Price-Based Distributed Algorithms for Rate-Reliability Tradeoff in Network Utility Maximization

Jang-Won Lee, *Member, IEEE*, Mung Chiang, *Member, IEEE*, and A. Robert Calderbank, *Fellow, IEEE*

***Abstract*—The current framework of network utility maximization for rate allocation and its price-based algorithms assumes that each link provides a fixed-size transmission "pipe" and each user's utility is a function of transmission rate only. These assumptions break down in many practical systems, where, by adapting the physical layer channel coding or transmission diversity, different tradeoffs between rate and reliability can be achieved. In network utility maximization problems formulated in this paper, the utility for each user depends on both transmission rate and signal quality, with an intrinsic tradeoff between the two. Each link may also provide a higher (or lower) rate on the transmission "pipes" by allowing a higher (or lower) decoding error probability. Despite nonseparability and nonconvexity of these optimization problems, we propose new price-based distributed algorithms and prove their convergence to the globally optimal rate-reliability tradeoff under readily-verifiable sufficient conditions.**

We first consider networks in which the rate-reliability tradeoff is controlled by adapting channel code rates in each link's physical-layer error correction codes, and propose two distributed algorithms based on pricing, which respectively implement the "integrated" and "differentiated" policies of dynamic rate-reliability adjustment. In contrast to the classical price-based rate control algorithms, in our algorithms, each user provides an offered price for its own reliability to the network, while the network provides congestion prices to users. The proposed algorithms converge to a tradeoff point between rate and reliability, which we prove to be a globally optimal one for channel codes with sufficiently large coding length and utilities whose curvatures are sufficiently negative. Under these conditions, the proposed algorithms can thus generate the Pareto optimal tradeoff curves between rate and reliability for all the users. In addition, the distributed algorithms and convergence proofs are extended for wireless multiple-inpit–multiple-output multihop networks, in which diversity and multiplexing gains of each link are controlled to achieve the optimal rate-reliability tradeoff. Numerical examples confirm that there can be significant enhancement of the network utility by distributively trading-off rate and reliability, even when only some of the links can implement dynamic reliability.

***Index Terms*—Mathematical programming/optimization, network control by pricing, network utility maximization, physical-layer channel coding, rate allocation.**

Manuscript received February 15, 2005; revised January 15, 2006. This work was supported in part by Yonsei University Research Fund of 2005 and in part by the National Science Foundation (NSF) under Grant CCF-0440443, Grant CNS-0417607, Grant CNS-0427677, and Grant CCF-0448012. This paper was presented in part at IEEE Infocom 2006.

J.-W. Lee is with the Center for Information Technology, Department of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: jangwon@yonsei.ac.kr).

M. Chiang and A. R. Calderbank are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: chiangm@princeton.edu; calderbk@princeton.edu).

Digital Object Identifier 10.1109/JSAC.2006.872877

I. Introduction

SINCE the publication of the seminal paper [1] by Kelly *et al.* in 1998, the framework of Network Utility Maximization (NUM) has found many applications in network rate allocation algorithms, Internet congestion control protocols, user behavior models, and network efficiency-fairness characterization. By allowing nonlinear, concave utility objective functions, NUM substantially expands the scope of the classical Network Flow Problem based on linear programming. Moreover, there is an elegant economics interpretation of the dual-based distributed algorithm, in which the Lagrange dual variables can be interpreted as shadow prices for resource allocation, and each end user and the network maximize their net utilities and net revenue, respectively.

Consider a communication network with L logical links, wired or wireless, each with a fixed capacity of c_l b/s, and S sources (i.e., end users), each transmitting at a source rate of x_s b/s. Each source s emits one flow, using a fixed set $L(s)$ of links in its path, and has a utility function $U_s(x_s)$. Each link l is shared by a set $S(l)$ of sources. NUM, in its basic version, is the following problem of maximizing the network utility $\sum_s U_s(x_s)$, over the source rates $\mathbf{x}$, subject to linear flow constraints $\sum_{s\in S(l)} x_s \le c_l$ for all links l

$$\begin{aligned} \text{maximize} \quad & \sum_s U_s(x_s) \\ \text{subject to} \quad & \sum_{s\in S(l)} x_s \le c_l, \quad \forall l \\ & \mathbf{x} \succeq 0. \end{aligned} \tag{1}$$

Making the standard assumption on concavity of the utility functions, problem (1) is a simple concave maximization of decoupled terms under linear constraints, which has long been studied in optimization theory as a monotropic program [2].

Among many of its recent applications to communication networks, the basic NUM (1) has been extensively studied as a model for distributed network rate allocation (e.g., [1], [3], and [4]), TCP congestion control protocol analysis (e.g., [5] and [6]), and design (e.g., FAST TCP [7]). Primal and dual-based distributed algorithms have been proposed to solve for the global optimum of (1) (e.g., [8]–[12]). Various extensions of the basic NUM problem (1) have been recently investigated, e.g., for joint congestion control and power control [13], and joint opportunistic scheduling and congestion control [14] in wireless networks, for TCP/IP interactions [15], for medium access

control [16]–[20], for joint congestion control and medium access [21], for joint multicommodity flow and resource allocation [22], for nonconcave utility functions [23]–[25], and for heterogeneous congestion control protocols [26].

Despite the above extensions, the current NUM framework for rate control and its price-based algorithms still assumes that each link provides a fixed coding and modulation scheme in the physical layer, and that each user's utility is only a function of local source rate. These assumptions break down in many practical systems. On some communication links, physical-layer's adaptive channel coding (i.e., error correction coding) can change the information "pipe" sizes and decoding error probability, e.g., through adaptive channel coding in Digital Subscriber Loop (DSL) broadband access networks or adaptive diversity-multiplexing control in multiple-input–multiple-output (MIMO) wireless systems. Then, each link capacity is no longer a fixed number but a function of the signal quality (i.e., decoding reliability) attained on that link.[1] A higher throughput can be obtained on a link at the expense of lower decoding reliability, which in turn lowers the end-to-end signal quality for sources traversing the link and reduces users' utilities, thus leading to an intrinsic tradeoff between rate and reliability. This tradeoff also provides an additional degree-of-freedom for improving each user's utility, as well as system efficiency. For example, if we allow lower decoding reliability, thus higher information capacity, on the more congested links, and higher decoding reliability, thus lower information capacity, on the less congested links, we may improve the end-to-end rate and reliability performance of each user. Clearly, rate-reliability tradeoff is globally coupled across the links and users.

How can we provide a price-based framework to exploit the degree of freedom in utility maximization offered by rate-reliability tradeoff and make the above intuitions rigorous? The basic NUM framework in (1) is inadequate because it completely ignores the concept of communication reliability. In particular, it takes link capacity as a fixed number rather than a function of decoding error probabilities, does not discriminate the transmission data rate from the information data rate,[2] and formulates the utility function as a function of rate only.

In this paper, we study the rate-reliability tradeoff by extending the NUM framework. In the basic NUM, convexity and separability properties of the optimization problem readily lead to a distributed algorithm that converges to the globally optimal rate allocation. However, our new optimization formulations for the rate-reliability tradeoff are neither separable problems nor convex optimization. The constraints become nonlinear, nonconvex, and globally coupled. Despite such difficulties, we develop simple and distributed algorithms based on network pricing that converge to the optimal rate-reliability tradeoff with readily verifiable sufficient conditions. In contrast to standard price-based rate control algorithms for the basic NUM, in which each link provides the *same* congestion price to each of its users and/or each user provides its *willingness to pay for rate allocation* to the network, in our algorithms each link provides a *possibly different* congestion price to each of its users and each user also provides its *willingness to pay for its own reliability* to the network.

In addition to extending the approach of price-based distributed algorithms, we also consider the specifics of controlling the rate-reliability tradeoff in different types of networks. We first study the case where the rate-reliability tradeoff is controlled through the code rate of each source on each link. We proposes a new optimization framework and price-based distributed algorithms by considering two policies: *integrated dynamic reliability policy* and *differentiated dynamic reliability policy*. In the integrated policy, a link provides the same error probability (i.e., the same code rate) to each of the sources traversing it. In this policy, since a link provides the same code rate to each of its sources, it must provide the lowest code rate that satisfies the requirement of the source with the highest reliability. This motivates a more general approach called the differentiated policy that we propose to fully exploit the rate-reliability tradeoff when there exist multiclass sources (i.e., sources with different reliability requirements) in the network. Under the differentiated dynamic reliability policy, a link can provide a different error probability (i.e., a different code rate) to each of the sources using this link. We also prove convergence of the proposed algorithms to the globally optimal rate-reliability tradeoff for sufficiently strong channel codes and sufficiently elastic applications.

We then consider wireless MIMO multihop networks. In MIMO networks, each logical link has multiple transmit antennas and multiple receive antennas, which can provide multiple independent wireless links between a transceiver pair. If independent information streams are transmitted in parallel, data rate can be increased. This is often referred to as the multiplexing gain. On the other hand, if the same information streams are transmitted in parallel, decoding error probability can be lowered. This is often referred to as the diversity gain. A larger value of the diversity gain implies a lower decoding error probability and a larger value of the multiplexing gain implies a higher information transmission rate. As in adaptive channel coding, there exists a tradeoff between information data rate and error probability on each MIMO link [27]–[29]. By appropriately controlling diversity and multiplexing gains on each individual link, we can achieve the optimal *end-to-end* rate-reliability tradeoff for all sources in the network. We show that our techniques developed for networks with adaptive channel coding can be extended to wireless MIMO networks with adaptive diversity-multiplexing gain.

The rest of this paper is organized as follows. In Section II, we provide the system model for the adaptive channel coding problem that we consider in this paper. In Sections III and IV, we present algorithms for the integrated dynamic reliability policy and the differentiated dynamic reliability policy, respectively. We provide numerical examples for the proposed algorithms in Section V to illustrate how our algorithms can trace the Pareto

[1]Some recent work like [13] studies how adaptive transmit power control in the physical layer interacts with congestion control. Note that the dimension of end-to-end signal reliability is still missing in such work. However, this paper can indeed be viewed as a new case in the general framework of "layering as optimization decomposition," as discussed in Section VII.

[2]In this paper, the information data rate refers to the data rate before adding the error control code and the transmission data rate refers to the data rate after adding the error control code.

optimal curves between rate and reliability for all users, and highlight the enhancement to the network utility through dynamic rate-reliability adjustments, even when only some of the links can implement dynamic reliability. In Section VI, we show how to control the diversity-multiplexing gains to attain the optimal rate-reliability tradeoff in wireless MIMO multihop networks. Finally, we conclude in Section VII.

