Network Resource Allocation for Competing Multiple Description Transmissions

Ying Li Chao Tian Suhas Diggavi

Mung Chiang A. Robert Calderbank

Abstract

Providing real-time multimedia services over a best-effort network is challenging due to the stringent delay requirements in the presence of complex network dynamics. Multiple description (MD) coding is one approach to transmit the media over diverse (multiple) paths to reduce the detrimental effects caused by path failures or delay. The novelty of this work is to investigate the resource allocation in a network, where there are several competing MD coded streams. This is done by considering a framework that chooses the operating points for asymmetric MD coding to maximize total quality of the users, while these streams are sent over multiple routing paths. The framework is based on the theoretical modeling where we consider two descriptions and high source coding rate region approximated within small constants.

We study the joint optimization of multimedia (source) coding and congestion control in wired networks. These ideas are extended to joint source coding and channel coding in wireless networks. In both situations, we propose distributed algorithms for optimal resource allocation.

In the presence of path loss and competing users, the service quality to any particular MD stream could be uncertain. In such circumstances it might be tempting to expect that we need greater redundancy in the MD streams to protect against such failures. However, one surprising aspect of our study reveals that for large number of users who compete for the same resources, the overall system could benefit through opportunistic (hierarchical) strategies. In general networks, our studies indicate that the user composition varies from conservative to opportunistic operating points, depending on the number of users and their network vantage points.

Index Terms

This work is presented in part at IEEE Globecom 2008 [21].

Y. Li, M. Chiang and A. R. Calderbank are with Department of Electrical Engineering, Princeton University, NJ, USA.

C. Tian is with AT&T Shannon Labs, NJ, USA.

S. Diggavi is with School of Computer and Communication Sciences, EPFL, Switzerland.

Mathematical programming/optimization, Multiple description, Data compression, Multipath routing, Congestion control, Network control by pricing, Network utility maximization, Resource allocation, Queueing.

I. INTRODUCTION

The proliferation of multimedia applications has greatly increased the demand for real-time communication network services. Such services are challenging since existing protocols deployed in the Internet mainly cater for delay insensitive applications. The problem is more challenging in wireless networks because link failure may occur in addition to network congestion. It is well understood that new network architecture and protocols have to be developed to support real-time multimedia communications, such as real-time video.

Without link failures, the performance of multimedia applications (such as video, images etc.) can be succinctly described through a rate versus quality (distortion) trade-off. However, in best-effort traffic, traditional single stream coding may fail if the stream encounters loss or excessive delay. Multiple description (MD) coding is one technique that can be used to reduce the detrimental effects due to packet loss or link failure [10]. In an MD system, a data source, such as a video sequence, is coded into multiple streams (or descriptions) in such a way that each stream is independently decodable; furthermore, any subset of these streams can be used to reconstruct the source with improved quality (lower distortion). MD coding can be understood as adding *redundancy*[1] among descriptions for better protection against the uncertainty in communication [10] [27].

The following question is our main interest in this work: given a resource-limited network, namely limited link capacity, how should we allocate resources to different sources who share the network, such that the total distortion can be minimized? We are motivated to ask this question since with MD coding, it is natural to send the descriptions through different paths. For such a case, the encoders should choose their rate-distortion operating points such that the network resource is utilized efficiently. Then a natural choice to quantify efficiency is the total distortion of the source-destination pairs in the network. Clearly, if a source chooses high coding rates (bits per sample), the distortion may be small if the transmission is successful. However, high rates mean more network traffic, which may result in larger packet loss due to congestion or unreliable wireless links, hence the overall performance may suffer. Optimizing network resource allocation for competing multiple-description transmissions is a challenging and important task.

[1]This redundancy is more general than that seen in packet erasure error correcting codes since each packet needs to be individually useful and needs to refine each other.

There has been some research into network resource allocation where the sources use MD coding, when there is unreliability due to congestion or link losses; see [1]–[3], [16], [18], [22] and references therein. Paper [1] studies the multiple description coding in point-to-point networks with congestion problem, where there is a single source and a single destination with multiple single-link paths between source and destination. The advantage of MD over single description in such setting is shown to be related to the rate-distortion operating point of the MD source via numerical results only, but there does not seem to be attempt to solve the related optimization problem. Papers [3] [22] [18] focus on how to select paths for each description of a source such that the distortion of the video is minimized. These papers consider multiple paths each with several links, but only *a single source-destination pair* is considered. The problems are conjectured as NP-hard, so heuristics is used to find the solution. In [16], the authors propose an algorithm on how to jointly adjust the rate of MD source coding and the redundancy of channel coding to minimize the distortion, but only a single receiver is considered, and the multiple source-destination pairs sharing the limited link bandwidth are not taken into account. In [2], the author proposes an algorithm on how to split the traffic from the source with symmetric multiple descriptions (a very special MD where the rates of the all the descriptions are the same) to all the available paths, and adjust the coding at each source, for multiple source-destination pairs in a *uniform* sensor network (a special network), and the general asymmetric MD and general network topology are not considered.

There are other papers that study MD and congestion control, but on a single path. Paper [20] proposes protocols for the congestion control and adaptive multiple description source coding in a network with multiple source-destination pairs, but the possibility of utilizing multiple paths is not considered.

In summary, network resource allocation for competing MD sources over unreliable network is still not well understood. The design of distributed resource allocation algorithms (with provable guarantee) of MD rate-distortion tuples (with two degrees of freedom, rate and distortion) over multiple paths, is still an open problem. In addition, to the best of our knowledge such resource allocation algorithm in the presence of multiple competing MD flows have not been studied. Therefore our formulation of combining multipath routing with competing MD flows and distributed algorithm for optimizing the total distortion performance is to the best of our knowledge, unique.

Along another thread, some researchers have studied the joint network congestion control and multipath routing, see [12], [14], [15], [26] and the references therein. Joint optimization of the routing and congestion control was considered, such that the network can be stable, robust and the users can have better QoS for the applications. In [14], [15] congestion control and multipath routing were studied and it was demonstrated that there are significant advantages when each source randomly selects multiple paths from all its available choices.

In this work, we propose a framework to combine multipath routing with the asymmetric MD coding (the general MD coding) in a general network with multiple competing source-destination pairs, such that the total distortion of all users can be minimized. Motivated by [14], [15], we adopt random path selection for descriptions of each source-destination pair in a large time scale. Over a small time scale, the rate-distortion operating points are found for the given paths, such that the total distortion is minimized. We consider the packet loss due to end-to-end delay exceeding some deadline and unreliable wireless links in wired and wireless networks, respectively. The framework can be further generalized to include the packet loss due to both factors without difficulty. It is worth noting at this point that our proposed framework reflects the intrinsic tradeoff between the rates and distortions at source coding, as well as the tradeoff between the rates at the source coding and packet loss due to congestion or unreliable links. This is inherently a design problem involving many degrees of freedom normally residing in different layers in a protocol stack.

Within the proposed framework, we study distributed algorithms for the optimal resource allocation, with techniques similar to those for network utility maximization (NUM) [5], [13]. In the basic NUM [13], convexity properties of the optimization problem readily lead to a distributed algorithm that converges to the globally optimal transmission rate allocation. In certain regimes, our problem is convex, and therefore, using the NUM techniques, we can develop distributed algorithms based on network pricing. In contrast to standard pricing-based algorithms for the basic NUM, in which each user communicates its willingness to pay for rate allocation to the network, in our algorithms each user provides willingness to pay for the reduction of end-to-end packet loss to the network.

Our framework provides full flexibility in the design space of rate-distortion. In the literature of the resource allocation in a network with MD coding, the full flexibility in the design space has not been fully investigated. For example, with fixed coding rates, the encoder can still adjust distortions for each subset of descriptions to minimize the overall distortion; similarly, with fixed distortions, the source can choose different coding rate pair. In this work, we investigate the importance of this design flexibility.

Hierarchical coding, which is also called layered or successive-refinement (SR) coding, can be considered as an extreme case of MD coding, where streams are useful only after successful reception of higher priority layers. Such schemes, which are part of standard image/video coding standards, add almost no redundancy to each stream since they need not be individually useful, but only refine previously received streams [9]. Therefore, this can be understood as an opportunistic scheme since it assumes reliable delivery of higher priority layers. In contrast, when the service is unreliable, descriptions need to be individually useful, and hence more redundancy is needed, leading to conservative approaches. Therefore, it is of interest to measure the amount of redundancy in the optimal solution in a network, and to determine whether the solution operates closer to SR rate-tuple or MD coding with more redundancy. The question has important engineering implication, since adaptive SR coding is less involved than adaptive MD coding.

In order to better understand whether the SR rate-tuples are sufficient to achieve the optimal solution of network resource allocation, we examine the operating points for the source coding in two cases: (i) Firstly, the case where the *same* links (representing paths) are shared by *all* the users; (ii) Secondly, the case where the resources of the network represented by a graph are disparately shared by users. We investigate how the number of users, the packet size and the channel coding block length influence the resource allocation.

The main contributions of this work are as follows:

1) We propose a general framework to combine multipath routing with the asymmetric MD coding for a network with competing flows for different source-destination pairs. The framework is based on the theoretical modeling where we consider two descriptions and high source coding rate region approximated within small constants.
2) Based on the framework, we propose distributed algorithms for resource allocation to maximize end-to-end utilities, for both wired and wireless networks.
3) We investigate the redundancy of the resulting operating points of the MD rate-tuples. Our results suggest that when many users are sharing the resources, it might be better to be opportunistic rather than conservative.

The rest of this paper is organized as follows. In Section II, we provide the system model for the network with asymmetric MD sources. In Section III and Section IV, we investigate the optimal resource allocation problem. The distributed algorithms are also given. We provide analysis and numerical examples in Section V and VI to illustrate how our framework and algorithms can be used to achieve optimal resource allocation, and investigate whether the sources operate on the SR rate-tuple. Section VII concludes the paper.

II. System Model

Consider a communication network with L logical links, wired or wireless, each link l with a capacity of C_l (bps), and S source-destination pairs (or S users), each using asymmetric MD source coding. Though general MD coding can produce more than two descriptions, the two description problem is the most well understood [8], [23], and therefore we focus on two descriptions in this work. Each source s has two descriptions, denoted as si for $i = 1, 2$, each using a fixed set $L(si)$ of links in its path. Each link l is shared by a set $\cup_{i=1}^{2} Si(l)$ of descriptions of sources. We abuse the notation of the number of links or users and the set of links or users.

