Alternative Distributed Algorithms for Network Utility Maximization: Framework and Applications

Daniel P. Palomar, *Member, IEEE*, and Mung Chiang, *Member, IEEE*

Abstract—Network utility maximization (NUM) problem formulations provide an important approach to conduct network resource allocation and to view layering as optimization decomposition. In the existing literature, distributed implementations are typically achieved by means of the so-called dual decomposition technique. However, the span of decomposition possibilities includes many other elements that, thus far, have not been fully exploited, such as the use of the primal decomposition technique, the versatile introduction of auxiliary variables, and the potential of multilevel decompositions. This paper presents a systematic framework to exploit alternative decomposition structures as a way to obtain different distributed algorithms, each with a different tradeoff among convergence speed, message passing amount and asymmetry, and distributed computation architecture. Several specific applications are considered to illustrate the proposed framework, including resource-constrained and direct-control rate allocation, and rate allocation among QoS classes with multipath routing. For each of these applications, the associated generalized NUM formulation is first presented, followed by the development of novel alternative decompositions and numerical experiments on the resulting new distributed algorithms. A systematic enumeration and comparison of alternative vertical decompositions in the future will help complete a mathematical theory of network architectures.

Index Terms—Congestion control, distributed algorithm, mathematical programming/optimization, network control by pricing, network utility maximization (NUM), rate control, resource allocation.

I. INTRODUCTION

A. Motivation

Why would one care about a systematic theory of alternative decompositions for variants of network utility maximization (NUM) problems? There are two main reasons: it leads to the most appropriate distributed algorithm for a given network resource allocation problem, and it quantifies the comparison across architectural alternatives of distributed, layered network control.

First, since the publication of the seminal paper [1] by Kelly *et al.* in 1998, the framework of NUM has found many applications in network resource allocation algorithms and Internet congestion control protocols, e.g., [2]–[7]. The key innovation from this series of work is to interpret source rates as primal variables, link congestion prices as dual variables, and a TCP-active queue management (AQM) protocol as a distributed algorithm over the Internet to solve an implicit, global utility maximization and its Lagrange dual problem. Different TCP-AQM protocols optimize different concave utility functions using different link prices. Equilibrium and dynamic properties of a large network under TCP/AQM control, such as throughput, fairness, queue lengths, loss probabilities, and stability, can be readily understood by studying the underlying nonlinear utility maximization problem. In addition to this reverse engineering direction, allocation of limited resources in wired and wireless networks, such as power, bandwidth, and rate, among competing users can also be designed by solving generalizations of the basic NUM in [1].

Almost all the papers in the vast, recent literature on NUM use a standard dual-based distributed algorithm or a primal penalty function approach. Contrary to the apparent impression that such decompositions are the only possibilities, there are, in fact, many alternatives to solve a given network utility problem in different, but all distributed, manners. Each of the alternatives provides a possibly different tradeoff among important considerations such as convergence speed, the amount and asymmetry of message passing's communication overhead, and the architecture of distributed computation. There is no universally "best" way to distribute the solution process across a network: which alternative is the most desirable depends on the specific problem formulation and application needs. Thus, motivated, we develop a systematic framework of alternative decompositions in this paper and apply it to four network rate allocation problems motivated by practical needs and constraints.

Second, the framework of NUM has recently been substantially extended from an analytic tool of reverse-engineering TCP congestion control to a general approach of understanding interactions across layers in the network protocol stack. One possible perspective to rigorously and holistically understand layering is to integrate the various protocol layers into a single coherent theory, by regarding them as carrying out an asynchronous distributed computation over the network to implicitly solve a global optimization problem. Different layers iterate on different subsets of the decision variables using local information to achieve individual optimality. These local algorithms

collectively attempt to achieve a global objective. This approach provides a unifying view and rigorous methodology to study performance and architectural issues in protocol layering.

Since the design of a complex system will always be broken down into simpler modules, a "layering as optimization decomposition" theory [8], [9] will allow us to systematically carry out this layering process and explicitly trade off design objectives. Each different decomposition represents a new possibility of network architecture. However, to develop such a theory, alternative decompositions must be fully explored to enable understanding of architectural possibilities, both "vertically" across functional modules, i.e., the layers, and "horizontally" across disparate network elements. This paper primarily studies alternatives of horizontal decompositions, although some results are directly applicable to vertical decompositions as well, e.g., the results in Section VI can be readily applied to the joint TCP and MAC design in [10]. A more comprehensive summary of alternative vertical decompositions can be found in the next subsection.

There are three stages of conceptual understanding of an optimization/decomposition view of network architectures.

- First, layered and distributed network architectures can be rigorously understood as decompositions of an underlying optimization problem.
- Second, there are, in fact, many alternatives of decompositions and, therefore, alternatives of network architectures. Furthermore, we can systematically explore and compare such alternatives.
- Third, there may be a methodology to exhaustively enumerate all alternatives, to quantify various comparison metrics, and even to determine a priori which alternative is the best according to any given combination of comparison metrics.

An extensive literature in the last eight years have accomplished the first stage. This paper aims at the second stage in the above list. Many issues in the third stage remain open for future research, as will be discussed in Section VIII.

We conclude this subsection with a clarification. There are, in general, three major steps going from an engineering problem to a distributed algorithm. First is the process of problem formulation, second is the decomposition from a single optimization problem into multiple subproblems, and third is a distributed algorithm solving the subproblems and the original problem. In this paper, we are primarily interested in step 2, although it is affected by step 1 and, in turn, influences step 3, both of which will be briefly discussed, as well.

B. Existing Work

There are at least three levels of understanding of a utility maximization problem. The first level is about theoretical properties, such as the size of duality gap and the implication of global optimality from local optimality [11]. The second level is about computationally efficient algorithms, e.g., there are provably polynomial-time and practically efficient, but centralized, algorithms to solve a convex optimization. The third level is about distributed algorithms that converge to the global optimum. It is the third level with which we concern ourselves in this paper.

There is, indeed, a large body of results on distributed computation, some of which are summarized in standard textbooks such as [12]–[15] and others. Our goal here is certainly *not* to survey these known results in linear programming, graph theory, or decomposable problems. Instead we focus on the engineering problems of network rate allocation through problems in the form of nonlinear, coupled NUM, and develop novel distributed algorithms through a systematic method of alternative decompositions.

There are at least five different meanings of the phrase "primal–dual" in optimization theory. First, primal–dual interior point method is an efficient but centralized algorithm commonly used for many convex optimization problems. Second, primal and dual decompositions in this paper refer to techniques of decoupling constraints. Third, there is an alternative approach to obtain a distributed solution through the primal penalty function method. Fourth, [1] outlines two major classes of approaches to solve the basic version of NUM: primal-based and dual-based, depending on where the iterative updates of variables is carried out. Both approaches in [1] are analyzed through penalty function and Lyapunov argument, thus different from the language of primal and dual decomposition in this paper. Fifth, the term "primal–dual algorithm" is used in [3] to mean that both the primal problem and the Lagrange dual problem are solved simultaneously, even though the algorithm is based on dual decomposition.

Similar to one of the first publications in reverse-engineering TCP congestion control [6], many recent papers on distributed resource allocation with optimization models are based on Lagrangian relaxation and one-level, full, dual decomposition. In fact, as illustrated in this paper, this standard dual decomposition is only one of the many choices one can make, including mutlilevel, indirect, and hybrid primal–dual decompositions. Despite its popularity, the standard dual decomposition may not be the best choice for all applications. While the basic reverse engineering results for TCP show that the current protocol is doing a dual-based distributed algorithm, it is also known that dual decomposition has major drawbacks when the application is inelastic and utility functions are nonconcave, which leads to divergence of congestion control. While most of the recent publications on "layering as optimization decomposition" are based on dual decomposition and use congestion price as "layering price," it is also known that congestion price on its own can be a poor coordination across layers such as TCP and IP. It remains to be seen how primal decomposition, which has recently started to be used in MIMO transceiver design [16], [17] and wireless resource allocation [18], [19], may help in these scenarios where the simplest dual decomposition does not work well.

This paper mainly focuses on alternative horizontal decomposition, by which we mean distributed control across geographically disparate network elements. Recent results on alternative vertical decomposition (i.e., modularized control over multiple functional modules or layers) scatter in an increasingly large research literature. For example, on the topic of joint congestion control, routing, and scheduling, different decompositions have

TABLE I
SUMMARY OF THE DECOMPOSITIONS CONSIDERED IN THE APPLICATIONS
(● DENOTES EXISTING ALGORITHM AND ✓ NEW ALGORITHM)

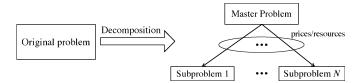| Section	Primal	Full Dual	Partial Dual	Primal-Dual	Dual-Dual
II-E	-	●	-	-	-
III	-	-	-	●	●
IV	-	-	✓	✓	-
V	●	-	✓	-	-
VI	-	✓	●	✓	-

Fig. 1. Decomposition of a problem into several subproblems controlled by a master problem through prices (dual decomposition) or by direct resource allocation (primal decomposition). The master problem may also be distributedly solved.

been obtained in [20]–[24], on the topic of joint congestion control and random access, different decompositions have been obtained in [25], [26], and on the topic of rate control for network coding based multicast, different decompositions have been obtained in [27]–[32]. A systematic treatise on this variety of vertical decompositions is an interesting future research direction that will contribute to a rigorous understanding of the architectural choices of allocating functionalities to control modules.

C. Summary of Results

We first present a systematic framework in Section II for alternative decompositions and show how that would lead to an array of choices of distributed algorithms. Section II thus serves both as a review of the necessary background and a summary of our new extensions in decomposition theorems. In particular, Lemmas 1 and 3 extend existing results on subgradients, and the techniques of multilevel and indirect primal/dual decompositions are systematically introduced in the context of NUM problems.

