An Integrated Source Transcoding and Congestion Control Paradigm for Video Streaming in the Internet

Rohit Puri, Kang-Won Lee, Kannan Ramchandran, *Member, IEEE*, and Vaduvur Bharghavan

Abstract—In this work we present an end-to-end optimized video streaming system comprising of synergistic interaction between a source packetization strategy and an efficient and responsive, *TCP-friendly* congestion control protocol [Linear Increase Multiplicative Decrease with History (LIMD/H)]. The proposed source packetization scheme transforms a scalable/layered video bitstream so as to provide graceful resilience to network packet drops. The congestion control mechanism provides low variation in transmission rate in steady state and at the same time is reactive and provably TCP-friendly.

While the two constituent algorithms identified above are novel in their own right, a key aspect of this work is the integration of these algorithms in a simple yet effective framework. This "application-transport" layer interaction approach is used to maximize the expected delivered video quality at the receiver. The integrated framework allows our system to gracefully tolerate and quickly react to sudden changes in the available connection capacity due to the onset of congestion, as verified in our simulations.

Index Terms—Application-transport layer, end-to-end congestion, multimedia transport protocols, multiple description coding.

I. INTRODUCTION

IN THE current deployment of the Internet, switches in the network are typically oblivious to the structure or content of the packets and treat all packets equally. While this simple network design has enabled the Internet to scale to its present size, it is ill-matched to, and adversely impacts multimedia flows. Typical multimedia streams are highly structured, i.e., characterized by a hierarchy of importance layers (e.g., I, P versus B frames in MPEG) and hence are fundamentally mismatched to the existing switching strategies. Furthermore, some of the emerging multimedia compression standards for image/video coding are evolving toward a multiresolution (MR) or layered representation of the coded bitstreams, JPEG-2000 and MPEG-4[1] being two popular examples. Thus we see that there is an inherent mismatch between the source coding algorithms that are multireso-

lution in character, and the network mechanisms in the Internet that do not discern prioritized classes.

Multiple description (MD) source coding has recently emerged as an attractive framework for robust transmission over unreliable channels. The essential idea is to generate multiple independent descriptions (N in number) of the source such that each description independently describes the source with a certain desired fidelity; when more than one description is available, they can be synergistically combined to enhance the quality. Note that while the MR bitstream is sensitive to the *position* of losses in the bitstream (e.g., a loss early on in the bitstream can render the rest of the bitstream useless to the decoder) the MD stream is insensitive to them and thus has the desired feature that the delivered quality is dependent only on the *fraction of packets* delivered reliably. While most of the initial work in this area has focused on the special case of $N = 2$ descriptions [4], [5], there has been recent interest in the $N > 2$ case also [6]–[9]. For example, [9] presents an efficient algorithm that transforms an MR-based prioritized bitstream into a nonprioritized MD packet stream. The proposed strategy uses a purely channel coding based paradigm based on forward error correction (FEC) channel codes and exhibits a departure from traditional signal processing based approaches that employ scalar quantizers, correlating transforms etc. [4], [5] to approach the MD problem. It incorporates the Priority Encoding Transmission (PET) [10] philosophy in formulating a systematic, fast and end-to-end rate-distortion optimized algorithm. Details of this are presented in Section II.

The "smart end-hosts/simple network" philosophy which has enabled the Internet to scale to its present size, entrusts to the end user, the responsibility of maintaining transmission rates that are fair to other connections. Two extreme transport services that have emerged under this model are the reliable transmission control protocol (TCP) [11] that has a sophisticated transmission rate control mechanism and the unreliable, best effort service called the User Datagram Protocol (UDP) [12] that can potentially be unfair to the "well-behaved" TCP flows. TCP is not well-suited for multimedia transmission (typically loss-tolerant and delay sensitive) for multiple reasons. First, the feature of reliability that it offers is unnecessary for multimedia applications and furthermore obtained at the expense of large round trip delays since reliability is attained through repeated retransmissions (ARQ).

Secondly, TCP employs an end-to-end congestion control paradigm based on the linear increase multiplicative decrease (LIMD) philosophy. The source continuously probes the network by incrementing its transmission rate (Linear Increase) when it encounters no loss. However, upon loss, it throttles

[1]There has been considerable effort for instance, within the MPEG-4 video coding standard on the issue of supporting Fine-Granular Scalability (FGS) [1]–[3].
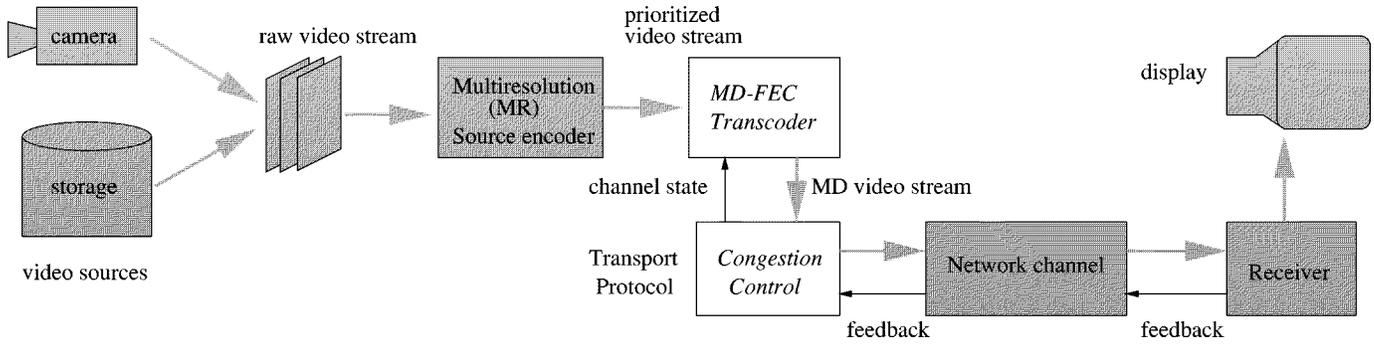
Fig. 1. Block diagram of the end-to-end video transmission system.

down its transmission rate by a large scaling factor (multiplicative decrease). The LIMD approach reacts aggressively to packet loss which enables quick response to the onset of congestion; however it leads to undesirably large variations in the delivered transmission rate even in steady state. Such an approach can have an adverse impact on multimedia flows since for such applications, "smoothness" in delivered quality is an important consideration from the end user's point of view.[2] Note that the variations in sender's transmission rate could theoretically be smoothed out with application-level buffering, but this could translate into huge client buffers and unacceptable end-to-end delays depending on the extent of fluctuations.

The design of multimedia protocols thus dictates that transmission rate variations be small, while also requiring the congestion control policy to be TCP-friendly, namely, that it should not grab more bandwidth from the sharing link that a conforming TCP flow would. Equation-based rate control which can be construed as being at the opposite end of the spectrum from standard LIMD approaches, has been recently proposed, [13], as an attractive option for multimedia Internet transmission. It has the desirable feature of delivering maximally smooth, TCP-fair transmission rate. However, it is characterized by very slow responsiveness to network dynamics. In this work we present a new congestion control algorithm, based on Linear Increase and graded Multiplicative Decrease, that can be conceptualized as generalizing both the LIMD and equation based approach. It explores an extra degree of freedom which continuously trades off "smoothness" for "responsiveness." We dub our approach LIMD with history or LIMD/H.

Recent work [14]–[16] on multimedia streaming has highlighted the benefits of joint design of the source coding schemes and the transport layer protocols. Our end-to-end system design subscribes to this inter-layer interaction philosophy. See Fig. 1 for the end-to-end block diagram of our system. While the source transcoding algorithm and the congestion control algorithm are novel in their own right, a key feature of our approach is their simple and synergistic coordination in an integrated architecture.

The remainder of this paper is organized as follows. Section II describes the FEC-based robust source transcoding (MD-FEC)

scheme. Section III presents the congestion control algorithm and the interaction between the application and the transport layer. In Section IV, we evaluate the performance of our system using the ns-2 network simulator [17]. In Section V, we place our work in the context of related work, and Section VI concludes the paper and offers directions for future work.

II. Forward Error Correction Codes Based Multiple Description Coding

In this section, we briefly describe the mechanics of the packetization strategy that converts the prioritized MR[3] bitstream into an N-packet unprioritized MD packet stream using efficient erasure channel codes. Each description in the MD stream occupies an entire network packet, thus the terms "description" and "packet" are used interchangeably. Later, we present an efficient algorithm based on Lagrangian optimization principles for allocating the total rate between source and channel codes tailoring the MD stream to the transmission channel at hand.

A. Transcoding Mechanism

The quality profile reflects the target quality (or equivalently distortion d: lower distortion implies higher quality and vice versa) when any k out of N descriptions are received. We will use the notation $\mathbf{d}(k)$ to describe the quality profile where the ith entry in $\mathbf{d}(k)$ represents the target quality when i descriptions are received.

Given N, $\mathbf{d}(k)$ and a progressive bitstream, the stream is marked at N positions (see Fig. 2) which corresponds to the attainment of distortion levels $\mathbf{d}(k)$, and is thus partitioned into N sections or resolution layers.

The goal is to enable the ith layer to be decodable when i or more descriptions arrive at the receiver (i.e., when the number of erasures does not exceed $N - i$). This can be attained using the Reed–Solomon family of erasure-correction block codes, which are characterized by the optimal code parameters (N, i, $N - i + 1$) and can correct any ($N - i$) erasures out of N descriptions. We

---

[2]Subjectively, it has been found that large deviation in delivered picture quality leads to a bad user experience.

[3]Bitstreams which have the property of successive refinability, i.e., new blocks of data refine the previous blocks of data are referred to as multiresolution or scalable bitstreams. Those with a large refinement block size are typically referred to as layered bitstreams while bitstreams for which the refinement block size is theoretically, of the order of a bit are referred to as "bit-progressive" bitstreams. The analysis presented in this section applies directly to the latter category of bitstreams, but it can be generalized to the case of layered bitstreams such as those that can be generated by H.263+/MPEG video encoders.

split the ith layer into i equal parts, and apply the $(N, i, N-i+1)$ Reed–Solomon code to get the contribution from the ith layer or section to each of the N descriptions. The contributions from each of the N levels or sections are then concatenated to form the N descriptions (see Fig. 3). Thus, every description contains all N layers, and all N descriptions are "equal" in information content. This packetization strategy thus provides the property that the greater the number of packets received, the better the received quality.

