{end}}$. This actually removes some protection candidates from optimization procedure. According to previous discussion, such truncation really occurs only when τ_{d} is small, in which case such truncation only reduces the granularity of the effective channel code.

We conclude that, in any case, the redundancy embedding property can be achieved in the constructed effective protection convex hull by enforcing (II), without incurring nontrivial suboptimality. Thus, the primary redundancy constraints can be always satisfied.

C. Complexity in Constructing Convex Hull $\mathcal{H}_{\mathrm{P}}$

The construction of the $P - R$ convex hull $\mathcal{H}_{\mathrm{P}}$ seems very complicated at a first glance. However, it can be constructed efficiently, following the same development in [25]. The key idea is to take advantage of the monotonicity in redundancy indices r_q and $\{\tilde{s}^h_{q,k}\}$. Instead of evaluating the function $P^2(r, \tilde{\lambda})$ and $R^2(r, \tilde{\lambda})$ for all combinations of r and $\tilde{\lambda}$, we only need to evaluate them for a sequences of λ values, $\lambda_1, \ldots, \lambda_W$, and for each λ value, only a small range of candidate values for r and $\{\tilde{s}^h_k\}$ needs to be considered. More importantly, the $P - R$ convex hull $\mathcal{H}_{\mathrm{P}}$ is independent of the media stream elements to be delivered. We can reduce the complexity by performing the convex-hull construction episodically—e.g., only when nontrivial changes in channel statistics are detected. Therefore, the proposed method is practical in real applications.

VII. Experimental Results

In this section, we perform some experiments to investigate performance of the LR-PET procedure and optimization algorithm proposed in this paper.

We consider four variants of LR-PET, combining two strategies ("greedy" and "hypothetical") for the primary transmission and two strategies ("RACA" and "RICA") for the secondary transmission. The "greedy" strategy is just plain PET optimization, considering the primary transmission alone. The "hypothetical" strategy considers the expected outcome of the whole LR-PET procedure; in particular, the effect of possible future retransmission is hypothesized in the optimization objective. The "RACA" and "RICA" retransmission strategies are described in Section IV. For comparison, plain PET is also evaluated.

For source content, we use the MPEG CIF 30 Hz sequences, *Mobile*, *Bus*, *Foreman* and *Football*. Each sequence is compressed with Motion JPEG2000, using 5 levels of DWT decomposition and 20 quality layers. For each sequence, we cycle through the first 30 frames 10^4 times, effectively creating a much larger sequence. Since the results for these sequences are highly similar, we only report results of *Mobile*.

For transmission simulation, we first consider the channel model described in Section III-C. For example, we consider one such channel with parameters $p_F = 0.4$, $\alpha_F = 5$, $\beta_F = 1.2$ and $p_B = 0.01$, $\alpha_B = 5$, $\beta_B = 0.6$. The forward channel has a mean delay of 6 while the time for 80% and 95% nonlost packets to get through the forward channel is 6.7 and 9.1, respectively.[6] The mean delay in the backward channel is set to half of that in the forward channel. The backward channel is assumed to be much reliable than the forward channel.

[5]Proof for this is exactly the same as the proof given in [25]. In fact, (I) is the direct result of Property 2 and 3.

[6]Time in Section VII is measured by number of slots. One transmission slot corresponds to one frame. For 30 Hz video, 1 slot = 33 ms.