The core of this paper then consists of Sections III to VI, covering four applications of distributed rate allocation: power-constrained rate allocation in Section III, rate allocation among different quality-of-service (QoS) groups in Section IV, hybrid rate-based and pricing-based rate allocation in Section V, and rate allocation with multipath routing in Section VI. In particular, the distributed algorithms obtained in Sections IV-B, IV-C, V-C, VI-B, and VI-D are new. The types of decompositions developed in each application are summarized in Table I (when there are two levels of decompositions, they are separated by a dash, and, for simplicity of terminology, we differentiate in the name between full and partial dual decomposition only for decompositions with one level).

In all application sections, after the optimization formulation is explained, we develop alternative decompositions and show the benefits of fully exploring the space of possible distributed algorithms. In particular, in some cases, the distribution of computational load and asymmetry of message passing are much more desirable in one of the possible alternatives, and in other cases the convergence can be accelerated as confirmed in the numerical examples in Section VII.

II. SYSTEMATIC FRAMEWORK FOR DECOMPOSITIONS: REVIEW AND EXTENSIONS

We first present a systematic framework to decompose a given optimization problem. In the following sections, we will see how different combinations of the basic elements in

Sections II-A to II-C lead to different distributed algorithms in network utility problems, among which one will typically be preferable to the others, depending on the specific application.

While most of the concepts in this section are quick summaries of known results (e.g., Sections II-D and II-E), a couple of extensions are also carried out (e.g., Lemmas 1 and 3) and some new techniques that will be very useful later in this paper are introduced (e.g., Sections II-B and II-C).

The basic idea of a decomposition is to decompose the original large problem into distributively solvable subproblems, which are then coordinated by a master problem by means of some kind of signaling (see Fig. 1) [12], [15], [33]. Most of the existing decomposition techniques can be classified into *primal decomposition* and *dual decomposition* methods. The former is based on decomposing the original primal problem, whereas the latter based on decomposing the Lagrangian of the problem [33], [34]. Primal decomposition methods have the interpretation that the master problem directly gives each subproblem the amount of resources that it can use; the role of the master problem is then to properly allocate the existing resources. In dual decomposition methods, the master problem sets the price for the resources, and each subproblem has to decide the amount of resources to be used depending on the price; the role of the master problem is then to obtain the best pricing strategy. The difference between primal and dual decompositions is one between direct resource allocation and pricing-feedback-based allocation.

Structures of coupled constraint sets often naturally lead to a particular combination of primal and dual decomposition, as illustrated below.

A. Direct Primal and Direct Dual Decompositions

A primal decomposition is appropriate when the problem has a coupling variable such that, when fixed to some value, the rest of the optimization problem decouples into several subproblems. Consider, for example, the following problem:

$$\begin{aligned}
\underset{\mathbf{y},\{\mathbf{x}_i\}}{\text{maximize}} \quad & \sum_i f_i(\mathbf{x}_i) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i \quad \forall i \\
& \mathbf{A}_i \mathbf{x}_i \leq \mathbf{y} \quad \forall i \\
& \mathbf{y} \in \mathcal{Y}.
\end{aligned} \quad (1)$$

Clearly, if variable $\mathbf{y}$ were fixed, then the problem would decouple. Therefore, it makes sense to separate the optimization in (1) into two levels of optimization. At the lower level, we have

the subproblems, one for each i, into which (1) decouples when $\mathbf{y}$ is fixed

$$
\begin{aligned}
\underset{\mathbf{x}_i}{\text{maximize}} \quad & f_i(\mathbf{x}_i) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i \\
& \mathbf{A}_i \mathbf{x}_i \le \mathbf{y}.
\end{aligned}
\tag{2}
$$

At the higher level, we have the master problem in charge of updating the coupling variable $\mathbf{y}$ by solving

$$
\begin{aligned}
\underset{\mathbf{y}}{\text{maximize}} \quad & \sum_i f_i^\star(\mathbf{y}) \\
\text{subject to} \quad & \mathbf{y} \in \mathcal{Y}
\end{aligned}
\tag{3}
$$

where $f_i^\star(\mathbf{y})$ is the optimal objective value of problem (2) for a given $\mathbf{y}$. If the original optimization problem (1) is convex (here meaning that the objective function is concave and the feasible set is convex), then the subproblems as well as the master problem are all convex.

If the function $\sum_i f_i^\star(\mathbf{y})$ is differentiable, then the master problem (3) can be solved with a gradient method. In general, however, the objective function $\sum_i f_i^\star(\mathbf{y})$ may be nondifferentiable, and the subgradient method is a convenient approach that only requires the knowledge a subgradient (see Section II-D for details on subgradients) for each $f_i^\star(\mathbf{y})$, as given by [15, Ch. 9], [33, Sec. 6.4.2]

$$
\mathbf{s}_i(\mathbf{y}) = \boldsymbol{\lambda}_i^\star(\mathbf{y})
\tag{4}
$$

where $\boldsymbol{\lambda}_i^\star(\mathbf{y})$ is the optimal Lagrange multiplier corresponding to the constraint $\mathbf{A}_i \mathbf{x}_i \le \mathbf{y}$ in problem (2). The global subgradient is then $\mathbf{s}(\mathbf{y}) = \sum_i \mathbf{s}_i(\mathbf{y}) = \sum_i \boldsymbol{\lambda}_i^\star(\mathbf{y})$. The subproblems in (2) can be locally and independently solved with the knowledge of $\mathbf{y}$.

A dual decomposition is appropriate when the problem has a coupling constraint such that, when relaxed, the optimization problem decouples into several subproblems. Consider, for example, the following problem:

$$
\begin{aligned}
\underset{\{\mathbf{x}_i\}}{\text{maximize}} \quad & \sum_i f_i(\mathbf{x}_i) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i \qquad \forall i \\
& \sum_i \mathbf{h}_i(\mathbf{x}_i) \le \mathbf{c}.
\end{aligned}
\tag{5}
$$

Clearly, if the constraint $\sum_i \mathbf{h}_i(\mathbf{x}_i) \le \mathbf{c}$ were absent, then the problem would trivially decouple. Therefore, it makes sense to relax the coupling constraint in (5) as

$$
\begin{aligned}
\underset{\{\mathbf{x}_i\}}{\text{maximize}} \quad & \sum_i f_i(\mathbf{x}_i) - \boldsymbol{\lambda}^T \left(\sum_i \mathbf{h}_i(\mathbf{x}_i) - \mathbf{c} \right) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i \qquad \forall i
\end{aligned}
\tag{6}
$$

such that the optimization separates into two levels of optimization. At the lower level, we have the subproblems, one for each i, in which (6) decouples

$$
\begin{aligned}
\underset{\mathbf{x}_i}{\text{maximize}} \quad & f_i(\mathbf{x}_i) - \boldsymbol{\lambda}^T \mathbf{h}_i(\mathbf{x}_i) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i.
\end{aligned}
\tag{7}
$$

At the higher level, we have the master dual problem updating the dual variable $\boldsymbol{\lambda}$ to solve the dual problem

$$
\begin{aligned}
\underset{\boldsymbol{\lambda}}{\text{minimize}} \quad & g(\boldsymbol{\lambda}) = \sum_i g_i(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \mathbf{c} \\
\text{subject to} \quad & \boldsymbol{\lambda} \ge \mathbf{0}
\end{aligned}
\tag{8}
$$

where $g_i(\boldsymbol{\lambda})$ is the dual function obtained as the maximum value of the Lagrangian solved in (7) for a given $\boldsymbol{\lambda}$. This approach is, in fact, solving the dual problem instead of the original primal one. Hence, it will only give appropriate results if strong duality holds (e.g., when the original problem is convex and there exists a strictly feasible solution to the nonlinear constraints).

If the dual function $g(\boldsymbol{\lambda})$ is differentiable, then the master dual problem in (8) can be solved with a gradient method. In general, however, it may not be differentiable, and the subgradient method is again a convenient approach that only requires knowledge of a subgradient for each $g_i(\boldsymbol{\lambda})$, as given by [33, Sec. 6.1]

$$
\mathbf{s}_i(\boldsymbol{\lambda}) = -\mathbf{h}_i\left(\mathbf{x}_i^\star(\boldsymbol{\lambda})\right)
\tag{9}
$$

where $\mathbf{x}_i^\star(\boldsymbol{\lambda})$ is the optimal solution of problem (7) for a given $\boldsymbol{\lambda}$. The global subgradient is then $\mathbf{s}(\boldsymbol{\lambda}) = \sum_i \mathbf{s}_i(\mathbf{y}) + \mathbf{c} = \mathbf{c} - \sum_i \mathbf{h}_i(\mathbf{x}_i^\star(\boldsymbol{\lambda}))$. The subproblems in (7) can be locally and independently solved with knowledge of $\boldsymbol{\lambda}$.

General Results. We now present the following new result to be used later in the paper, which generalizes the known result in [15, Ch. 9], [33, Sec. 6.4.2], [where the particular result in (4) is obtained] and gives the subgradient for a more general case of primal decomposition.

Lemma 1: Consider the following function defined as the optimal value of a maximization problem:

$$
f^\star(\mathbf{x}) \triangleq \sup_{\mathbf{y}: f_i(\mathbf{x}, \mathbf{y}) \le 0} f_0(\mathbf{x}, \mathbf{y})
\tag{10}
$$

where $f_0 : \Re^n \times \Re^m \longmapsto \Re$ is (strictly) concave, $f_i : \Re^n \times \Re^m \longmapsto \Re$ are convex, the supremum is achieved by $\mathbf{y}^\star(\mathbf{x})$ (assumed to be a continuous mapping), and strong duality holds for any given $\mathbf{x}$. Let $\mathcal{F} = \{\mathbf{x} \,|\, f^\star(\mathbf{x}) > -\infty\}$ be the domain of $f^\star$.