For source s, denote d_{s0} as the central distortion when both descriptions are received, d_{si} for $i = 1, 2$ as the side distortion if description si is received, and d_{s3} as the distortion if none of the description is received. Denote r_{si} (bits/sample) for $i = 1, 2$ as the rates for description si.

Assume the number of the samples transmitted per second is fixed to be b (samples/second). Suppose data from a source is transmitted in packets and each packet consists of K bits. Denote p_{si} for $i = 1, 2$ as the packet loss probability of the path for description si. We consider the packet loss due to the end-to-end packet delay exceeding some threshold for wired networks, and due to the unreliable wireless links for wireless networks.

Assume in a large time scale, a path is randomly selected for each description of each user. In a small time scale, given the paths, users find the optimal rate-distortion operating points to minimize the total distortion,

$$\text{minimize} \quad \sum_s D_s, \tag{1}$$

where D_s is the distortion of each user s,

$$\begin{aligned} D_s = \; & d_{s0}(1 - p_{s1})(1 - p_{s2}) + d_{s1}(1 - p_{s1})p_{s2} \\ & + d_{s2}p_{s1}(1 - p_{s2}) + p_{s1}p_{s2}d_{s3}. \end{aligned} \tag{2}$$

Note that to have (2), we assume that the portion of description si corresponding to a single sample (or a block of samples) is contained entirely in a single packet, in other words it is not distributed over several packets.

It can be seen in the following subsections that the distortions d_{si} are related to the rates r_{si}. A higher rate yields a lower distortion in general. Furthermore, the packet loss p_{si} is related to the rates of all the users who share the links on its path. For a fixed number of samples per second and a fixed link capacity, having more bits per sample of the descriptions sharing the link with description si will result in larger p_{si}, hence larger distortion D_s.

A. Distortion

Assuming Gaussian sources[2] with mean zero and unit variance, the distortion d_{s3} is always 1. The distortion-rate region is given by [8], [23]

$$\begin{cases} d_{s0} \geq 2^{-2(r_{s1}+r_{s2})} \frac{1}{1-(\sqrt{\Pi}-\sqrt{\Gamma})^2} \\ d_{si} \geq 2^{-2r_{si}}, \; i=1,2 \end{cases}, \tag{3}$$

where $\Pi = (1-d_{s1})(1-d_{s2})$ and $\Gamma = d_{s1}d_{s2} - 2^{-2(r_{s1}+r_{s2})}$.

A special case of MD is successive refinement, for which the distortion-rate region is [9]

$$\begin{cases} d_{s0} \geq 2^{-2(r_{s1}+r_{s2})} \\ d_{s1} \geq 2^{-2r_{s1}}, \; d_{s2} = 1 \end{cases}. \tag{4}$$

B. High-Rate Regime

If the link capacity is large, for a fixed number of samples per second at the sources, a large number of bits per sample is expected, which means the rate r_{si} can be high.

In high-rate regime, the distortion-rate (3) can be simplified. A simplified expression for symmetric MD is given in [25], and it is straightforward to extend it to asymmetric MD. By ignoring some small factors, this regime yields the approximation to the distortion-rate tuple as follows,

$$\begin{cases} d_{s0} \geq 2^{-2(r_{s1}+r_{s2}-\min(E_{s1},E_{s2}))} \\ d_{si} \geq 2^{-2E_{si}}, \; E_{si} \in [0, r_{si}], \; i=1,2 \end{cases}, \tag{5}$$

where E_{si} is the exponent characterizing the distortion for only the description i of user s being successfully received. Note that this approximation is accurate up to small constants, which may not be uniform for all the distortion triples, as shown in Appendix A. This implies that the theoretical result in this work is accurate only in this limited (though reasonably accurate) framework. When interpreting the result within the most general framework, it should be understood as accurate up to certain constants, which also depend on the accuracy of the delay tail distribution approximation in later section.

[2] It should be clarified that the memoryless Gaussian source we use in this work provides a very concrete model to capture the underlying tradeoff in MD coding. Therefore, we believe treating the Gaussian source, though maybe limited, does capture the most fundamental aspects of the problem, and is important in understanding theoretically the performance comparison of MDC and successive refinement in various scenarios.

In high-rate regime, for either wired or wireless networks, the optimal packet loss is small, otherwise the distortion would be large and the entire minimization would force the rate to be reduced until the total distortion is minimized. In this regime, the distortion of user s in (2) can be simplified to $d_{s0} + d_{s1}p_{s2} + d_{s2}p_{s1} + p_{s1}p_{s2}$. Hence the objective function (1) can be simplified to

$$\sum_s d_{s0} + d_{s1}p_{s2} + d_{s2}p_{s1} + p_{s1}p_{s2}. \tag{6}$$

C. End-to-End Packet Loss

1) Wired networks: For wired networks, the communication channels can be assumed to be noise-free. We consider the packet loss due to the end-to-end delay of the packet exceeds some deadline Δ (in second).

At each link, assume the packets from sources are stored in a queue and transmitted in a first-in-first-out (FIFO) fashion. Assume only the destination drops the packet whose end-to-end delay exceeds deadline. The end-to-end delay is dominated by the delay at the bottleneck links [11]. We focus on the core network, where it is reasonable to assume Poisson arrivals of packets at each link. Assume that at each link l, in addition to the MD traffic we are interested in, there is ω_l bps background traffic. Assume all the packets are of the same length, then the queue is $M/D/1$ queue; assume the packet size is various, then the queue can be assumed as $M/M/1$ queue [17].

The end-to-end delay tail probability for description si is

$$p_{si} = F(y_l, l \in L(si)) \approx \max_{l \in L(si)} \exp[f(y_l)], \tag{7}$$

where y_l is the MD traffic on link l, F and f are functions. The end-to-end delay tail probability is a function in the traffic load of all the links on the path. Note that in equation (7) we use the bottleneck delay tail probability $\max_{l \in L(si)} \exp[f(y_l)]$ to approximate the delay tail of the entire path because the end-to-end delay is dominated by the delay at the bottleneck links [11]. For $M/M/1$ queue, function f takes the form of $f = -\mu\Delta(1-\rho)$ and $\exp(f)$ is the delay tail probability of the queue [17], and for $M/D/1$ queue $f \approx -2\mu\Delta(1-\rho)(1+(1-\rho)/3)$ [24], where μ is the service rate C_l/K (in packets/second, where K is the average packet length) and ρ is the load $(y_l + \omega_l)/C_l$ at link l. We have done extensive numerics which are not presented in here that suggest such delay tail approximation is reasonable.

2) Wireless networks: For wireless networks, we consider the packet loss due to unreliable wireless links.

Assume that the adaptive channel coding is used, similar to [19]. After link l receives the data of sources from the upstream link, it first decodes it to extract the information data of the source and encodes it again with its own coding rate θ_l, where the coding rate is defined by the ratio of the information data rate $\sum_i \sum_{s\in Si(l)} r_{s,i}b$ at the input of the encoder to the transmission data rate C_l at the output of the encoder,

$$\theta_l = \sum_i \sum_{s\in Si(l)} r_{si}b/C_l, \tag{8}$$

where b is the number of samples per second.

The receiver decoding bit error probability of description si at link l can be defined as an increasing function of the coding rate θ_l. An exact characterization of this function is difficult. However, an upper bound using an error exponent function can be found.[3] The receiver decoding error probability of description si at link l is [19] $\hat{p}_{sil} \geq \frac{1}{2}2^{-N(R_0-\theta_l)}$ where N is the block length and R_0 is the information-theoretic capacity of the link.

For a packet of length K bits, assuming the bits are independent, the packet error probability of description si at link l is $p_{sil} = 1-(1-\hat{p}_{sil})^K$, where for small p_{sil}, it can be approximated as $p_{sil} \approx K\hat{p}_{sil}$. The end-to-end packet error probability for description si is $p_{si} = 1 - \prod_{l\in L(si)}(1-p_{sil})$. Assuming that the error probability of each link is small, we can approximate p_{si} as

$$p_{si} \approx \sum_{l\in L(si)} p_{sil} \approx \sum_{l\in L(si)} \frac{K}{2}2^{-N(R_0-\theta_l)} \tag{9}$$

where θ_l is defined in (8).

III. Joint Rate-Distortion Adaptation and Congestion Control for Wired Networks

This section develops the optimal resource allocation in wired networks. To minimize the total distortion, each source s may increase its rate r_{si} and decrease distortion d_{si}, but the increased rate may increase the load of the links on its path, hence the link may be congested, the packet loss may be increased and the total distortion may be increased. There is an intrinsic tradeoff between the rate-distortion adaptation at the sources and the congestion control at the links. We investigate the optimal resource allocation by jointly optimizing the rate-distortion adaptation and congestion control.

[3]This bound is applicable for the fast fading case when the transmission is over several fade intervals. In the high SNR regime, this bound can be used for the case where we have a random quasi-static fading channel as well. Note that it is very challenging to model packet losses over wireless links. In our formulation we use decoding error exponents to bound the packet loss. We think this is a way to capture the main characteristics to model the packet loss due to decoding errors.

A. Optimization Problem

The optimization problem is

$$\begin{array}{ll} \text{minimize} & \sum_s D_s \\ \text{subject to} & p_{si} = \max_{l \in L(si)} \exp(f(y_l)), \forall s, i = 1, 2 \\ & y_l = \sum_{i=1}^{2} \sum_{si \in Si(l)} r_{si} \cdot b < C_l - \omega_l, \forall l \\ \text{variables} & d_{s0}, d_{si}, r_{si}, p_{si}, y_l, \forall s, i = 1, 2, \forall l, \end{array} \tag{10}$$

together with the constraints of distortion-rate region (3), where D_s is in (2). The first constraint is for the packet loss, and the second is the flow constraint.

This problem is a non-convex optimization problem. All the constraints on the distortions are in convex form (since the distortion-rate region is convex), but the constraint for packet loss may not be convex because the function f may not be convex in y, the objective function is in non-convex form because there are products of variables with negative coefficients. In the following, we discuss how to tackle this non-convex optimization problem.