We have so far shown the mechanics of our transcoding algorithm but have not yet addressed the question of how to optimize the rate partitions in the framework so as to match it to transmission channel at hand. We address these questions in the following section.

B. Optimization Problem Statement

In this section, we formulate the optimization problem. We assume that our model is characterized by $q_i(N)$, $i = -1, \cdots, N-1$ denoting the probability that $i+1$ out of N packets are delivered to the destination. We coin the term *transmission profile* to refer to the above channel state information.

From operational rate-distortion theory [18], it follows that distortion is a one-to-one function ($D(r)$) of the rate r. Hence determining the quality profile $\mathbf{d}(k)$ of order N corresponds to finding the rate partition $\{R_0, R_1, \cdots, R_{N-1}\}$ of the bitstream (see Fig. 2). Define the expected distortion ED:

$$ED = q_{-1} \cdot E + \sum_{j=0}^{N-1} q_j \cdot D(R_j) \qquad (1)$$

where E equals the source variance, the distortion encountered when the source is represented by zero bits. When the codes as outlined in the above section are used, the total rate used R_t equals

$$R_t = \frac{R_0}{1} \cdot N + \frac{(R_1 - R_0)}{2} \cdot N + \cdots + \frac{(R_{N-1} - R_{N-2})}{N} \cdot N \quad (2)$$

or equivalently

$$R_t = \sum_{j=0}^{N-1} \alpha_j \cdot R_j \qquad (3)$$

where

$$\alpha_j = \frac{N}{(j+1) \cdot (j+2)}, \qquad j = 0, \cdots, N-2 \qquad (4)$$

and $\alpha_{N-1} = 1$.

Problem Statement: Given the number of packets N, each packet of size L (i.e., a total rate budget $R^* = NL$), an embedded bit stream with rate-distortion curve $D(r)$, and the transmission profile q_i, find the rate partition $\{R_0, R_1, \cdots, R_{N-1}\}$ that minimizes ED subject to

$$(a) \ R_t = \sum_{j=0}^{N-1} \alpha_j \cdot R_j \leq R^*, \qquad (5)$$

$$(b) \ R_0 \leq R_1 \leq \cdots \leq R_{N-2} \leq R_{N-1}, \qquad (6)$$

$$(c) \ R_1 - R_{i-1} = k_i \cdot (i+1), \qquad k_i \geq 0, \ i = 1, \cdots, N-1. \qquad (7)$$
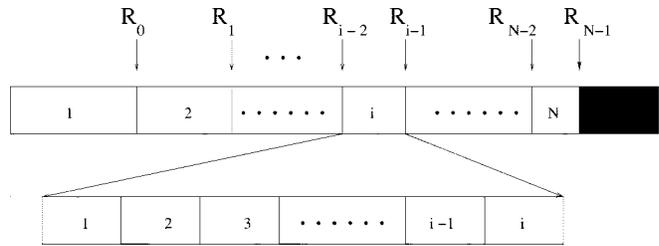

Fig. 2. Progressive bitstream from the source coder partitioned into N layers or quality levels.
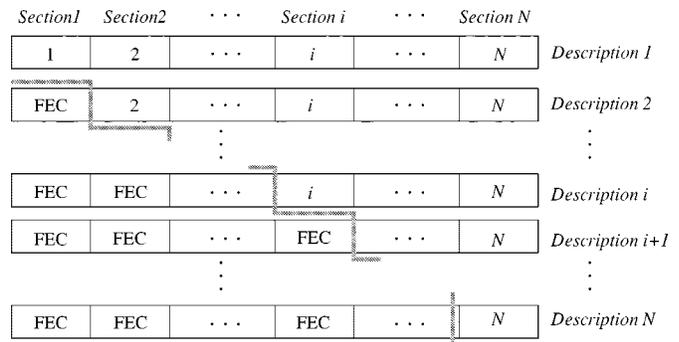

Fig. 3. N-description generalized MD codes using forward error correction codes.
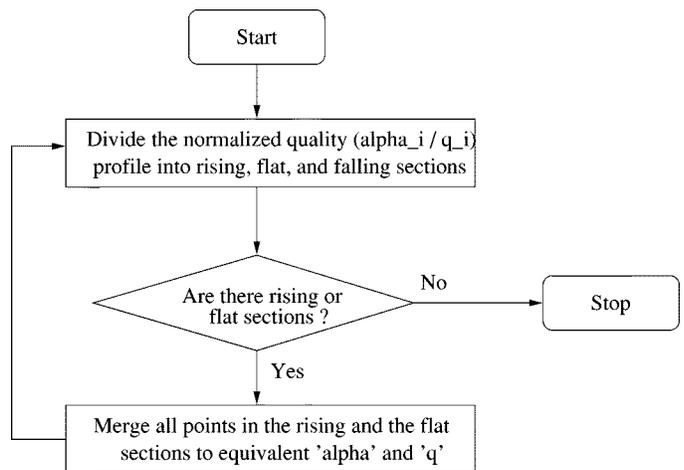

Fig. 4. The monotonizing algorithm that gives the "form" of the optimal solution.

It turns out that the optimal solution to this constrained[4] problem can be obtained by using a computationally intensive dynamic program $O(R^2 N)$ [9]. However, if a "fluid" model were to hold true for the source i.e., the discretization artifacts arising because of constraint (7) are nonexistent, the optimal solution can be obtained in time $O(N)$ with the algorithm proposed in Section II-C! Our approach hence consists of first solving the problem optimally under the "fluid" approximation

---

[4]The total rate constraint (5) is referred as the "resource" constraint. The constraint (6) is referred to as the "embedding" constraint and it arises due to the physical nature of the bitstream i.e., a greater number of source layers would use a greater amount of rate (see Fig. 2). The constraint (7) is referred to as the "channel coding" constraint and arises from the fact that in order to apply an $(n, i, n - i + 1)$ Reed–Solomon code to the ith layer, the ith layer should contain a number of source symbols that is an integral multiple of i.
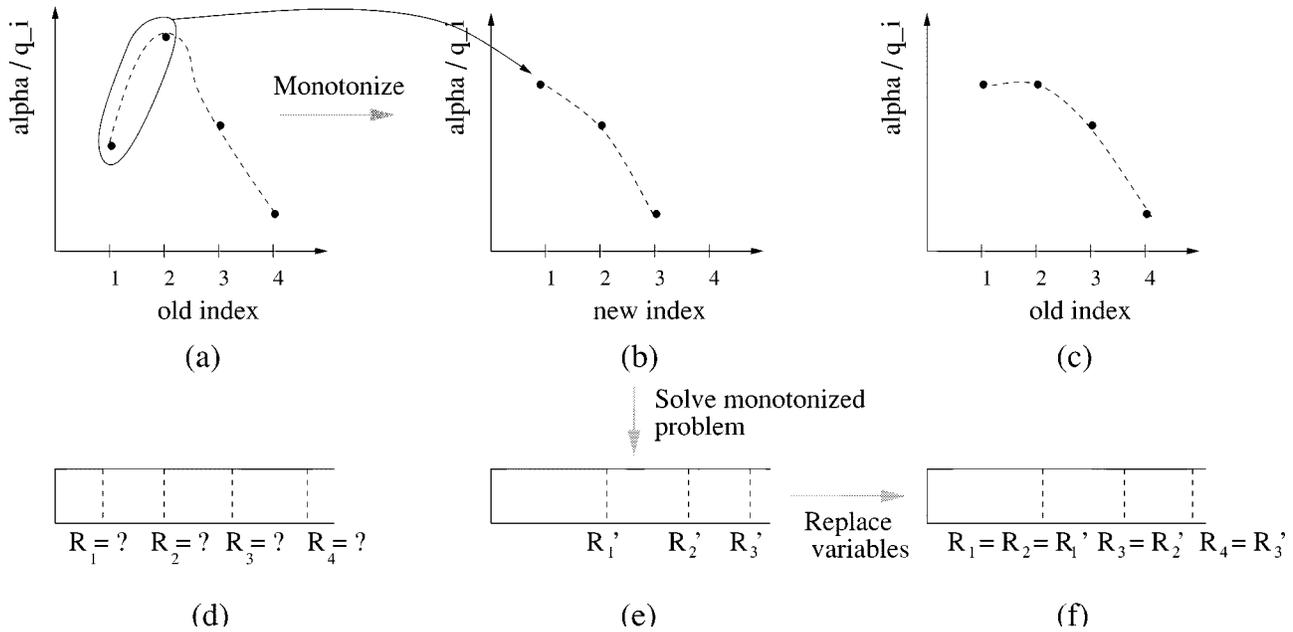
Fig. 5. The optimal allocation algorithm example: (a) $(\alpha_1/q_2) = (0.05/0.5), (\alpha_2/q_2) = (0.1/0.5), (\alpha_3/q_3) = (0.03/0.25), (\alpha_4/q_4) = (0.01/0.25)$. (b) After applying the algorithm, we get in terms of new variables: $(\alpha_1'/q_1') = (0.075/0.5), (\alpha_2'/q_2') = (0.03/0.25), (\alpha_3'/q_3') = (0.01/0.25)$. (c) Expressing in terms of the original variables: $(\alpha_1/q_1) = (0.075/0.5), (\alpha_2/q_2) = (0.075/0.5), (\alpha_3/q_3) = (0.03/0.25), (\alpha_4/q_4) = (0.01/0.25)$. (d) Unknown rate variables. (e) Solution to the monotonized problem in terms of new variables. (f) Solution in terms of the original variables.
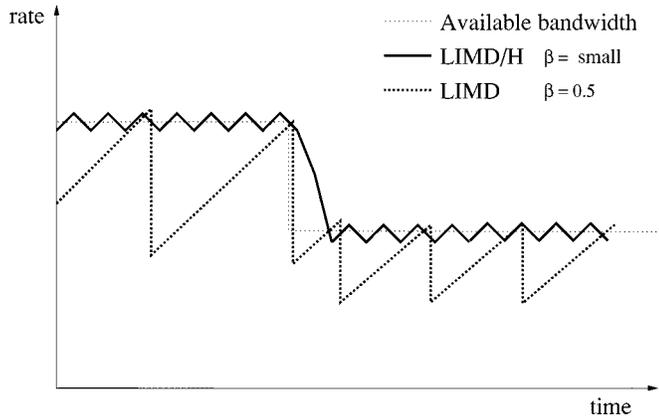

Fig. 6. Illustration of the adaptations of LIMD and LIMD/H when the available bandwidth of the network has suddenly decrerased.

and then perturbing the solution so as to satisfy the constraints (7).