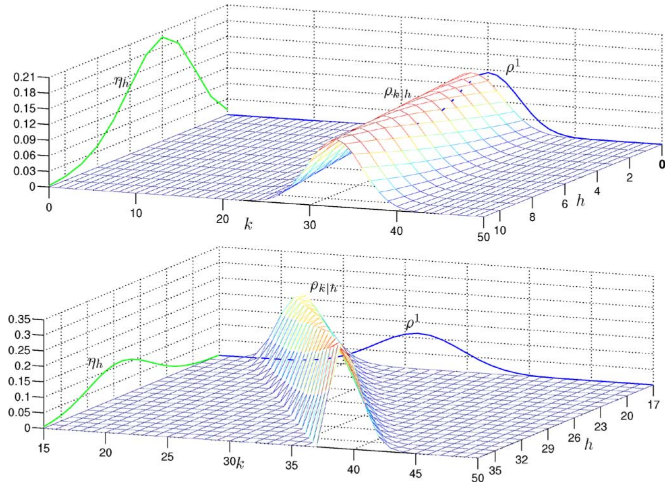


Fig. 9. Probabilities $\boldsymbol{\rho}^1$, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}^1_{|h}$ for LR-PET procedure over the simulation channel, using different retransmission intervals $\tau_{\rm d}$. The channel is based upon the model described in Section III-C, setting the channel parameters to $p_F = 0.4$, $\alpha_F = 5$, $\beta_F = 1.2$ and $p_B = 0.01$, $\alpha_B = 5$, $\beta_B = 0.6$. The total delivery time constraint is set to 20. $N = 50$. Left: $\tau_{\rm d} = 5$. Right: $\tau_{\rm d} = 15$.

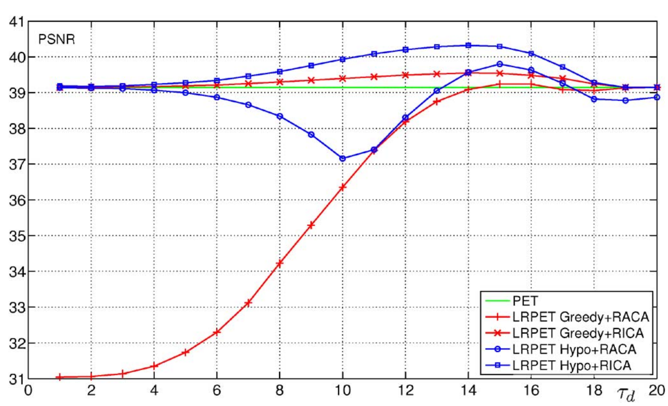


Fig. 10. Performance comparison between PET and four LR-PET variant procedures using different retransmission intervals. $p_F = 0.2, \alpha_F = 5, \beta_F = 1.2$ and $p_B = 0.01$, $\alpha_B = 5$, $\beta_B = 0.6$. $\tau = 20$. $N = 50$. $L_{\max} = 5.0 \times 10^4$ bytes.

Fig. 9 illustrates statistics of this channel for two different $\tau_{\rm d}$ values. It is evident that $\boldsymbol{\eta}$ and $\boldsymbol{\rho}^1_{|h}$ vary remarkably with retransmission interval $\tau_{\rm d}$. When $\tau_{\rm d}$ is small, the conditional probability $\boldsymbol{\rho}^1_{|h}$ is very similar to the marginal probability $\boldsymbol{\rho}^1$. In this case, the received feedback (i.e., the h value) is almost useless in estimating k — the number of successful packets. On the contrary, when $\tau_{\rm d}$ is large, $\boldsymbol{\rho}^1_{|h}$ highly depends upon h — the major possibility is that k exceeds h only slightly.

Now we evaluate the performance of the four LR-PET procedures over this channel. We first set the display deadline (DD), i.e., total allowed delivery time τ, to 20. The LR-PET procedures are tested using all possible retransmission intervals $\tau_{\rm d}$, ranging from 1 to 20. Fig. 10 illustrates the simulation results. Here, the result of plain PET is actually invariant with $\tau_{\rm d}$, because there is no retransmission.

It is clear from Fig. 10 that the two LR-PET+"RICA" procedures outperform plain PET, regardless of whether greedy or hypothetical strategy is used for the primary transmission. This conclusion applies to all retransmission intervals. In particular, the performance gain of LR-PET+"RICA" over plain PET is maximum when $\tau_{\rm d} \approx 0.7\tau$, while the gain diminishes as $\tau_{\rm d} \to 0$ or $\tau_{\rm d} \to \tau$. Of course, this gain comes from the opportunity to retransmit lost data based upon feedback. A smaller $\tau_{\rm d}$ tends to make feedback more incomplete so that the opportunity for retransmission only provides finer granularity in the effective channel codes. A larger $\tau_{\rm d}$, on the other hand, tends to make retransmission useless, because retransmitted data almost always arrive at the receiver too late (and the sender knows this). To obtain optimal performance, retransmission should be scheduled at a time at which the completeness of feedback and the usefulness of retransmission opportunities are balanced.[7] By comparing the two LR-PET+"RICA" procedures in Fig. 10, we further find that the "hypothetical" strategy for primary transmission consistently provides better performance. The performance gain obtained by considering hypotheses is especially remarkable when the retransmission interval is optimally chosen.