B. Distortion

We consider the distortion in high-rate regime given by (5). The distortion formula (5) is not convex in E_{si}. But if for every source s it is known whether $\min(E_{s1}, E_{s2})$ is E_{s1} or E_{s2}, the distortion formula is convex in E_{si}, and the optimization problem is a convex optimization if we apply log change of variable to variable p and if the function f for packet loss is convex in y. We propose to randomly choose the ordering of E_{s1} and E_{s2} for each user s and therefore specifically choose which path carries the individual description with lower distortion. From our numerical studies this choice does not seem to significantly affect the resulting total distortion.

C. Packet Loss

For $M/D/1$, function f in the packet loss expression (7) is not convex in y. To make it convex in y, we approximate it by $f(y_l) \approx -\frac{2\Delta}{K}(C_l - \omega_l - y_l)$ because bottleneck has a large load in general.

With the additional log change of variable to p,

$$p_{si} = \exp(m_{si}), \; m_{si} \le 0,$$

optimization problem (10) then becomes

$$\begin{aligned}
\text{minimize} \quad & \textstyle\sum_s 2^{-2E_{s1}} e^{m_{s2}} + 2^{-2E_{s2}} e^{m_{s1}} \\
& + e^{m_{s1}} e^{m_{s2}} + 2^{-2(r_{s1}+r_{s2}-\min(E_{s1},E_{s2}))} \\
\text{subject to} \quad & E_{si} \in [0, r_{si}], \forall s, i = 1, 2 \\
& m_{si} \geq f(y_l), \forall s, i = 1, 2, l \in L(si) \\
& \textstyle\sum_{i=1}^{2} \sum_{si \in Si(l)} r_{si} \cdot b \leq y_l - \epsilon_l y_l^2, \forall l \\
\text{variables} \quad & E_{si}, r_{si}, m_{si}, y_l, \forall s, i = 1, 2, \forall l,
\end{aligned} \tag{11}$$

where $y_l \leq C_l - \omega_l$. The inequality of the last constraint is because the objective function is decreasing in r. To make the problem (10) strictly convex in y, we add $-\epsilon_l y_l^2$ to the right hand side of the last constraint, where ϵ_l is a small number such that $\epsilon_l y_l^2$ is small compared with y_l.

Given the ordering of E_{s1} and E_{s2} for every user s, the optimization problem (11) is a convex optimization. By the standard dual decomposition approach [19], we propose the following distributed algorithm where each source and each link solve their own problem with only local information.

D. Distributed Algorithm

Distributed Algorithm 1:

In each iteration t, by solving the following problem (12) over (E_{si}, r_{si}, m_{si}), for each description i, each source s determines its distortion exponent $E_{si}(t)$, rate $r_{si}(t)$ and packet loss exponent $m_{si}(t)$ that maximize its net utility based on the prices $(\nu^{si}(t), \varphi_{sil}(t))$ in the current iteration. Furthermore, by price update equation (13), the source adjusts its offered prices per unit traffic load reduction for each description and each link on its path for the next iteration.

Source problem and delay price update at source s:

Source problem over (E_{si}, r_{si}, m_{si}):

$$\begin{aligned}
\text{minimize} \quad & 2^{-2(r_{s1}+r_{s2}-\min(E_{s1},E_{s2}))} + 2^{-2E_{s1}} e^{m_{s2}} \\
& + 2^{-2E_{s2}} e^{m_{s1}} + e^{m_{s1}} e^{m_{s2}} \\
& - \textstyle\sum_i \sum_{l \in L(si)} \varphi_{sil} m_{si} + b \sum_i r_{si} \nu^{si}(t) \\
\text{subject to} \quad & E_{si} \in [0, r_{si}], i = 1, 2
\end{aligned} \tag{12}$$

where $\nu^{si}(t) = \sum_{l \in L(si)} \nu_l(t)$ is the end-to-end congestion price at iteration t.

Price update:

$$\varphi_{sil}(t+1) = [\varphi_{sil}(t) + \alpha(t) f(y_l)]^+, \quad l \in L(si), \tag{13}$$

where $\alpha(t)$ is the step size, and $[a]^+ = \max(a, 0)$.

Concurrently at each iteration t, by solving problem (14) over y_l, each link l determines its total traffic $y_l(t)$ that maximize the 'net revenue' of the network based on the prices. In addition, by price update equation (15), the link adjusts its congestion price per unit rate for the next iteration.

Link problem and congestion price update at link l:

Link problem over y_l:

$$\begin{aligned} \text{minimize} \quad & \varphi^l(t) f(y_l) - \nu_l(y_l - \epsilon_l y_l^2) \\ \text{subject to} \quad & y_l < C_l - \omega_l, \end{aligned} \tag{14}$$

where $\varphi^l(t) = \sum_{i=1}^{2} \sum_{si \in Si(l)} \varphi_{sil}(t)$ is the aggregate traffic load reduction price paid by sources using link l.

Price update:

$$\nu_l(t+1) = [\nu_l(t) + \alpha(t)(r^l(t)b - y_l(t) + \epsilon_l y_l^2)]^+, \tag{15}$$

where $r^l(t) = \sum_{i=1}^{2} \sum_{si \in Si(l)} r_{si}$ is the aggregate number of bits per sample of all the sources on link l at iteration t, $\alpha(t)$ is the step size, and $[a]^+ = \max(a, 0)$.

To get the end-to-end aggregate price $\nu^{si}(t)$, a message passing procedure similar to the one in [19] is adopted. The price is added up along the path, hence the message passing will only need each link to add its own price to the one it received from previous hop and pass on to the next hop. The signaling overhead is rather limited and it does not scale up by the number of the links on the path. The step size $\alpha(t)$ satisfies $\lim_{t\to\infty} \alpha(t) = 0$ and $\lim_{t\to\infty} \sum_{i=1}^{t} \alpha(i) = \infty$.

After the above dual decomposition, the following result can be proved using standard techniques from the convergence analysis of the distributed gradient algorithm:

Theorem 1: Given the ordering of E_{s1} and E_{s2} for each user s, the optimization problem (11) is a convex optimization. By Algorithm 1, dual variables $\boldsymbol{\nu}(t)$ and $\boldsymbol{\varphi}(t)$ converge to the optimal dual solutions $\boldsymbol{\nu}^*$ and $\boldsymbol{\varphi}^*$ and the corresponding primal variables $\mathbf{E}^*$, $\mathbf{r}^*$, $\mathbf{m}^*$ and $\mathbf{y}^*$ are the globally optimal primal solutions of this convex optimization problem.

Proof: Since strong duality holds for the primal problem and its Lagrange dual, we solve the dual problem through distributed gradient method and recover the primal optimizers from the dual optimizers. The algorithm is a gradient projection algorithm for dual problem. The proof uses the results of [4] and is similar to that given in [19] and we omit the details. ■

IV. Joint Rate-Distortion Adaptation and Channel Coding Adaptation for Wireless Networks

This section discusses the optimal resource allocation in wireless networks. Different from wired networks, wireless networks can have link failure because of the unreliable radio communication environment such as path loss and channel fading, which adds on the challenges for the resource allocation in wireless networks. This brings another design knob, the channel coding which adds protection (redundancy) for the source information, into our design space. The packet loss due to exceeding the deadline as discussed in wired network can be incorporated, but for simplicity, here we mainly consider the packet loss due to unreliable wireless links. We assume the redundancy added at link to protect the source information is adaptive.

To minimize the total distortion, each source s may increase its rate r_{si} and decrease the distortion d_{si}, but the increased rate may increase the incoming information of the links on its path, hence the redundancy that the channel coding can add may be reduced, the end-to-end packet loss probability may be increased and the total distortion may be increased. There is an intrinsic tradeoff between the rate-distortion adaptation at the sources and the channel coding adaptation at the links. We try to utilize the diversity in networks [6], [7]. The optimal resource allocation by jointly optimize the rate-distortion adaptation and channel coding adaptation is investigated.

A. Optimization Problem

Similar to the wired case, we consider the distortion-rate in high-rate regime. The optimization problem is

$$\begin{array}{ll} \text{minimize} & \sum_s 2^{-2(r_{s1}+r_{s2}-\min(E_{s1},E_{s2}))} + 2^{-2E_{s1}} e^{m_{s2}} \\ & +2^{-2E_{s2}} e^{m_{s1}} + e^{m_{s1}} e^{m_{s2}} \\ \text{subject to} & E_{si} \in [0, r_{si}], \forall s, i = 1,2 \\ & m_{si} \geq \log[\sum_{l \in L(si)} \frac{K}{2} 2^{-N(R_0 - \theta_l)}], \forall s, i = 1,2 \\ & \sum_i \sum_{s \in Si(l)} r_{si} b \leq \theta_l C_l, \forall l \\ \text{variables} & E_{si}, r_{si}, m_{si}, \theta_l, \forall s, i = 1,2, \forall l. \end{array} \tag{16}$$

This problem is to minimize the total distortion. The first constraint is for the rate-distortion region, the second denotes a performance bound on the packet loss, and the third is the flow constraint.

Given the ordering of E_{s1} and E_{s2} for every user s, problem (16) is a convex optimization. The second constraint is convex because log-sum-exp function is convex. Hence problem (16) can be solved in a centralized way for its unique optima.

B. Distributed Algorithm

Although problem (16) is in convex form given the ordering of E_{si} for every user s, the distributed algorithm cannot be readily derived because the second constraint is not separable, though it can be written as $e^{m_{si}} \geq \sum_{l \in L(si)} \frac{K}{2} 2^{-N(R_0-\theta_l)}$ which is separable, but not in convex form.

To derive distributed algorithm, we consider relaxation of the second constraint. The packet error probability $\sum_{l \in L(si)} \frac{K}{2} 2^{-N(R_0-\theta_l)}$ can be upper bounded by $|L(si)| \frac{K}{2} 2^{-N(R_0 - \max_{l \in L(si)} \theta_l)}$, which may be conservative, but it still captures the major characteristic of the error probability. Therefore, the second constraint in problem (16) is replaced by

$$m_{si} \geq \beta_{si} - N(\log 2)(R_0 - \theta_l), \forall s, i = 1, 2, \forall l \tag{17}$$

where $\beta_{si} = \log(|L(si)| \frac{K}{2})$ are constants.

Compared with problem (11) in wired network, problem (16) with the second constraint replaced by (17) has the same structure as (11). The distributed algorithm is readily derived and proved to converge to the optima, which is similar to Algorithm 1 and the details are omitted.

V. A NETWORK WITH TWO PARALLEL LINKS

In this section we consider a simple topology of two parallel links each with capacity C shared by S users. This is done to build intuition for the structure of the resource allocation solution.