II. System Model

In this section, we present the system model for the adaptive channel coding problem. Introducing the concept of reliability into the NUM framework, each source s has a utility function $U_s(x_s, R_s)$, where x_s is an information data rate and R_s is the reliability of source s. We assume that the utility function is a continuous, increasing, and concave function[3] of x_s and R_s. Each source s has its information data rate bounded between a minimum and a maximum: $x_s^{\min}$ and $x_s^{\max}$, and has a minimum reliability requirement $R_s^{\min}$. The reliability of source s is defined as

$$R_s = 1 - p^s$$

where p^s is the end-to-end error probability of source s. Each logical link l has its maximum transmission capacity $C_l^{\max}$. After link l receives the data of source s from the upstream link, it first decodes it to extract the information data of the source and encodes it again with its own code rate $r_{l,s}$, where the code rate is defined by the ratio of the information data rate x_s at the input of the encoder to the transmission data rate $t_{l,s}$ at the output of the encoder [30], [31]. This allows a link to adjust the transmission rate and the error probability of the sources, since the transmission rate of source s at link l can be defined as

$$t_{l,s} = \frac{x_s}{r_{l,s}}$$

and the error probability of source s at link l can be defined as a function of $r_{l,s}$ by

$$p_{l,s} = E_l(r_{l,s})$$

which is assumed to be an increasing function of $r_{l,s}$. Rarely is there an analytic formula for $E_l(r_{l,s})$, and we will use various upper bounds on this function in this paper. The end-to-end error probability for each source s is

$$p^s = 1 - \Pi_{l\in L(s)}(1 - p_{l,s}) = 1 - \Pi_{l\in L(s)}\left(1 - E_l(r_{l,s})\right).$$

Assuming that the error probability of each link is small (i.e., $p_{l,s} \ll 1$), we can approximate the end-to-end error probability of source s as

$$p^s \approx \sum_{l\in L(s)} p_{l,s} = \sum_{l\in L(s)} E_l(r_{l,s}).$$

[3]Throughout this paper, a concave (convex) function means a *strictly* concave (convex) function.

Hence, the reliability of source s can be expressed as

$$R_s \approx 1 - \sum_{l\in L(s)} E_l(r_{l,s}).$$

Since each link l has a maximum transmission capacity $C_l^{\max}$, the sum of transmission rates of sources that are traversing the link cannot exceed $C_l^{\max}$, i.e.,

$$\sum_{s\in S(l)} t_{l,s} = \sum_{s\in S(l)} \frac{x_s}{r_{l,s}} \le C_l^{\max},\ \forall l.$$

III. Integrated Dynamic Reliability Policy

We first investigate a simpler, more restrictive policy where a link provides the *same* code rate to each of the sources that are using it, i.e.,

$$r_{l,s} = r_l,\ \forall s \in S(l), \quad \forall l$$

and an extended NUM problem is formulated as follows:

$$\begin{aligned}
\text{maximize} \quad & \sum_s U_s(x_s, R_s) \\
\text{subject to} \quad & R_s \le 1 - \sum_{l\in L(s)} E_l(r_l), \quad \forall s \\
& \sum_{s\in S(l)} \frac{x_s}{r_l} \le C_l^{\max}, \quad \forall l \\
& x_s^{\min} \le x_s \le x_s^{\max}, \quad \forall s \\
& R_s^{\min} \le R_s \le 1, \quad \forall s \\
& 0 \le r_l \le 1, \quad \forall l.
\end{aligned} \tag{2}$$

In this problem, the first constraint states that the network must provide reliability to each source that equals to its desired reliability. Notice that the inequality constraint will be satisfied with equality at optimality since the objective function is an increasing function of $\{R_s\}$. The second constraint states that the aggregate transmission rate of the sources at each link must be smaller than or equal to its maximum transmission capacity. The rest are simple, decoupled constraints on the minimum and maximum values for the data rate, reliability for each source, and code rate for each link.

In order to derive a distributed algorithm to solve problem (2) and to prove convergence to global optimum, the critical properties of separability and convexity need to be carefully examined. Because of the physical layer coding and the rate-reliability tradeoff that we introduce into the problem formulation, these two properties are no longer trivially held as in the basic NUM (1).

First, the integrated policy naturally leads to a decomposition of problem (2) across the links, since the second constraint can be written as

$$\sum_{s\in S(l)} x_s \le C_l^{\max} r_l, \quad \forall l. \tag{3}$$

The more complicated issue is the convexity of function $E_l(r_l)$. As an example, if random coding based on M-ary binary coded signals is used, a standard upper bound on the error probability is [31]

$$p_l < \frac{1}{2}2^{-N(R_0-r_l)}$$

where N is the block length and R_0 is the cutoff rate. Hence, if we let

$$E_l(r_l) = \frac{1}{2}2^{-N(R_0-r_l)}$$

then $E_l(r_l)$ is a convex function for given N and R_0. A more general approach is to use the random code ensemble error exponent [30] that upper bounds the decoding error probability

$$p_l \leq \exp\left(-NE_0(r_l)\right)$$

where N is the codeword block length and $E_0(r_l)$ is the random coding exponent function, which is defined for discrete memoryless channels as [30]

$$E_0(r_l) = \max_{0\leq\rho\leq 1}\max_{\mathbf{Q}}\left[E_o(\rho,\mathbf{Q}) - \rho r_l\right]$$

where

$$E_o(\rho,\mathbf{Q}) = -\log\sum_{j=0}^{J-1}\left[\sum_{k=0}^{K-1}Q(k)P(j|k)^{\frac{1}{(1+\rho)}}\right]^{1+\rho}$$

K is the size of input alphabet, J is the size of output alphabet, $Q(k)$ is the probability that input letter k is chosen, and $P(j|k)$ is the probability that output letter j is received given that input letter k is transmitted.

In general, $E_l(r_l) = \exp(-NE_0(r_l))$ may not be convex, even though it is known [30] that $E_0(r_l)$ is a convex function. However, the following lemma provides a sufficient condition for its convexity.

Lemma 1: If the absolute value of the first derivatives of $E_0(r_l)$ is bounded away from 0 and absolute value of the second derivative of $E_0(r_l)$ is upper bounded, then for a large enough codeword block length N, $E_l(r_l)$ is a convex function.

Proof: Assume that there exist positive constants ϵ_1 and ϵ_2 such that $\|dE_0(r_l)/dr_l\| \geq \epsilon_1$ and $\|d^2E_0(r_l)/dr_l^2\| \leq \epsilon_2$. We have

$$\begin{aligned}\frac{d^2E_l(r_l)}{dr_l^2} &= N\exp\left(-NE_0(r_l)\right)\\ &\quad\times\left(N\left(\frac{dE_0(r_l)}{dr_l}\right)^2 - \frac{d^2E_0(r_l)}{dr_l^2}\right)\\ &\geq N\exp\left(-NE_0(r_l)\right)\left(N\epsilon_1^2-\epsilon_2\right).\end{aligned}$$

If $N > (\epsilon_2/\epsilon_1^2)$, then $(d^2E_l(r_l)/dr_l^2) > 0$ and convexity is proved. ∎

Throughout this paper, we assume that $E_l(r_l)$ is a convex function, e.g., the conditions in Lemma 1 are satisfied if the random code ensemble error exponent is used. Hence, problem (2) is turned into a convex and separable optimization problem.

To solve problem (2), we use a dual decomposition approach. We write down the Lagrangian associated with problem (2) as follows, where λ_l and μ_s are the Lagrange multipliers on link l with an interpretation of "congestion price" and on source s with an interpretation of "reliability price," respectively

$$\begin{aligned}L(\mathbf{x},\mathbf{R},\mathbf{r},\boldsymbol{\lambda},\boldsymbol{\mu}) &= \sum_s U_s(x_s,R_s)\\ &\quad+\sum_l\lambda_l\left(r_lC_l^{\max}-\sum_{s\in S(l)}x_s\right)\\ &\quad+\sum_s\mu_s\left(1-\sum_{l\in L(s)}E_l(r_l)-R_s\right)\\ &=\sum_s\left\{U_s(x_s,R_s)-\sum_{l\in L(s)}\lambda_lx_s-\mu_sR_s\right\}\\ &\quad+\sum_l\left\{\lambda_lr_lC_l^{\max}-\sum_{s\in S(l)}\mu_sE_l(r_l)\right\}\\ &\quad+\sum_s\mu_s\\ &=\sum_s\left\{U_s(x_s,R_s)-\lambda^sx_s-\mu_sR_s\right\}\\ &\quad+\sum_l\left\{\lambda_lr_lC_l^{\max}-\mu^lE_l(r_l)\right\}+\sum_s\mu_s\end{aligned}$$

where $\lambda^s = \sum_{l\in L(s)}\lambda_l$ and $\mu^l = \sum_{s\in S(l)}\mu_s$ with interpretations of "end-to-end congestion price" on source s and "aggregate reliability price" on link l. The Lagrange dual function is

$$Q(\boldsymbol{\lambda},\boldsymbol{\mu}) = \max_{\substack{\mathbf{x}^{\min}\preceq\mathbf{x}\preceq\mathbf{x}^{\max}\\ \mathbf{R}^{\min}\preceq\mathbf{R}\preceq\mathbf{1}\\ \mathbf{0}\preceq\mathbf{r}\preceq\mathbf{1}}} L(\mathbf{x},\mathbf{R},\mathbf{r},\boldsymbol{\lambda},\boldsymbol{\mu}) \tag{4}$$

where $\mathbf{0}$ and $\mathbf{1}$ are vectors whose elements are all zeros and ones, respectively. The dual problem is formulated as

$$\begin{aligned}&\text{minimize} && Q(\boldsymbol{\lambda},\boldsymbol{\mu})\\ &\text{subject to} && \boldsymbol{\lambda}\succeq\mathbf{0}\\ &&& \boldsymbol{\mu}\succeq\mathbf{0}.\end{aligned} \tag{5}$$

To solve the dual problem, we first consider problem (4). Since the Lagrangian is separable, this maximization of the Lagrangian over $(\mathbf{x},\mathbf{R},\mathbf{r})$ can be conducted in parallel at each source s

$$\begin{aligned}&\text{maximize} && U_s(x_s,R_s)-\lambda^sx_s-\mu_sR_s\\ &\text{subject to} && x_s^{\min}\leq x_s\leq x_s^{\max}\\ &&& R_s^{\min}\leq R_s\leq 1\end{aligned} \tag{6}$$

and on each link l

$$\begin{aligned}&\text{maximize} && \lambda_lr_lC_l^{\max}-\mu^lE_l(r_l)\\ &\text{subject to} && 0\leq r_l\leq 1.\end{aligned} \tag{7}$$

Then, dual problem (5) can be solved by using the gradient projection algorithm as

$$\lambda_l(t+1) = \left[\lambda_l(t)-\beta(t)\left(r_l(t)C_l^{\max}-\sum_{s\in S(l)}x_s(t)\right)\right]^+,\ \forall l$$

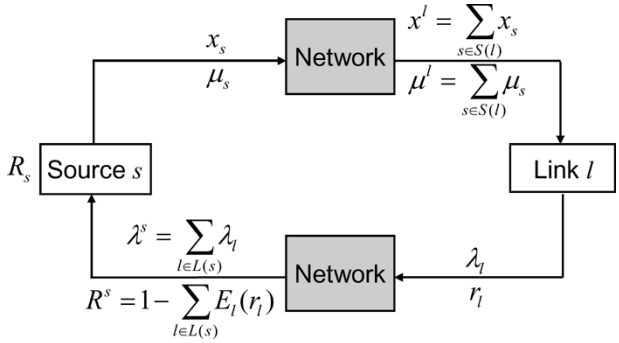

Fig. 1. Diagram for the distributed algorithm of the integrated dynamic reliability policy. Each link needs the aggregate information rate x^l and reliability price μ^l from the sources using it, and in turn provides the same congestion price λ_l and code rate r_l to these sources. Note that the source does not need to tell its desired reliability R_s to links.

and

$$\mu_s(t+1)=\left[\mu_s(t)-\beta(t)\left(1-\sum_{l\in L(s)} E_l(r_l(t))-R_s(t)\right)\right]^+, \forall s$$

where $[a]^+ = \max\{a, 0\}$, $\beta(t)$ is step size, $x_s(t)$ and $R_s(t)$ are solutions of problem (6), and $r_l(t)$ is a solution of problem (7) for a given $(\boldsymbol{\lambda}(t), \boldsymbol{\mu}(t))$.