Now we evaluate the two LR-PET+"RACA" procedures. Fig. 10 demonstrates that they both perform worse than the corresponding LR-PET+"RICA" procedures. This loss is due to over-retransmission caused by inaccurate estimation of previous channel behavior based upon incomplete feedback.

1) When the "greedy" strategy is used in the primary transmission, the loss can be up to 8 dB. In the extreme case of $\tau_{\rm d} \approx 0$, source frames are simply protected and transmitted twice, in the false belief that all the data sent in the primary transmission are lost. The loss is reduced as $\tau_{\rm d}$ increases, due to the receipt of more feedback.
2) When the "hypothetical" strategy is used instead, the streaming server refrains from sending too much data in the primary transmission. In the extreme case of $\tau_{\rm d} \approx 0$, the server sends nothing during the primary transmission of a frame, knowing that all transmitted data will be retransmitted again. For small $\tau_{\rm d}$, the secondary transmission dominates but it becomes progressively less efficient as $\tau_{\rm d}$ increases (i.e., $\tau_{\rm s}$ decreases). This explains a degradation in performance when $\tau_{\rm d}$ varies from 0 to 10. As $\tau_{\rm d}$ further increases, the primary transmission becomes progressively dominant and increasingly efficient, because feedback becomes reliable. This explains the improvement in performance when $\tau_{\rm d}$ varies from 10 to 15.

With the optimal choice of $\tau_{\rm d}$, the LR-PET+"RACA" procedures can outperform plain PET.

Next we consider the influence of delivery time constraint τ. Fig. 11 shows the performance of LR-PET procedures, with $\tau =$ 18, 20, 22 , over the same channel used in previous simulations. For all these LR-PET procedures in Fig. 11, a longer delivery time always produces a better performance, regardless of the

[7]For practical LR-PET transmission, we can choose the near-optimal value for retransmission interval $\tau_{\rm d}$, without really evaluating all possibility of protection for each value of $\tau_{\rm d}$ and each element. To achieve this, we only need the channel statistics on packet losses and delays, the distribution of stream element utility-length ratios, and the approximate λ value that is expected to be used in a certain number of subsequent transmission slots. We may first construct the $P - R$ convex hull $\mathcal{H}_{\rm P}$ for each case of $\tau_{\rm d}$. Then we compare the effectiveness of these $P - R$ convex hulls, by evaluating the ratio of total utility and total transmission length, based upon the expected λ value and the distribution of source element utility-length ratios. Thorough discussions and studies on this topic are out of the scope of this paper.

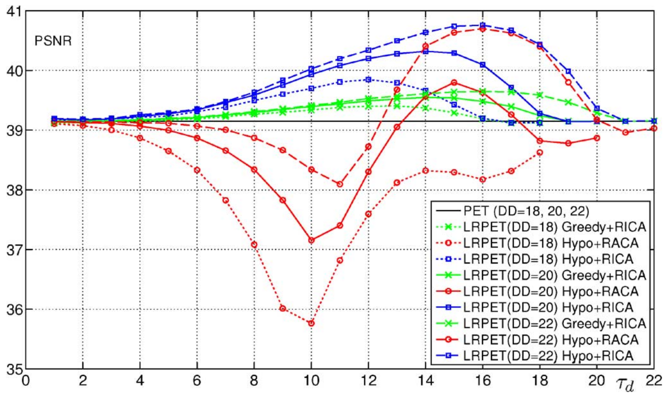


Fig. 11. Performance comparison between PET and three LR-PET variant procedures using different delivery time constraints. $p_F = 0.2, \alpha_F = 5, \beta_F = 1.2$ and $p_B = 0.01, \alpha_B = 5, \beta_B = 0.6$. $N = 50$. $L_{\max} = 5.0 \times 10^4$ bytes. $\tau = 18, 20, 22$.

choice of retransmission interval. Importantly, the range of $\tau_{\rm d}$ value in which LR-PET+"RICA" outperforms plain PET is very wide. However, the range of $\tau_{\rm d}$ in which LR-PET+"RACA" outperforms plain PET is much narrower. In other words, the LR-PET+"RICA" procedure is less sensitive to the choice of value $\tau_{\rm d}$ than the LR-PET+"RACA" procedure. Therefore, for the LR-PET+"RICA" procedure, it is easier to find a $\tau_{\rm d}$ value that yields a satisfactory performance. One more thing to note is that LR-PET+"RACA" may perform worse than plain PET or LR-PET+"RICA" even when $\tau_{\rm d} = \tau$, because even with $\tau_{\rm d} = \tau$ the feedback information can still be incomplete due to inadequate τ or an imperfect backward channel.