Then, $f^\star$ is (strictly) concave, provided that $f^\star(\mathbf{x}) < \infty$ for some $\mathbf{x} \in \mathcal{F}$, and a subgradient is given by

$$
\mathbf{s}_\mathbf{x}^\star(\mathbf{x}) = \mathbf{s}_{0,\mathbf{x}}\left(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})\right) - \mathbf{S}_\mathbf{x}\left(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})\right) \boldsymbol{\lambda}^\star(\mathbf{x})
\tag{11}
$$

where $\mathbf{s}_{0,\mathbf{x}}(\mathbf{x}, \mathbf{y})$ is a subgradient of $f_0(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{x}$, $\mathbf{S}_\mathbf{x}(\mathbf{x}, \mathbf{y})$ is a matrix containing in the ith column a subgradient of $f_i(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{x}$, and $\boldsymbol{\lambda}^\star(\mathbf{x})$ is the optimal Lagrange multiplier associated with the constraints $f_i(\mathbf{x}, \mathbf{y}) \le 0$, $\forall i$, of the maximization in (10).[1]

Proof: See the Appendix. ∎

Note that if there is a unique value of the primal–dual pair $(\mathbf{y}^\star(\mathbf{x}), \boldsymbol{\lambda}^\star(\mathbf{x}))$ for any given $\mathbf{x}$ and the functions f_i's are differentiable, then $\mathbf{s}_\mathbf{x}^\star(\mathbf{x})$ is unique and $f^\star(\mathbf{x})$ is differentiable. If the problem is instead the infimum of a convex function $f_0(\mathbf{x}, \mathbf{y})$, the previous result holds by using $-f_0(\mathbf{x}, \mathbf{y})$ instead of $f_0(\mathbf{x}, \mathbf{y})$; it then follows that $f^\star(\mathbf{x})$ is convex and a subgradient is given by $\mathbf{s}_{0,\mathbf{x}}(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})) + \mathbf{S}_\mathbf{x}(\mathbf{x}, \mathbf{y}^\star(\mathbf{x}))\boldsymbol{\lambda}^\star(\mathbf{x})$.

---

[1] The optimal Lagrange multiplier $\boldsymbol{\lambda}^\star(\mathbf{x})$ is obtained "for free" each time that $f^\star(\mathbf{x})$ is evaluated at some point.

We will also later need the following well-known result concerning a partial Lagrangian obtained by relaxing some constraints $f_i(\mathbf{x}) \leq 0$ while still retaining some others $g_i(\mathbf{x}) \leq 0$.

Lemma 2: [33, Sec. 6.1] Consider the following dual function defined as the supremum of a partial Lagrangian:

$$g(\boldsymbol{\lambda}) \triangleq \sup_{\mathbf{x}:g_i(\mathbf{x}) \leq 0} \left\{ f_0(\mathbf{x}) - \sum_i \lambda_i f_i(\mathbf{x}) \right\} \quad (12)$$

where the supremum is achieved by $\mathbf{x}^\star(\boldsymbol{\lambda})$.[2]

Then, $g(\boldsymbol{\lambda})$ is convex and a subgradient, denoted by $\mathbf{s}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$, is given by

$$s_{\lambda_i}(\boldsymbol{\lambda}) = -f_i(\mathbf{x}^\star(\boldsymbol{\lambda})). \quad (13)$$

Note that if there is a unique value $\mathbf{x}^\star(\boldsymbol{\lambda})$ that achieves the supremum in (12) for any given $\boldsymbol{\lambda}$, then $g(\boldsymbol{\lambda})$ is differentiable and $\nabla g(\boldsymbol{\lambda}) = \mathbf{s}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ (this happens, for example, if f_0 is strictly concave and the f_i's are linear) [33, Prop. 6.1.1]. If the original problem is instead a minimization of a convex function $f_0(\mathbf{x})$, problem (12) is defined as the infimum of $f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x})$, $g(\boldsymbol{\lambda})$ is then a concave function to be maximized, and a subgradient is given by $s_{\lambda_i}(\boldsymbol{\lambda}) = f_i(\mathbf{x}^\star(\boldsymbol{\lambda}))$.

B. Indirect Primal and Indirect Dual Decompositions

Often, the problem can be reformulated and more effective primal and dual decompositions can be indirectly applied. The introduction of auxiliary variables provides much flexibility in choosing a primal or a dual decomposition and the resulting distributed algorithm. In [35], an indirect dual decomposition is applied to a problem with coupling among the utilities, as well as in the constraints.

The basic techniques are illustrated as follows. Problem (1) contains a coupling variable and was decoupled in (2) and (3) via a primal decomposition approach. However, it can also be solved with an indirect dual decomposition by first introducing the additional variables $\{\mathbf{y}_i\}$

$$
\begin{aligned}
\underset{\mathbf{y},\{\mathbf{y}_i\},\{\mathbf{x}_i\}}{\text{maximize}} \quad & \sum_i f_i(\mathbf{x}_i) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i \quad && \forall i \\
& \mathbf{A}_i \mathbf{x}_i \leq \mathbf{y}_i \quad && \forall i \\
& \mathbf{y}_i = \mathbf{y} \quad && \forall i \\
& \mathbf{y} \in \mathcal{Y}.
\end{aligned} \quad (14)
$$

This way, we have transformed the coupling variable $\mathbf{y}$ into a set of coupling constraints $\mathbf{y}_i = \mathbf{y}$, which can be dealt with using a dual decomposition.

Consider now problem (5), which contains a coupling constraint and was decoupled in (7) and (8) via a dual decomposition. By introducing again additional variables $\{\mathbf{y}_i\}$, the problem becomes

$$
\begin{aligned}
\underset{\{\mathbf{y}_i\},\{\mathbf{x}_i\}}{\text{maximize}} \quad & \sum_i f_i(\mathbf{x}_i) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i \quad && \forall i \\
& \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{y}_i \quad && \forall i \\
& \sum_i \mathbf{y}_i \leq \mathbf{c}.
\end{aligned} \quad (15)
$$

[2]The optimal point $\mathbf{x}^\star(\boldsymbol{\lambda})$ is obtained "for free" each time that $g(\boldsymbol{\lambda})$ is evaluated at some point.

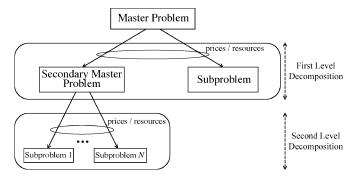Fig. 2. Example of a multilevel primal/dual decomposition with two levels.

This way, we have transformed the coupling constraint $\sum_i \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{c}$ into a coupling variable $\mathbf{y} = [\mathbf{y}_1^T, \cdots, \mathbf{y}_N^T]^T$, which can be dealt with using a primal decomposition.

C. Multilevel Primal and Dual Decompositions

An important technique that leads to alternatives of distributed architectures is to apply primal/dual decompositions recursively: The basic decompositions are repeatedly applied to a problem to obtain smaller and smaller subproblems as illustrated in Fig. 2. For example, consider the following problem which includes both a coupling variable and a coupling constraint:

$$
\begin{aligned}
\underset{\mathbf{y},\{\mathbf{x}_i\}}{\text{maximize}} \quad & \sum_i f_i(\mathbf{x}_i, \mathbf{y}) \\
\text{subject to} \quad & \mathbf{x}_i \in \mathcal{X}_i \quad && \forall i \\
& \sum_i \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{c} \\
& \mathbf{A}_i \mathbf{x}_i \leq \mathbf{y} \quad && \forall i \\
& \mathbf{y} \in \mathcal{Y}.
\end{aligned} \quad (16)
$$

One way to decouple this problem is by first taking a primal decomposition with respect to the coupling variable $\mathbf{y}$ and then a dual decomposition with respect to the coupling constraint $\sum_i \mathbf{h}_i(\mathbf{x}_i) \leq \mathbf{c}$. This would produce a two-level optimization decomposition: a master primal problem, a secondary master dual problem, and the subproblems. Observe that an alternative approach would be to first take a dual decomposition and then a primal one.

Another example that shows flexibility in terms of different decompositions is the following problem with two sets of constraints:

$$
\begin{aligned}
\underset{\mathbf{x}}{\text{maximize}} \quad & f_0(\mathbf{x}) \\
\text{subject to} \quad & f_i(\mathbf{x}) \leq 0 \\
& g_i(\mathbf{x}) \leq 0 \quad && \forall i.
\end{aligned} \quad (17)
$$

One way to deal with this problem is via the dual problem with a full relaxation of both sets of constraints to obtain the dual function $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$. At this point, instead of minimizing g directly with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, it can be minimized over only one set of Lagrange multipliers first and then over the remaining one: $\min_{\boldsymbol{\lambda}} \min_{\boldsymbol{\mu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu})$. This approach corresponds to first applying a full dual decomposition and then a primal one on the dual problem as illustrated in Fig. 3. The following new result characterizes the subgradient of the master problem at the top level.
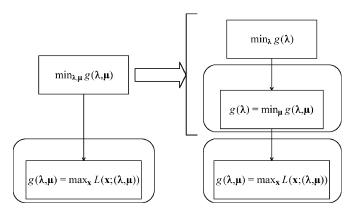
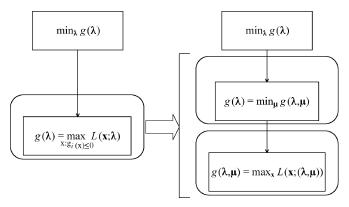Fig. 3. Illustration of the two-level dual-primal decomposition approach.

Fig. 4. Illustration of the two-level dual–dual decomposition approach.

Lemma 3: Consider the following partial minimization of the dual function:

$$g(\boldsymbol{\lambda}) = \inf_{\boldsymbol{\mu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad (18)$$

where the infimum is achieved by $\boldsymbol{\mu}^\star(\boldsymbol{\lambda})$ (assumed to be a continuous mapping) and $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is the dual function defined as

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq \sup_{\mathbf{x} \in \mathcal{X}} \left\{ f_0(\mathbf{x}) - \sum_i \lambda_i f_i(\mathbf{x}) - \sum_i \mu_i g_i(\mathbf{x}) \right\} \quad (19)$$

where the supremum is achieved by $\mathbf{x}^\star(\boldsymbol{\lambda}, \boldsymbol{\mu})$.