A. Model and analysis

In this network, since there are only two paths available to each source-destination pair, we assume for each source, each description takes one path respectively. We notate source s's description going to link 1 as description $s1$, and the description going to link 2 as description $s2$. Therefore our interest is in examining how this completely shared resource gets allocated among the users. We do this for wired networks in Section V-A1 and for wireless networks in V-A2.

1) Wired networks: We consider two cases, (i) the capacity scales with the number of users S, say, the capacity being SC; (ii) the capacity is fixed to be C.

We define a measure of how much *relative redundancy* is added in the source coding as,

$$\frac{\min(E_{s1}, E_{s2})}{r_{s1} + r_{s2}}, \ \forall s. \tag{18}$$

The following lemma (which is proved in the Appendix) is useful in understanding the behavior of the resource allocation with packet size and the number of users.

Lemma 1: Consider the following optimization problem,

$$\begin{aligned}
\text{minimize} \quad & e^{-(2\log 2)(r_1+r_2-B)} + e^{-(2\log 2)A}e^{-h+(2\log 2)\alpha r_2} \\
& +e^{-(2\log 2)B}e^{-h+(2\log 2)\alpha r_1} \\
& +e^{-2h+(2\log 2)\alpha(r_1+r_2)} \\
\text{subject to} \quad & A \in [0, r_1], B \in [0, r_2], B \leq A \\
& 0 \leq r_1 < R, 0 \leq r_2 < R \\
\text{variables} \quad & r_1, r_2, A, B,
\end{aligned} \tag{19}$$

where R is a positive constant, h and α are functions of R where $e^{-h+(2\log 2)\alpha r_i} \to 1$ as $r_i \to R$ for $i = 1, 2$. Denote the optima as (r_1^*, r_2^*, A^*, B^*).

(i) If the following condition (20) holds, then B^*=0,

$$\alpha \geq \tfrac{1+\sqrt{5}}{2}. \tag{20}$$

(ii) If $1 < \alpha < \frac{1+\sqrt{5}}{2}$, and if $B^* \in (0, \min(r_1^*, r_2^*))$, then

$$\begin{aligned}
B^* = \; & \tfrac{1+\alpha}{1+2\alpha}\log(1+\alpha-\alpha^2) - \log(\alpha) \\
& -\tfrac{1}{1+2\alpha}\log(1+\alpha) + \tfrac{1}{1+2\alpha}h,
\end{aligned} \tag{21}$$

$$\begin{aligned}
r_1^* + r_2^* = \; & \tfrac{2+3\alpha}{(1+\alpha)(1+2\alpha)}\log(1+\alpha-\alpha^2) - \tfrac{2}{1+\alpha}\log(\alpha) \\
& -\tfrac{2}{1+2\alpha}\log(1+\alpha) + \tfrac{3+4\alpha}{(1+\alpha)(1+2\alpha)}h,
\end{aligned} \tag{22}$$

where the parameters h and α are assumed to make B^* and $r_1^* + r_2^*$ feasible.

Observation: If h is not a function in α, if $\alpha > 1$ and $h \geq 2$, by (21) and (22), $B^*/(r_1^* + r_2^*)$ is decreasing in α. If $h = \alpha m$ where m is a positive constant, if $\alpha \geq 1$ and $m \geq 3$, by (21) and (22), $B^*/(r_1^* + r_2^*)$ is decreasing in α. This can be easily checked by plotting the curves.

For the case of S users sharing a two-parallel-link network and all the users having the same distortion ordering $E_{s1} \geq E_{s2}, \forall s$, the optimization problem (11) is the same as (19) in Lemma 1, with specific α and h. According to the packet loss model introduced in Section II-C-1, For the case of the link capacity scaling with S, say, being SC, α and h are both related to S,

$$\alpha = \tfrac{ab\Delta S}{K(2\log 2)}, \;\; h = \alpha \tfrac{C}{b}, \tag{23}$$

where a is some constant ($a = 2$ in this work), b is samples/sec, Δ is maximum allowable delay (sec) and K is bits/packet. While for the case of no capacity scaling, α keeps the same, but h is not related to S,

$$\alpha = \frac{ab\Delta S}{K(2\log 2)}, \ h = \frac{a\Delta C}{K}. \tag{24}$$

Applying Lemma 1 (i) leads to Proposition 1 which gives a sufficient condition on the parameters such that MD has the same performance as SR, under the high-rate regime assumption (5). Consider S users using MD coding, in the setting of point-to-point two parallel links each with capacity SC bits/second or capacity being C available for MD flows where C is a constant, where each description takes one link respectively. Assume the delay tail probability is $\exp(-a(\mu-\lambda)\Delta)$ where μ and λ are the packet service rate and arrival rate (in packets/second) respectively, a is some constant, Δ is the maximum allowable delay (in second). Assume all the users have the same distortion ordering $E_{s1} \geq E_{s2}, \forall s$.

Proposition 1: If

$$\frac{a\Delta bS}{K} \geq (1+\sqrt{5})\log 2, \tag{25}$$

where b is the number of samples transmitted per second and K is the number of bits per packet, then in high-rate regime, MD has the same distortion performance as SR.

Note that all the statement in Proposition 1 is under the high-rate assumption (5).

2) Wireless networks: When S users share the links, assuming all the users have the same distortion ordering $E_{s1} \geq E_{s2}, \forall s$, scaling the capacity with S does not change the formulation of problem (16), while if the capacity does not scale up with S, the packet loss is affected by S.

Similar to wired network, we have Proposition 2, under the high-rate assumption (5). Consider S users using MD coding, in the setting of point-to-point two parallel links each with capacity C bits/second available for MD flows, where each description takes one link respectively. Assume the packet loss probability is $2^{-N(1-\frac{r_i bS}{C})}$ where N is the block length of channel coding, b is the number of samples transmitted per second and r_i is the number of bits per sample for description i. Assume all the users have the same distortion ordering $E_{s1} \geq E_{s2}, \forall s$.

Proposition 2: If

$$\frac{NbS}{C} \geq 1+\sqrt{5}, \tag{26}$$

then in high-rate regime, MD has the same distortion performance as SR.

B. Numerical examples

Assume the background traffic is 50% of the total capacity and in the figures in this section, capacity refers to the capacity available for MD flows (total capacity minus background traffic).

1) Wired networks: Assume the number of samples transmitted per second is $b = 100,000$, the maximal allowable end-to-end queueing delay is $\Delta = 0.2$ s.

a) Single user: For single user case, the performance comparison of MD, SR and single description (SD) is shown in Fig. 1. For SD, the traffic is split equally to the two links, while for MD and SR, each description takes one link. Our main focus is to compare MD and SR. SD is here only for a baseline reference. Fig. 1 shows that the distortion decreases as capacity increases. In general, MD is better than SR and SR is better than SD. As packet size decreases, the distortion decreases, and SR performs more similarly to MD, and when packet size goes down to some point, SR has the same performance as MD, i.e., SR becomes the optimal solution among all MD instances. We denote such scenario as "SR=MD". Intuitively, when packet becomes small, the congestion is less likely to happen and the queueing delay is reduced, hence the packet loss becomes small and the extreme case is that for zero packet loss we have SR=MD.

More precisely, by the distortion formula and packet loss formula, the distortion decays at a speed of $r(2\log 2)$ exponentially, while delay tail increases at a speed of $r(\frac{a\Delta b}{K})$ exponentially where r is the rate, a is a positive constant (in this work, $a = 2$). By the tradeoff in rate r, these parameters can affect the operating point of MD to be close SR or away from SR.

The condition (25) yields that if $K < 17,833$, then SR=MD. Figure 1 shows that condition (25) can give a very good estimation on the packet size for SR=MD.

Note that the single description (SD) is in fact a special case of SR coding, with for one instance, the refinement description being given zero rate, or for another instance, the distortion of the base layer description being 1 (i.e., $d_{s1} = 1$). These two instances correspond to two scenarios, sending single description through one single path, or breaking the single description into two sub-streams and sending them over two paths, respectively. As such, the framework we consider in fact includes the single description case naturally. For this reason, in the following simulations, we do not show SD comparison any more, rather, we focus on the performance comparison of SR and general MD.

b) Multiple users: From (23) and (24), the delay tail increases at a speed of $r(\frac{a\Delta bS}{K})$ exponentially. Since we have seen smaller packet size K can push the operating point moving towards SR, we expect that larger S can have the same effect. To see the performance with respect to S, Fig. 2 and Fig. 3 show the distortion comparison between SR and MD for different S, for the case of capacity being SC and capacity being C, respectively. Indeed, the performance gap decreases as S increases.

For capacity being SC, in Fig. 4, assuming all the users use general MD, the rate-distortion operating points for users are investigated, by measuring the relative redundancy (18), which captures whether the sources are likely to be opportunistic or conservative. For SR, the relative redundancy is zero because $\min(E_{s1}, E_{s2})$ is zero. From Fig. 4, it is clear that as the number of users sharing the same resource increases, the optimal operating points are moving towards SR, which means opportunistic approach is becoming optimal.

Applying **Observation** based on Lemma 1 (ii), it can be easily understood that the relative redundancy decreases as the number of users S increases for both cases of scaling the capacity or no scaling.

Given packet size $K = 80,000$ bits, the condition (25) gives us SR=MD if $S > 4.47$, which can be checked by Fig. 2, 3 that SR=MD for $S = 5$. Again, the condition (25) gives a very good guideline.

2) Wireless networks: For a wireless network with two links, as the block length N of the channel coding increases, the packet loss decreases, the distortion decreases, and the performance gap between SR and MD decreases. Due to the space limitation, we omit the figure to show such. Suppose the capacity does not scale up with user number S, the packet loss increases at a speed of $r(\frac{NbS}{C})$ exponentially where r is the rate. Since a larger N can push the operating point moving towards SR, we expect that larger S can have the same effect. Indeed, the results in Fig. 5 show that as S increases the users are more likely to operate at SR tuples.

VI. Results for general networks

In order to understand the behavior for general networks we have run experiments on several networks. The numerical results presented in this section are representative of the observations made on them. In general networks, the multiple paths are not completely shared among the users as was done in Section V. Therefore, the operating points become more complicated with a distribution of user population having low and high redundancy. However, for when the network grows and each user shares resources with *many* other users, our numerics suggest that the population shifts towards lower redundancy, suggesting a similar behavior as Section V.