The simulated channel adopted for the previously mentioned results is not completely realistic, because individual packet delays are modeled as statistically independent. We now complete our experimental investigation by considering a more realistic channel, which exhibits correlation in the losses and delays of consecutive packets. Rather than attempting to capture these dependencies through an analytical model, we perform a numerical simulation, in which the communication channel consists of a set of concatenated links, each shared by the LR-PET procedure and side traffic generated by other applications. The side traffic is randomly generated, based upon a long-range dependency model. Incoming packets are dropped when the queue in a link is full. In our experimental setting, both the forward channel and the backward channel consist of five links, each having a capacity of 100 Mbps. The rate of side traffic on each link independently varies in the range of 40 to 200 Mbps. Fig. 12 illustrates the statistical distributions for the forward-trip time and round-trip time of a packet in our queue-based simulation channel. The probabilities $\boldsymbol{\rho}^1$, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}^1_{|h}$ in this simulation channel have similar characteristics to those illustrated in Fig. 9.

Fig. 13 illustrates the performance of PET, LR-PET+"RICA" and LR-PET+"RACA" procedures on this queue-based simulation channel, using a total delivery time constraint of $\tau = 22$. We ignore "greedy" strategies since they are always inferior to the corresponding "hypothetical" strategies. Similar to the results fromthe previous simulation, LR-PET+"RICA" outperforms plain PET for a large range of $\tau_{\rm d}$ values (i.e., $\tau_{\rm d} = 7 \sim 17$) while LR-PET+"RACA" outperforms plain PET only for a

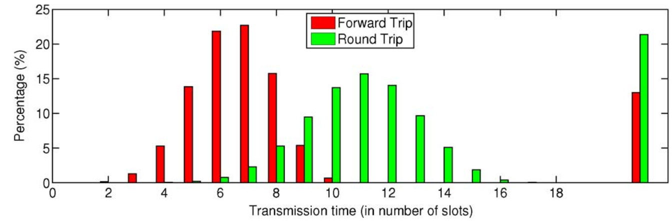


Fig. 12. Distribution of forward-trip time and round-trip time in the queue-based simulation channel. The last two columns at the right-most of the figure correspond to packet loss (i.e., trip time is infinite).

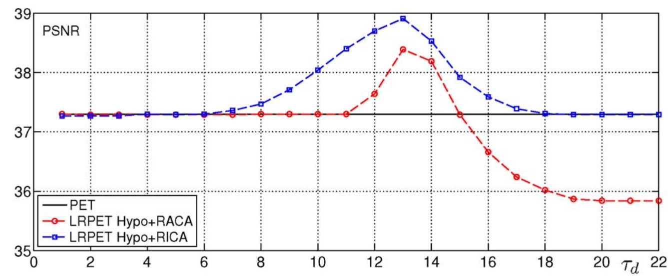


Fig. 13. Performance of PET, LR-PET+RICA and LR-PET+RACA procedures. $N = 50$, $\tau = 22$, $L_{\max} = 4.0 \times 10^4$ bytes.

small range of $\tau_{\rm d}$ values (i.e., $\tau_{\rm d} = 12 \sim 14$). Moreover, the best performance that the LR-PET+"RICA" procedure can achieve is about 0.5 dB higher than that of the LR-PET+"RACA" procedure.