C. The Allocation Algorithm

We propose a fast, nearly-optimal algorithm based on Lagrange multipliers for choosing the rate markers $\mathbf{d}(k)$. The analysis presented here deals with the case of a continuous R-D curve, although it can be extended in a straightforward manner to the discrete case [19] too up to a convex hull approximation. To justify the intuition for the proposed algorithm we initially confine ourselves to solving the problem when the "channel coding" constraint (7), which represents a discretization artifact, is not present. In what follows, the "optimal" solution refers to the best solution corresponding to the situation when only the constraints (5) and (6) are present.

We note that the R-D curve for any source is theoretically a convex curve with the property that $\lambda(R_i) > \lambda(R_j)$ for $i < j$

where $\lambda(R)$ denotes the absolute value of the slope at point $(R, D(R))$. We observe that if the "embedding" constraints (6) were not present, the problem stated above would default to a standard bit resource allocation problem [19], generalized to include the notion of "weights" in the form of α_i and q_i. The optimal solution subject only to (5) is easily found using the theory of Lagrange Multipliers [20] and we briefly illustrate the procedure. Introducing the Lagrangian [20] for this problem, we get

$$L_c(R_1, \cdots, R_{N-1}, \lambda)$$
$$= q_{-1} \cdot E + \sum_{j=0}^{N-1} q_j \cdot D(R_j) + \lambda \left(\sum_{j=0}^{N-1} \alpha_j \cdot R_j - R_t^* \right). \tag{8}$$

At optimality, the partial derivative of the Lagrangian function with respect to R_i, $i = 0, \cdots, N-1$ and λ equals zero. This yields

$$\frac{q_i}{\alpha_i} \cdot \frac{dD(R_i)}{dR_i} + \lambda = 0, \qquad i = 0, \cdots, N-1. \tag{9}$$

In other words, the optimal solution is obtained by locating the points on the R-D curve whose slope is $(\alpha_i/q_i) \cdot \lambda$. [$(\alpha_i/q_i)$ has the interpretation of normalized quality: lower (α_i/q_i) implies higher R_i hence higher quality, and vice versa.] To find the "absolute" solution, the variable λ or the Lagrange multiplier needs to be eliminated (i.e., matched to the total budget constraint) which can be done iteratively by using bisection search.

Since the product of the slope of the rate-distortion curve at the point R_i with (q_i/α_i) is a constant, it is clear that for monotonically decreasing (α_i/q_i), the absolute value of the slope of the rate distortion curve at the point R_i is a monotonically decreasing sequence in i. Since the rate distortion curve is strictly

TABLE I
COMPARISON OF DIFFERENT ROBUSTNESS SCHEMES

	MR	EEP	FUEP	MD-FEC (proposed)
Robustness	none	fixed FEC	fixed FEC	adaptive FEC
Priority notion	yes (first-k)	no (any-k)	no (any-k)	no (any-k)
Quality profile	-	cliff	staircase	staircase
Channel Coding Overhead	none	prespecified	prespecified	adaptive
Summary of performance	jittery quality; highly susceptible to losses	two step quality; fails beyond the pre-specified correction capability	multi-step quality; fails beyond the pre-specified correction capability	multi-step quality; suffers during sudden network changes but adapts quickly to the new conditions

convex, it follows that for this case, (6) is satisfied automatically. However, monotonicity of (α_i/q_i) cannot be guaranteed in general as we have no control over the channel state information q_i. Hence, in general the constraints of (6) cannot be ignored. We now prove a key result that sheds insight into the nature or "form" of the optimal solution.

Fact: If $(\alpha_i/q_i) \leq (\alpha_{i+1}/q_{i+1})$, then in the optimal solution, $R_i = R_{i+1}$.

Proof: Assume that there exists an optimal solution such that $R_i < R_{i+1}$. Then take away ΔR bits from R_{i+1} and give $((\alpha_{i+1}/\alpha_i) \cdot \Delta R)$ bits to R_i (so that the rate constraints (5), (6) are satisfied). Now, the net decrease in our cost is

$$Decrease = \Delta R \cdot \alpha_{i+1} \cdot \left[\frac{q_i}{\alpha_i} \cdot \lambda(R_i) - \frac{q_{i+1}}{\alpha_{i+1}} \cdot \lambda(R_{i+1}) \right]. \tag{10}$$

Since $(\alpha_i/q_i) \leq (\alpha_{i+1}/q_{i+1})$ and $\lambda(R_i) > \lambda(R_{i+1})$, the decrease is positive, which is a contradiction because if a solution with $R_i < R_{i+1}$ were optimal, we could still improve on the solution by successively decreasing R_{i+1} and increasing R_i while satisfying the conditions (5) and (6). Hence, $R_i = R_{i+1}$ in the optimal solution.

The above result serves to characterize the form of the optimal solution based on the nature of the (α_i/q_i) profile. We make two key observations from the above result.

- We have shown above that if $(\alpha_i/q_i) \leq (\alpha_{i+1}/q_{i+1})$, the optimal solution to the original problem is the same as that to a reduced problem where $R_i = R_{i+1}$ is replaced by R_i' so that $\alpha_i' = \alpha_i + \alpha_{i+1}$ and $q_i' = q_i + q_{i+1}$ (all other variables are the same).
- Also we see that if $(\alpha_i/q_i) \leq (\alpha_{i+1}/q_{i+1})$ then $(\alpha_i/q_i) \leq (\alpha_i'/q_i') = ((\alpha_i+\alpha_{i+1})/q_i+q_{i+1})) \leq (\alpha_{i+1}/q_{i+1})$. That is, the above observation can be successively applied to a monotonically increasing or flat section in the (α_i/q_i) profile, and all the corresponding rate variables are equal in the optimal solution thus reducing the dimensionality of the problem.
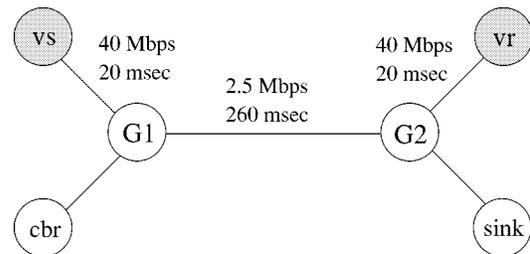

Fig. 7.   One-hop network topology.

Based on the above observations, we propose an $O(N)$ algorithm (Fig. 4) which terminates giving the simplest possible "form" for the optimal solution to the original problem. We note that, when the algorithm terminates, the final (α_i/q_i) will be a monotonic decreasing curve. For this reduced problem, with a monotonic decreasing (α_i/q_i) curve, we already have a mechanism for obtaining the solution! This action is illustrated in the simple example shown in Fig. 5.

If the "channel coding" constraint (7) were not present, then the above method provides the "optimal" solution to within a convex hull approximation. We conjecture that the overall optimal solution, i.e., including the discretization constraint (7) is NP-complete since it can be formulated as a mixed integer program. However, the following simple heuristic can be used to solve the problem using the solution $\{R_0, R_1, \cdots, R_{N-1}\}$ given by the above algorithm. The approach consists of decreasing R_i until $R_i - R_{i-1}$ becomes an integral multiple of $i + 1$, for each value of i. This way all the extra bits are "rippled" to the last quality level. The heuristic thus biases the solution toward receiving increased peak PSNR at the receiver. In the experiments, [9], conducted by us, we observed that the heuristic solution comes extremely close (within 0.1 dB) to the optimal solution almost all the time.