We now describe the following distributed algorithm, where each source and each link solve their own problems with only local information, as in Fig. 1. In this case, we can interpret λ_l as the price per unit rate to use link l, and μ_s as the price per unit reliability that the source s must pay to the network.

Algorithm 1 for Integrated Dynamic Reliability Policy

Source problem and reliability price update at source s:

- Source problem

$$\begin{aligned} \text{maximize} \quad & U_s(x_s, R_s) - \lambda^s(t)x_s - \mu_s(t)R_s \\ \text{subject to} \quad & x_s^{\min} \le x_s \le x_s^{\max} \\ & R_s^{\min} \le R_s \le 1 \end{aligned} \tag{8}$$

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

- Price update

$$\mu_s(t+1) = [\mu_s(t) - \beta(t)\,(R^s(t) - R_s(t))]^+ \tag{9}$$

  where $R^s(t) = 1 - \sum_{l\in L(s)} E_l(r_l(t))$ is the end-to-end reliability at iteration t.

Link problem and congestion price update at link l:

- Link problem

$$\begin{aligned} \text{maximize} \quad & \lambda_l(t) r_l C_l^{\max} - \mu^l(t) E_l(r_l) \\ \text{subject to} \quad & 0 \le r_l \le 1 \end{aligned} \tag{10}$$

  where $\mu^l(t) = \sum_{s\in S(l)} \mu_s(t)$ is the aggregate reliability price paid by sources using link l at iteration t.

- Price update

$$\lambda_l(t+1) = \left[\lambda_l(t) - \beta(t)\left(r_l(t) C_l^{\max} - x^l(t)\right)\right]^+ \tag{11}$$

  where $x^l(t) = \sum_{s\in S(l)} x_s(t)$ is the aggregate information rate on link l at iteration t.

In each iteration t, by locally solving the following problem (8) over (x_s, R_s), each source s determines its information data rate and desired reliability [i.e., $x_s(t)$ and $R_s(t)$] that maximize its net utility based on the prices $(\lambda^s(t), \mu_s(t))$ in the current iteration. Furthermore, by price update (9), the source adjusts its offered price per unit reliability for the next iteration. Intuitively, from the end-to-end principle of network architecture design, reliability prices should indeed be updated at the sources.

Concurrently, in each iteration t, by locally solving problem (10) over r_l, each link l determines its code rate [i.e., $r_l(t)$] that maximizes the "net revenue" of the network based on the prices in the current iteration. This pricing interpretation holds because maximizing $\lambda_l(t) r_l C_l^{\max} - \mu^l(t) E_l(r_l)$ over r_l, $\forall l$, i.e., maximizing $\sum_l \{\lambda_l(t) r_l C_l^{\max} - \sum_{s\in S(l)} \mu_s E_l(r_l)\}$ over $\mathbf{r}$, is equivalent to maximizing $\sum_l \lambda_l(t) r_l C_l^{\max} + \sum_s \mu_s(t)(1 - \sum_{l\in L(s)} E_l(r_l))$ over $\mathbf{r}$. Since $r_l C_l^{\max}$ is the available (anticipated) aggregate information data rate at link l, we can interpret that $\lambda_l(t) r_l C_l^{\max}$ as the revenue obtained at link l through the congestion price. We can also interpret $\mu_s(t)(1 - \sum_{l\in L(s)} E_l(r_l))$ as the revenue obtained from source s through the reliability price. In addition, by price update (11), the link adjusts its congestion price per unit rate for the next iteration.

In Algorithm 1, to solve problem (8), source s needs to know $\lambda^s(t)$, the sum of $\lambda_l(t)$'s of links that are along its path $L(s)$. This can be obtained by the notification from the links, e.g., through the presence or timing of acknowledgment packets in TCP [5]. To carry out price update (9), the source needs to know the sum of error probabilities of the links that are along its path [i.e., its own reliability that are offered by the network $R^s(t)$]. This can be obtained by the notification from the destination that can measure its end-to-end reliability. To solve link problem (10), each link l needs to know $\mu^l(t)$, the sum of $\mu_s(t)$'s from sources $s \in S(l)$ using this link l. This can be obtained by the notification from these sources. To carry out price update (11), the link needs to know $x^l(t)$, the aggregate information data rate of the sources that are using it. This can be measured by the link itself.

After the above dual decomposition, the following result can be proved using standard techniques in distributed gradient algorithm's convergence analysis.

Theorem 1: By Algorithm 1, dual variables $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\mu}(t)$ converge to the optimal dual solutions $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$, and the corresponding primal variables $\mathbf{x}^*$, $\mathbf{R}^*$, and $\mathbf{r}^*$ are the globally optimal primal solutions of (2).

Outline of the Proof: Since strong duality holds for problem (2) and its Lagrange dual problem (5), we solve the dual problem through distributed gradient method and recover the primal optimizers from the dual optimizers. By Danskin's Theorem [32]

$$\frac{\partial Q\left(\boldsymbol{\lambda}(t), \boldsymbol{\mu}(t)\right)}{\partial \lambda_l(t)} = r_l(t) C_l^{\max} - x^l(t), \quad \forall l$$

and

$$\frac{\partial Q\left(\boldsymbol{\lambda}(t), \boldsymbol{\mu}(t)\right)}{\partial \mu_s(t)} = R^s(t) - R_s(t), \quad \forall s.$$

Hence, the algorithm in (9) and (11) is a gradient projection algorithm for dual problem (5). Since the dual objective function

$Q(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is a convex function, there exists a step size $\beta(t)$ that guarantees $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\mu}(t)$ to converge to the optimal dual solutions $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ [32]. Also, if $Q(\boldsymbol{\lambda}, \boldsymbol{\mu})$ satisfies a Lipschitz continuity condition, i.e., there exists a constant $L > 0$ such that

$$\|\nabla Q(\boldsymbol{\nu}_1) - \nabla Q(\boldsymbol{\nu}_2)\| \le L\|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|, \quad \forall \boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \{\boldsymbol{\nu} = (\boldsymbol{\lambda}, \boldsymbol{\mu}) | \boldsymbol{\lambda} \succeq \mathbf{0}, \boldsymbol{\mu} \succeq \mathbf{0}\}$$

then $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\mu}(t)$ converges to the optimal dual solutions $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ with a constant step size $\beta(t) = \beta, 0 < \beta < 2/L$ [32]. The Lipschitz continuity condition is satisfied if the curvatures of the utility functions are bounded away from zero, see [10] for further details.

Furthermore, since problem (2) is a convex optimization problem and problems (8) and (10) have unique solutions, $\mathbf{x}^*$, $\mathbf{R}^*$, and $\mathbf{r}^*$ are the globally optimal primal solutions of (2) [33]. ■

IV. Differentiated Dynamic Reliability Policy

We now investigate the more general differentiated dynamic reliability policy in which a link may provide a *different* code rate $r_{l,s}$ to each of the sources traversing it.[4] To this end, the extended NUM formulation in (2) needs to be further generalized to the following problem:

$$\begin{aligned} \text{maximize} \quad & \sum_s U_s(x_s, R_s) \\ \text{subject to} \quad & R_s \le 1 - \sum_{l \in L(s)} E_l(r_{l,s}), \quad \forall s \\ & \sum_{s \in S(l)} \frac{x_s}{r_{l,s}} \le C_l^{\max}, \quad \forall l \\ & x_s^{\min} \le x_s \le x_s^{\max}, \quad \forall s \\ & R_s^{\min} \le R_s \le 1, \quad \forall s \\ & 0 \le r_{l,s} \le 1, \quad \forall l,\ s \in S(l). \end{aligned} \tag{12}$$

The objective function and constraints of problem (12) are the same as those of problem (2), except that we have $r_{l,s}$ here instead of r_l. Due to this critical difference, we may not modify the second constraint in this problem as in (3), and problem (12) in general is neither a convex problem nor a separable one, even when $E_l(r_{l,s})$ is assumed to be a convex function.

These issues can be resolved through a series of problem transformations. First, we modify problem (12) by introducing the auxiliary variables $c_{l,s}$, which can be interpreted as the allocated transmission capacity to source s at link l. Then, the above problem can be reformulated as

$$\begin{aligned} \text{maximize} \quad & \sum_s U_s(x_s, R_s) \\ \text{subject to} \quad & R_s \le 1 - \sum_{l \in L(s)} E_l(r_{l,s}), \quad \forall s \\ & \frac{x_s}{r_{l,s}} \le c_{l,s}, \ \forall l,\ s \in S(l) \\ & \sum_{s \in S(l)} c_{l,s} \le C_l^{\max}, \quad \forall l \\ & x_s^{\min} \le x_s \le x_s^{\max}, \quad \forall s \\ & R_s^{\min} \le R_s \le 1, \quad \forall s \\ & 0 \le r_{l,s} \le 1, \quad \forall l,\ s \in S(l) \\ & 0 \le c_{l,s} \le C_l^{\max}, \quad \forall l,\ s \in S(l). \end{aligned} \tag{13}$$

The second constraint in problem (12) is now decomposed into two constraints in problem (13): the second and third constraints. Here, the second constraint states that the transmission data rate of each source at each link must be smaller than or equal to its allocated transmission capacity at the link, and the third constraint states that the aggregate allocated transmission capacity to the sources at each link must be smaller than or equal to its maximum transmission capacity. In this formulation, each link explicitly allocates a transmission capacity to each of its sources. We can easily show that problem (13) is equivalent to problem (12), since at the optimal solution of problem (13), the second constraint must be satisfied with the equality.

Note that $c_{l,s}$ is a parameter in the link layer. Hence, even though we started from the TCP/PHY layer problem, here we introduce another layer, i.e., link layer, by using a vertical decomposition of the optimization problem. At the first sight, it may seem that this makes the problem more complicated. However, it in fact enables us to implement a simple and distributed algorithm.

The next step of problem transformation is to take the log of both sides of the second constraint in problem (13) and a change of variable: $x_s' = \log x_s$ (i.e., $x_s = e^{x_s'}$). This reformulation turns the problem into

$$\begin{aligned} \text{maximize} \quad & \sum_s U_s'(x_s', R_s) \\ \text{subject to} \quad & R_s \le 1 - \sum_{l \in L(s)} E_l(r_{l,s}), \quad \forall s \\ & x_s' - \log r_{l,s} \le \log c_{l,s}, \quad \forall l,\ s \in S(l) \\ & \sum_{s \in S(l)} c_{l,s} \le C_l^{\max}, \quad \forall l \\ & x_s'^{\min} \le x_s' \le x_s'^{\max}, \quad \forall s \\ & R_s^{\min} \le R_s \le 1, \quad \forall s \\ & 0 \le r_{l,s} \le 1, \quad \forall l,\ s \in S(l) \\ & 0 \le c_{l,s} \le C_l^{\max}, \quad \forall l,\ i \in S(l) \end{aligned} \tag{14}$$

where $U_s'(x_s', R_s) = U_s(e^{x_s'}, R_s)$, $x_s'^{\min} = \log x_s^{\min}$, and $x_s'^{\max} = \log x_s^{\max}$.