For instance, consider a network with 14 links as shown in Fig. 6. Suppose each link has same capacity. Suppose there are 10 users (S=10), using paths (in terms of link order) (1 4; 2 5), (3 4; 5), (1; 2 3), (5

7; 3 6), (9 10, 6 11), (10, 8 11), (11 12; 13), (7 13, 14), (6 13; 4 14), (6; 9 8), respectively. For $S = 20$ and $S = 30$, each user in 10 users case is replaced by 2 and 3 users, respectively. In the experiments we fix the number of samples transmitted per second $b = 100,000$.

Figure 7 shows the distortion comparison, assuming all the users use SR or general MD, for different S. As S increases, the gap between SR and MD reduces. Assuming all the users use general MD, Fig. 8 shows the fraction of the users operating at SR, where by saying that a user operates at SR we mean that its relative redundancy (18) is less than 0.001. As the packet size decreases, or as S increases, more and more users are operating at SR. We have investigated other networks and similar phenomenon can be seen for general networks.

Figure 9 and Fig. 10 are for the 14-link wireless network. Again, it can be seen that when many users share the resource, the user population tends towards lower redundancy (opportunistic) operating points.

VII. Conclusions

In this paper we have studied how competing MD coded streams interact in a resource constrained network. The framework is based on the availability of multiple routing paths and NUM formulation with end-to-end distortion measuring user utility function.

Though the exact delay dynamics of networks is hard to model/analyze, numerics suggest that our approximation captures the behavior of the asymptotic delay tail. This modeling then allows us to develop distributed algorithms, which can be implemented in networks. This in itself is an important step in deploying real-time MD service with competing flows in a best-effort network. Note that no matter whether the bottleneck is at the access network or the core network, we believe that the practitioners can interpret our results in their specific problems. As long as the bandwidth is limited, our work provides a guideline to understand the tradeoff among various resources.

In addition to these distributed algorithms, we also gained architectural insight into the value of MD coding when many users share a network. In particular if the *same* resources are shared by *all* the users, both analysis and numerics suggest that SR coding may be sufficient to achieve the full flexibility of general MD coding. However, when the users have disparate access to network resources, we could have groups of users using different operating points of various redundancy. Our distributed algorithms automatically adapt to these diverse scenarios and choose the correct operating points. We believe that this is the first step in understanding the important question of how competing MD flows can co-exist in an unreliable network. Several extensions can be investigated including use of arbitrary number of MD descriptions per flow, implementation in real networks, and so on.

Acknowledgement

This work was partially supported by Swiss National Science Foundation NCCR-MICS and United States NSF CCF-0448012 and DARPA grant HR0011-06-1-0008. The authors thank reviewers and editors for their insightful reviews, comments, and suggestions.

Appendix

A. Appendix A

Denote $d_{si} = 2^{-2\sigma_i r_{si}}$, $i = 1, 2$ and assume that $\sigma_1 r_{s1} < \sigma_2 r_{s2}$, where following [25] we assume that σ_1, σ_2 are constants in the interval (0, 1). Then, we have

$$\frac{2^{-2(r_{s1}+r_{s2})}}{1-(\sqrt{\Pi}-\sqrt{\Gamma})^2} = \frac{2^{-2(r_{s1}+r_{s2}-\sigma_1 r_{s1})}}{1+\frac{d_{s2}}{d_{s1}}-2d_{s2}+\frac{2^{-2(r_{s1}+r_{s2})}}{d_{s1}}+2\sqrt{\frac{d_{s2}}{d_{s1}}}\times\sqrt{(1-d_{s1})(1-d_{s2})(1-\frac{2^{-2(r_{s1}+r_{s2})}}{d_{s1}d_{s2}})}}$$

Let Ω denote the denominator in the right hand side. The assumption of high rate regime means that $r_{si} \to \infty$, $i = 1, 2$. It is readily checked that terms 3 and 4 of Ω approach 0, and the second square root in the last term approaches 1. Denote $\xi = \frac{d_{s2}}{d_{s1}}$, which can take any value in the interval (0, 1) (by supposition that $d_{s2} < d_{s1}$). Assuming that ξ is a constant, and letting $r_{si} \to \infty$, $i = 1, 2$, we obtain $\Omega \to (1+\sqrt{\xi})^2$, which leads to the approximation

$$d_{s0} \geq \frac{2^{-2(r_{s1}+r_{s2})}}{1-(\sqrt{\Pi}-\sqrt{\Gamma})^2} \approx \frac{2^{-2(r_{s1}+r_{s2}-\sigma_1 r_{s1})}}{(1+\sqrt{\xi})^2}$$

We ignore the constant $\frac{1}{(1+\sqrt{\xi})^2}$ in high resolution formula (5), i.e., we approximate $\frac{1}{(1+\sqrt{\xi})^2}$ as ‘1’, for simplicity.

B. Appendix B

Proof of Lemma 1.

Proof:

(i) It is clear that $A^* = r_1^*$. It is readily checked that if $r_1^* = 0$, then $A^* = 0$, $B^* = 0$; if $r_2^* = 0$, then $B^* = 0$. Denote the objective function as $F(r_1, r_2, B)$. It is straightforward to verify that $\frac{\partial F}{\partial r_1}|_{r_1 \to R} > 0$

and $\frac{\partial F}{\partial r_2}|_{r_2\to R} > 0$ if $\alpha > 1$. Hence the optima $r_i^*, \forall i \in \{1,2\}$ is not on the boundary $r_i = R$ of the feasible set of problem (19) if $\alpha > 1$.

Letting $z = r_1 + r_2$, the objective function can be written as a function in (r_1, z, B), denoted as $F(r_1, z, B)$. Since $A^* = r_1^*$, the first three constraints in problem (19) become $B \le r_1$, $B \le z - r_1, 0 \le B$. The Lagrangian is denoted as $L(r_1, z, B, \lambda_1, \lambda_2, \lambda_3) = F(r_1, z, B) - \lambda_1(r_1 - B) - \lambda_2(z - r_1 - B) - \lambda_3 B$. The following KKT (Karush-Kuhn-Tucker) conditions are the sufficient and necessary conditions for optima (r_1^*, z^*, B^*) [4],

$$\frac{\partial F(r_1, z, B)}{\partial r_1}|_{(r_1^*, z^*, B^*)} - \lambda_1 + \lambda_2 = 0,$$

$$\frac{\partial F(r_1, z, B)}{\partial z}|_{(r_1^*, z^*, B^*)} - \lambda_2 = 0,$$

$$\frac{\partial F(r_1, z, B)}{\partial B}|_{(r_1^*, z^*, B^*)} + \lambda_1 + \lambda_2 - \lambda_3 = 0,$$

$$\lambda_1(r_1^* - B^*) = 0, \quad \lambda_1 \ge 0, \quad r_1^* \ge B^*,$$

$$\lambda_2(z^* - r_1^* - B^*) = 0, \quad \lambda_2 \ge 0, \quad z^* - r_1^* \ge B^*,$$

$$\lambda_3 B^* = 0, \quad \lambda_3 \ge 0, \quad B^* \ge 0.$$

Based on the above KKT conditions, in the following we will show that $B^* = 0$, in four cases: 1) $\lambda_1 = 0, \lambda_2 = 0$; 2) $\lambda_1 > 0, \lambda_2 > 0$; 3) $\lambda_1 > 0, \lambda_2 = 0$; 1) $\lambda_1 = 0, \lambda_2 > 0$.

To simplify notation, we denote $H = h/(2\log 2)$ and $E = e^{2\log 2}$. We have

$$\frac{\partial F(r_1, z, B)}{\partial r_1} = 2\log 2\left[-(\alpha+1)E^{-r_1 - H + \alpha(z - r_1)} + \alpha E^{-B - H + \alpha r_1}\right]$$

$$\frac{\partial F(r_1, z, B)}{\partial z} = 2\log 2\left[-E^{-(z-B)} + \alpha E^{-r_1 - H + \alpha(z - r_1)} + \alpha E^{-2H + \alpha z}\right]$$

$$\frac{\partial F(r_1, z, B)}{\partial B} = 2\log 2\left[E^{-(z-B)} - E^{-B - H + \alpha r_1}\right].$$

Case 1: $\lambda_1 = 0, \lambda_2 = 0$. For this case, we have $\frac{\partial F(r_1,z,B)}{\partial r_1}|_{(r_1^*, z^*, B^*)} = 0$, which yields $E^{-r_1^* + \alpha(z^* - r_1^*)} = \frac{\alpha}{\alpha+1}E^{-B^* + \alpha r_1^*}$. Plugging this to $\frac{\partial F(r_1,z,B)}{\partial z}|_{(r_1^*, z^*, B^*)} = 0$ yields $-E^{-(z^* - B^*)} + \frac{\alpha^2}{\alpha+1}E^{-H - B^* + \alpha r_1^*} + \alpha E^{-2H + \alpha z^*} = 0$. Since the last term is positive, and (20) gives us $\frac{\alpha^2}{\alpha+1} \ge 1$, we have $E^{-(z^* - B^*)} - E^{-B^* - H^* + \alpha r_1^*} > 0$, which is exactly the same as $\frac{\partial F(r_1,z,B)}{\partial B}|_{(r_1^*,z^*,B^*)} > 0$. Since $\lambda_3 = \frac{\partial F(r_1,z,B)}{\partial B}|_{(r_1^*,z^*,B^*)} > 0$ and $\lambda_3 B^* = 0$, we have $B^* = 0$.

Case 2: $\lambda_1 > 0, \lambda_2 > 0$. For this case, from $\lambda_1(r_1^* - B^*) = 0$ and $\lambda_2(z^* - r_1^* - B^*) = 0$, we have $2r_1^* = 2B^* = z^*$. We have $\lambda_2 = \frac{\partial F(r_1,z,B)}{\partial z}|_{(r_1^*,z^*,B^*)}$ and $\lambda_1 = \lambda_2 + \frac{\partial F(r_1,z,B)}{\partial r_1}|_{(r_1^*,z^*,B^*)} = (2\log 2)[-E^{-B^*} + \alpha E^{-2H+2\alpha B^*} + (\alpha-1)E^{-B^*-H+\alpha B^*}]$. We also have $\lambda_3 = \frac{\partial F(r_1,z,B)}{\partial B}|_{(r_1^*,z^*,B^*)} + \lambda_1 + \lambda_2 = (2\log 2)[-E^{-B^*} + 2\alpha E^{-2H+2\alpha B^*} + 2(\alpha-1)E^{-B^*-H+\alpha B^*}]$. Comparing λ_1 and λ_3, we have $\lambda_3 = 2\lambda_1 + (2\log 2)E^{-B^*} > 0$ since $\lambda_1 > 0$. From $\lambda_3 B^* = 0$ we have $B^* = 0$. Actually from $\lambda_1(r_1^* - B^*) = 0$ and $\lambda_2(z^* - r_1^* - B^*) = 0$, we have $r_1^* = B^* = z^* = 0$.