To summarize, we thus presented a nonprioritizing transcoding strategy that is ideally suited for multimedia transmission over the Internet. In the following section, we deal with the second aspect of the problem, namely, designing of a
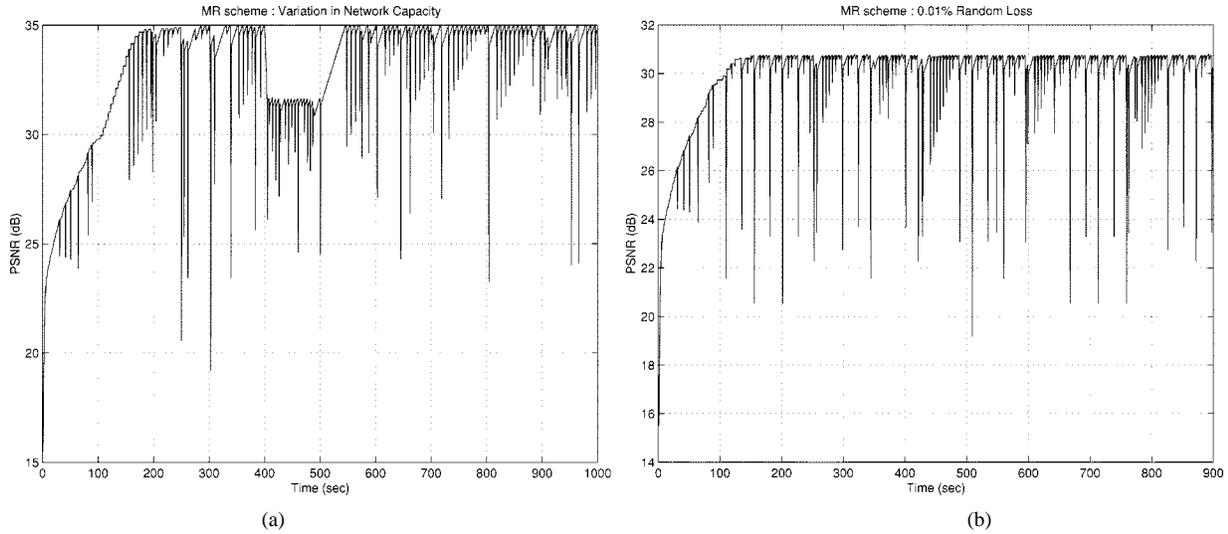
Fig. 8.   Performance of MR: (a) In steady state and upon the introduction of a CBR flow and (b) 0.01% random loss on the bottleneck link.
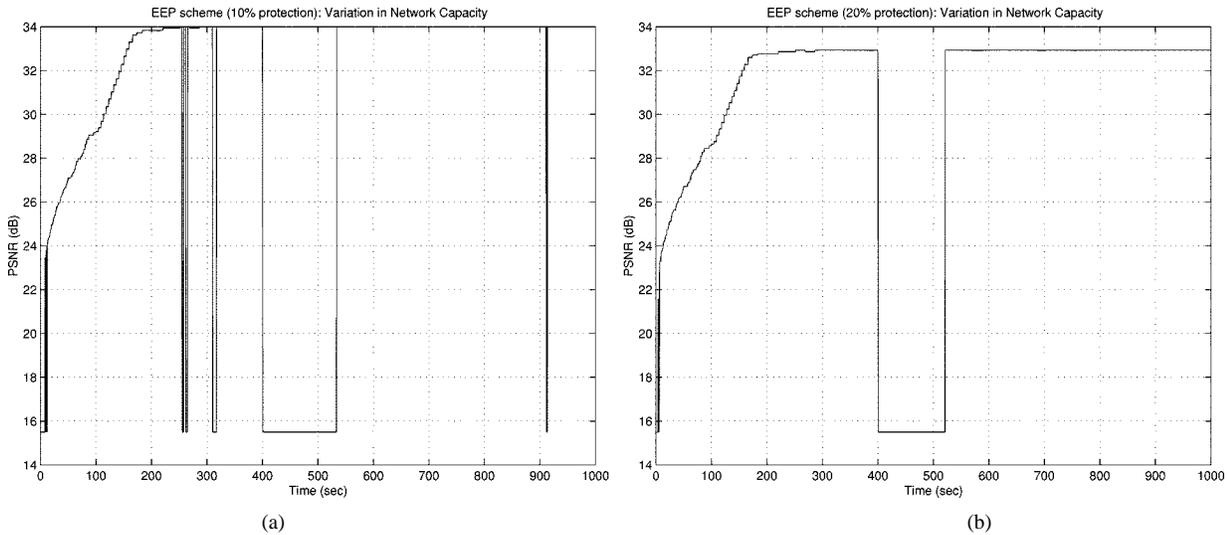

Fig. 9.   Performance of equal error protection (EEP) in a one hop topology network with no random loss depicted in Fig. 7 for varying degrees of redundancy: (a) 10% redundancy and (b) 20% redundancy.

TCP friendly congestion control protocol given that the target application is multimedia transmission.

III. CONGESTION CONTROL

In the target environment, a congestion control algorithm should provide a smooth variation of transmission rate so as to ensure lesser variation in delivered multimedia quality. In addition it should possess features such as quick response to network dynamics, efficiency, fairness, and TCP-friendliness.

A. *Motivation for LIMD/H Algorithm*

Congestion control algorithms that are deployed in the current Internet typically employ the linear increase multiplicative decrease (LIMD) paradigm for adapting the transmission rate of a connection to match the available bandwidth whether the target application is data-oriented [11] or multimedia [13], [21]. Briefly, LIMD periodically adapts the sending rate of a connection by gently increasing the rate by an additive constant α

upon observing no packet losses (in order to probe for additional bandwidth), and aggressively decreasing the sending rate by a large multiplicative constant β upon observing packet losses in order to alleviate congestion (see Fig. 6).[5]

The tradeoffs of conventional LIMD (with a large decrease factor as implemented in TCP) are quite well-known. While the LIMD paradigm has been shown to be robust and provably convergent to fairness [22], it also reacts identically and aggressively to all packet losses, both congestion induced and noncongestion induced. When there is actual congestion, LIMD throttles the sending rate sufficiently so that congestion can be cleared out quickly. But when the network bandwidth is invariant, periodic packet losses due to channel probing still induce large rate fluctuations, which is undesirable especially given the target application.

The congestion control algorithm we present in this paper called LIMD/H (LIMD with history) is an augmentation of the basic LIMD approach. It features a simple loss differentiation

---

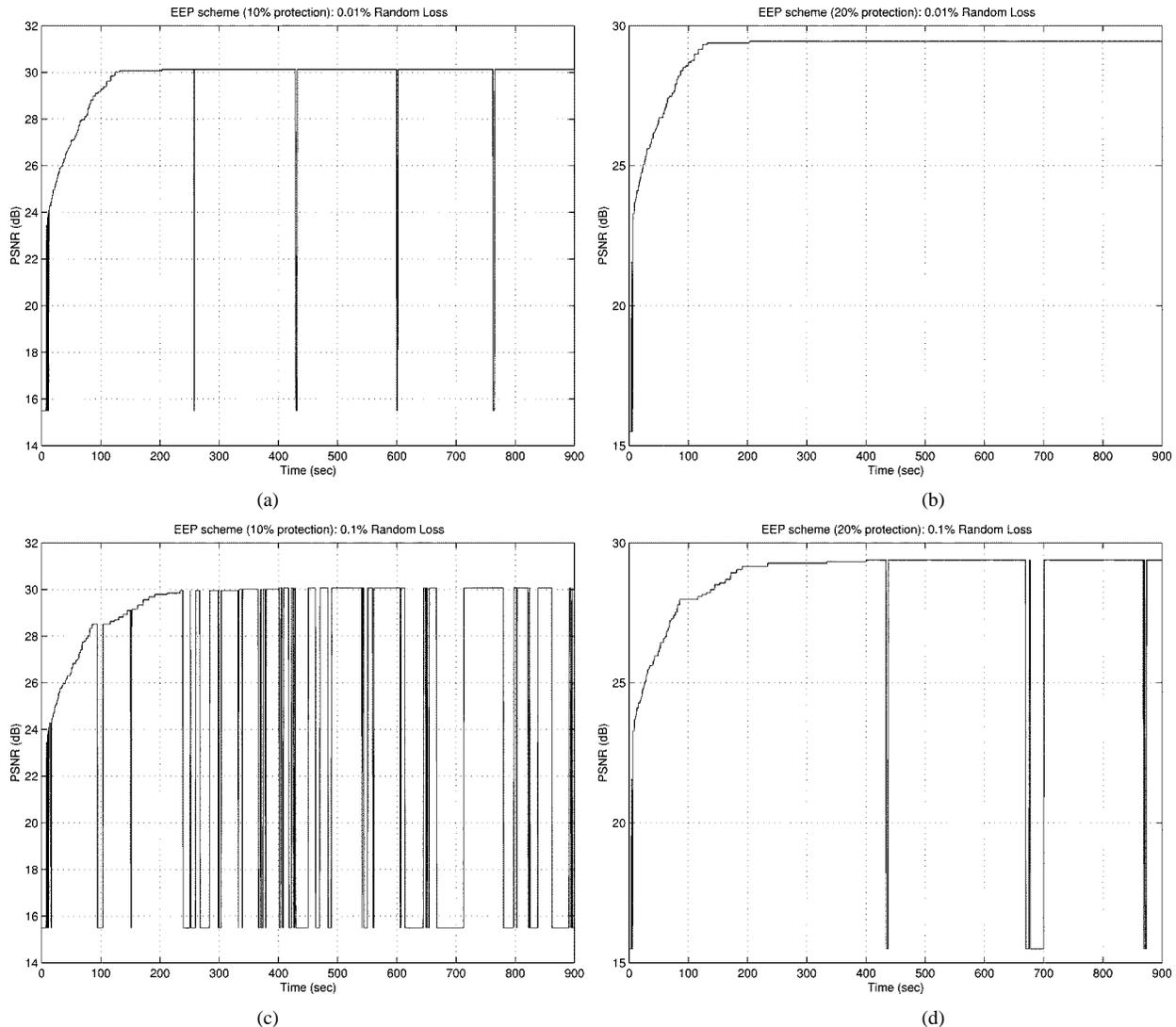[5]For TCP, the decrease factor β is set to a half.

Fig. 10. Performance of equal error protection (EEP) in a one hop topology network depicted in Fig. 7 for varying degrees of redundancy and random loss (a) 10% redundancy, 0.01% random loss, (b) 20% redundancy, 0.01% random loss, (c) 10% redundancy, 0.1% random loss, (d) 20% redundancy, 0.1% random loss.

mechanism to predict the cause of packet loss and react accordingly. The LIMD/H congestion control algorithm has the following salient attributes. First, LIMD/H uses the *history* of packet losses in order to distinguish between congestion-induced and non-congestion-induced packet losses. Secondly, it reacts gently to noncongestion-induced losses in order to keep the sending rate variation to minimum, but reacts quickly to the onset of the congestion. Thirdly, in the steady state it delivers the same average rate as any LIMD paradigm based algorithm would do and hence is TCP friendly.

We present the LIMD/H algorithm in three parts. We first describe the framework for end-to-end congestion control, then describe the LIMD/H algorithm, and finally discuss some of its properties.

B. Framework for Congestion Control

In our framework, the evolution of the transmission rate occurs over discrete time periods called *epochs*. At the end of each epoch, the receiver computes the number of received packets

during that epoch, and then sends a *congestion feedback* message to the sender containing this number. The *loss fraction* $0 \leq f \leq 1$ can thus be calculated at the sender. Upon receiving the loss feedback, the sender executes the LIMD/H congestion control algorithm to adjust its sending rate.

C. LIMD/H Congestion Control

Let r denote the sending rate, f denote the loss fraction, α denotes the linear increase constant, and β denotes the multiplicative decrease constant.

For LIMD:

- if $f = 0$, $r \leftarrow r + \alpha$.
- if $f > 0$, $r \leftarrow r \times (1 - \beta)$.