Note that problem (14) is now separable but still may not be a convex optimization problem since the objective $U_s'(x_s', R_s)$

[4] An example of such coding techniques recently proposed is coding with embedded diversity [34], which allows data streams with different rate-reliability tradeoffs be embedded within each other. An interesting next step is to study how to use this technique as a practical mechanism to implement the differentiated dynamic reliability policy. Other adaptive resource allocation in the physical layer, such as adaptive modulation or interleaver depth, can also be used to the implement dynamic reliability policy.

may not be a concave function, even though $U_s(x_s, R_s)$ is a concave function. However, the following lemma provides a sufficient condition for its concavity. Define

$$g_s(x_s,R_s)=\frac{\partial^2 U_s(x_s,R_s)}{\partial x_s^2}x_s+\frac{\partial U_s(x_s)}{\partial x_s}$$

$$h_s(x_s,R_s)=\left(\left(\frac{\partial^2 U_s(x_s,R_s)}{\partial x_s\partial R_s}\right)^2 - \frac{\partial^2 U_s(x_s,R_s)}{\partial x_s^2}\frac{\partial^2 U_s(x_s,R_s)}{\partial R_s^2}\right)x_s - \frac{\partial^2 U_s(x_s,R_s)}{\partial R_s^2}\frac{\partial U_s(x_s,R_s)}{\partial x_s}$$

and

$$q_s(x_s,R_s)=\frac{\partial^2 U_s(x_s,R_s)}{\partial R_s^2}.$$

Lemma 2: If $g_s(x_s,R_s)<0$, $h_s(x_s,R_s)<0$, and $q_s(x_s,R_s)<0$, then $U_s'(x_s',R_s)$ is a concave function of x_s' and R_s.

Proof: This can be easily verified using the fact that $U_s'(x_s',R_s)$ is a concave function if and only if

$$\mathbf{y}^T\mathbf{H}_s'\mathbf{y}<0,\quad \forall \mathbf{y}\neq 0$$

where $\mathbf{y}=[y_1,y_2]^T$ and $\mathbf{H}_s'$ is a Hessian matrix of $U_s'(x_s',R_s)$, which is defined as

$$\mathbf{H}_s'=\begin{bmatrix}\frac{\partial^2 U_s(x_s,R_s)}{\partial x_s^2}x_s^2+\frac{\partial U_s(x_s,R_s)}{\partial x_s}x_s & \frac{\partial^2 U_s(x_s,R_s)}{\partial x_s\partial R_s}x_s\\ \frac{\partial^2 U_s(x_s,R_s)}{\partial x_s\partial R_s}x_s & \frac{\partial^2 U_s(x_s,R_s)}{\partial R_s^2}\end{bmatrix}.$$

It can be readily verified that if the conditions in the Lemma are satisfied, then indeed

$$y_1^2\left(\frac{\partial^2 U_s(x_s,R_s)}{\partial x_s^2}x_s^2+\frac{\partial U_s(x_s,R_s)}{\partial x_s}x_s\right)+2y_1y_2\frac{\partial^2 U_s(x_s,R_s)}{\partial x_s\partial R_s}x_s+y_2^2\frac{\partial^2 U_s(x_s,R_s)}{\partial R_s^2}<0,\ \forall \mathbf{y}\neq 0$$

i.e., $\mathbf{H}_s'$ is negative definite and $U_s'(x_s',R_s)$ is a concave function of (x_s',R_s). ■

Furthermore, if $U_s(x_s,R_s)$ is additive in each variable, $U_s(x_s,R_s)=U_s^x(x_s)+U_s^R(R_s)$ (i.e., $U_s'(x_s',R_s)=U_i^{x'}(x_s')+U_s^R(R_s)$), the following lemma provides the sufficient condition. Define

$$g_s^x(x_s)=\frac{d^2U_s^x(x_s)}{dx_s^2}x_s+\frac{dU_s^x(x_s)}{dx_s}. \tag{15}$$

Lemma 3: Suppose $U_s(x_s,R_s)=U_s^x(x_s)+U_s^R(R_s)$ and $U_s^R(R_s)$ is a concave function. Then, if $g_s^x(x_s)<0$, $U_s^{x'}(x_s')$ is a concave function of x_s' and $U_s'(x_s',R_s)$ is a concave function of x_s' and R_s.

Proof: Since $U_s^R(R_s)$ is a concave function, $q_s(x_s,R_s)<0$. Furthermore, since $U_s(x_s,R_s)$ is additive, $g_s^x(x_s)<0$ implies $g_s(x_s)<0$, and $\partial^2U_s(x_s,R_s)/\partial x_s\partial R_s=0$. This implies $h_s(x_s,R_s)=-q_s(x_x,R_s)g_s(x_s,R_s)<0$. Hence, the conditions in Lemma 2 are all satisfied. ■

The conditions of $g_s(x_s,R_s)<0$ and $h_s(x_s,R_s)<0$ are equivalent to

$$\frac{d^2U_s(x_s,R_s)}{\partial x_s^2}<-\frac{\partial U_s(x_s,R_s)}{x_s\partial x_s}$$

and

$$\frac{\partial^2U_s(x_s,R_s)}{\partial x_s^2}<\left(\frac{\partial^2U_s(x_s,R_s)}{\partial R_s^2}\right)^{-1}\left(\frac{\partial^2U_s(x_s,R_s)}{\partial x_s\partial R_s}\right)^2-\frac{\partial U_s(x_s,R_s)}{x_s\partial x_s}.$$

Since $dU_s(x_s)/dx_s$ is positive and $\partial^2U_s(x_s,R_s)/\partial R_s^2$ is negative, the above inequalities state that the utility function needs to be more than just concave. In particular, if it is additive in x_s and R_s, its curvature needs to be not just negative but bounded away from 0 by as much as $-(dU_s^x(x_s)/x_sdx_s)$, i.e., the application represented by this utility function must be elastic enough.

For example, consider the following utility function parameterized by $\alpha\geq 0$ [12]:

$$U_s(x_s,R_s)=\begin{cases}\log x_sR_s, & \text{if } \alpha=1\\ (1-\alpha)^{-1}(x_sR_s)^{1-\alpha}, & \text{otherwise}\end{cases}$$

where x_sR_s can be interpreted as the correctly decoded throughput of source s. Then, we can show that

$$\begin{cases}g_s(x_s,R_s)>0, h_s(x_s,R_s)>0, q_s(x_s,R_s)<0 & \text{if } \alpha<1\\ g_s(x_s,R_s)=0, h_s(x_s,R_s)=0, q_s(x_s,R_s)<0 & \text{if } \alpha=1\\ g_s(x_s,R_s)<0, h_s(x_s,R_s)<0, q_s(x_s,R_s)<0 & \text{if } \alpha>1\end{cases}.$$

Hence, in this type of utility functions, if $\alpha>1$, $U_s'(x_s',R_s)$ becomes a concave function.

Throughout this section, we will assume that the conditions in Lemma 2 are satisfied. Hence, problem (14) is turned into a convex and separable optimization problem.

To solve problem (14), we use a dual decomposition approach again. We first write down the Lagrangian function associated with problem (14) as

$$\begin{aligned}&L(\mathbf{x}',\mathbf{R},\mathbf{r},\mathbf{c},\boldsymbol{\lambda},\boldsymbol{\mu})\\ &=\sum_s U_s'(x_s',R_s)+\sum_s\mu_s\left(1-\sum_{l\in L(s)}E_l(r_{l,s})-R_s\right)+\sum_l\sum_{s\in S(l)}\lambda_{l,s}(\log c_{l,s}+\log r_{l,s}-x_s')\\ &=\sum_s\left(U_s'(x_s',R_s)-\sum_{l\in L(s)}\lambda_{l,s}x_s'-\mu_sR_s\right)+\sum_l\left(\sum_{s\in S(l)}(\lambda_{l,s}(\log c_{l,s}+\log r_{l,s})-\mu_sE_l(r_{l,s})\right)+\sum_s\mu_s\\ &=\sum_s(U_s'(x_s',R_s)-\lambda^sx_s'-\mu_sR_s)+\sum_l\left(\sum_{s\in S(l)}(\lambda_{l,s}(\log c_{l,s}+\log r_{l,s})-\mu_sE_l(r_{l,s})\right)+\sum_s\mu_s\end{aligned}$$

where $\lambda^s = \sum_{l \in L(s)} \lambda_{l,s}$. As in the previous case, we can interpret $\lambda_{l,s}$, λ^s, and μ^s as "congestion price" on link l, end-to-end congestion price on source s, and "reliability price" on source s, respectively. Note that in this Lagrangian function, we do not relax the third constraint in problem (14). The Lagrange dual function is

$$Q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \max_{\substack{\mathbf{x}'^{\min} \preceq \mathbf{x}' \preceq \mathbf{x}'^{\max} \\ \mathbf{R}^{\min} \preceq \mathbf{R} \preceq \mathbf{1} \\ \mathbf{0} \preceq \mathbf{r} \preceq \mathbf{1} \\ \mathbf{c} \in C}} L(\mathbf{x}', \mathbf{R}, \mathbf{r}, \mathbf{c}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad (16)$$

where $C = \{(c_{l,s})_{\forall l, s \in S(l)} | \sum_{s \in S(l)} c_{l,s} \le C_l^{\max}, \forall l, 0 \le c_{l,s} \le C_l^{\max}, \forall l, s \in S(l)\}$. The dual problem is formulated as

$$\begin{aligned} \text{minimize} \quad & Q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \succeq \mathbf{0} \\ & \boldsymbol{\mu} \succeq \mathbf{0}. \end{aligned} \quad (17)$$

To solve the dual problem, we first consider problem (16). This maximization of the Lagrangian over $\mathbf{x}'$, $\mathbf{R}$, $\mathbf{r}$, $\mathbf{c}$ can be conducted in parallel at each source s

$$\begin{aligned} \text{maximize} \quad & U_s'(x_s', R_s) - \sum_{l \in L(s)} \lambda_{l,s} x_s' - \mu_s R_s \\ \text{subject to} \quad & x_s'^{\min} \le x_s' \le x_s'^{\max} \\ & R_s^{\min} \le R_s \le 1 \end{aligned} \quad (18)$$

and on each link l

$$\begin{aligned} \text{maximize} \quad & \sum_{s \in S(l)} \lambda_{l,s} (\log c_{l,s} + \log r_{l,s}) - \mu_s E_l(r_{l,s}) \\ \text{subject to} \quad & \sum_{s \in S(l)} c_{l,s} \le C_l^{\max} \\ & 0 \le c_{l,s} \le C_l^{\max}, \ s \in S(l) \\ & 0 \le r_{l,s} \le 1, \ s \in S(l). \end{aligned} \quad (19)$$

Then, dual problem (17) can be solved by using the gradient projection algorithm as

$$\lambda_{l,s}(t+1) = [\lambda_{l,s}(t) - \beta(t)(\log c_{l,s}(t) + \log r_{l,s}(t) - x_s'(t))]^+, \quad \forall l, s \in S(l)$$

and

$$\mu_s(t+1) = \left[\mu_s(t) - \beta(t)\left(1 - \sum_{l \in L(s)} E_l(r_{l,s}(t)) - R_s(t)\right)\right]^+ \quad \forall s$$

where $x_s'(t)$ and $R_s(t)$ are solutions to problem (18), and $r_{l,s}(t)$ and $c_{l,s}(t)$ are solutions to problem (19) for a given $(\boldsymbol{\lambda}(t), \boldsymbol{\mu}(t))$.