Case 3: $\lambda_1 > 0, \lambda_2 = 0$. For this case, from $\lambda_1 > 0$ and $\lambda_1(r_1^* - B^*) = 0$ we have $r_1^* = B^*$. Together with $\lambda_1 = \frac{\partial F(r_1,z,B)}{\partial r_1}|_{(r_1^*,z^*,B^*)} > 0$ we have $E^{\alpha B^*} > E^{\alpha(z^*-B^*)}\frac{\alpha+1}{\alpha}$. It is readily checked that $\lambda_3 = \frac{\partial F(r_1,z,B)}{\partial B}|_{(r_1^*,z^*,B^*)} + \lambda_1 = (2\log 2)[\alpha E^{-2H+\alpha z^*} + (\alpha-1)E^{-B^*-H+\alpha B^*} - E^{-B^*-H+\alpha(z^*-B^*)}]$. The first term is positive, and since $E^{\alpha B^*} > E^{\alpha(z^*-B^*)}\frac{\alpha+1}{\alpha}$ we have $(\alpha-1)E^{\alpha B^*} > \frac{\alpha^2-1}{\alpha}E^{\alpha(z^*-B^*)} > E^{\alpha(z^*-B^*)}$ where the latter inequality comes from that $\frac{\alpha^2-1}{\alpha} > 1$ if (20) holds, hence $\lambda_3 > 0$. Since $\lambda_3 B^* = 0$, we have $B^* = 0$.

Case 4. $\lambda_1 = 0, \lambda_2 > 0$. For this case, from $\lambda_2 > 0$ we have $z^* - B^* = r_1^*$. From $\lambda_1 = 0$, we have $\lambda_2 = -\frac{\partial F(r_1,z,B)}{\partial r_1}|_{(r_1^*,z^*,B^*)} = \frac{\partial F(r_1,z,B)}{\partial z}|_{(r_1^*,z^*,B^*)}$, where the latter equality gives $\alpha E^{(\alpha+1)r_1^*} = \frac{E^{\alpha B^*}+E^H}{E^{-B^*}+E^{\alpha B^*-H}}$. We also have $\lambda_2 = (2\log 2)[-E^{-r_1^*} + \alpha E^{-r_1^*-H+\alpha B^*} + \alpha E^{-2H+\alpha r_1^*+\alpha B^*}] > 0$. Multiplying $E^{r_1^*}$ to both sides and plugging in $\alpha E^{(\alpha+1)r_1^*} = \frac{E^{\alpha B^*}+E^H}{E^{-B^*}+E^{\alpha B^*-H}}$ yields $(1+\alpha)E^{-2H+2\alpha B^*} + \alpha E^{-H+\alpha B^*-B^*} - E^{-B^*} > 0$. We also have $\lambda_3 = \frac{\partial F(r_1,z,B)}{\partial B}|_{(r_1^*,z^*,B^*)} + \lambda_2 = (2\log 2)\frac{E^{-r_1^*}}{\alpha(E^{-2H+\alpha B}+E^{-B})}[(1+\alpha)E^{-2H+2\alpha B^*} + \alpha E^{-H+\alpha B^*-B^*} - E^{-B^*} + (\alpha^2-1)E^{-2H+2\alpha B^*} + (\alpha^2-\alpha-1)E^{\alpha B^*-B^*-H} + \alpha E^{-H+\alpha B^*}] > 0$ if (20) holds because the first three terms in $[\cdot]$together are positive as shown, and $(\alpha^2-1) > 0, (\alpha^2-\alpha-1) > 0$ if (20) holds. Since $\lambda_3 B^* = 0$, we have $B^* = 0$.

(ii) By the hypothesis, problem (19) has a solution (r_1^*, z^*, B^*) which is in the interior of the feasible set. Therefore, since the objective function is continuous and differentiable, all the partial derivatives at the point $((r_1^*, z^*, B^*)$ have to be equal to 0. Equations $\frac{\partial F}{\partial B} = 0$ and $\frac{\partial F}{\partial r_1} = 0$ give us $-z+B = -h-B+\alpha r_1$ and $-r_1+\alpha(z-r_1) = -B+\alpha r_1+\log\frac{\alpha}{1+\alpha}$. Plugging these to $\frac{\partial F}{\partial z} = 0$ yields $-B+\alpha r_1 = \alpha z+\log\frac{\alpha(1+\alpha)}{1+\alpha-\alpha^2}-h$. Solving these for (B, z, r_1) yields (21) and (22). ■

References

[1] M. Alasti, K. Sayrafian-Pour, A. Ephremides, and N. Farvardin. Multiple description coding in networks with congestion problem. *IEEE Trans. on Information Theory*, 47(3):891–902, March 2001.

[2] G. Barrenechea. Distributed routing algorithm for sensor networks. PhD thesis, 2006.

[3] A. Begen, Y. Altunbasak, O. Ergun, and M. Ammar. Multipath selection for multiple description video streaming over overlay networks. *Signal Processing: Image Communication*, 20(1):39–60, Jan. 2005.

[4] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[5] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle. Layering as optimization decomposition: A mathematical theory of network architectures. *Proceedings of the IEEE*, 95(1):255–312, 2007.

[6] S. N. Diggavi. Diversity in communication: From source coding to wireless networks. *Brain and Systems: New Directions in Statistical Signal Processing*, 2006.

[7] S. N. Diggavi, N. Al-Dhahir, A. Stamoulis, and A. R. Calderbank. Great expectations: the value of spatial diversity in wireless networks. *Proceedings of the IEEE*, 92(2):217–290, 2004.

[8] A. El-Gamal and T. M. Cover. Achievable rates for multiple descriptions. *IEEE Trans. Information Theory*, 28(6):851–857, Nov. 1982.

[9] W. H. R. Equitz and T. M. Cover. Successive refinement of information. *IEEE Trans. on Information Theory*, 37(2):269–275, Mar. 1991.

[10] V. Goyal. Multiple description coding: compression meets the network. *IEEE Signal Processing Mag.*, 18(5):74–93, Sept. 2001.

[11] D. Gunarwardena, P. Key, and L. Massoulie. Network characteristic: modeling, measurements and admission control. In *Proc. IEEE/IFIP/ACM IWQoS*, June 2003.

[12] J. He, M. Bresler, M. Chiang, and J. Rexford. Towards robust multi-layer traffic engineering: optimization of congestion control and routing. *IEEE J. on Selected Areas in Communications*, 25(5):868–880, June 2007.

[13] F. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *J. of Operational Research Society*, 49(3):237–252, March 1998.

[14] P. Key, L. Massoulie, and D. Towsley. Multipath routing and congestion control: a robust Internet achitecture. In *UMass-Amherst Computer Science Technical Report 05-55*, Aug. 2005.

[15] P. Key, L. Massoulie, and D. Towsley. Combining multipath routing and congestion control for robustness. In *Proc. IEEE CISS*, March 2006.

[16] J. Kim, R. M. Mersereau, and Y. Altunbasak. Distributed video streaming using multiple description coding and unequal error protection. *IEEE Trans. on Image Processing*, 14(7):849–861, 2005.

[17] L. Kleinrock. *Queueing Systems, Vol.1: Theory, Vol.2: Computer*. Wiley, 1976.

[18] S. Kompella, S. Mao, Y. T. Hou, and H. D. Sherali. Cross-layer optimized multipath routing for video communications in wireless networks. *IEEE J. on Selected Areas in Communications*, 25(4):831–840, 2007.

[19] J. W. Lee, M. Chiang, and R. A. Calderbank. Price-based distributed algorithm for optimal rate-reliability tradeoff in network utility maximization. *IEEE J. on Selected Areas in Communications*, 24(5):962–976, May 2006.

[20] 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*, Mar 2000.

[21] Y. Li, C. Tian, S. N. Diggavi, M. Chiang, and A. R. Calderbank. Optimal network resource allocation for competing multiple description transmissions. In *Proc. IEEE Globecom*, Nov. 2008.

[22] S. Mao, Y. T. Hou, X. Cheng, H. D. Sherali, and S. F. Midkiff. Multipath routing for multiple description video in wireless ad hoc networks. In *Proc. IEEE INFOCOM*, Mar 2005.

[23] L. Ozarow. On a source-coding problem with two channels and three receivers. *The Bell System Technical Journal*, 59(10):1909–1921, Dec. 1980.

[24] J. M. Pitts, J. A. Schormans, E. M. Scharf, and A. J. Pearmain. Tight bounds for the tail of the packet waiting time distribution in buffered networks. *International Journal of Commun. Sys.*, 14:715–723, June 2001.

[25] V. A. Vaishampayan and J.-C. Batllo. Asymtotic analysis of multiple description quantizers. *IEEE Trans. on Information Therory*, 44(1):278–284, Jan. 1998.

[26] J. Wang, S. H. L. L. Li, and J. C. Doyle. Cross-layer optimization in TCP/IP networks. *IEEE/ACM Trans. on Networking*, 13(3):582–595, June 2005.

[27] Y. Wang, A. R. Reibman, M. T. Orchard, and H. Jafarkhani. An improvement to multiple description transform coding. *IEEE Trans. on Signal Processing*, 50(11):2843–2854, Nov. 2002.