Our goal is to set the value of the rate throttling factor β based on the cause of packet loss. An intuitive heuristic to predict congestion is the following: if the packet losses observed during an epoch are caused by congestion but the sender does not throttle aggressively enough, then the loss will persist; on the other hand, if the packet loss is a probe loss, then so long as

Fig. 11. Performance comparison between MD-FEC scheme and fixed unequal error protection (FUEP) schemes with rates (60%, 80%, 95%) and (40%, 70%, 95%) respectively for varying degrees of random loss (a) MD-FEC, no loss, (b) MD-FEC, 0.01% loss, (c) FUEP (60%, 80% 95%), no loss (d) FUEP (60%, 80% 95%), 0.01% loss, (e) FUEP (40%, 70%, 95%), no loss (f) FUEP (40%, 70%, 95%), 0.01% loss.

the sender throttles enough to account for the loss, there will be no loss in the next epoch.

Taking these factors into account, we propose the LIMD/H algorithm that keeps track of a very simple *history* parameter,
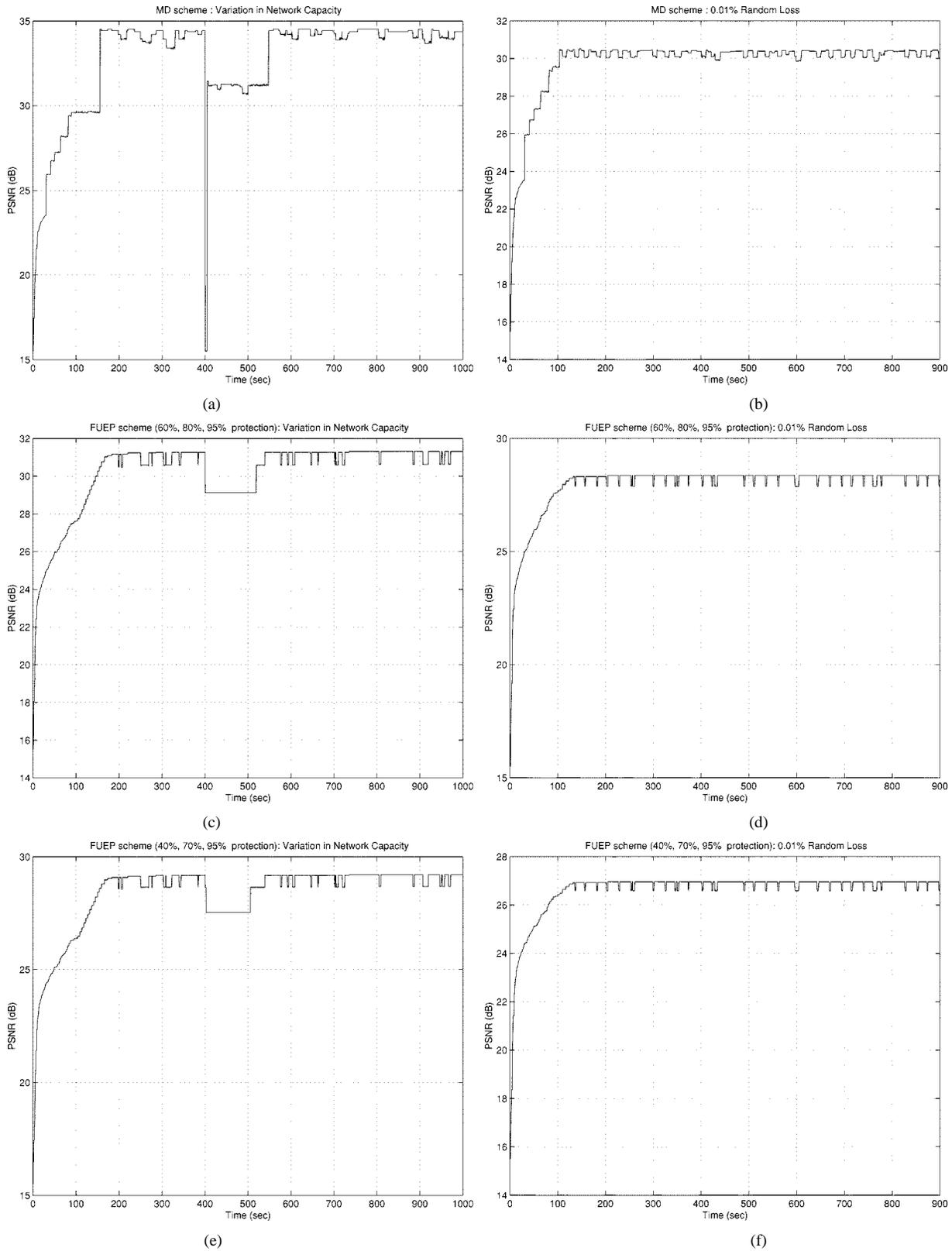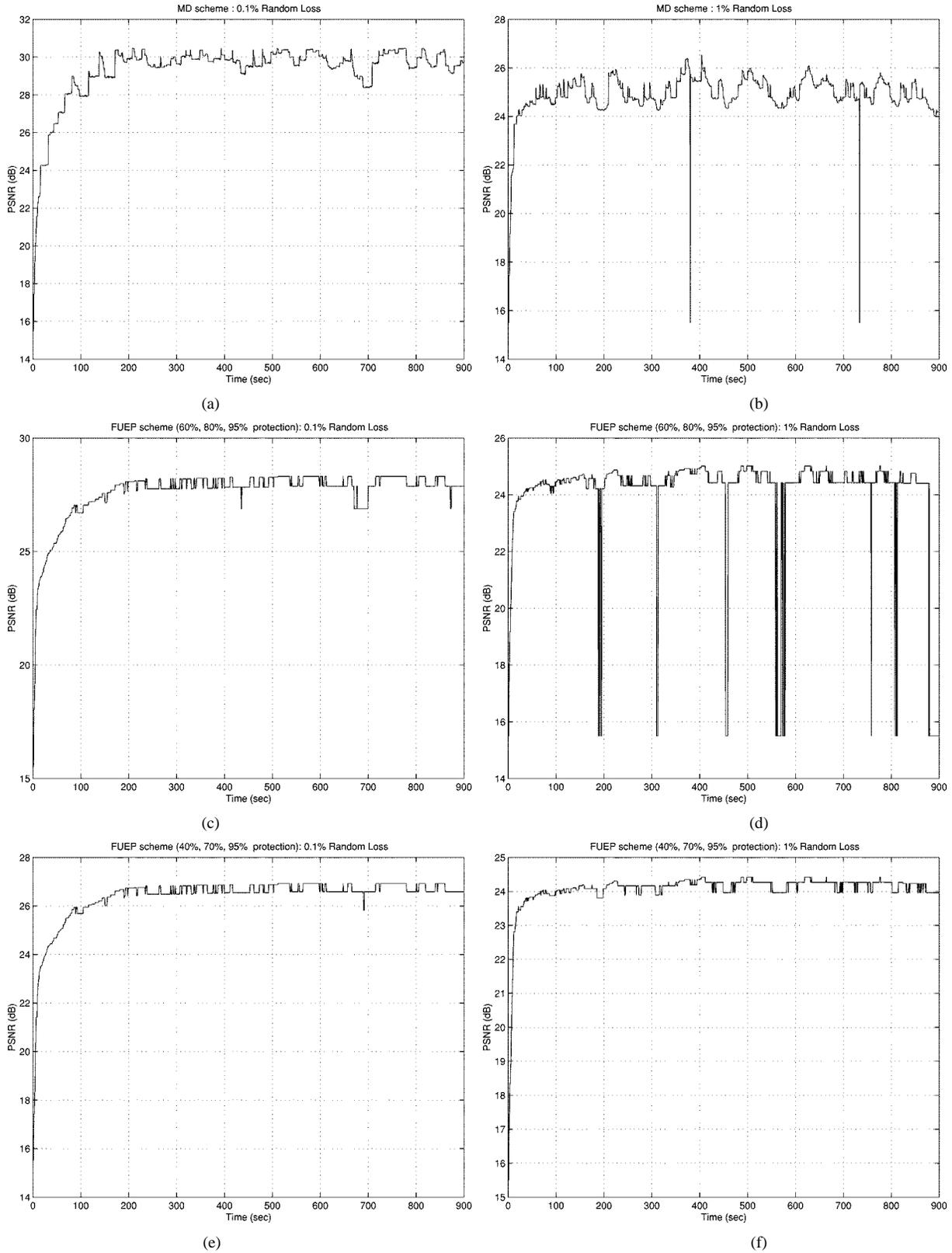
Fig. 12.   Performance comparison between MD-FEC scheme and fixed unequal error protection (FUEP) schemes with rates (60%, 80%, 95%) and (40%, 70%, 95%) respectively for varying degrees of random loss (a) MD-FEC, 0.1% loss, (b) MD-FEC, 1% loss, (c) FUEP (60%, 80% 95%), 0.1% loss (d) FUEP (60%, 80% 95%), 1% loss, (e) FUEP (40%, 70%, 95%), 0.1% loss (f) FUEP (4-%, 70%, 95%), 1% loss.

h. The h variable is initially set to 1, is doubled if there is a packet loss in an epoch, and reset to 1 if there is not. Thus, the h variable is a very simple mechanism to capture the history of

packet loss in the previous epochs. Based on the value of the h variable, LIMD/H exercises a graded multiplicative decrease upon packet loss.

- If $f = 0$, $r \leftarrow r + \alpha$; and $h \leftarrow 1$.
- If $f > 0$, $r \leftarrow r \times (1 - \beta' \times h)$; and $h \leftarrow 2h$.

In case of LIMD/H, we typically set the multiplicative decrease constant β' to be small values, e.g., 0.1, in order to achieve small variations of sending rate when the available bandwidth is invariant. LIMD/H thus throttles its transmission rate gently when there has been no packet loss in the previous epoch, and progressively more aggressively when previous epochs have also experienced packet loss (see Fig. 6).