Then, $g(\boldsymbol{\lambda})$ is convex, and a subgradient, denoted by $\mathbf{s}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$, is given by

$$s_{\lambda_i}(\boldsymbol{\lambda}) = -f_i\left(\mathbf{x}^\star\left(\boldsymbol{\lambda}, \boldsymbol{\mu}^\star(\boldsymbol{\lambda})\right)\right). \qquad (20)$$

Proof: It follows from Lemmas 1 and 2. ∎

Alternatively, problem (17) can be approached via the dual problem but with a partial relaxation of only one set of constraints, say $f_i(\mathbf{x}) \leq 0, \forall i$, obtaining the dual function $g(\boldsymbol{\lambda})$ to be minimized by the master problem. Observe now that in order to compute $g(\boldsymbol{\lambda})$ for a given $\boldsymbol{\lambda}$, the partial Lagrangian has to be maximized subject to the remaining constraints $g_i(\mathbf{x}) \leq 0, \forall i$, for which yet another relaxation can be used. This approach corresponds to first applying a partial dual decomposition and then, for the subproblem, another dual decomposition, as illustrated in Fig. 4.

Interestingly, in this particular case, both dual-primal and dual–dual decompositions lead to the same final decomposition

as can be observed from Figs. 3 and 4; however, the conceptual way by which the final decomposition is obtained still indicates a different implementation of the distributed algorithm (e.g., different allocation of the optimization levels to different protocol layers).

When there is more than one level of decomposition and all levels conduct some type of iterative algorithms, such as the subgradient method, convergence and stability are guaranteed if the lower level master problem is solved on a faster timescale than the higher level master problem, so that, at each iteration of a master problem, all the problems at a lower level have already converged.

D. Review: Gradient/Subgradient Methods

Consider the following general concave maximization over a convex set:

$$\begin{aligned} \underset{\mathbf{x}}{\text{maximize}} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}. \end{aligned} \qquad (21)$$

Both the gradient and subgradient projection methods generate a sequence of feasible points $\{\mathbf{x}(t)\}$, indexed by t, as

$$\mathbf{x}(t+1) = [\mathbf{x}(t) + \alpha(t)\mathbf{s}(t)]_{\mathcal{X}} \qquad (22)$$

where $\mathbf{s}(t)$ is a gradient of f_0 at $\mathbf{x}(t)$ if f_0 is differentiable, and a subgradient otherwise, $[\cdot]_{\mathcal{X}}$ denotes the projection onto the feasible set $\mathcal{X}$, and $\alpha(t)$ is a positive scalar stepsize. It is interesting to point out that each iteration of the subgradient method may not improve the objective value as happens with a gradient method. What makes the subgradient method work is that for sufficiently small stepsize $\alpha(t)$, the distance of the current solution $\mathbf{x}(t)$ to the optimal solution $\mathbf{x}^\star$ decreases [33].

There are many results on convergence of the gradient/subgradient method with different choices of stepsize [33], [34], [36]. For example, for a diminishing stepsize rule $\alpha(t) = (1 + m)/(t + m)$, where m is a fixed nonnegative number, the algorithm is guaranteed to converge to the optimal value (assuming bounded gradients/subgradients) [36]. For a constant stepsize $\alpha(t) = \alpha$, more convenient for distributed algorithms, the gradient algorithm converges to the optimal value provided that the stepsize is sufficiently small (assuming that the gradient is Lipschitz) [33], whereas for the subgradient algorithm the best value converges to within some neighborhood of the optimal value (assuming bounded subgradients) [36].

E. Review: Standard Dual-Based Algorithm for Basic NUM

Before concluding this section on a systematic framework of alternative decompositions, we briefly review the standard way [3] to solve the basic NUM problem [1] for distributed end-to-end rate allocation. It illustrates a simple application of the one-level, full dual decomposition. In the rest of this paper, we will see a number of more sophisticated NUM formulations motivated by new application contexts and a much richer array of decomposition alternatives, beyond the well-known problem and solution method in this subsection.

Consider a communication network with L links, each with a fixed capacity of c_l, and S sources or nodes, each transmitting at a source rate of x_s. Each source s emits one flow, using a fixed set of links $L(s)$ in its path, and has a utility function

$U_s(x_s)$. Basic NUM is the problem of maximizing the total utility $\sum_s U_s(x_s)$, over the source rates $\mathbf{x}$, subject to linear flow constraints $\sum_{s:l\in L(s)} x_s \leq c_l$ for all links l (note that $\{s : l \in L(s)\}$ denotes the set of users that use link l)

$$\begin{aligned} \underset{\mathbf{x}\geq\mathbf{0}}{\text{maximize}} \quad & \sum_s U_s(x_s) \\ \text{subject to} \quad & \sum_{s:l\in L(s)} x_s \leq c_l \quad \forall l \end{aligned} \qquad (23)$$

where the utilities U_s are twice-differentiable, increasing, and strictly concave functions.

The standard distributed solution to (23) is based on dual decomposition. We first form the Lagrangian of (23)

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= \sum_s U_s(x_s) + \sum_l \lambda_l \left(c_l - \sum_{s:l\in L(s)} x_s \right) \\ &= \sum_s \left[U_s(x_s) - \left(\sum_{l\in L(s)} \lambda_l \right) x_s \right] + \sum_l c_l \lambda_l \\ &= \sum_s L_s(x_s, \lambda^s) + \sum_l c_l \lambda_l \end{aligned} \qquad (24)$$

where $\lambda_l \geq 0$ is the Lagrange multiplier (link price) associated with the linear flow constraint on link l, $\lambda^s = \sum_{l\in L(s)} \lambda_l$ is the aggregate path congestion price of those links used by source s, and $L_s(x_s, \lambda^s) = U_s(x_s) - \lambda^s x_s$ is the sth Lagrangian to be maximized by the sth source.

Dual decomposition results in each source s, for the given λ^s, solving

$$x_s^\star(\lambda^s) = \arg\max_{x_s\geq 0} [U_s(x_s) - \lambda^s x_s] \quad \forall s \qquad (25)$$

which is unique due to the strict concavity of U_s. The master dual problem is

$$\begin{aligned} \underset{\boldsymbol{\lambda}}{\text{minimize}} \quad & g(\boldsymbol{\lambda}) = \sum_s g_s(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \mathbf{c} \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0} \end{aligned} \qquad (26)$$

where $g_s(\boldsymbol{\lambda}) = L_s(x_s^\star(\lambda^s), \lambda^s)$. Since the solution in (25) is unique, it follows that the dual function $g(\boldsymbol{\lambda})$ is differentiable and the following gradient method can be used

$$\lambda_l(t+1) = \left[\lambda_l(t) - \alpha \left(c_l - \sum_{s:l\in L(s)} x_s^\star(\lambda^s(t)) \right) \right]^+ \quad \forall l \qquad (27)$$

where t is the iteration index, $\alpha > 0$ is a positive stepsize, and $[\cdot]^+$ denotes the projection onto the nonnegative orthant.

The dual variable $\boldsymbol{\lambda}(t)$ will converge to the dual optimal $\boldsymbol{\lambda}^\star$ as $t \to \infty$ and, since the duality gap for (23) is zero and the solution to (25) is unique, the primal variable $\mathbf{x}^\star(\boldsymbol{\lambda}(t))$ will also converge to the primal optimal variable $\mathbf{x}^\star$ (cf. [2], [6]).

Summarizing, the basic NUM problem in (23) can be distributively solved with the gradient update in (27) carried out independently by each the link and the maximization in (25) solved independently by each source. Notice that there is no need for explicit message passing since λ^s can be measured by each source s as queuing delay and $\sum_{s:l\in L(s)} x_s$ can be measured by each link l as the total traffic load.

In practice, feedback of dual variables suffers from errors. The case of deterministic feedback error has been treated, e.g., in [8] and [37], where the key methodologies used include ensuring that feedback errors are sufficiently small compared to the true values, and constructing contraction regions. The case of stochastic feedback error has also been treated, e.g., in [38], where the key methodology used is the theory of stochastic approximation.

III. APPLICATION 1: POWER-CONSTRAINED RATE ALLOCATION

We start the applications sections with the simplest and recently studied extension of the basic NUM, before moving on to more involved formulations and novel solutions in Sections IV, V, and VI.

A. Problem Formulation

In some applications, such as wireless broadcast or DSL broadband access, distributed rate allocation can be carried out over transmission "pipes" of different sizes, with the help of adaptive resource allocation in the physical layer. This is an example of balancing "supply" of resources and "demand" of link capacities "built" from the limited resources.

Consider now the basic NUM in (23) but with variable link capacities $\{c_l(p_l)\}$, each of which depends on the allocated resource p_l, such as transmit power, with a constraint on the maximum total resource P_T associated with downlink transmissions. For many models such as TDMA or FDMA, c_l is a strictly concave function.[3] Consider the following NUM formulation:

$$\begin{aligned} \underset{\mathbf{x},\mathbf{p}\geq\mathbf{0}}{\text{maximize}} \quad & \sum_s U_s(x_s) \\ \text{subject to} \quad & \sum_{s:l\in L(s)} x_s \leq c_l(p_l) \quad \forall l \\ & \sum_l p_l \leq P_T. \end{aligned} \qquad (28)$$

Although only slightly more sophisticated than the basic NUM, this problem already contains sufficient elements such that one can try different decompositions. We will consider two decompositions: a primal decomposition with respect to the power allocation, and a dual decomposition with respect to the flow constraints.