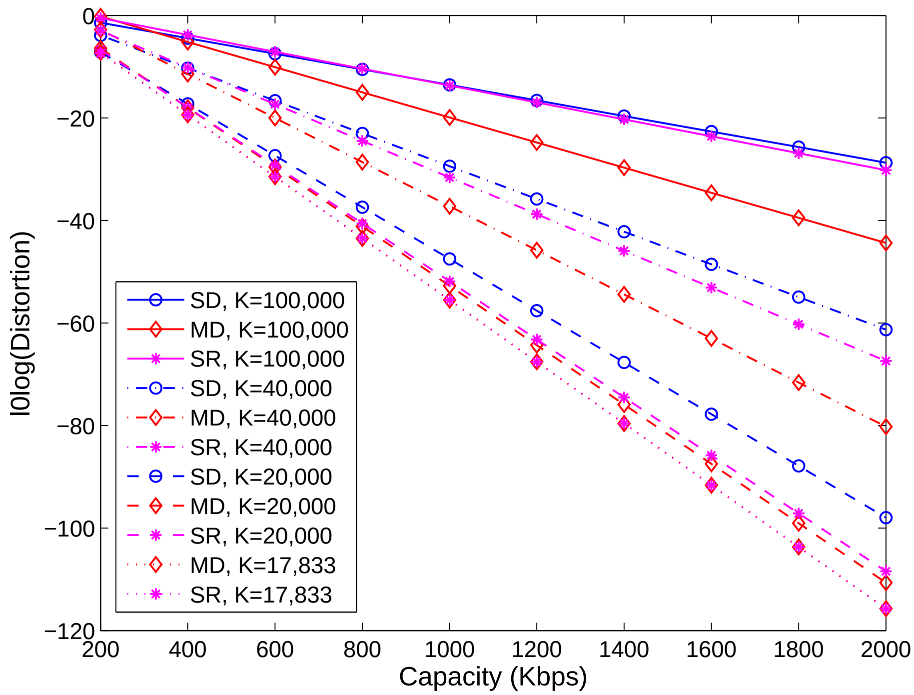

Fig. 1. Distortion comparison, in 2-link wired network with single user, w.r.t different capacity available.

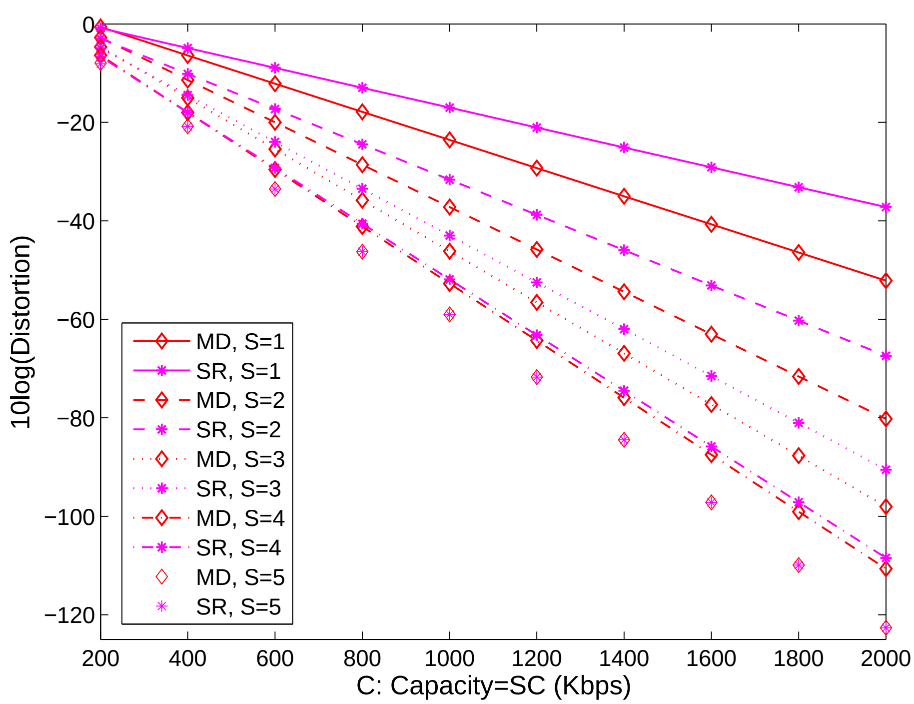

Fig. 2. Distortion comparison, in 2-link S-user wired network with capacity SC available for MD/SR flows.

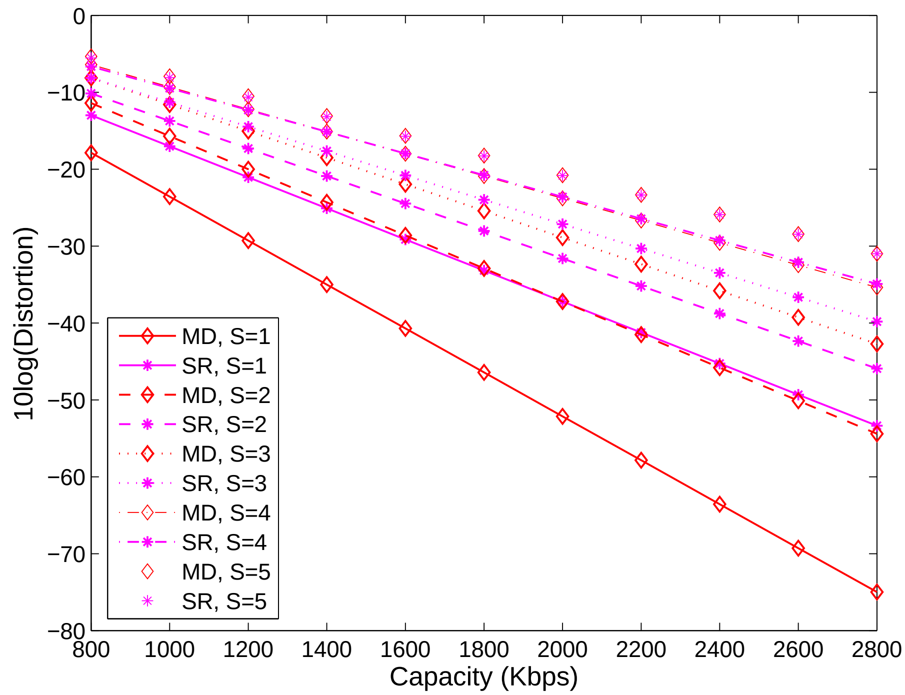

Fig. 3. Distortion comparison, in 2-link S-user wired network with capacity C available for MD/SR flows.

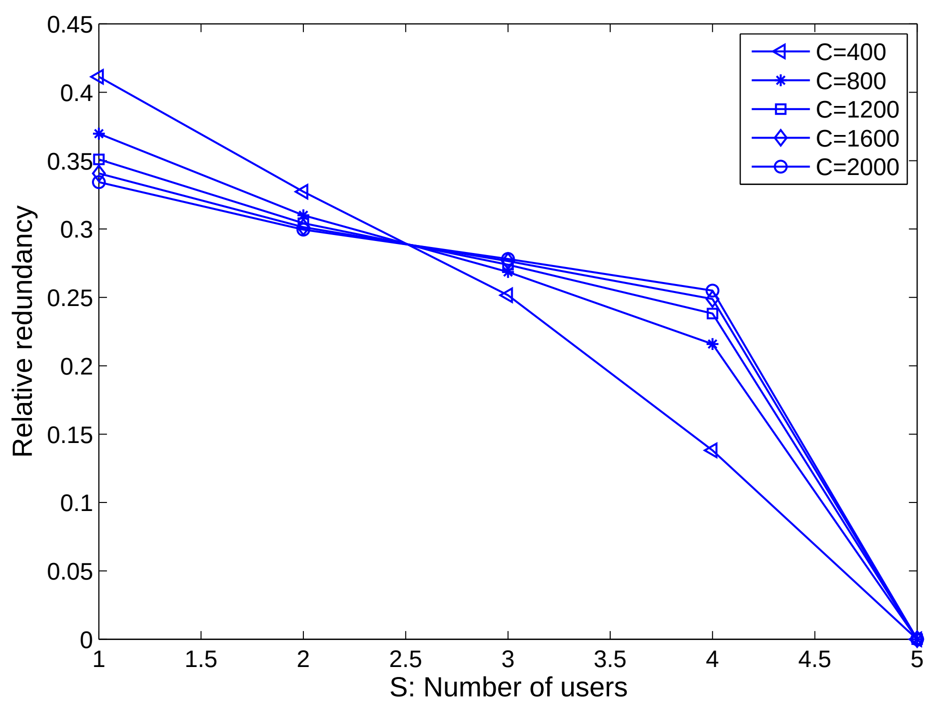

Fig. 4. Relative redundancy, in 2-link S-user wired network with capacity SC available for MD flows.

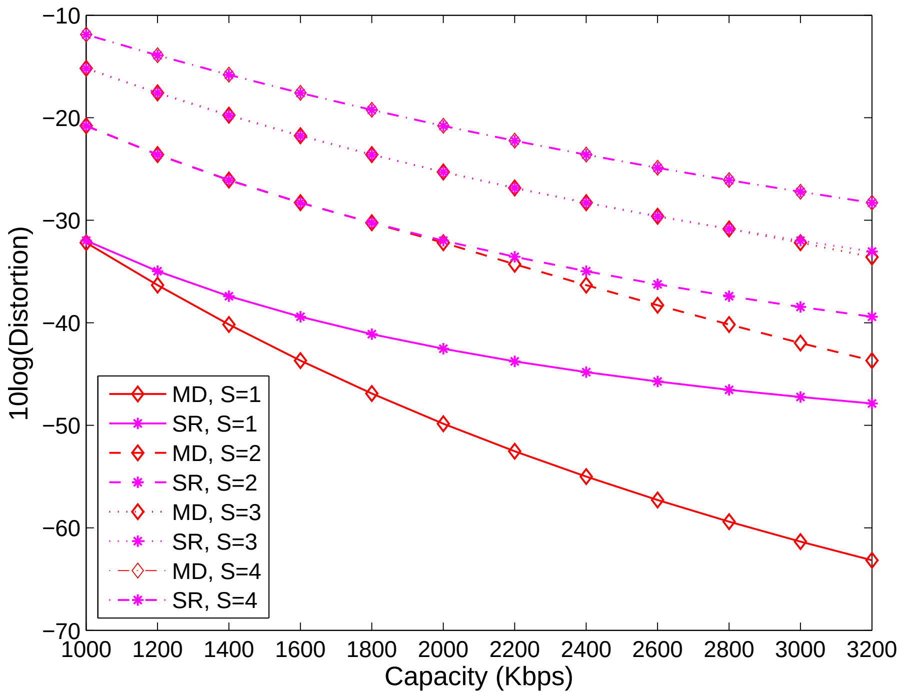

Fig. 5. Distortion comparison, in 2-link S-user wireless network with capacity C available for MD/SR flows.