The most noticeable difference between the results in Fig. 13 and those in Figs. 10 and 11 is that the performance loss associated with the LR-PET+"RACA" procedure compared to plain PET, when $\tau_{\rm d} \approx \tau$, is much larger in this simulation. As we already pointed out, this loss is due to the incompleteness in acknowledgment information, as the LR-PET+"RACA" procedure relies heavily on these ACK messages for its protection decisions. As shown in Fig. 12 the feedback channel has a loss ratio as high as 9% in this simulation, which causes about 1.5 dB loss for the LR-PET+"RACA" procedure. The LR-PET+"RICA" procedure avoids this performance loss by deriving a better estimate of the current channel behaviour, using channel statistics $\boldsymbol{\rho}^1_{|h}$ which can be obtained from earlier transmission history. This clearly demonstrates the advantages of the LR-PET+"RICA" scheme.

VIII. Conclusion

This paper proposes an extended LR-PET framework for streaming scalably compressed video streams over unreliable networks, in which retransmission is scheduled and optimized based upon incomplete feedback information. To assist in formulating protection decisions, the incomplete feedback information for a current frame is enhanced by statistics obtained from earlier transmissions to derive a better estimate of current channel behaviour. As a key contribution of the paper, we have developed a method to efficiently derive the effective $P - R$ characteristic of the LR-PET procedure, which facilitates fast protection assignment. Importantly, we have shown that the redundancy constraints in both the primary and the secondary transmission can be easily guaranteed in practical implementation of our proposed optimization algorithm. The advantage of the proposed scheme is clearly demonstrated. In particular, the scheme provides effective protection of the content, with comparatively low sensitivity to selection of the retransmission point, τ_d. Further investigation may be performed to find an efficent way to learn the required channel statistics and to adapt the LR-PET procedure to dynamically varying channel statistics.

APPENDIX A

Theorem 1: If $k_0 < k_1 < \cdots < k_m$ is an enumeration of superscripts for the elements $\{\phi_h^k\}$ on the lower convex hull of $\Phi_h^{k_{\min}(r_q)}$, the optimal indices $\{\tilde{s}_{q,k}^h\}$ which maximize (21) under the constraint $\tilde{s}_{q,k}^h \leqslant \tilde{s}_{q,k+1}^h$, $\forall k \in [h, k_{\min}(r_q) - 1)$ must satisfy $\tilde{s}_{q,k}^h = \tilde{s}_{q,k_j}^h$, $\forall k \in (k_{j-1}, k_j]$.

Proof: Recall that $\Phi_h^M = \{\phi_h^{h-1}, \phi_h^h, \ldots, \phi_h^{M-1}\}$, $\phi_h^k = \sum_{i=h}^k \rho_{i|h}^1$, $\phi_h^{h-1} = 0$. The lower convex hull of Φ_h^M can be constructed as follows. (1) Initialize an ordered set $\mathbb{I}$ as $\{h-1, h, \ldots, M-1\}$. (2) Pick up any three neighboring indices $i_1 < i_2 < i_3$ from $\mathbb{I}$. (3) Check the slope values $s_2 = (\phi_h^{i_2} - \phi_h^{i_1})/(i_2 - i_1)$ and $s_3 = (\phi_h^{i_3} - \phi_h^{i_2})/(i_3 - i_2)$. If $s_2 > s_3$, remove the index i_2 from $\mathbb{I}$. (4) Go back and repeat (2) and (3) iteratively, until $((\phi_h^{i_2} - \phi_h^{i_1})/(i_2 - i_1)) \leqslant ((\phi_h^{i_3} - \phi_h^{i_2})/(i_3 - i_2))$ is satisfied for any three neighboring indices $i_1 < i_2 < i_3$ in $\mathbb{I}$. The indices ultimately remained in $\mathbb{I}$ is $k_0, k_1, \cdots, k_m$.

To prove the theorem, we only need to generally prove that each time we remove index i_2 from $\mathbb{I}$ due to $s_2 > s_3$, we must have $\tilde{s}_{q,k}^h = \tilde{s}_{q,i_3}^h$, $\forall k \in (i_1, i_3]$. Suppose that $i_1 < i_2 < i_3$ are the three neighboring indices from $\mathbb{I}$ we are going to consider and we have proved $\tilde{s}_{q,k}^h = \tilde{s}_{q,i_2}^h$, $\forall k \in (i_1, i_2]$ and $\tilde{s}_{q,k}^h = \tilde{s}_{q,i_3}^h$, $\forall k \in (i_2, i_3]$. The contribution of $\{\tilde{s}_{q,k}^h\}$, $k \in (i_1, i_3]$ to $J_q(\lambda)$ is