B. Primal–Dual Decomposition

Consider first a primal decomposition of (28) by fixing the power allocation $\mathbf{p}$. Clearly, the link capacities become fixed numbers and problem (28) becomes a basic NUM like (23), which can be solved via a dual decomposition as explained in Section II-E. The master primal problem is

$$\begin{aligned} \underset{\mathbf{p}\geq\mathbf{0}}{\text{maximize}} \quad & U^\star(\mathbf{p}) \\ \text{subject to} \quad & \sum_l p_l \leq P_T \end{aligned} \qquad (29)$$

where $U^\star(\mathbf{p})$ is the optimal objective value of (28) for a given $\mathbf{p}$. Since a subgradient of $U^\star(\mathbf{p})$ with respect to c_l is given

[3]A related and different model has been recently studied in [8]. The primal–dual solution in Section III-B was first proposed in [19], and that in Section III-C was first proposed in [39].

by the Lagrange multiplier λ_l associated with the constraint $\sum_{s:l\in L(s)} x_s \leq c_l$ in (28), it follows that a subgradient of $U^\star(\mathbf{p})$ with respect to p_l is given by $\lambda_l c_l'(p_l)$. Therefore, the master primal problem (29) can be solved with a subgradient method by updating the powers as

$$\mathbf{p}(t+1) = \left[\mathbf{p}(t) + \alpha \begin{bmatrix} \lambda_1^\star(\mathbf{p}(t)) c_1'(p_1(t)) \\ \vdots \\ \lambda_L^\star(\mathbf{p}(t)) c_L'(p_L(t)) \end{bmatrix} \right]_{\mathcal{P}} \quad (30)$$

where $[\cdot]_{\mathcal{P}}$ denotes the projection onto the feasible convex set $\mathcal{P} \triangleq \{\mathbf{p} : \mathbf{p} \geq \mathbf{0}, \sum_l p_l \leq P_T\}$, which is a simplex. Due to the projection, this subgradient update cannot be performed independently by each link and requires some centralized approach. The projection of a point $\mathbf{p}_0$ (the expression inside the outer bracket in (30)) onto the simplex $\mathcal{P}$, i.e., $\mathbf{p} = [\mathbf{p}_0]_{\mathcal{P}}$, can be easily obtained in the following waterfilling form [17]:

$$p_l = \left(p_l^0 - \gamma\right)^+ \quad \forall l \quad (31)$$

where the waterlevel γ is chosen as the minimum nonnegative value such that $\sum_l p_l \leq P_T$. Observe that only the computation of γ requires a central node since the update of each power p_l can be done at each link.

C. Dual–Dual Decomposition

Consider now a dual decomposition of (28) by relaxing the flow constraints $\sum_{s:l\in L(s)} x_s \leq c_l(p_l)$

$$\begin{aligned} \underset{\mathbf{x},\mathbf{p}\geq 0}{\text{maximize}} \quad & \sum_s \left[U_s(x_s) - \left(\sum_{l\in L(s)} \lambda_l\right) x_s \right] + \sum_l c_l(p_l)\lambda_l \\ \text{subject to} \quad & \sum_l p_l \leq P_T. \end{aligned} \quad (32)$$

This problem decomposes into one maximization for each source, as (25) in the basic NUM, plus the following additional maximization to update the power allocation:

$$\begin{aligned} \underset{\mathbf{p}\geq 0}{\text{maximize}} \quad & \sum_l \lambda_l c_l(p_l) \\ \text{subject to} \quad & \sum_l p_l \leq P_T \end{aligned} \quad (33)$$

which can be further decomposed via a second-level dual decomposition, yielding the following subproblems:

$$\underset{p_l \geq 0}{\text{maximize}} \quad \lambda_l c_l(p_l) - \gamma p_l \quad (34)$$

with solution given by

$$p_l = \left(c_l'\right)^{-1} (\gamma/\lambda_l) \quad (35)$$

and a secondary master dual problem that updates the dual variable γ as

$$\gamma(t+1) = \left[\gamma(t) - \alpha\left(P_T - \sum_l p_l^\star(\gamma(t))\right)\right]^+. \quad (36)$$

The master dual problem is updated as in the basic NUM (27).

D. Summary

We have obtained two different distributed algorithms for power-constrained rate allocation in (28):

- Primal–dual decomposition: The master problem (29) is solved with the subgradient power update in (30) carried out by the links with a small central coordination (due to the projection on the simplex) and then, for a given set of powers, the resulting basic NUM is solved via the standard dual-based decomposition in (25) and (27). This implies two levels of decompositions: on the highest level there is a master primal problem, on a second level there is a secondary master dual problem, and on the lowest level the subproblems.

- Dual–dual decomposition: The master dual problem is solved with the standard price update in (27), which is carried out independently by each link, and then, for a given set of prices, each source solves its own subproblem as in (25), and subproblem (33) is solved with some central node updating the price with (36) and each link obtaining the optimal power with (35). This approach contains two levels of decompositions: on the highest level there is a master dual problem, on a second level there are rate subproblems and a secondary master dual problem, and on the lowest level the power subproblems.

In both approaches, the only explicit signaling required is the power-price γ from the central unit to the links and possibly the powers from the links back to the central node.

E. Special Case: Cellular Downlink Power/Rate Control

An interesting special case of the signal model in (28) arises in cellular downlink power/rate control with the flow constraints on each downlink connection modeled in the high SNR regime (where $\log(1 + g_s p_s) \simeq \log(g_s p_s)$) of a CDMA system with orthogonal codes

$$\begin{aligned} \underset{\mathbf{x},\mathbf{p}\geq 0}{\text{maximize}} \quad & \sum_s U_s(x_s) \\ \text{subject to} \quad & x_s \leq \log(g_s p_s) \quad \forall s \\ & \sum_s p_s \leq P_T \end{aligned} \quad (37)$$

where g_s is the channel gain of the sth user. This problem can be solved in many different combinations of multilevel primal–dual decompositions, each with a different signaling scheme and convergence speed. An empirical comparison of the convergence of several methods can be found in Section VII-A.

IV. APPLICATION 2: QOS RATE ALLOCATION

A. Problem Formulation

Sometimes a rate allocation mechanism needs to differentiate users in different QoS classes. For example, the total link capacity received by each QoS class must lie within a range prescribed in the service level agreement. Such constraints introduce new coupling to the basic NUM problem and lead to alternative decomposition possibilities. We will see in this section

two different distributed algorithms to solve this type of QoS rate allocation problem, both with a differential pricing interpretation for the new set of Lagrange multipliers introduced. Therefore, these algorithms provide an intuitive pricing alternative to the recent proposals of NUM-based rate allocation among different QoS classes in [40] and [41]. They also complement the results in [42] and [43] where the authors considered welfare maximization for loss networks, with deterministic optimization problems presented from the limiting regime of stochastic network optimization, and characterized the structures of asymptotically optimal static-pricing-based coordination.

Consider now the basic NUM in (23), but with different classes of users that will be treated differently. The idea behing having several classes of users is, for example, to impose limits on the maximum rate and to guarantee a minimum rate for each class. To simplify the exposition, we consider only two classes of users, and the results extend straightforwardly to more classes of users. Denoting by $y_l^{(1)}$ and $y_l^{(2)}$ the aggregate rates of classes 1 and 2, respectively, along the lth link, the NUM formulation is as follows:

$$
\begin{aligned}
\underset{\mathbf{x},\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} \quad & \sum_s U_s(x_s) \\
\text{subject to} \quad & \sum_{s\in S_i:l\in L(s)} x_s \leq y_l^{(i)} \quad \forall l, i=1,2 \\
& \mathbf{y}^{(1)} + \mathbf{y}^{(2)} \leq \mathbf{c} \\
& \mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)} \quad i=1,2.
\end{aligned} \tag{38}
$$

Observe that in the absence of the constraints $\mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)}$, problem (38) becomes the basic NUM in (23). We will consider two decompositions: a primal decomposition with respect to the aggregate rate of each class, and a dual decomposition with respect to the total aggregate rate constraints from both classes.

B. Primal–Dual Decomposition

Consider first a primal decomposition of (38) by fixing the aggregate rates $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$. Problem (38) becomes two independent subproblems, for $i=1,2$, identical to the basic NUM in (23)

$$
\begin{aligned}
\underset{\mathbf{x}\geq\mathbf{0}}{\text{maximize}} \quad & \sum_{s\in S_i} U_s(x_s) \\
\text{subject to} \quad & \sum_{s\in S_i:l\in L(s)} x_s \leq y_l^{(i)} \quad \forall l
\end{aligned} \tag{39}
$$

where the fixed aggregate rates $y_l^{(i)}$ play the role of the fixed link capacities in the basic NUM of (23). These two independent basic NUMs can be solved as explained in Section II-E.

The master primal problem is

$$
\begin{aligned}
\underset{\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} \quad & U_1^\star\left(\mathbf{y}^{(1)}\right) + U_2^\star\left(\mathbf{y}^{(2)}\right) \\
\text{subject to} \quad & \mathbf{y}^{(1)} + \mathbf{y}^{(2)} \leq \mathbf{c} \\
& \mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)} \quad i=1,2
\end{aligned} \tag{40}
$$

where $U_i^\star(\mathbf{y}^{(i)})$ is the optimal objective value of (39) for a given $\mathbf{y}^{(i)}$, with a subgradient given by the Lagrange multiplier $\boldsymbol{\lambda}^{(i)}$ associated to the constraints $\sum_{s\in S_i:l\in L(s)} x_s \leq y_l^{(i)}$ in (39).

Observe that $\boldsymbol{\lambda}^{(i)}$ is the set of differential prices for the QoS class i. The master primal problem (40) can now be solved with a subgradient method by updating the aggregate rates as

$$
\begin{bmatrix} \mathbf{y}^{(1)}(t+1) \\ \mathbf{y}^{(2)}(t+1) \end{bmatrix} = \left[\begin{bmatrix} \mathbf{y}^{(1)}(t) \\ \mathbf{y}^{(2)}(t) \end{bmatrix} + \alpha \begin{bmatrix} \boldsymbol{\lambda}^{\star(1)}\left(\mathbf{y}^{(1)}(t)\right) \\ \boldsymbol{\lambda}^{\star(2)}\left(\mathbf{y}^{(2)}(t)\right) \end{bmatrix} \right]_{\mathcal{Y}} \tag{41}
$$

where $[\cdot]_{\mathcal{Y}}$ denotes the projection onto the feasible convex set $\mathcal{Y} \triangleq \{(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}) : \mathbf{y}^{(1)} + \mathbf{y}^{(2)} \leq \mathbf{c}, \mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)}, i=1,2\}$. Nicely enough, this feasible set enjoys the property that it already naturally decomposes into a Cartesian product for each of the links: $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_L$. Therefore, this subgradient update can be performed independently by each link simply with the knowledge of its corresponding Lagrange multipliers $\lambda_l^{(1)}$ and $\lambda_l^{(2)}$, which, in turn, are also updated independently by each link as in Section II-E.