We now describe the following distributed algorithm where each source and each link solve their own problems with only local information, as in Fig. 2. As before, we can interpret $\lambda_{l,s}$ as the price per unit rate (in terms of $x_s' = \log x_s$) for source s to use link l, and μ_s as the price per unit reliability that the source s must pay to the network.

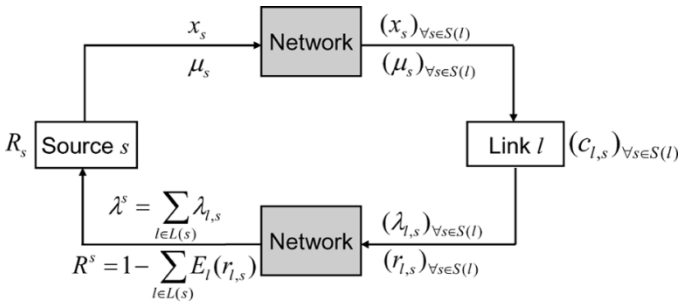

Fig. 2. Diagram for the distributed algorithm of the differentiated dynamic reliability policy. Each link needs the individual information rate x_s and reliability price μ_s from each of the sources using it, and in turn provides a different congestion price $\lambda_{l,s}$ and code rate $r_{l,s}$ to each of these sources. Note that the source does not need to tell its desired reliability R_s to links, and the link does not need to tell the allocated capacity $c_{l,s}$ to sources.

Algorithm 2 for Differentiated Dynamic Reliability Policy

Source problem and reliability price update at source s:

- Source problem

$$\begin{aligned} \text{mazimize} \quad & U_s(x_s', R_s) - \lambda^s(t) x_s' - \mu_s(t) R_s \\ \text{subject to} \quad & x_s'^{\min} \le x_s' \le x_s'^{\max} \\ & R_s^{\min} \le R_s \le 1 \end{aligned} \quad (20)$$

  where $\lambda^s(t) = \sum_{l \in L(s)} \lambda_{l,s}(t)$ is the end-to-end congestion price at iteration t.

- Price update

$$\mu_s(t+1) = [\mu_s(t) - \beta(t)(R^s(t) - R_s(t))]^+ \quad (21)$$

  where $R^s(t) = 1 - \sum_{l \in L(s)} E_l(r_{l,s}(t))$ is the end-to-end reliability at iteration t.

Link problems and congestion price update at link l:

- Link problems

  Link-layer problem

$$\begin{aligned} \text{maximize} \quad & \sum_{s \in S(l)} \lambda_{l,s}(t) \log c_{l,s} \\ \text{subject to} \quad & \sum_{s \in S(l)} c_{l,s} \le C_l^{\max} \\ & 0 \le c_{l,s} \le C_l^{\max}, \ s \in S(l). \end{aligned} \quad (22)$$

  Physical-layer problem for source s, $s \in S(l)$

$$\begin{aligned} \text{maximize} \quad & \lambda_{l,s}(t) \log r_{l,s} - \mu_s(t) E_l(r_{l,s}) \\ \text{subject to} \quad & 0 \le r_{l,s} \le 1. \end{aligned} \quad (23)$$

- Price update

$$\begin{aligned} & \lambda_{l,s}(t+1) \\ & = [\lambda_{l,s}(t) - \beta(t)(\log c_{l,s}(t) + \log r_{l,s}(t) - x_s'(t))]^+ \\ & = [\lambda_{l,s}(t) - \beta(t)(\log c_{l,s}(t) + \log r_{l,s}(t) - \log x_s(t))]^+, \\ & \quad s \in S(l). \end{aligned} \quad (24)$$

In each iteration t, by locally solving (20) over (x_s', R_s), each source s determines its information data rate and desired reliability [i.e., $x_s'(t)$, or equivalently, $x_s(t) = e^{x_s'(t)}$, and $R_s(t)$]

that maximize its net utility based on the prices in the current iteration. Furthermore, by price update (21), the source adjusts its offered price per unit reliability for the next iteration.

Concurrently, in each iteration t, by locally solving problems (22) over $c_{l,s}, \forall s \in S(l)$ and (23) $r_{l,s}$, for each $s \in S(l)$, which are decomposed from (19), each link l determines the allocated transmission capacity $c_{l,s}(t)$ and the code rate $r_{l,s}(t)$, respectively, of each of the sources using the link so as to maximize the "net revenue" of the network based on the prices in the current iteration. In addition, by price update (24), the link adjusts its congestion price per unit rate for source s during the next iteration.

Even though Algorithm 2 (i.e., the differentiated dynamic reliability policy) can be implemented and interpreted in a similar way to Algorithm 1 (i.e., the integrated dynamic reliability policy), there are important differences. In Algorithm 2, the link differentiates each of its sources and provides a different code rate $r_{l,s}$ and a different price $\lambda_{l,s}$, while in Algorithm 1, the link provides the same code rate r_l and the same price λ_l to all of its sources. In Algorithm 2, a link provides an explicit capacity allocation $c_{l,s}$ to each of its sources and the price $\lambda_{l,s}$ is determined based on the expected information data rate $c_{l,s}r_{l,s}$ and the actual information data rate x_s of each individual source, while in Algorithm 1, a link does not explicitly allocated the capacity to each individual source, and the price λ_l is determined based on the expected aggregate information rate $r_l C_l^{\max}$ and the actual aggregate information rate $x^l = \sum_{s \in S(l)} x_s$ of its sources.

The advantages of allowing different code rates for different sources on a shared link and more message passing overhead include a more dynamic rate-reliability tradeoff and a higher network utility, as confirmed through numerical examples below. However, this performance improvement is achieved at the expense of keeping per-flow states on the links and more restrictive conditions on utility functions for convergence proof (i.e., the conditions in Lemma 2).

After the above decomposition and by Lemma 2, the following result can be proved similar to Theorem 1.

Theorem 2: By Algorithm 2, dual variables $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\mu}(t)$ converge to the optimal dual solutions $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ and the corresponding primal variables $\mathbf{x}'^*$, $\mathbf{R}^*$, $\mathbf{c}^*$, and $\mathbf{r}^*$ are the globally optimal primal solutions of problem (14), i.e., $\mathbf{x}^* = (e^{x'^*})_{\forall s}$, $\mathbf{R}^*$, $\mathbf{c}^*$, and $\mathbf{r}^*$ are the globally optimal primal solutions of problem (12).

V. Numerical Examples

A. Basic Examples

In this section, we present numerical examples for the proposed algorithms by considering a simple network, shown in Fig. 3, with a linear topology consisting of four links and eight users. Utility function for user s is $U_s(x_s, R_s)$ in the following standard form of concave utility parameterized by α [12], shifted such that $U_s(x_s^{\min}, R_s^{\min}) = 0$ and $U_s(x_s^{\max}, R_s^{\max}) = 1$, and with utility on rate and utility on reliability summed up with a given weight a_s between rate and reliability utilities

$$U_s(x_s, R_s) = a_s \frac{x_s^{1-\alpha} - x_s^{\min(1-\alpha)}}{x_s^{\max(1-\alpha)} - x_s^{\min(1-\alpha)}} + (1 - a_s) \frac{R_s^{(1-\alpha)} - R_s^{\min(1-\alpha)}}{R_s^{\max(1-\alpha)} - R_s^{\min(1-\alpha)}}.$$

The decoding error probability on each link l is assumed to be of the following form:

$$P_l^L = \frac{1}{2} \exp\left(-N(1 - r_l)\right)$$

where N is the channel code block length and r_l the code rate for link l.

We trace the globally optimal tradeoff curve between rate and reliability using Algorithms 1 and 2, and then compare the network utility achieved by the following three schemes.

- Static reliability (by the standard dual-based algorithm in the literature e.g., [5], [10]): Each link provides a fixed error probability 0.025. Only rate control is performed to maximize the network utility.
- Integrated dynamic reliability (by Algorithm 1): Each link provides the same adjustable error probability to each of its users.
- Differentiated dynamic reliability (by Algorithm 2): Each link provides a possibly different error probability to each of its users.

In Figs. 4–9, the constant parameters have the following values: $\alpha = 1.1$, $x_i^{\min} = 0.1$ (Mb/s), $x_i^{\max} = 2$ (Mb/s), $C_l^{\max} = 2$ (Mb/s), $R_i^{\max} = 1$, and $R_i^{\min} = 0.9$. Each point on the curves in all the graphs represents the result of running a corresponding distributed algorithm until convergence to a provably global optimum.

We first investigate the case where all the users have the same $a_s = a$, and vary the value of a from 0 to 1 in step size of

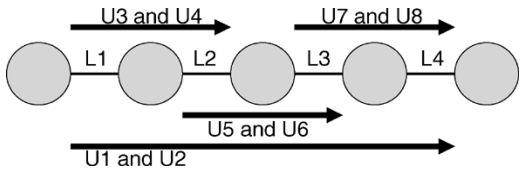

Fig. 3. Network topology and flow routes for numerical examples.

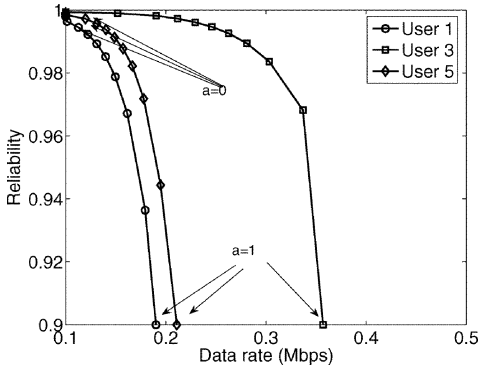

Fig. 4. Optimal tradeoff between rate and reliability for different users under the integrated dynamic reliability (Algorithm 1).

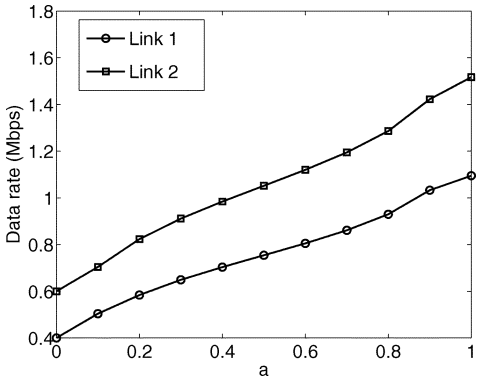

Fig. 5. Aggregate information data rate on each link if the integrated dynamic reliability is used (Algorithm 1).