$$\Delta J_q(\lambda) = \left(\phi_h^{i_2} - \phi_h^{i_1}\right) P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_2}^h\right) - \lambda_q \cdot \frac{i_2 - i_1}{k_{\min}(r_q)} R\left(\tilde{s}_{q,i_2}^h\right) + \left(\phi_h^{i_3} - \phi_h^{i_2}\right) P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_3}^h\right) - \lambda_q \cdot \frac{i_3 - i_2}{k_{\min}(r_q)} R\left(\tilde{s}_{q,i_3}^h\right). \quad (43)$$

By hypothesis, $\Delta J_q(\lambda)$ is at least as large as the solution obtained if $\tilde{s}_{q,i_2}^h$ or $\tilde{s}_{q,i_3}^h$ are used for all the indices $\{\tilde{s}_{q,k}^h\}$, $k \in (i_1, i_3]$. That is, we must have

$$\Delta J_q(\lambda) \geqslant \left(\phi_h^{i_3} - \phi_h^{i_1}\right) P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_2}^h\right) - \lambda_q \cdot \frac{i_3 - i_1}{k_{\min}(r_q)} R\left(\tilde{s}_{q,i_2}^h\right) \quad (44)$$

$$\Delta J_q(\lambda) \geqslant \left(\phi_h^{i_3} - \phi_h^{i_1}\right) P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_3}^h\right) - \lambda_q \cdot \frac{i_3 - i_1}{k_{\min}(r_q)} R\left(\tilde{s}_{q,i_3}^h\right). \quad (45)$$

Substituting (43) in (44) and (45) respectively, we easily get

$$\frac{\phi_h^{i_3} - \phi_h^{i_2}}{i_3 - i_2}\left(P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_3}^h\right) - P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_2}^h\right)\right) \geqslant \frac{R\left(\tilde{s}_{q,i_3}^h\right) - R\left(\tilde{s}_{q,i_2}^h\right)}{k_{\min}(r_q)/\lambda_q} \quad (46)$$

$$\frac{\phi_h^{i_2} - \phi_h^{i_1}}{i_2 - i_1}\left(P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_2}^h\right) - P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_3}^h\right)\right) \geqslant \frac{R\left(\tilde{s}_{q,i_2}^h\right) - R\left(\tilde{s}_{q,i_3}^h\right)}{k_{\min}(r_q)/\lambda_q}. \quad (47)$$

Combining (46) and (47), we obtain

$$\left(\frac{\phi_h^{i_3} - \phi_h^{i_2}}{i_3 - i_2} - \frac{\phi_h^{i_2} - \phi_h^{i_1}}{i_2 - i_1}\right) \cdot \left(P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_3}^h\right) - P_{\boldsymbol{\rho}^2}\left(\tilde{s}_{q,i_2}^h\right)\right) \geqslant 0. \quad (48)$$

If $\tilde{s}_{q,i_2}^h < \tilde{s}_{q,i_3}^h$, we have $P_{\boldsymbol{\rho}^2}(\tilde{s}_{q,i_3}^h) \geqslant P_{\boldsymbol{\rho}^2}(\tilde{s}_{q,i_2}^h)$, so that (48) must lead to $((\phi_h^{i_3} - \phi_h^{i_2})/(i_3 - i_2)) \geqslant ((\phi_h^{i_2} - \phi_h^{i_1})/(i_2 - i_1))$. Equivalently, if $((\phi_h^{i_3} - \phi_h^{i_2})/(i_3 - i_2)) < ((\phi_h^{i_2} - \phi_h^{i_1})/(i_2 - i_1))$, i.e., index i_2 is removed from $\mathbb{I}$, we must have $\tilde{s}_{q,i_2}^h = \tilde{s}_{q,i_3}^h$. ■

REFERENCES

[1] A. Albanese, J. Blömer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoding transmission," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pt. 1, pp. 1737–1744, Nov. 1996.

[2] R. Puri and K. Ramchandran, "Multiple description source coding using forward error correction codes," in *Proc. 33rd Asilomar Conf. Signals, Systems, and Computers*, Oct. 1999, vol. 1, pp. 342–346.