C. Partial Dual Decomposition

Consider now a dual decomposition of (38) by relaxing the flow constraints $\sum_{s\in S_i:l\in L(s)} x_s \leq y_l^{(i)}$

$$
\begin{aligned}
\underset{\mathbf{x},\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} \quad & \sum_{s\in S_1}\left[U_s(x_s) - \left(\sum_{l\in L(s)}\lambda_l\right)x_s \right] \\
& + \sum_{s\in S_2}\left[U_s(x_s) - \left(\sum_{l\in L(s)}\lambda_l\right)x_s \right] \\
& + \boldsymbol{\lambda}^{(1)T}\mathbf{y}^{(1)} + \boldsymbol{\lambda}^{(2)T}\mathbf{y}^{(2)} \\
\text{subject to} \quad & \mathbf{y}^{(1)} + \mathbf{y}^{(2)} \leq \mathbf{c} \\
& \mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)} \quad i=1,2.
\end{aligned} \tag{42}
$$

This problem decomposes into one maximization for each source, as (25) in the basic NUM, plus the following additional maximization to update the aggregate rates:

$$
\begin{aligned}
\underset{\mathbf{y}^{(1)},\mathbf{y}^{(2)}\geq\mathbf{0}}{\text{maximize}} \quad & \boldsymbol{\lambda}^{(1)T}\mathbf{y}^{(1)} + \boldsymbol{\lambda}^{(2)T}\mathbf{y}^{(2)} \\
\text{subject to} \quad & \mathbf{y}^{(1)} + \mathbf{y}^{(2)} \leq \mathbf{c} \\
& \mathbf{c}_{\min}^{(i)} \leq \mathbf{y}^{(i)} \leq \mathbf{c}_{\max}^{(i)} \quad i=1,2
\end{aligned} \tag{43}
$$

which can be solved independently by each link with knowledge of its corresponding Lagrange multipliers $\lambda_l^{(1)}$ and $\lambda_l^{(2)}$. These multipliers, in turn, are also updated independently by each link as in Section II-E.[4]

The master dual problem corresponding to this dual decomposition is updated with the following subgradient method, similarly to (27)

$$
\lambda_l^{(i)}(t+1) = \left[\lambda_l^{(i)}(t) - \alpha\left(y_l^{(i)}(t) - \sum_{s\in S_i:l\in L(s)} x_s^\star\left(\lambda^{(i)s}(t)\right) \right) \right]^+
$$
$$
\forall l, i=1,2. \tag{44}
$$

---

[4]Observe that the optimum solution of the linear problem in (43) may oscillate in practice as $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{\lambda}^{(2)}$ are updated by the master problem [the same applies to (47) and (55)]. To avoid this behavior, one can add a small quadratic term to the objective function; alternatively, one can introduce some memory in the solution by averaging the current solution of (43) with the values in previous iterations.

D. Summary

We have obtained two different distributed algorithms for rate allocation among QoS classes in (38):

- Primal–dual decomposition: the master problem (40) is solved with the subgradient update of the aggregate rate in (41) carried out independently by each of the links. Then, for a given set of aggregate rates, the two resulting basic NUMs are independently solved via the standard dual-based decomposition in (25) and (27). This implies two levels of decompositions: on the highest level there is a master primal problem, on a second level there is a secondary master dual problem, and on the lowest level the subproblems. There is no explicit signaling required.
- Partial dual decomposition: the master dual problem is solved with the standard price update for each class in (44) which is carried out independently by each link. Then, for a given set of prices, each source solves its own subproblem as in (25) and subproblem (43) is independently solved by each link. This approach contains only one level of decomposition and no explicit signaling is required.

Observe that in the primal–dual decomposition approach, each link updates the aggregate rates on a slower timescale and the prices on a faster timescale, whereas in the partial dual decomposition approach each link updates the prices on a slower timescale and the aggregate rates on a faster timescale (actually in one shot). Therefore, the speed of convergence of the partial dual approach should be faster in general. In both cases, the users are ignorant of the existence of classes and only the links have to take this into account by having one price for each class. In other words, this is a way to give each class of users a different price than the one based on the standard dual-based algorithm so that they can be differentiated into different QoS classes. The next application hinges on this observation.

V. APPLICATION 3: HYBRID RATE-BASED AND PRICE-BASED RATE ALLOCATION

A. Problem Formulation

One extreme way to control the rate allocation process is to directly give each source the rate they can use, at the expense of a centralized computation. At the other extremum, we can optimize the system in a fully distributed way via pricing feedback, as in the basic NUM of Section II-E, at the expense of trusting the sources even though they can be noncooperative and try to obtain more bandwidth by using a more aggressive utility function. Neither of these two extreme approaches is completely satisfactory in all applications, and hybrid solutions between rate-based and window-based rate allocation are desirable for both robustness of fair allocation against aggressive users and speed of converging to the correct rate allocation equilibrium.

New congestion control protocols using direct rate allocation have recently been proposed, such as XCP [44] and RCP [45], which are based on a heuristic computation of the processor-sharing type of rate allocation by each router that a flow traverses. We now describe a systematic method using primal decomposition to perform distributed and direct rate allocation

to each user. It turns out that direct rate control is a special case of alternative NUM decompositions. The key idea is to use the approach of Section IV but with one class for each user.

The NUM formulation becomes

$$
\begin{aligned}
\underset{\mathbf{x}, \{\mathbf{y}^{(s)}\} \geq \mathbf{0}}{\text{maximize}} \quad & \sum_s U_s(x_s) \\
\text{subject to} \quad & x_s \leq y_l^{(s)} \qquad \forall s, l \in L(s) \\
& \sum_s \mathbf{y}^{(s)} \leq \mathbf{c} \\
& \mathbf{c}_{\min}^{(s)} \leq \mathbf{y}^{(s)} \leq \mathbf{c}_{\max}^{(s)} \qquad \forall s. \qquad (45)
\end{aligned}
$$

Note that if a source s does not use a path l, then $y_l^{(s)}$ is taken as zero in the constraint $\sum_s \mathbf{y}^{(s)} \leq \mathbf{c}$.

B. Primal Decomposition

If we now take a primal decomposition approach, then the master primal problem will be in charge of the update of $y_l^{(s)}$ and each user will simply choose x_s equal to the minimum of the $y_l^{(s)}$ along its path in order to satisfy $x_s \leq y_l^{(s)}, \forall l \in L(s)$. This approach constitutes one of the extreme methods in which each user is directly given the amount of bandwidth it can use.

C. Partial Dual Decomposition

We may also take a dual decomposition approach by relaxing the flow constraints

$$
\begin{aligned}
\underset{\mathbf{x}, \{\mathbf{y}^{(s)}\} \geq \mathbf{0}}{\text{maximize}} \quad & \sum_s \left[U_s(x_s) - \left(\sum_{l \in L(s)} \lambda_l^{(s)} \right) x_s \right] \\
& + \sum_s \sum_{l \in L(s)} \lambda_l^{(s)} y_l^{(s)} \\
\text{subject to} \quad & \sum_s \mathbf{y}^{(s)} \leq \mathbf{c} \\
& \mathbf{c}_{\min}^{(s)} \leq \mathbf{y}^{(s)} \leq \mathbf{c}_{\max}^{(s)} \qquad \forall s. \qquad (46)
\end{aligned}
$$

This problem decomposes into one maximization for each source, as (25) in the basic NUM, with $\lambda^s = \sum_{l \in L(s)} \lambda_l^{(s)}$ being the aggregate path price specific for user s, plus the following additional rate-bounding maximization to obtain the $y_l^{(s)}$, for each link l:

$$
\begin{aligned}
\underset{\{y_l^{(s)}\} \geq 0}{\text{maximize}} \quad & \sum_{s: l \in L(s)} \lambda_l^{(s)} y_l^{(s)} \\
\text{subject to} \quad & \sum_{s: l \in L(s)} y_l^{(s)} \leq c_l \\
& c_{l,\min}^{(s)} \leq y_l^{(s)} \leq c_{l,\max}^{(s)} \qquad \forall s: l \in L(s). \quad (47)
\end{aligned}
$$

This problem can be solved independently by each link as a way to distribute its capacity c_l among the sources using the link according to the weights given by the prices $\lambda_l^{(s)}$, which are different for each source.

The master dual problem corresponding to this dual decomposition is solved with the following subgradient price update step, similarly to (27):

$$
\lambda_l^{(s)}(t+1) = \left[\lambda_l^{(s)}(t) - \alpha \left(y_l^{(s)}(t) - x_s^\star(\lambda^s(t)) \right) \right]^+
$$
$$
\forall l, s: l \in L(s). \quad (48)
$$

D. Summary

We have explored different decompositions for the hybrid rate/price-based rate allocation in (45):

- Primal decomposition: It leads to a direct rate allocation and is based on one level of decomposition. This approach requires the signaling to inform each user what rate to transmit at.
- Partial dual decomposition: The master dual problem is solved with the price update in (48), which is carried out independently by each link. Then, for a given set of prices, each source solves its own subproblem as in (25) and the bounding rates of subproblem (47) are also obtained independently by each link. This approach only shows one level of decomposition and does not require any explicit signaling. It is a hybrid of rate-bounding and pricing-feedback mechanisms.

VI. APPLICATION 4: MULTIPATH-ROUTING RATE ALLOCATION

A. Problem Formulation

Consider now a more general setup of the basic NUM of Section II-E where each source can choose among several possible paths (possibly using a weighted combination of them).[5] The structure of a network with S sources, L links, and J paths can be summarized with the $L \times J$ path availability $0 - 1$ matrix $\mathbf{H}$ defined by

$$[\mathbf{H}]_{l,j} = \begin{cases} 1, & \text{if the } j\text{th path uses the } l\text{th link} \\ 0, & \text{otherwise} \end{cases}$$

together with the $J \times S$ path choice nonnegative matrix $\mathbf{W}$ defined by[6]

$$[\mathbf{W}]_{j,s} = \begin{cases} w_{js}, & \text{if the } s\text{th source uses the } j\text{th path} \\ 0, & \text{otherwise} \end{cases}$$

where w_{js} indicates the percentage of the rate of the sth user allocated to the jth path and has to satisfy $w_{js} > 0$ and $\sum_j w_{js} = 1$. These two matrices can be combined into the routing matrix $\mathbf{R} = \mathbf{HW}$ that tells how much each source is using each link.