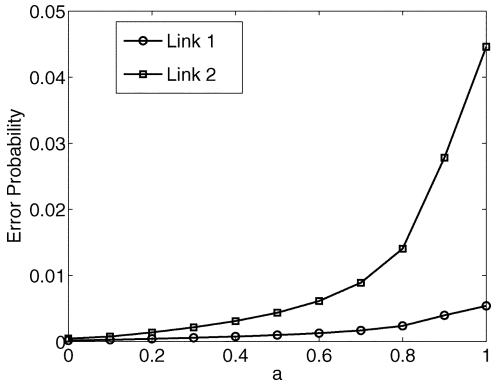

Fig. 6. Error probability on each link if the integrated dynamic reliability is used (Algorithm 1).

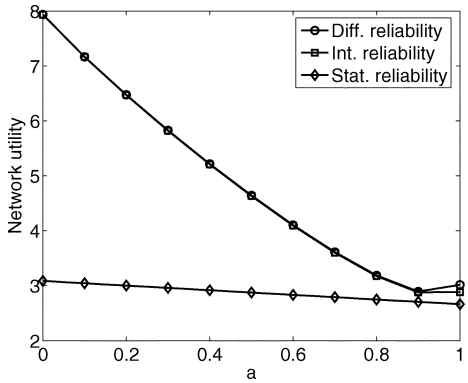

Fig. 7. Comparison of the achieved network utility attained by the differentiated dynamic policy (Algorithm 2), the integrated dynamic policy (Algorithm 1), and the static policy (current practice).

0.1. The resulting tradeoff curve, which is globally optimal, is shown in Fig. 4.[5] As expected, a larger a (i.e., rate utility is given a heavier weight) leads to a higher rate at the expense of lower reliability. The tradeoff curve has a steeper slope at a larger a (i.e., compared with low rate regime, in high rate regime further rate increment is achieved at the expense of a larger drop in reliability). The more congested links a user's flow passes through, the steeper the tradeoff curve becomes. For each user, the area

[5]Note that by symmetry, the tradeoff curves for users 1 and 2 are the same, those for users 5 and 6 are the same, and those for the other fours users are the same. Therefore, only users 1, 3, and 5 are shown in this figure.

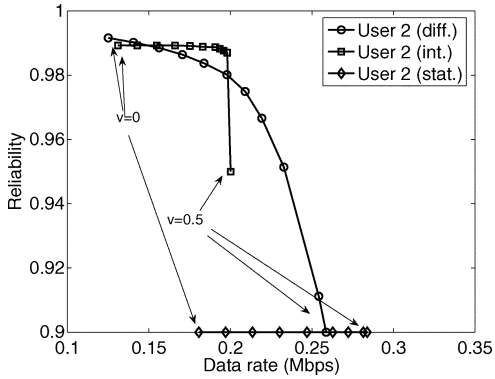

Fig. 8. Comparison of optimal tradeoff between rate and reliability for user 2 in each policy, when a_s are changed according to (25).

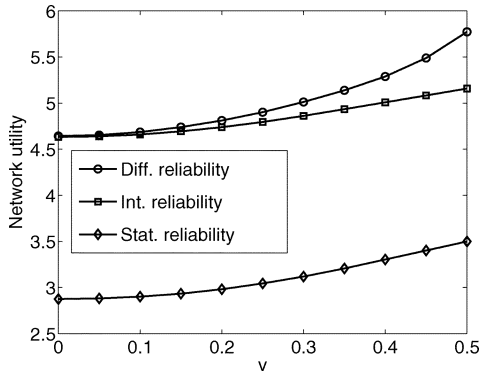

Fig. 9. Comparison of the achieved network utility attained by the differentiated dynamic policy (Algorithm 2), the integrated dynamic policy (Algorithm 1), and the static policy (current practice), when a_s are changed according to (25).

to the left and below of the tradeoff curve is the achievable region [i.e., every (rate, reliability) point in this region can be obtained], and the area to the right and above of the tradeoff curve is the infeasible region [i.e., it is impossible to have any combination of (rate, reliability) represented by points in this region]. It is impossible to operate in the infeasible region. Moreover, we can always find a better operating point on the boundary of the achievable region than a operating point in the interior of the achievable region. Hence, operating on the boundary of the achievable region, i.e., the Pareto optimal tradeoff curve, is the best. In the Pareto optimal tradeoff curve, which point is better depends on the relative weight between rate utility and reliability utility given by each user.

Figs. 5 and 6 show the aggregate information data rate and decoding error probability on each link as the weight a varies.[6] As expected, as a increases, data rates increase but decoding error probabilities also rise. Since link 2 is more congested than link 1, link 2 provides a higher data rate, also a higher error probability than link 1.

For the same experimental setup, Fig. 7 shows the network utility achieved (i.e., sum of utilities of all the users) as the weight a varies. It is clear that the performance of the NUM algorithm that takes into account the rate-reliability tradeoff is significantly better than the standard distributed algorithms for

[6]By symmetry, link 1 and 4's curves are the same, and link 2 and 3's curves are the same. Hence, the figure shows only links 1 and 2.

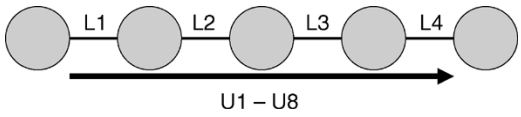

Fig. 10. Network topology and flow routes for the partial dynamic reliability example.

the basic NUM that ignores the possibility of jointly optimizing rate and reliability.[7] Within the dynamic approach, the performance gap between the integrated and differentiated policies is small because all the users are set to have the same weight a.

We now give different weights a_s to the eight users indexed by s as follows:

$$a_s = \begin{cases} 0.5 - v, & \text{if } s \text{ is an odd number} \\ 0.5 + v, & \text{if } s \text{ is an even number} \end{cases} \tag{25}$$

and vary v from 0 to 0.5 in step size of 0.05. No two users have the same utility function and flow route now. Fig. 8 shows the globally optimal tradeoff curves between rate and reliability for user 2, under the three policies of static reliability, integrated dynamic reliability, and differentiated dynamic reliability, respectively. The differentiated scheme shows a much larger dynamic range of tradeoff than both the integrated and static schemes.

Fig. 9 shows the relative performance in terms of the network utility achieved by the three policies as v changes. As expected, dynamic reliability policies are much better than the static reliability policy, and in this case where users have different weights a_s, the gap between differentiated and integrated policies is wider than that in Fig. 7.

B. Partially Dynamic Reliability

In some networks, only a subset of links have adaptive channel coding while other links have fixed code rates. For example, in the DSL access networks, edge links can provide adaptive channel coding while links in the backbone may not. Hence, it is useful to know what happens if only a subset of links adopt the dynamic reliability policy. Here, we present the performance of the differentiated dynamic reliability policy for such situations considering the network in Fig. 10, in which each link is symmetric in terms of the demand from users.

We fix the capacity of links 1 and 4 as $C_1^{\max} = C_4^{\max} = 2$ (Mb/s) and compute the network utility while varying the capacity of links 2 and 3 as $C_2^{\max} = C_3^{\max} = c^{\max}$ (Mb/s), $1.2 \leq c^{\max} \leq 3$. Hence, if $c^{\max} < 2$, links 2 and 3 are bottleneck links. Otherwise, links 1 and 4 are bottleneck links. We compare the performance of four scenarios.

- *Dynamic*: All links use the differentiated dynamic reliability policy.
- *Partial1*: Only links 2 and 3 use the differentiated dynamic reliability policy, while links 1 and 4 still use the static reliability policy.
- *Partial2*: Only links 1 and 4 use the differentiated dynamic reliability policy, while links 2 and 3 still use the static reliability policy.
- *Static*: All links still use the static reliability polity.

[7]In this example, this performance gap narrows as the weight a on rate utility increases. This happens since the error probability of each link in the static scheme is set to barely satisfy the minimum reliability requirement of the users. With a different parameter setting, the performance gap may widen as parameter a increases.

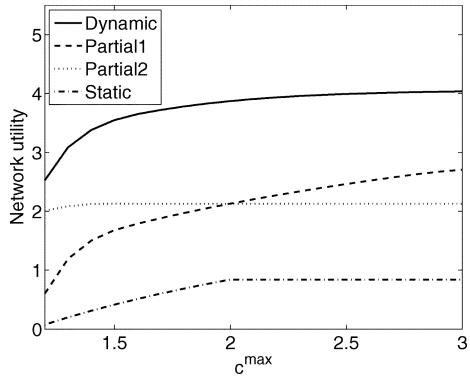

Fig. 11. Comparison of the achieved network utility in each system with the partially differentiated dynamic reliability policy. Dynamic: all links have the differentiated dynamic reliability policy; Partial1: only links 2 and 3 have the differentiated reliability policy; Partial2: only links 1 and 4 have the differentiated reliability policy; Static: all links have the static reliability policy.

Fig. 11 shows that we can still obtain significant performance gain over the static policy even though only a subset of links have adaptive channel coding. As expected, this performance gain is reduced compared with the case where all links have the dynamic policy. It also shows that we can obtain a higher performance gain by applying dynamic reliability to more congested links (usually the access links, precisely where adaptive coding is possible) than less congested ones. Furthermore, in the static policy, increasing link capacities beyond a certain threshold does not provide any further increase in network utility, while with the dynamic policy we can continue to improve network utility (with a diminishing marginal return) by increasing link capacities.

VI. Rate-Reliability Tradeoff in Wireless MIMO Multihop Networks Through Diversity-Multiplexing Gain Control

In the previous sections, we have developed the optimization framework and distributed algorithms for the optimal rate-reliability tradeoff by controlling adaptive channel coding. They can be applied to other types of networks such as wireless MIMO multihop networks, where the optimal rate-reliability tradeoff can be achieved by appropriately controlling diversity and multiplexing gains on each MIMO link [27]–[29]. In this section, we briefly present how our framework can be extended to wireless MIMO multihop networks. The detailed algorithms are then seen to be readily derived from Algorithms 1 and 2, and their details are skipped here for brevity.

In MIMO networks, each logical link l has multiple transmit and multiple receive antennas, m_l and n_l, respectively, enabling both diversity and spatial multiplexing gains, which are defined as follows [29].

Definition 1: Link l has diversity gain d_l and multiplexing gain r_l if its data rate $c_l(\gamma_l)$ (b/s/Hz) and average error probability $p_l(\gamma_l)$ satisfy the following conditions:

$$\lim_{\gamma_l \to \infty} \frac{c_l(\gamma_l)}{\log \gamma_l} = r_l \tag{26}$$

and

$$\lim_{\gamma_l \to \infty} \frac{\log p_l(\gamma_l)}{\log \gamma_l} = -d_l \tag{27}$$

where γ_l is signal-to-noise ratio (SNR).

Hence, for a given (high) SNR, we can achieve a higher data rate with a larger multiplexing gain or a lower error probability with a larger diversity gain. There exists a tradeoff between these two gains, and the optimal tradeoff curve is proved for a point-to-point link in [29]. For given multiplexing gain r, define $d_l^*(r)$ to be the supremum of the diversity gain that can be achieved over all schemes at link l.