[3] A. Mohr, R. Ladner, and E. Riskin, "Approximately optimal assignment for unequal loss protection," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2000, vol. 1, pp. 367–370.

[4] T. Stockhammer and C. Buchner, "Progressive texture video streaming for lossy packet networks," in *Proc. 11th Int. Packet Video Workshop*, May 2001.

[5] S. Dumitrescu, X. Wu, and Z. Wang, "Globally optimal uneven error-protected packetization of scalable code streams," in *Proc. IEEE Data Compression Conf.*, Apr. 2002, pp. 73–82.

[6] V. Sankovic, R. Hamzaoui, and Z. Xiong, "Packet loss protection of embedded data with fast local search," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2002, vol. 2, pp. 165–168.

[7] J. Thie and D. Taubman, "Optimal protection assignment for scalable compressed images," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2002, vol. 3, pp. 713–716.

[8] M. Podolsky, M. Vetterli, and S. McCanne, "Limited retransmission of real-time layered multimedia," in *Proc. IEEE 2nd Workshop Multimedia Signal Processing*, Dec. 1998, pp. 591–596.

[9] M. Podolsky, S. McCanne, and M. Vetterli, "Soft arq for layered streaming media," *J. VLSI Signal Process. Signal, Image, Video Tech.*, vol. 27, pp. 81–97, Feb. 2001.

[10] V. Chande, N. Farvardin, and H. Jafarkhani, "Image communication over noisy channels with feedback," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1999, vol. 2, pp. 540–544.

[11] C.-Y. Hsu and A. Ortega, "Rate control for robust video transmission over burst-error wireless channels," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 5, pp. 756–773, May 1999.

[12] P. Chou and Z. Miao, Rate-Distortion Optimized Streaming of Packetized Media Microsoft Tech. Rep. MSR-TR-2001-35, 2001.

[13] G. Cheung, W.-T. Tan, and T. Yoshimura, "Rate-distortion optimized application-level retransmission using streaming agent for video streaming over 3 g wireless network," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2002, vol. 1, pp. 529–532.

[14] R. Puri and K. Ramchandran, "Joint source channel coding with hybrid fec/arq for buffer constrained video transmission," in *Proc. IEEE 2nd Workshop Multimedia Signal Processing*, Dec. 1998, pp. 567–572.

[15] P. Chou, A. Mohr, A. Wang, and S. Mehrotra, "Fec and pseudo-arq for receiver-driven layered multicast of audio and video," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2000, pp. 440–449.

[16] J. Chakareski, P. Chou, and B. Aazhang, "Computing rate-distortion optimized policies for streaming media to wireless clients," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Apr. 2002, pp. 53–62.

[17] G. Wang, Q. Zhang, Q. Zhu, and Y.-Q. Zhang, "Channel-adaptive error control for scalable video over wireless channel," in *Proc. 7th Int Workshop on Mobile Multimedia Communications (MoMuC)*, Oct. 2000.

[18] T. Stockhammer, H. Jenak, and C. Weiss, "Feedback and error protection strategies for wireless progressive video transmission," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 465–482, Jun. 2002.

[19] X. Zheng, S.-H. Chan, Q. Zhang, W.-W. Zhu, and Y.-Q. Zhang, "Feedback-free packet loss recovery for video multicast," in *Proc. IEEE Int. Conf. Communications*, Apr. 2003, vol. 2, pp. 870–874.

[20] T. Gan and K.-K. Ma, "Sliding-window packetization for forward error correction based multiple description transcoding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 2003, vol. 5, pp. 756–759.

[21] D. Taubman and J. Thie, “Optimal erasure protection for scalably compressed video streams with limited retransmission,” *IEEE Trans. Image Process.*, vol. 14, no. 8, pp. 1006–1019, Aug. 2005.

[22] J. Thie and D. Taubman, “Optimal erasure protection for scalably compressed video streams with limited retransmission on channels with iid and bursty loss characteristics,” *Signal Process. Image Commun.*, vol. 20, no. 8, pp. 697–709, May 2005.

[23] M. Durigon and D. Taubman, “Lr-pet optimization strategy for protection of scalable video with unreliable acknowledgement,” in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2006, pp. 9–12.