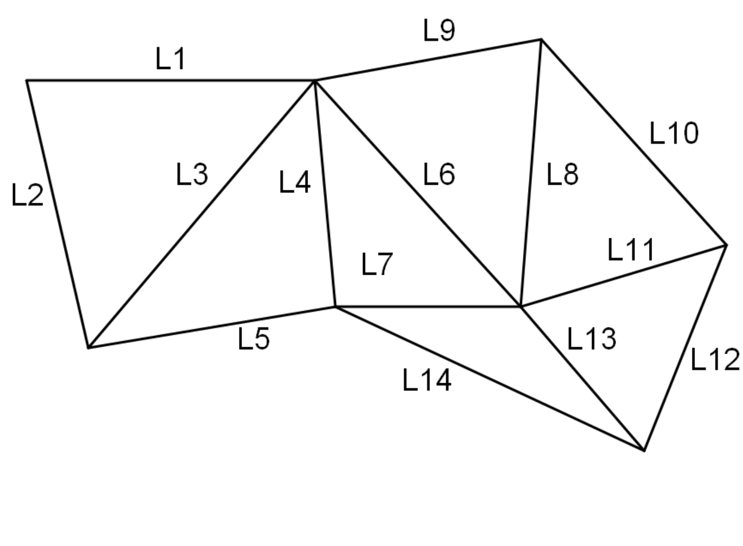

Fig. 6. A 14-link network. For 10 users (S=10), they use paths (in terms of link order) (1 4; 2 5), (3 4; 5), (1; 2 3), (5 7; 3 6), (9 10, 6 11), (10, 8 11), (11 12; 13), (7 13, 14), (6 13; 4 14), (6; 9 8), respectively. For $S = 20$ and $S = 30$, each user in 10 users case is replaced by 2 and 3 users, respectively.

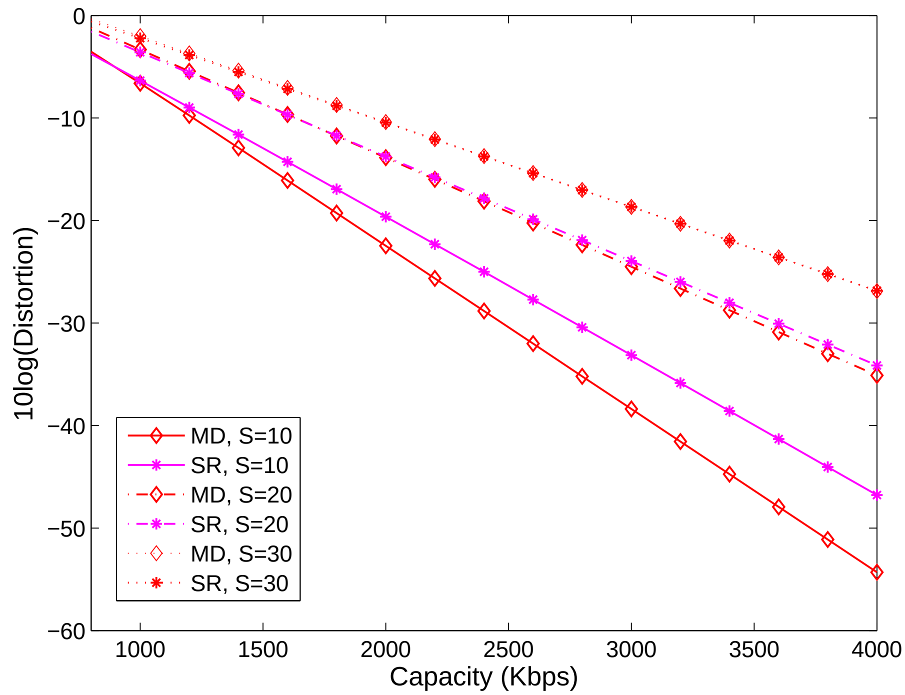

Fig. 7. Distortion comparison of MD and SR, in 14-link wired network, w.r.t different capacity available for MD/SR flows.

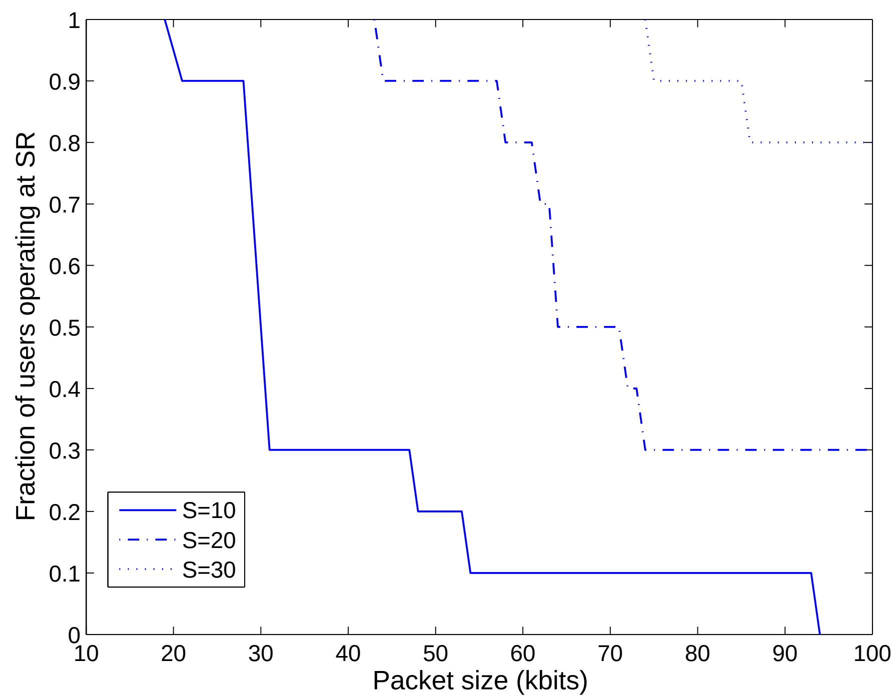

Fig. 8. Fraction of users operating at SR, in 14-link wired network.

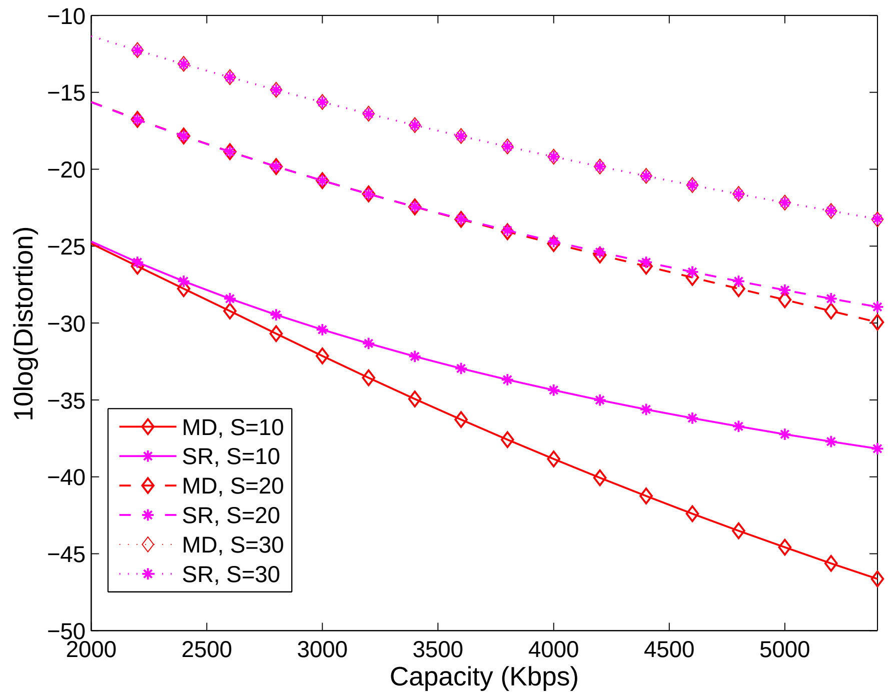

Fig. 9. Distortion comparison of MD and SR, in 14-link wireless network., w.r.t different capacity available for MD/SR flows.

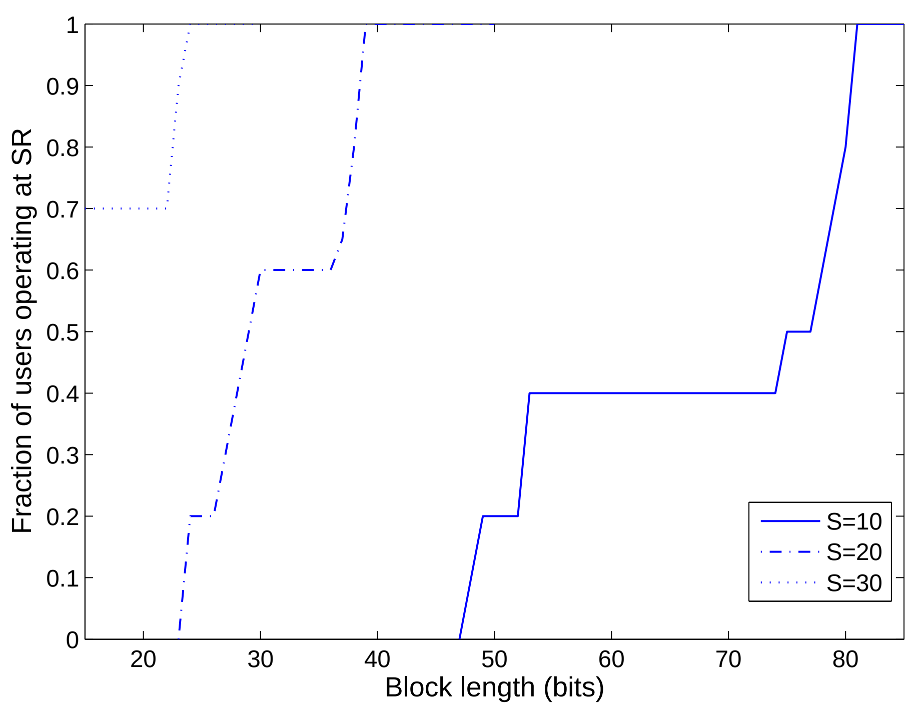

Fig. 10. Fraction of users operating at SR, in 14-link wireless network.