Lemma 4: [29] Suppose that the channel gain is constant for a duration of i symbols. If $i \geq m_l + n_l - 1$,[8] $d_l^*(r)$ is given by the piecewise-linear function connecting the point $(k, d_l^*(k))$, $k = 1, 2, \cdots, \min\{m_l, n_l\}$, where

$$d_l^*(k) = (m_l - k)(n_l - k). \tag{28}$$

By using the result in Lemma 4, we study the *end-to-end* rate-reliability tradeoff in wireless MIMO multihop networks, i.e., we assume that each link uses a multiple-antenna coding scheme that achieves the optimal diversity-multiplexing tradeoff given in Lemma 4, and that there is no joint coding across point-to-point logical links. To turn the piecewise linear tradeoff curve into a differentiable one, we approximate $d_l^*(r_l)$ with a differentiable function as

$$d_l^*(r_l) = (m_l - r_l)(n_l - r_l),\ 0 \leq r_l \leq \min\{m_l, n_l\}. \tag{29}$$

The approximated tradeoff curve (29) is always below the optimum one (28), thus more conservative than the optimum curve. As the number of antennas increases, the gap between (29) and (28) decreases. From (26) and (27), we assume that for each source s traversing link l, its data rate $c_{l,s}(\gamma_l)$ (b/s) and error probability $p_{l,s}(\gamma_l)$ are given by

$$c_{l,s}(\gamma_l) = k_c r_{l,s} \log \gamma_l$$

and

$$p_{l,s}(\gamma_l) = k_p \gamma_l^{-d_{l,s}}$$

where k_c and k_p are positive constants, and $r_{l,s}$ and $d_{l,s}$ are multiplexing and diversity gains of source s at link l, respectively.

We assume that each link transmits data from each of its sources in a round-robin manner and the fraction of the transmission time of source s on link l is given by $t_{l,s}$. Hence, the average data rate $x_{l,s}$ of source s at link l is obtained by

$$x_{l,s} = k_c t_{l,s} r_{l,s} \log \gamma_l$$

and the reliability R_s of source s is obtained by

$$R_s = 1 - \sum_{l \in L(s)} k_p \gamma_l^{-d_{l,s}^*(r_{l,s})}.$$

SNR γ_l at each link l is fixed, and as before, each source s has its utility function $U_s(x_s, R_s)$.

We consider two policies: static and dynamic scheduling policies. In the static scheduling policy, each link transmits data of each of its sources for a fixed fraction of time, i.e., $t_{l,s}$ is fixed (e.g., $t_{l,s} = 1/|S(l)|$, $\forall s \in S(l)$). In the dynamic scheduling policy, $t_{l,s}$ can be adjusted based on network conditions under the following constraint:

$$\sum_{s \in S(l)} t_{l,s} \leq t_l^{\max}, \quad \forall l$$

where $t_l^{\max}$ is the maximum fraction of the time that link l can transmit data.[9]

A. Static Scheduling Policy

The NUM problem for the static scheduling policy is formulated as

$$\begin{aligned} \text{maximize} \quad & \sum_s U_s(x_s, R_s) \\ \text{subject to} \quad & R_s \leq 1 - \sum_{l \in L(s)} k_p \gamma_l^{-d_{l,s}^*(r_{l,s})}, \quad \forall s \\ & x_s \leq k_c t_{l,s} r_{l,s} \log \gamma_l, \quad \forall s,\ l \in L(s) \\ & x_s^{\min} \leq x_s \leq x_s^{\max}, \quad \forall s \\ & R_s^{\min} \leq R_s \leq 1, \quad \forall s \\ & 0 \leq r_{l,s} \leq \min\{m_l, n_l\}, \quad \forall l,\ s \in S(l). \end{aligned} \tag{30}$$

In the above problem, $t_{l,s}$ is a constant, and if $p_{l,s}(r_{l,s})$ $(= k_p \gamma_l^{-d_{l,s}^*(r_{l,s})})$ were a convex function, it would be a separable and convex problem. In general, $p_{l,s}(r_{l,s})$ is not convex. However, the following lemma shows that in most cases that are interesting in practice, it is a convex function or can be closely approximated with a convex function. Define

$$r_{l,s}^o = \frac{m_l + n_l}{2} - \frac{1}{2}\sqrt{\frac{2}{\ln \gamma_l}}.$$

Lemma 5: If $r_{l,s}^o > \min\{m_l, n_l\}$, then $p_{l,s}(r_{l,s})$ is a convex function, i.e., $(d^2 p_{l,s}(r_{l,s})/dr_{l,s}^2) > 0$, $0 \leq r_{l,s} \leq \min\{m_l, n_l\}$. Otherwise, it has an inflection point $r_{l,s}^o$ such that $(d^2 p_{l,s}(r_{l,s})/dr_{l,s}^2) > 0$ for $0 \leq r_{l,s} < r_{l,s}^o$ and $(d^2 p_{l,s}(r_{l,s})/dr_{l,s}^2) < 0$ for $r_{l,s}^o < r_{l,s} \leq \min\{m_l, n_l\}$.

Proof:

$$\frac{d^2 p_{l,s}(r_{l,s})}{dr_{l,s}^2} = \gamma_l^{-(m_l - r_{l,s})(n_l - r_{l,s})} \times \ln \gamma_l \left\{ \ln \gamma_l (-2r_{l,s} + m_l + n_l)^2 - 2 \right\}.$$

Hence, $p_{l,s}(r_{l,s})$ is convex if and only if

$$(-2r_{l,s} + m_l + n_l)^2 > \frac{2}{\ln \gamma_l}.$$

Since $0 \leq r_{l,s} \leq \min\{m_l, n_l\}$, $-2r_{l,s} + m_l + n_l \geq 0$. Therefore, $p_{l,s}(r_{l,s})$ is convex if

$$r_{l,s} < \min\left[\frac{m_l + n_l}{2} - \frac{1}{2}\sqrt{\frac{2}{\ln \gamma_l}}, \min\{m_l, n_l\} \right]$$

which completes the proof. ■

Since $\min\{m_l, n_l\} < (m_l + n_l)/2$ if $m_l \neq n_l$, and $\min\{m_l, n_l\} = (m_l + n_l)/2$ if $m_l = n_l$, the above lemma implies that $p_{l,s}(r_{l,s})$ converges to a convex function if one of the following conditions is satisfied: 1) $m_l + n_l \to \infty$;

[8] For the case of $i < m_l + n_l - 1$, the upper and lower bounds of $d^*(r)$ are also proved in [29]. For the brevity of the paper, we only consider the case that $i \geq m_l + n_l - 1$ here, and the other case can be studied in a similar manner.

[9] Ideally, $t_l^{\max} = 1$, but if we considered collisions under random-access MAC protocols, it would be less than one.

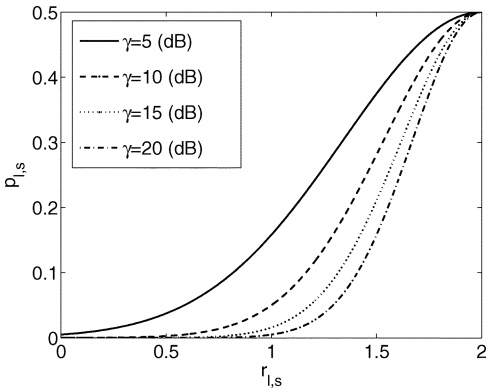

Fig. 12. Error probability function $(p_{l,s}(r_{l,s}))$ with $m_l = n_l = 2$ and $k_p = 0.5$.

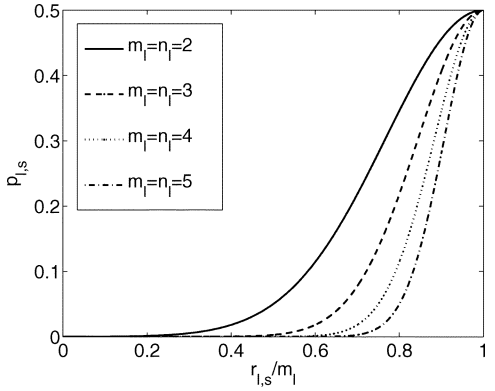

Fig. 13. Error probability function $(p_{l,s}(r_{l,s}))$ with $\gamma_l = 10$ (dB) and $k_p = 0.5$.

2) $|m_l - n_l| \to \infty$; and 3) $\gamma_l \to \infty$. Furthermore, from the above lemma, we have the following.

Corollary 1: If $m_l \neq n_l$ and $\gamma_l > 20\log(e)$ (dB) [≈8.69 (dB)], then $p_{l,s}$ is a convex function.

This corollary implies that if the number of transmit antennas is different from that of receive antennas, $p_{l,s}$ is a convex function even when SNR is not too high. Also, the required SNR for the convexity decreases as the difference between the numbers of transmit and receive antennas increases. In Figs. 12 and 13, $p_{l,s}(r_{l,s})$ with $k_p = 0.5$ is plotted when the number of transmit antennas is the same as that of receive antennas, varying SNR and the number of antennas, respectively. As indicated in Lemma 5, the figures show that as either SNR or the number of antennas increases, $p_{l,s}(r_{l,s})$ can be approximated by a convex function with higher and higher accuracy.

In the rest of this section, we assume that $p_{l,s}(r_{l,s})$ is a convex function. Hence, problem (30) is turned into a convex and separable problem. Moreover, since its structure is similar to that of problem (2), we can achieve the optimal rate-reliability tradeoff to problem (30) by using the same dual decomposition approach to that in Section III. Due to space limit, we do not repeat the details here.

B. Dynamic Scheduling Policy

The problem formulation for the dynamic scheduling policy is similar to problem (30). The crucial difference is that, in this more general policy, each link can dynamically adjust the fraction $t_{l,s}$ of the transmission time for each of its sources. The extended NUM problem is formulated as

$$\begin{aligned}
\text{maximize} \quad & \sum_s U_s(x_s, R_s) \\
\text{subject to} \quad & R_s \le 1 - \sum_{l \in L(s)} k_p \gamma_l^{-d^*_{l,s}(r_{l,s})}, \quad \forall s \\
& x_s \le k_c t_{l,s} r_{l,s} \log \gamma_l, \quad \forall s,\ l \in L(s) \\
& \sum_{s \in S(l)} t_{l,s} \le t_l^{\max}, \quad \forall l \\
& x_s^{\min} \le x_s \le x_s^{\max}, \quad \forall s \\
& R_s^{\min} \le R_s \le 1, \quad \forall s \\
& 0 \le r_{l,s} \le \min\{m_l, n_l\}, \quad \forall l,\ s \in S(l) \\
& 0 \le t_{l,s} \le t_l^{\max}, \quad \forall l,\ s \in S(l). \qquad (31)
\end{aligned}$$

In contrast to the static scheduling policy where $t_{l,s}$ is a constant, it is now a variable to be determined in the above problem. Due to the second constraint, problem (31) is neither convex nor separable. However, using the same technique as in Section IV, we can turn problem (31) into a convex and separable problem. We take the log of both sides of the second constraint in problem (31) and a log change of variable: $x'_s = \log x_s$. The problem is turned into

$$\begin{aligned}
\text{maximize} \quad & \sum_s U'_s(x'_s, R_s) \\
\text{subject to} \quad & R_s \le 1 - \sum_{l \in L(s)} k_p \gamma_l^{-d^*_{l,s}(r_{l,s})}, \quad \forall s \\
& x'_s \le \log t_{l,s} + \log r_{l,s} + \log(k_p \log \gamma_l), \\
& \forall s,\ l \in L(s) \\
& \sum_{s \in S(l)} t_{l,s} \le t_l^{\max}, \quad \forall l \\
& \log x_s^{\min} \le x'_s \le \log x_s^{\max}, \quad \forall s \\
& R_s^{\min} \le R_s \le 1, \quad \forall s \\
& 0 \le r_{l,s} \le \min\{m_l, n_l\}, \quad \forall l,\ s \in S(l) \\
& 0 \le t_{l,s} \le t_l^{\max}, \quad \forall l,\ s \in S(l) \qquad (32)
\end{aligned}$$

where $U'_s(x'_s, R_s) = U_s(e^{x'_s}, R_s)$. The structure of the above problem is similar to that of problem (14). Hence, if the conditions on utility curvature in Lemma 2 are satisfied, it is convex and separable, and the optimal rate-reliability tradeoff can be achieved by using the same dual decomposition approach as in Section IV. Algorithm 2 in that section then readily extends to solve problem (31).