To start with, the problem can be directly formulated with the routing matrix $\mathbf{R}$ like the basic NUM in (23)

$$\begin{aligned} \underset{\mathbf{x} \geq \mathbf{0}}{\text{maximize}} \quad & \sum_s U_s(x_s) \\ \text{subject to} \quad & \mathbf{Rx} \leq \mathbf{c} \end{aligned} \qquad (49)$$

and then the standard dual-based decomposition algorithm can be used. We will later see that it may be more flexible to

---

[5]Recently, a similar formulation of multipath routing utility maximization is considered in [46], where a proximal optimization method followed by full dual decomposition is taken.

[6]This notation follows that in [47]. However, the problem being considered here is to design rate allocation algorithm with a fixed $\mathbf{H}$ and $\mathbf{W}$, whereas the problem considered in [47] is to analyze the effect of joint routing and rate allocation with $\mathbf{W}$ also being a variable.

formulate the problem alternatively in terms of $\mathbf{H}$ and $\mathbf{W}$ as follows:

$$\begin{aligned} \underset{\mathbf{x}, \mathbf{y} \geq \mathbf{0}}{\text{maximize}} \quad & \sum_s U_s(x_s) \\ \text{subject to} \quad & \mathbf{Wx} \leq \mathbf{y} \quad \text{(path constraint)} \\ & \mathbf{Hy} \leq \mathbf{c} \quad \text{(link constraint)} \end{aligned} \qquad (50)$$

where y_l contains the aggregate rate along the lth path.

B. Primal–Dual Decomposition

We can now consider a primal decomposition approach of (50) by fixing the path rates $\mathbf{y}$. Problem (50) then becomes a basic NUM where $\mathbf{y}$ plays the role of the link capacities in (23). This problem can be solved via the standard dual-based algorithm as reviewed in Section II-E.

The master primal problem is

$$\begin{aligned} \underset{\mathbf{y} \geq \mathbf{0}}{\text{maximize}} \quad & U^\star(\mathbf{y}) \\ \text{subject to} \quad & \mathbf{Hy} \leq \mathbf{c} \end{aligned} \qquad (51)$$

where $U^\star(\mathbf{y})$ is the optimal objective value of (50) for a given $\mathbf{y}$, with subgradient given by the Lagrange multiplier $\boldsymbol{\lambda}$ associated to the constraints $\mathbf{Wx} \leq \mathbf{y}$ in (50). As usual, the master primal problem (51) can be solved with a subgradient method by updating the path rates as follows:

$$\mathbf{y}(t+1) = \left[\mathbf{y}(t) + \alpha \boldsymbol{\lambda}^\star (\mathbf{y}(t)) \right]_{\mathcal{Y}} \qquad (52)$$

where $[\cdot]_{\mathcal{Y}}$ denotes the projection onto the feasible convex set $\mathcal{Y} \triangleq \{ \mathbf{y} : \mathbf{y} \geq \mathbf{0}, \mathbf{Hy} \leq \mathbf{c} \}$. In principle, this subgradient update cannot be performed independently by each path due to the projection onto $\mathcal{Y}$, which makes it impractical.

C. Partial Dual Decomposition

We can also take a partial dual decomposition of (50) by relaxing only the constraint $\mathbf{Wx} \leq \mathbf{y}$ (similar to [10])

$$\begin{aligned} \underset{\mathbf{x}, \mathbf{y} \geq \mathbf{0}}{\text{maximize}} \quad & \sum_s U_s(x_s) + \boldsymbol{\gamma}^T (\mathbf{y} - \mathbf{Wx}) \\ \text{subject to} \quad & \mathbf{Hy} \leq \mathbf{c}. \end{aligned} \qquad (53)$$

This problem decomposes into one maximization for the sources as in (25) for the basic NUM

$$\underset{\mathbf{x} \geq \mathbf{0}}{\text{maximize}} \quad \sum_s \left[U_s(x_s) - \gamma^s x_s \right] \qquad (54)$$

where $\gamma^s = \boldsymbol{\gamma}^T \mathbf{W}_{:,s} = \sum_{j \in J(s)} \gamma_j w_{js}$ is the aggregate price for the sth source, plus one maximization for the path rates

$$\begin{aligned} \underset{\mathbf{y} \geq \mathbf{0}}{\text{maximize}} \quad & \boldsymbol{\gamma}^T \mathbf{y} \\ \text{subject to} \quad & \mathbf{Hy} \leq \mathbf{c} \end{aligned} \qquad (55)$$

which has to be solved in a centralized way.

The master dual problem updates the prices as

$$\boldsymbol{\gamma}(t+1) = \left[\boldsymbol{\gamma}(t) - \alpha \left(\mathbf{y} - \mathbf{Wx} \left(\boldsymbol{\gamma}(t) \right) \right) \right]^+. \qquad (56)$$

D. Full Dual Decomposition

Yet another different way to solve problem (50) is with a full dual decomposition by relaxing both constraints $\mathbf{Wx} \leq \mathbf{y}$ and $\mathbf{Hy} \leq \mathbf{c}$

$$\underset{\mathbf{x} \geq 0, \mathbf{y}}{\text{maximize}} \quad \sum_s U_s(x_s) + \boldsymbol{\gamma}^T(\mathbf{y} - \mathbf{Wx}) + \boldsymbol{\lambda}^T(\mathbf{c} - \mathbf{Hy}) \quad (57)$$

which can be rewritten as

$$\underset{\mathbf{x} \geq 0, \mathbf{y}}{\text{maximize}} \quad \sum_s [U_s(x_s) - x_s \gamma^s] + \sum_j y_j(\gamma_j - \lambda^j) + \boldsymbol{\lambda}^T \mathbf{c} \quad (58)$$

where $\lambda^j = \boldsymbol{\lambda}^T \mathbf{H}_{:,j} = \sum_{l \in L(j)} \lambda_l$ is the aggregate price of the jth path and $\gamma^s = \boldsymbol{\gamma}^T \mathbf{W}_{:,s} = \sum_{j \in J(s)} \gamma_j w_{js}$ is the aggregate price for the sth source. This problem separates into maximization over $\mathbf{x}$, as in (25) for the basic NUM, and maximization over $\mathbf{y}$, which is unbounded unless $\gamma_j = \lambda^j$. Therefore, the optimal choice for the master dual problem is $\gamma_j = \lambda^j$ and then $\gamma^s = \sum_{j \in J(s)} \lambda^j w_{js} = \sum_{j \in J(s)} w_{js} \sum_{l \in L(j)} \lambda_l$. Hence, this approach reduces to the standard dual-based algorithm applied to problem (49).

Now consider a variant of this rate allocation problem with multipath-routing, where the objective of the Internet service provider (ISP) is combined with the end user utility objective. In today's operating environment of the Internet, the ISP controlling each autonomous system tries to minimize a total convex cost function of the link utilizations [48]. Suppose the cost function is quadratic, and the network utility maximization is now formulated as maximizing the weighted difference between end user utility and ISP cost

$$\sum_s U_s(x_s) - \theta \mathbf{y}^T \mathbf{y} \quad (59)$$

where θ is the weight. Observe that by taking θ sufficiently small the quadratic term becomes negligible and we are closely approximating the original problem (50).

Repeating the same full relaxation as before, one gets the following maximization problem:

$$\underset{\mathbf{x}, \mathbf{y} \geq 0}{\text{maximize}} \quad \sum_s [U_s(x_s) - x_s \gamma^s] + \sum_j y_j(\gamma_j - \lambda^j) - \theta \mathbf{y}^T \mathbf{y} + \boldsymbol{\lambda}^T \mathbf{c}. \quad (60)$$

This problem separates as before into maximization over $\mathbf{x}$, as in (25) for the basic NUM, and maximization over $\mathbf{y}$ with optimal solution given by

$$y_j = \frac{1}{2\theta}(\gamma_j - \lambda^j) \qquad \forall j. \quad (61)$$

Then, the master dual problem has to update two sets of prices

$$\boldsymbol{\lambda}(t+1) = [\boldsymbol{\lambda}(t) - \alpha(\mathbf{c} - \mathbf{Hy}(t))]^+ \quad (62)$$

$$\boldsymbol{\gamma}(t+1) = [\boldsymbol{\gamma}(t) - \alpha(\mathbf{y}(t) - \mathbf{Wx}^\star(\boldsymbol{\gamma}(t)))]^+ \quad (63)$$

where $\mathbf{y}(t) \triangleq \mathbf{y}^\star(\boldsymbol{\lambda}(t), \boldsymbol{\gamma}(t))$ is the optimal $\mathbf{y}$ for the given $\boldsymbol{\lambda}(t)$ and $\gamma(t)$ as in (61). Note that the matrix-vector products can be conveniently written as $[\mathbf{Hy}]_l = \sum_{j:l \in L(j)} y_j$ and $[\mathbf{Wx}]_j = \sum_{s:j \in J(s)} w_{js} x_s$.

E. Summary

We have explored several possibilities for distributed algorithms for rate allocation with multipath-routing possibilities in (50):

- Standard dual decomposition: By reformulating the problem as in (49), we recover the basic NUM formulation and the standard dual-based algorithm can be readily used.
- Primal–dual decomposition: The master primal problem (51) is solved with the path rate subgradient update in (52), and then, for a given set of path rates, the resulting basic NUMs is solved via the standard dual-based decomposition in (25) and (27). Unfortunately, due to the projection in (52), a centralized computation is required, which makes this approach impractical.
- Partial dual decomposition: The master dual problem is solved with the price update in (56), and then, for a given set of prices, each source solves its own subproblem as in (54). Subproblem (55) is solved in a centralized way, making this approach also inconvenient.
- Full dual decomposition: The master dual problem is solved with the price updates in (62), (63), and then, for a given set of prices, each source solves its own subproblem as in (54) and the path rates are obtained as in (61). This approach contains one level of decomposition: on the higher level the master dual problem and on the lower level the source-rate and path-rate subproblems. Explicit signaling is required for the update of the price $\boldsymbol{\gamma}(t)$ in (63), and for the computation of the path rate in (61).