Of course, the LIMD/H algorithm presented above is a very simple augmentation to LIMD. A detailed treatment of this algorithm is presented in [23]. Also, in [1], it has been shown that for a range of values of α, there exist values of β such that the strategy delivers the same average rate as TCP ($\alpha = 1$, $\beta = 0.5$). Hence, it is long term fair to TCP. Essentially, to reduce the variation in the transmission rate, typically small values of β are used in LIMD/H. So as to ensure same average rate as say TCP, α needs to be made correspondingly smaller. This approach to congestion control contrasts that proposed in [13] where the TCP throughput equation has been used to perform rate based congestion control. Equation-based congestion control, as it is known, measures the round-trip time and observed packet loss probability to compute the nearly constant TCP-friendly rate for a flow. The advantage of such an approach lies in the fact that it can provide very smooth rate control which can be well-matched to multimedia applications such as video streaming over the network—to this extent, its properties are similar to LIMD/H. However, it is extremely slow to respond to network dynamics. Another drawback of equation-based congestion control is that its success depends on accurate measurements of the connection round-trip time and the packet loss probability; the latter is very difficult to calibrate accurately in a responsive manner since it requires large observation intervals.

The LIMD and the equation based paradigm in fact, represent two extreme operating points in the "space" of TCP-fair congestion control mechanisms. The former offers quick responsiveness to network dynamics but results in wide rate fluctuations while the latter provides very smooth rate control but slow reaction to network dynamics. The LIMD/H approach offers the designer an entire space of all intermediate solutions as well that retain most of the vaunted features of both the paradigms. A quantitative comparison of these approaches, is however, outside the scope of this paper.

D. Coordination Between Application and Transport Layer

To summarize, so far, we have advocated a state-of-the-art algorithm at the application layer (MD-FEC) which offers a robust source transcoding option and an efficient algorithm at the transport layer (LIMD/H) which offers smooth transmission rate variations that are amenable to multimedia transmission. A key aspect of the end-to-end system is the synergistic interaction between the application and the transport layer.

This interaction is effected by the efficient transfer of the channel state information from the transport layer to the application layer (see Fig. 1). The congestion feedback from the receiver (that is in the form of the number of packets successfully received in a particular epoch) is available at the transport layer

and can be used to infer the channel state and hence the transmission profile which can then be operated on by MD-FEC algorithm to deduce the optimal encoding strategy.

While a wealth of literature is available on sophisticated traffic models for the Internet, capturing the dynamics of the Internet channel has proved to be a difficult task in general. Driven by our goal of designing a simple but effective interface for transfer of channel state information, we resort to the proven histogram based adaptive estimation principles that blend long-term averaging and short-term updates, that have been deployed with great success in various research areas (ranging from control systems to adaptive filtering to arithmetic coding). Our approach for channel estimation is measurement-based and simply involves collecting statistics available from reverse channel feedback and then conveying them to the application layer (MD-FEC channel profile). This is achieved by maintaining a profile of the relative frequency $f_i(t)$ of occurrence of each rate sample r_i (the number of packets successfully received by the receiver). For the actual implementation, we maintain a histogram of the rate samples r_i. For updating the histogram upon availability of new feedback information, on receiving new feedback information, we add the "new" information to the "old" information in a weighted fashion where the weight for the old information has the connotation of a "forgetting factor" which enables faster adaptation to network dynamics. The histogram is interpreted as the transmission profile by the MD-FEC transcoder. Based on our experiments, we fixed the forgetting factor at 0.875 which worked reasonably well for all situations.

To summarize, our current mechanisms incur very low complexity and despite their simple nature perform well in simulations. Since our architecture is modular in design, it can easily accommodate sophisticated channel modeling machineries also.

IV. SIMULATION RESULTS

In this section, we present simulation-based performance results of the proposed video transmission scheme. The ns-2 [17] network simulator was used to implement the LIMD/H congestion control algorithm as a part of the transport layer protocol. For our simulations, we used a parameterized version of the "Football" video sequence (frame size 352×240 pixels for luminance component) as encoded by the 3-D SPIHT embedded video encoder [24]. Each GOP comprises of 16 frames. The transmission speed chosen was 1 GOP per epoch, where each epoch has a duration equal to 650 ms, which corresponds to approximately 26.7 fps.

The details of the performance of the individual components namely the MD-FEC and the LIMD/H algorithm are presented in [1] and [9], respectively wherein the two respective components were shown to significantly outperform state-of-the-art reference systems. (For example, the MD-FEC algorithm was shown to outperform reference systems [6], [7] that used comparable source coders by rate savings of the order of 30% while delivering identical image qualities). In this paper, we present the performance of the end-to-end system as a whole especially highlighting the effectiveness of the simple interaction between the application and the transport layer. Our
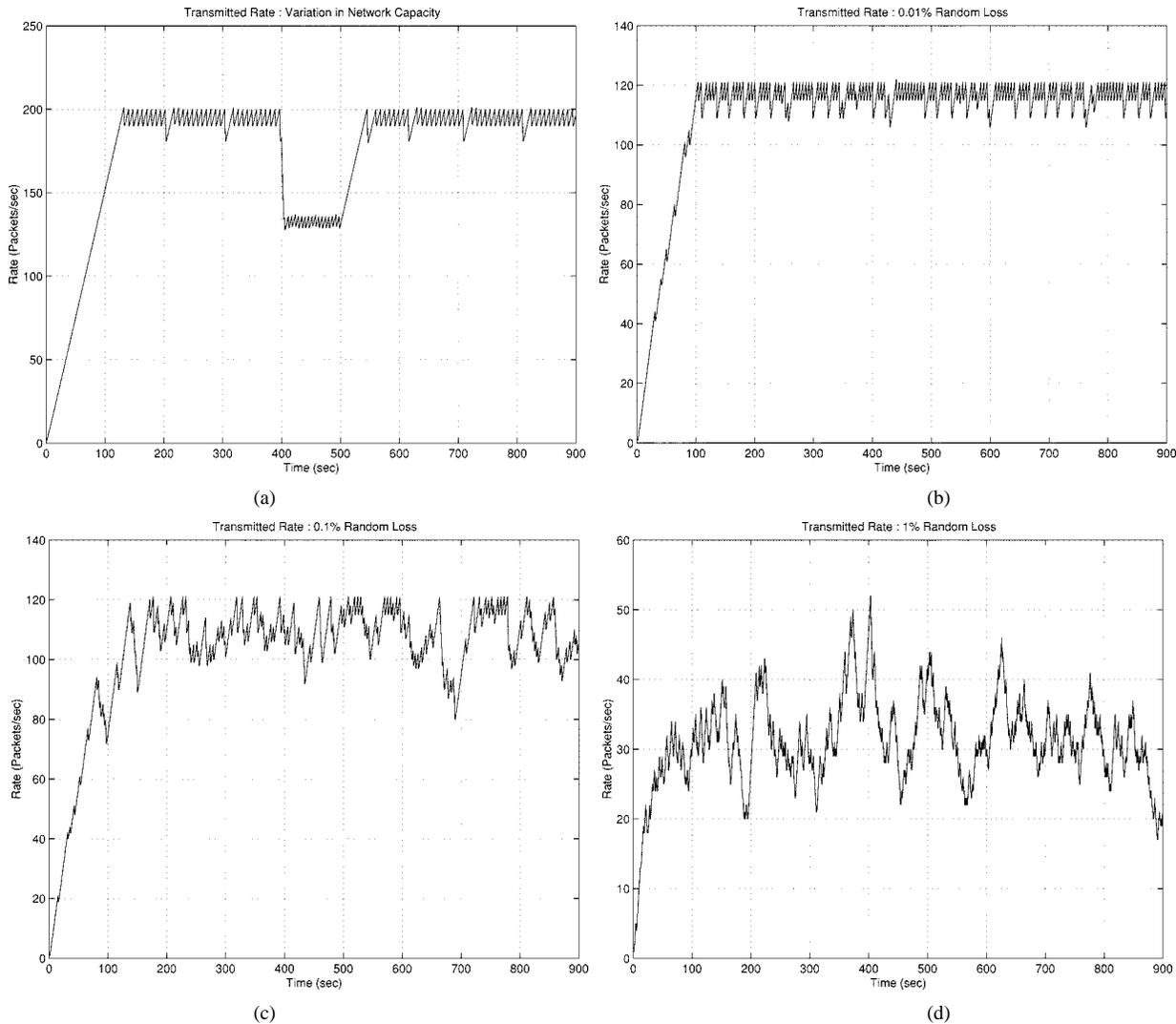
Fig. 13. LIMD/H transmission rate evolution for the one hop network topology: (a) no loss, CBR flow, (b) 0.01% random loss, (c) 0.1% random loss, and (d) 1% random loss.

reference transmission systems represent a gradual evolution from the *dumb* MR scheme with no robustness at all to the proposed MD-FEC scheme. They are

1) multiresolution (MR) source encoder;
2) constant FEC or equal error protection (EEP);
3) fixed unequal error protection (FUEP).

MR source encoder corresponds to no robustness against packet loss. EEP is the case when different layers of the prioritized bit stream are assigned the same level of error protection. FUEP (based on priority encoding transmission (PET) [10]) corresponds to the case when different layers of the prioritized bit stream are assigned unequal error protection that is *fixed*, and therefore does not adapt to the varying network conditions. The performance and characteristics of these systems are summarized in Table I.

The performance of the video transmission systems was compared in terms of the measured delivered quality at the receiver (or peak signal-to-noise ratio (PSNR) in dB) versus time (seconds). We provide performance results of the candidate robustness mechanisms in a variety of network scenarios ranging from single hop to multihop topologies to varying degrees of random loss probabilities to sudden variations in network capacity. The random loss model captures the commonly accepted connection model for a flow that shares a large network with a large number of other flows. The scenario where the network capacity suddenly changes illustrates the behavior of the system for networks with heavy dynamics or upon the occurrence of major network events such as the NASA pathfinder webcast.

A. Comparison of the Source Robustness Mechanisms

In this section, we focus on the performance of various robustness mechanisms with LIMD/H congestion control algorithm in a simple network topology shown in Fig. 7. In all the cases, we tested the performance of each system when there is random loss in the network channel, and when there is a sudden congestion due to the introduction of 800 Kbps CBR on $cbr \rightarrow sink$ during 400–500 s period. The bottleneck link has a bandwidth of nearly 2.5 Mbps.

The performance of the MR encoder is shown in Fig. 8. As expected the MR encoder exhibits extreme fragility and delivers wildly fluctuating quality even in the situation of no random

loss, the amplitude of the fluctuations being more than 5 dB (Fig. 8).