[24] R. Xiong and D. Taubman, “Optimal pet protection for scalably compressed video with multiple retransmissions,” in *Proc. Visual Information Engineering 2008*, Jul. 29.–Aug. 1 2008, pp. 795–800.

[25] R. Xiong, D. Taubman, and V. Sivaraman, “Optimal pet protection for streaming scalably compressed source data with multiple retransmissions,” *IEEE Trans. Information Theory*, submitted for publication.

[26] J. Shapiro, “An embedded hierarchical image coder using zerotrees of wavelet coefficients,” in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, 1993, pp. 214–223.

[27] A. Said and W. Pearlmanc, “A new, fast and efficient image codec based on set partitioning in hierarchical trees,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, Jun. 1996.

[28] D. Taubman, “High performance scalable image compression with ebcot,” *IEEE Trans. Image Process.*, vol. 9, no. 7, pp. 1158–1170, Jul. 2000.

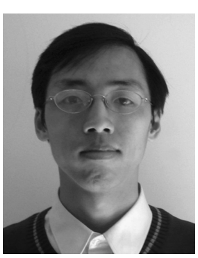
[29] J. Ohm, “Three dimensional subband coding with motion compensation,” *IEEE Trans. Image Process.*, vol. 3, no. 4, pp. 559–571, Apr. 1994.

[30] D. Taubman and A. Zakhor, “Highly scalable low-delay video compression,” in *Proc. IEEE Int. Conf. Image Processing*, Nov. 1994, vol. 1, no. 4, pp. 740–744.

[31] S. Choi and J. Woods, “Motion compensated 3-d subband coding of video,” *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[32] B. Pesquet-Popescu and V. Bottreu, “Three-dimensional lifting schemes for motion compensated video compression,” in *Proc. Int. Conf. Acoustic, Speech, Signal Processing*, 2001, pp. 1793–1796.

[33] A. Secker and D. Taubman, “Lifting-based invertible motion adaptive transform (limat) framework for highly scalable video compression,” *IEEE Trans. Image Process.*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.



Ruiqin Xiong (M’08) received the B.S. degree from the University of Science and Technology of China (USTC), Hefei, in 2001, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences (ICT,CAS), Beijing, in 2007.

He was a Research Intern at Microsoft Research Asia from August 2002 to July 2007, and a Senior Research Associate at the University of New South Wales, Sydney, Australia, from September 2007 to Aug 2009. He joined Peking University, Beijing, China, in September 2009. His research interests include image and video compression, multimedia transmission, channel coding, and distributed coding.

David Taubman (M’95–SM’06) received the B.S. and B.E. (Electrical) degrees from the University of Sydney, Australia, in 1986 and 1988, respectively, and the M.S. and Ph.D. degrees from the University of California at Berkeley, in 1992 and 1994, respectively.

From 1994 to 1998, he worked at Hewlett-Packard’s Research Laboratories in Palo Alto, CA, joining the University of New South Wales, Sydney, Australia, in 1998, where he is a Professor in the School of Electrical Engineering and Telecommunications. His research interests include highly scalable image and video compression, inverse problems in imaging, perceptual modeling, joint source/channel coding, and multimedia distribution systems.

Dr. Taubman is author with M. Marcellin of the book, *JPEG2000: Image compression fundamentals, standards and practice*. He was awarded the University Medal from the University of Sydney, the Institute of Engineers, Australia, Prize and the Texas Instruments Prize for Digital Signal Processing, all in 1998. He has received two Best Paper awards: from the IEEE Circuits and Systems Society for the 1996 paper, “A Common Framework for Rate and Distortion Based Scaling of Highly Scalable Compressed Video;” and from the IEEE Signal Processing Society for the 2000 paper, “High Performance Scalable Image Compression with EBCOT.”

Vijay Sivaraman (M’94) received the B. Tech. degree from the Indian Institute of Technology, Delhi, India, in 1994, the M.S. degree from North Carolina State University, Raliegh, in 1996, and the Ph.D. degree from the University of California at Los Angeles in 2000.

He has worked at Bell-Labs and a Silicon Valley startup manufacturing optical switch-routers. He is currently a Senior Lecturer at the University of New South Wales, Sydney, Australia, and works part-time at the CSIRO. His research interests include optical networking, packet switching and routing, network design, and QoS.