VII. Concluding Remarks

Motivated by needs from the application-layer utilities and possibilities at the physical-layer coding, this paper removes the rate-dependency assumption on utility functions and allows

the physical-layer adaptive channel coding (or diversity-multiplexing gain control) to tradeoff rate and reliability. The basic network utility maximization is thus extended to concave maximization problems over nonlinear, nonconvex, and coupled constraints, which are much more difficult problems to be solved by distributed and globally optimal algorithms. In particular, the standard price-based distributed algorithm cannot be applied since the entire dimension of reliability is absent from the original formulation of network utility maximization.

We present two new price-based distributed algorithms for two possible formulations of the adaptive channel coding problem: the integrated policy where each link maintains the same code rate for all flows traversing it, and the differentiated policy where each link can assign different code rates for each of the flows traversing it. We also provide sufficient conditions under which convergence to the globally optimal rate-reliability tradeoff can be proved. A key idea is that, in addition to link-updated congestion prices for distributed rate control, we need to introduce source-updated signal quality prices for distributed reliability control. We also show that the optimization framework and distributed algorithms of network-wide rate-reliability optimal tradeoff can be extended from tuning the "knobs" of adaptive channel coding to tuning the "knobs" of diversity-multiplexing gain in wireless MIMO networks.

In the context of "layering as optimization decomposition," as outlined in [13] (see also, e.g., [15], [21], [22], [35], and [36]), this paper shows that in the case of joint congestion control in the transport layer and channel coding or MIMO signal processing in the physical layer, the new reliability prices, in addition to the standard congestion prices, become the "layering variables" controlling the optimal interaction between the two layers. The technique of introducing auxiliary flow variables, using a log change of variables, and proving convergence under additional conditions on utility curvatures may also be applicable (e.g., [37]) to solve other nonconvex and coupled NUM problems in distributed network resource allocation.

REFERENCES

[1] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar. 1998.

[2] R. T. Rockafellar, *Network Flows and Monotropic Programming*. Nashua, NH: Athena Scientific, 1998.

[3] F. P. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[4] H. Yäiche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing of elastic connections in broadband networks: Theory and algorithms," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 667–678, October 2000.

[5] S. H. Low, "A duality model of TCP and queue management algorithms," *IEEE/ACM Trans. Netw.*, vol. 11, no. 4, pp. 525–536, August 2003.

[6] R. Srikant, *The Mathematics of Internet Congestion Control*. Cambridge, MA: Birkhauser, 2004.

[7] C. Jin, D. X. Wei, and S. H. Low, "TCP FAST: Motivation, architecture, algorithms, and performance," in *IEEE INFOCOM*, vol. 4, Hong Kong, China, Mar. 2004, pp. 2490–2501.

[8] S. Kunniyur and R. Srikant, "End-to-end congestion control: Utility functions, random losses and ECN marks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 5, pp. 689–702, Oct. 2003.

[9] R. J. La and V. Anantharam, "Utility-based rate control in the Internet for elastic traffic," *IEEE/ACM Trans. Netw.*, vol. 9, no. 2, pp. 272–286, Apr. 2002.

[10] S. H. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 861–874, Dec. 1999.

[11] S. H. Low, L. Peterson, and L. Wang, "Understanding Vegas: A duality model," *J. ACM*, vol. 49, no. 2, pp. 207–235, Mar. 2002.

[12] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[13] M. Chiang, "Balancing transport and physical layer in wireless multihop networks: Jointly optimal congestion control and power control," *IEEE J. Sel. Area Commun.*, vol. 23, no. 1, pp. 104–116, Jan. 2005.

[14] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Opportunistic resource scheduling for wireless ad-hoc networks," in *Proc. BroadWISE'04*, Oct. 2004.

[15] J. Wang, L. Li, S. H. Low, and J. C. Doyle, "Can TCP and shortest path routing maximize utility," in *Proc. IEEE INFOCOM*, vol. 3, Apr. 2003, pp. 2049–2056.

[16] T. Nandagopal, T.-E. Kim, X. Gao, and V. Bharghavan, "Achieving MAC layer fairness in wireless packet networks," in *Proc. ACM MobiCom*, 2000, pp. 87–98.

[17] Z. Fang and B. Bensaou, "Fair bandwidth sharing algorithms based on game theory frameworks for wireless ad-hoc networks," in *Proc. IEEE INFOCOM*, vol. 2, 2004, pp. 1284–1295.

[18] X. Wang and K. Kar, "Distributed algorithms for max-min fair rate allocation in ALOHA networks," in *Proc. 41th Annual Allerton Conf.*, 2003.

[19] K. Kar, S. Sarkar, and L. Tassiulas, "Achieving proportional fairness using local information in Aloha networks," *IEEE Trans. Automatic Control*, vol. 49, no. 10, pp. 1858–1862, Oct. 2004.

[20] J.-W. Lee, M. Chiang, and A. R. Calderbank, "Utility-optimal medium access control: Reverse and forward engineering," in *Proc. IEEE INFOCOM*, 2006.

[21] L. Chen, S. Low, and J. Doyle, "Joint congestion control and media access control design for ad hoc wireless networks," in *Proc. IEEE INFOCOM*, 2005.

[22] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1136–1144, Jul. 2004.

[23] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Non-convexity issues for internet rate control with multi-class services: Stability and optimality," in *Proc. IEEE INFOCOM*, vol. 1, Hong Kong, China, Mar. 2004, pp. 24–34.

[24] M. Chiang, S. Zhang, and P. Hande, "Distributed rate allocation for inelastic flows: Optimization frameworks," in *Proc. IEEE INFOCOM*, Miami, FL, Mar. 2005, pp. 2679–2690.

[25] M. Fazel and M. Chiang, "Network utility maximization with nonconcave utilities using sum-of-squares method," in *Proc. IEEE Conf. Dec. Contr.*, Dec. 2005, pp. 1867–1874.

[26] A. Tang, J. Wang, S. H. Low, and M. Chiang, "Network equilibrium of heterogeneous congestion control protocols," in *Proc. IEEE INFOCOM*, Miami, FL, Mar. 2005, pp. 1338–1349.

[27] D. Gesbert, M. Shafi, D. Shiu, P. J. Smith, and A. Naguib, "From theory to practice: An overview of MIMO space-time coded wireless systems," *IEEE J. Sel. Area Commun.*, vol. 21, no. 3, pp. 281–302, Apr. 2003.

[28] S. N. Diggavi, N. Al-Dhahir, A. Stamoulis, and A. R. Calderbank, "Great expectations: The value of spatial diversity in wireless networks," *Proc. IEEE*, vol. 92, no. 2, pp. 219–270, Feb. 2004.

[29] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003.

[30] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[31] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 1989.

[32] D. P. Bertsekas, *Nonlinear Programming*. Nashua, NH: Athena Scientific, 1999.

[33] M. Minoux, *Mathematical Programming: Theory and Algorithms*. New York: Wiley, 1986.

[34] A. R. Calderbank, S. N. Diggavi, and N. Al-Dhahir, "Space-time signalling based on Kerdock and Delsarte-Goethals codes," in *Proc. IEEE ICC*, vol. 1, Paris, France, Jun. 2004, pp. 483–487.

[35] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer rate control in wireless networks," in *Proc. IEEE INFOCOM*, Miami, FL, Mar. 2005, pp. 1804–1814.

[36] L. Chen, S. H. Low, M. Chiang, and J. C. Doyle, "Optimal cross-layer congestion control, routing, and scheduling design in ad hoc wireless networks," in *Proc. IEEE INFOCOM*, 2006.

[37] J.-W. Lee, M. Chiang, and A. R. Calderbank, "Jointly optimal congestion and contention control in wireless ad hoc networks," *IEEE Commun. Lett.*, vol. 10, no. 3, pp. 216–218, Mar. 2006.

Jang-Won Lee (S'02–M'04) received the B.S. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1994, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Taejon, Korea, in 1996, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2004.

From 1997 to 1998, he was with Dacom R&D Center, Taejon. From 2004 to 2005, he was a Postdoctoral Research Associate in the Department of Electrical Engineering, Princeton University, Princeton, NJ. Since September 2005, he has been an Assistant Professor in the School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea. His research interests include resource allocation, quality-of-service (QoS) and pricing issues, optimization, and performance analysis in communication networks.

Mung Chiang (S'00–M'03) received the B.S. (Hon.) degree in electrical engineering and mathematics, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1999, 2000, and 2003, respectively.

He is an Assistant Professor of Electrical Engineering at Princeton University, Princeton, NJ. He conducts research in the areas of nonlinear optimization of communication systems, distributed network control algorithms, broadband access network architectures, and information theory and stochastic analysis of communication systems.

Prof. Chiang has been awarded a Hertz Foundation Fellow. He received the Stanford University School of Engineering Terman Award, the SBC Communications New Technology Introduction Contribution Award, the NSF CAREER Award, and the Princeton University Howard B. Wentz Junior Faculty Award. He is the Lead Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (Special Issue on Nonlinear Optimization of Communication Systems), a Guest Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY, and the IEEE/ACM TRANSACTIONS ON NETWORKING (Special Issue on Networking and Information Theory). He is Program Co-Chair of the 38th Conference on Information Science and Systems.

A. Robert Calderbank (M'89–SM'97–F'98) received the B.Sc. degree from Warwick University, Warwick, U.K., in 1975, the M.Sc. degree from Oxford University, Oxford, U.K., in 1976, and the Ph.D. degree from the California Institute of Technology, Pasadena, in 1980, all in mathematics.

He is currently a Professor of Electrical Engineering and Mathematics at Princeton University, Princeton, NJ, where he directs the Program in Applied and Computational Mathematics. He joined Bell Telephone Laboratories as a Member of Technical Staff in 1980, and retired from AT&T in 2003 as Vice President of Research. He has research interests that range from algebraic coding theory and quantum computing to the design of wireless and radar systems.

Dr. Calderbank was honored by the IEEE Information Theory Prize Paper Award in 1995 for his work on the Z4 linearity of Kerdock and Preparata Codes (joint with A. R. Hammons, Jr., P. V. Kumar, N. J. A. Sloane, and P. Sole), and again in 1999 for the invention of space–time codes (joint with V. Tarokh and N. Seshadri). He is a recipient of the IEEE Millennium Medal, and was elected to the National Academy of Engineering in 2005. He served as Editor-in-Chief of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1995 to 1998, and as Associate Editor for *Coding Techniques* from 1986 to 1989. He was a member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996.