Fig. 9 depicts the performance of the EEP scheme for the case of variation in network capacity. As expected, as the strength of the channel codes used is increased, smoother albeit lower quality is delivered to the destination. In this case, the quality profile exhibits a *cliff* effect as can be seen for the case corresponding to 10% redundancy via channel codes. In Fig. 10 we observe the performance of the EEP scheme for situations of random loss.[6]

The FUEP scheme assigns prespecified protection to prespecified source resolution layers. For instance, in [25], the Priority Encoding Transmission scheme [10] was used for MPEG encoding. Here the I-frames were encoded with priority 60%, P-frames with priority 85% and B-frames with priority 95%.[7] As in [25], in FUEP, we partitioned the source into three layers. Figs. 11 and 12 compare the performance of the FUEP scheme and the MD-FEC scheme. We observe that, in general, the MD-FEC scheme delivers a superior and smoother quality in comparison with the two chosen FUEP schemes. The LIMD/H transmission rate evolution for the four scenarios of no loss/CBR flow, 0.01% random loss, 0.1% random loss and 1.0% random loss is plotted in Fig. 13.

We conclude this section with the following remark. A realistic network scenario would comprise of all the simulation scenarios presented above. The MD-FEC/LIMD-H approach owing to its robustness and fast adaptivity properties and its reasonable performance in all the simulation scenarios described above is well suited for such varying network situations.

B. Performance in a Large Network

In this section, we present results for the performance of the FUEP and MD-FEC robustness mechanisms in a complex multihop network topology (see Fig. 14). There is a single video stream ($vs \rightarrow vr$) and seven TCP connections ($S_i \rightarrow R_i$) in the network. The link capacity and the one-way delay is annotated on the graph. All the links without annotation were set to 40 Mbps and 20 msec.

Fig. 15 shows the performance results of the video stream when encoded by the FUEP and MD-FEC methods. We observe that in general the MD-FEC scheme delivers a smoother and marginally higher end-to-end quality than the FUEP scheme. Moreover, MD-FEC emerges faster from the interval of congestion than the FUEP scheme once again highlighting its adaptivity features. To summarize, MD-FEC scheme responds reasonably well to varying network dynamics and is more adaptive than either of the reference systems.

---

[6]It is worth pointing out the plots of the transmission rate evolution under the LIMD/H approach in situations of random loss (see Fig. 13). There is a delay in feedback, which in turn is due to the network channel delay and the fact that some amount of buffering is deployed at the application level. There can be a difference in the rate at which the source stream is coded (which depends on the transmission profile at that instant) and the rate at which it is eventually transmitted (which depends on the LIMD/H transmission rate value at that instant). This difference could be large depending upon the random loss percentage and the end-to-end delays involved and is infact depicted in Fig. 10 as the need for stronger channel codes to counteract these differences.

[7]Protecting a resolution layer with priority x% means that the receiver can recover it from x% of the encoded packets. In channel coding jargon, x is referred to as the *code rate*.
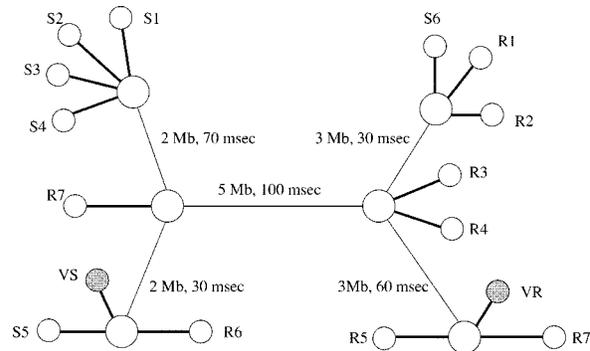

Fig. 14. Multihop network topology with a single video flow. Links without annotation have a bandwidth of 40 Mbps and a delay of 20 ms.

V. RELATED WORK

The related work for this problem falls into the two broad categories of robust source coding and design of multimedia transport protocols. While significant work has been done in the individual areas, not much literature, with the exception of [14], [16] is available that emphasizes the total end-to-end system design.

Prior work on transmission of prioritized data over error/erasure channels has focused largely on still image transmission with the exception of [26], [27] where the authors deal with the issue of video transmission using an H.263 video codec. In [28], RCPC (rate compatible punctured convolutional) codes were used to protect images compressed with the 2D-SPIHT progressive image encoder [29] over binary symmetric noisy channels with an empirical choice of code rates. In [8], the issue of image transmission over packet erasure channels was considered. The proposed methodology is based on the PET [10] scheme. Priority Encoding Transmission is an algorithm that assigns forward error correction (FEC) codes, according to priorities specified by the user, to message fragments (also specified by the user) transmitted over packet networks. The PET algorithm, however does not specify how to assign priorities or how to fragment the message. Besides answering these questions, [8] offers finer granularity in number of delivered quality levels than PET. The actual algorithm, however, is based on greedy, iterative descent techniques and is suboptimal in terms of both complexity and performance.

Signal Processing based methods for introducing robustness into the source include the classical approaches to the MD problem that are based on quantizers [4], correlating transforms [5] and newer approaches based on wavelet polyphase decompositions [6]. Most of these approaches suffer from the issue of implementation complexity and the fact that they are difficult to generalize to more than two descriptions except for in [6], [7]. The MD-FEC scheme represents a systematic approach for the construction of generalized (more than two descriptions) multiple description codes using purely a channel coding based paradigm.

We now focus on the literature pertaining to the design of multimedia transmission protocols. Park *et al.* proposed the framework of an adaptive FEC (AFEC) for delay-sensitive real-time traffic [30]. AFEC adopts a simple control algorithm which decreases the degree of the redundancy when the net-
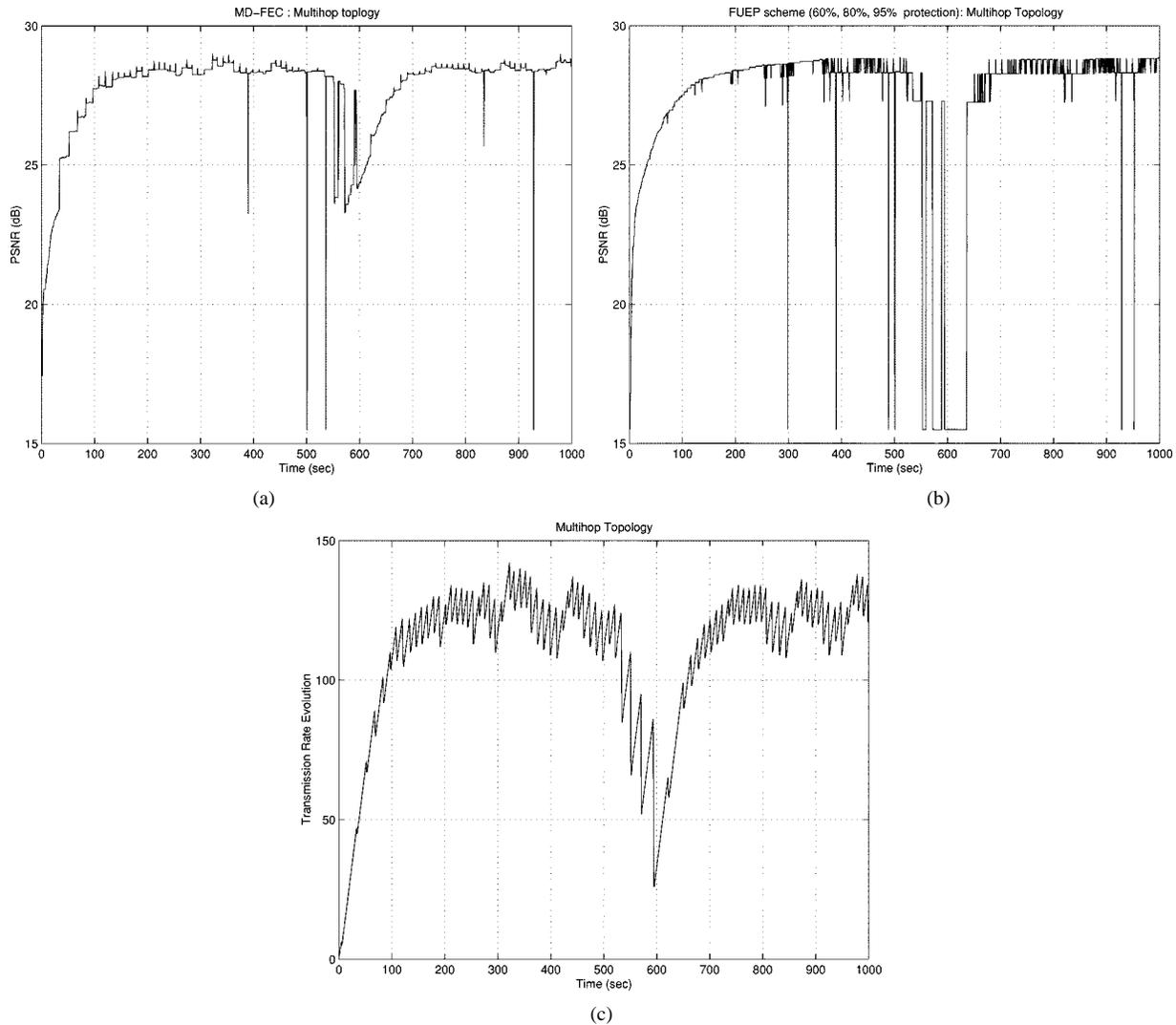
Fig. 15.   Multihop network performance: (a) MD-FEC, (b) FUEP with code rates 60%, 80%, 95%, (c) LIMD/H transmission rate evolution.

work is well-behaved and increases the redundancy otherwise. However, the choice of FEC seems to be empirical.

RAP (rate adaptation protocol) [21] adopts a rate-based LIMD rate adaptation paradigm in order to provide fairness among the multimedia streams as well as TCP-friendliness. While the rate based paradigm serves to space out packets uniformly in a given unit of time instead of bursting them out all at once, there is a large variation in the number of packets transmitted across different units of time, owing to the usage of a TCP kind of rate mechanism. In order to deliver smooth video quality, large variations in the transmission rate need to be smoothed out and that could potentially demand extensive buffering and consequently delays.

In [16] the end-to-end multimedia streaming problem has been addressed with focus on the video source compression and packetization and also the transport protocol that is based on the usage of the TCP-friendly rate equation [13] the limitations of which have been discussed in Section III.

To summarize, to the best of our knowledge there is not much work available that tackles the *end-to-end* problem with synergistic coordination between the source layer and the transport layer.

VI. CONCLUSIONS AND FUTURE WORK

In recent years, multiresolution coding has become a very popular paradigm for image/video source coding. However, the current network model of the Internet is not well suited for the transmission of MR bitstreams because most routers are oblivious to any application-specific priorities. Furthermore, the predominantly-used linear increase multiplicative decrease congestion control paradigm results in large rate variations at the source and a suboptimal use of the available connection capacity.

In order to address these issues, we have presented two components: 1) the MD-FEC transcoding algorithm that converts MR streams optimally to nonprioritized MD streams that are able to better tolerate variations in the frequency and relative position of packet loss; and 2) the LIMD/H congestion control algorithm that reduces the fluctuation in the sending rate when available connection capacity is invariant, and at the same time, is highly responsive to congestion. Finally, we have introduced a simple mechanism to coordinate the congestion control and source transcoding algorithms closely in order to best use the varying network resources. Our preliminary performance eval-

uation has shown that our approach is both viable and better than the traditional approach.

A possible direction for future research could be the evaluation of the role of the MD-FEC coding strategy in multimedia multicast applications, a problem we have started investigating recently [2], [31]. Another possibility is the design of stochastic principles based channel modeling and estimation strategies and their comparison with the current measurement based approach. It would also be interesting to study the impact of the source decoder level error resilience mechanisms on the decoded quality. Our current mechanisms that do not deploy error resilience mechanisms would only stand to gain from this. Another challenge could be the design of a scalable video coder that is efficient from the source coding standpoint.

REFERENCES

[1] K. W. Lee, R. Puri, T. Kim, K. Ramchandran, and V. Bharghavan, "An integrated source coding and congestion control framework for video streaming in the internet," in *Proc. IEEE INFOCOM*, Tel-Aviv, March 2000.

[2] R. Puri, K. W. Lee, K. Ramchandran, and V. Bharghavan, "Application of FEC based multiple description coding for internet video streaming and multicast," in *Proc. Packet Video Workshop*, Cagliari, Italy, May 2000.

[3] H. Radha and Y. Chen, "Fine-granular scalable video for packet networks," in *Proc. Packet Video Workshop*, New York, Apr. 1999.

[4] V. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Information Theory*, vol. 3, pp. 821–834, May 1993.

[5] V. K. Goyal, J. Kovǎcević, R. Arean, and M. Vetterli, "Multiple description transform coding of images," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, Chicago, IL, Oct. 1998, pp. 674–678.

[6] W. Jiang and A. Ortega, "Multiple description coding via polyphase transform and selective quantization," in *Proc. SPIE Visual Communications and Image Processing Conf. (VCIP)*, San Jose, CA, January 1999.

[7] S. Servetto, K. Ramchandran, V. A. Vaishampayan, and K. Nahrstedt, "Multiple description wavelet based image coding," *IEEE Trans. Image Processing*, vol. 9, pp. 813–826, May 2000.

[8] A. E. Mohr, E. A. Riskin, and R. E. Ladner, "Unequal loss protection: Graceful degradation of image quality over packet erasure channels through forward error correction," in *Proc. Data Compression Conf. (DCC)*, Snowbird, UT, March 1999.

[9] R. Puri and K. Ramchandran, "Multiple description source coding through forward error correction codes," in *33rd Asilomar Conf. Signals, Systems and Computers*, CA, Oct. 1999.

[10] J. Albanese, A. amd Blomer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoded transmission," *IEEE Trans. Information Theory*, vol. 42, pp. 1737–1744, Nov. 1996.

[11] W. Stevens, "TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms," RFC 2001, 1997.

[12] J. Postel, "User datagram protocol," RFC 768, 1980.

[13] S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation based Congestion Control for Unicast Applications," in *Proc. ACM SIGCOMM*, Stockholm, Sweden, Aug. 2000.

[14] S. D. Servetto and K. Nahrstedt, "Video streaming over the public internet: Multiple description codes and adaptive transport protocols," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Kobe, Japan, October 1999.

[15] S. D. Servetto, "Compression and reliable transmission of digital image and video signals," Ph.D. dissertation, Dept. Comput. Sci., Univ. Illinois, Urbana-Champaign, May 1999.

[16] W.-T. Tan and A. Zakhor, "Real-time internet video using error resilient scalable compression and TCP-friendly transport protocol," *IEEE Trans. Multimedia*, vol. 1, pp. 172–186, June 1999.

[17] S. Mccanne and S. Floyd, "The LBL/UCB network simulator," Lawrence Berkeley Laboratory, Univ. California, Berkeley, http://www-nrg.ee.lbl.gov/ns.

[18] M. T. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[19] Y. Shoham and A. Gersho., "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.

[20] D. P. Bertsekas, *Non Linear Programming*. Belmont, MA: Athena Scientific, 1995.

[21] R. Rejaie, M. Handley, and D. Estrin, "RAP: An end-to-end rate-based congestion control mechanism for realtime streams in the internet," in *Proc. IEEE INFOCOM*, New York, March 1999.

[22] R. Jain and D.-M. Chiu, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *J. Comput. Networks ISDN Syst.*, June 1989.

[23] T. Kim, S. Lu, and V. Bharghavan, "Improving congestion control through loss differentiation," in *Proc. Int. Conf. Computers and Communication Networks*, Boston, MA, Oct. 1999.

[24] B. A. Kim, Z. Xiong, and W. A. Pearlman, "Very low bit-rate embedded videocoding with 3-D set partitioning in hierarchical trees 3D-SPIHT," *IEEE Trans. Circuits Syst. Video Technol.*.

[25] C. Leicher, "Hierarchical encoding of MPEG sequences using priority encoding transmission," ICSI, Tech. Rep. TR-94-058, 1994.

[26] U. Horn, K. W. Stuhlmüller, M. Link, and B. Girod, "Robust internet video transmission based on scalable coding and unequal error protection," *Image Commun.*, vol. 15, no. 1–2, pp. 77–94, September 1999.

[27] M. Gallant and F. Kossentini, "Robust and Efficient layered H.263 internet video based on rate-distortion optimized joint source/channel coding," in *Proc. Packet Video Workshop*, Cagliari, Italy, May 2000.

[28] P. G. Sherwood and K. Zeger, "Progressive image coding for noisy channels," in *Proc. Data Compression Conference (DCC)*, Salt Lake City, UT, March 1997.

[29] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees.," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.

[30] K. Park and W. Wang, "AFEC: An adaptive forward error correction protocol for end-to-end transport of real-time traffic," in *Proc. Int. Conf. Computer Communications and Networks*, Lafayette, LA, Oct. 1998.

[31] P. A. Chou and K. Ramchandran, "Clustering source/channel rate allocations for receiver-driven multicast with error control under a limited number of streams," in *IEEE Int. Conf. Multimedia and Expo*, New York, Aug. 2000.

Rohit Puri received the B.Tech. degree from the Indian Institute of Technology (IIT), Bombay, in 1997 and the M.S. degree from the University of Illinois, Urbana–Champaign, in 1999, both in electrical engineering. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering and Computer Science, University of California, Berkeley.

His research interests include video coding/transmission, multimedia networking and multiple description coding.

Mr. Puri was awarded the Institute Silver medal at the IIT Bombay for being the best student in electrical engineering in 1997.

Kang-Won Lee received the B.S. and M.S. degrees in computer engineering from the Seoul National University, Seoul, Korea, in 1992 and 1994, respectively, and the Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign in 2000.

From 1996 to 2000, he was a Research Assistant at the TIMELY Research Group, University of Illinois. Since August 2000, he has been a Research Staff Member at IBM T. J. Watson Research, Hawthorne, NY. His research interests include multimedia streaming over the Internet and multicasting heterogeneous packet flows.

Kannan Ramchandran (M'93) received the M.S. and Ph.D. degrees in electrical engineering from Columbia University, New York, in 1984 and 1993, respectively.

From 1984 to 1990, he was Member of Technical Staff at AT&T Bell Labs in the telecommunications R&D area. From 1990 to 1993, he was a Graduate Research Assistant at the Center for Telecommunications Research at Columbia University. From 1993 to 1999, he was a Research Assistant Professor in the ECE Department, University of Illinois at Urbana-Champaign, and a Research Assistant Professor at the Beckman Institute and the Coordinated Science Laboratory. Since Fall 1999, he has been an Associate Professor in the EECS Department, University of California, Berkeley. His research interests include image and video compression and communication, wavelet theory and multirate signal processing, multimedia networking, distributed signal processing, and unified algorithms for signal processing, communications, and networking.

Dr. Ramchandran was the recipient of the 1993 Elaihu I. Jury Award at Columbia University for the best doctoral thesis in the area of systems, signal processing, or communications. He received the NSF Research Initiation Award in 1994, the Army Research Office Young Investigator Award in 1996, the NSF CAREER Award in 1997, the Office of Naval Research Young Investigator Award in 1997, and the Okawa Foundation Award in 2000. In 1998, he was selected as a Henry Magnusky Scholar by the ECE Department at the University of Illinois, chosen to recognize excellence among junior faculty. He is the co-recipient of two Best Paper awards from the IEEE Signal Processing Society: the 1996 Senior Paper Award and the 1998 Signal Processing Magazine Paper Award. He is a member of the technical committees of the IEEE Image and Multidimensional Signal Processing Committee and the IEEE Multimedia Signal Processing Committee, and serves as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING

Vaduvur Bharghavan received the B.Tech. degree from the Indian Institute of Technology, Madras, in 1992 and the Ph.D. degree from the University of California, Berkeley, in 1995.

He is an Assistant Professor in the Electrical and Computer Engineering Department at the University of Illinois, Urbana-Champaign. His research interests include computer networking and mobile computing.