VII. NUMERICAL EXAMPLES

A. Downlink Power/Rate Control

The purpose of this subsection is to illustrate the convergence behavior of different decomposition approaches, which can be quite different. Note, however, that in order to properly compare different methods one should also consider other aspects such as computational complexity and amount of signaling. Consider the downlink power/rate control formulated in (37) with logarithmic utilities $U_s(x_s) = \beta_s \log x_s$. Fig. 5 shows the evolution of the dual variable corresponding to one of the constraints $x_s \leq \log(g_s p_s)$ for seven methods based on various combinations of primal/dual multilevel decompositions solved by combinations of Gauss–Seidel and Jacobi iterations [12] and the subgradient method. The seven methods are based on the following combinations of decomposition/method (see [18] for more details):

1) Full dual decomposition on (37), obtaining the dual function $g(\boldsymbol{\lambda}, \gamma)$ (the λ_i's and γ are the dual variables corresponding to the constraints $x_s \leq \log(g_s p_s)$ and $\sum_s p_s \leq P_T$, respectively), plus a subgradient method to minimize $g(\boldsymbol{\lambda}, \gamma)$ by updating the dual variables $(\boldsymbol{\lambda}, \gamma)$.
2) Full dual decomposition on (37) as in 1), obtaining the dual function $g(\boldsymbol{\lambda}, \gamma)$, plus a Gauss–Seidel algorithm to minimize $g(\boldsymbol{\lambda}, \gamma)$ by updating alternatively the dual variables $\boldsymbol{\lambda}$ and γ: $\boldsymbol{\lambda} \to \gamma \to \boldsymbol{\lambda} \to \gamma \to \cdots$.
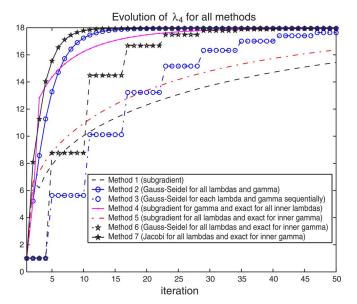
Fig. 5.   Evolution of λ_4 for the seven methods based on a dual decomposition.

3) Full dual decomposition on (37) as in 1) and 2), obtaining the dual function $g(\boldsymbol{\lambda}, \gamma)$, plus a Gauss–Seidel algorithm to minimize $g(\boldsymbol{\lambda}, \gamma)$ by updating alternatively each dual variable λ_i and γ: $\lambda_1 \to \gamma \to \lambda_2 \to \gamma \to \cdots$.

4) Dual-primal decomposition on (37), obtaining $g(\boldsymbol{\lambda}, \gamma)$ after the dual decomposition and then $g(\gamma) \triangleq \inf_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}, \gamma)$ after an additional primal decomposition by fixing the dual variable γ and minimizing only with respect to $\boldsymbol{\lambda}$, plus a subgradient method to minimize $g(\gamma)$ by updating γ.

5) Dual-primal decomposition on (37), obtaining $g(\boldsymbol{\lambda}, \gamma)$ after the dual decomposition and then $g(\boldsymbol{\lambda}) \triangleq \inf_{\gamma} g(\boldsymbol{\lambda}, \gamma)$ after an additional primal decomposition by fixing the dual variable $\boldsymbol{\lambda}$ and minimizing only with respect to γ, plus a subgradient method to minimize $g(\boldsymbol{\lambda})$ by updating $\boldsymbol{\lambda}$.

6) Dual-primal decomposition on (37) as in 5), obtaining $g(\boldsymbol{\lambda})$, plus a Gauss–Seidel algorithm to minimize $g(\boldsymbol{\lambda})$ by updating alternatively each λ_i : $\lambda_1 \to \lambda_2 \to \lambda_3 \to \cdots$.

7) Dual-primal decomposition on (37) as in 5) and 6), obtaining $g(\boldsymbol{\lambda})$, plus a Jacobi algorithm to minimize $g(\boldsymbol{\lambda})$ by updating simultaneously all the λ_i's.

B. QoS Rate Allocation

To illustrate the distributed algorithms for a rate allocation among QoS classes (as in Section IV), we consider a simple example consisting of four sources transmitting to the same destination and sharing a common link as shown in Fig. 6. Users in class 1 are aggressive, with utility functions $U_1(x) = 12 \log(x)$ and $U_2(x) = 10 \log(x)$, whereas users in class 2 are not aggressive, with utility functions $U_3(x) = 2 \log(x)$ and $U_4(x) = \log(x)$. If no QoS control is included in the design and the standard dual-based distributed algorithm of Section II-E is used, then the aggressive users of class 1 get most of the available capacity in the common link. In particular, class 1 gets a rate of 4.5 out of the total available rate of 5, leaving class 2 with a rate of only 0.5. This is precisely the kind of unfair behavior that can be avoided with QoS control.
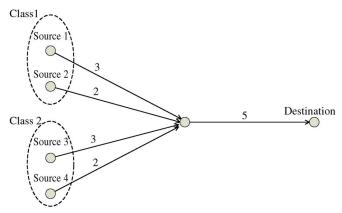

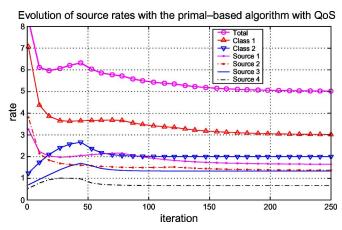Fig. 6.   Block diagram of the considered example of NUM with priorities.

Fig. 7.   Evolution of the rates with the primal-based algorithm for a NUM with QoS control.
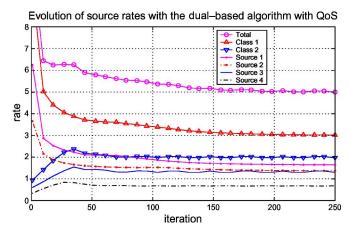

Fig. 8.   Evolution of the rates with the dual-based algorithm for a NUM with QoS control.

Figs. 7 and 8 show the evolution of the rates of the sources when QoS control is included in the distributed algorithms based on a primal decomposition and on a dual decomposition, respectively (as described in Section IV). In particular, the rate for each class has been limited to 3. As can be observed, the rate of class 1 now converges to the limit of 3 and, since the link capacity is 5, class 2 is left with a rate of 2 (as opposed to 0.5 obtained without QoS control). Hence, the distribution of the total rate between the two classes is more balanced.
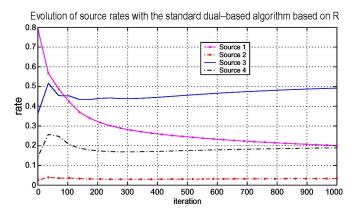
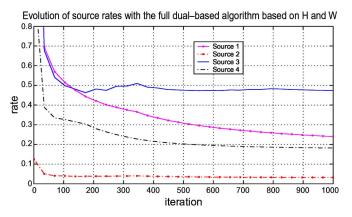Fig. 9. Evolution of the rates with the standard dual-based algorithm for a NUM based directly on the routing matrix $\mathbf{R}$.

Fig. 10. Evolution of the rates with the full-dual-based algorithm for a NUM based on the path-link $\mathbf{H}$ and link-source $\mathbf{W}$ matrices.

Both primal-based and dual-based algorithms show a similar convergence (a constant stepsize of 0.05 was used for all subgradient updates). Note that the primal-based algorithm contains two levels of subgradient updates and, in principle, the inner subgradient algorithm should run until convergence before updating the outer subgradient. In practice, however, this is not necessary and both subgradients can run simultaneously (in general, using a smaller stepsize for the outer subgradient so that it works on a slower timescale).

C. Multipath-Routing Rate Allocation

We now consider a NUM with different groupings of the path and link constraints as described in Section VI. In particular, we generate a random network topology with $S = 4$ sources, $J = 12$ paths, and $L = 36$ links, such that each user uses three paths and each path uses five links.

Fig. 9 shows the evolution of the rates of the sources for the standard dual-based algorithm based directly on the routing matrix $\mathbf{R} = \mathbf{HW}$. Fig. 10 shows the evolution of the rates of the sources with a full dual-based algorithm (including the quadratic term $-\theta \mathbf{y}^T \mathbf{y}$ with $\theta = 0.001$), which follows closely the performance of the standard algorithm. In practice, the optimal solution for the path rates in (61) leads to a large dynamic range that can lead to instability; this can be easily avoided

by providing the update with some memory (controlled by the weighting factor β)

$$y_j(t+1) = \beta \frac{1}{2\theta} \left(\gamma_j(t) - \lambda^j(t) \right) + (1 - \beta) y_j(t) \quad \forall j. \quad (64)$$

The other two methods described in Section VI, based on a primal decomposition and on a partial dual decomposition, provide similar convergence trajectories. However, their complexity and the need for centralized computation make them impractical [due to the projection in (52) and to the resolution of problem (55), respectively].

VIII. CONCLUSION AND FUTURE RESEARCH

Despite the recent focus in the literature on the standard dual-based method, there are more than one way to solve a network utility maximization in a distributed manner. A systematic framework is developed in this paper to explore alternative decompositions, and four specific rate allocation applications are presented. Implications of these results include designing faster distributed algorithms with the right distribution of computation and communication load across network elements, developing five new congestion control algorithms under various practical constraints, and understanding architectural tradeoffs in distributed network control.

An intellectually bold direction for future research is to explore if both the *enumeration* and *comparison* of alternative decompositions, horizontally and vertically, can be carried out systematically or even be automated.

To enumerate the set of possible decompositions, one has to take into account that transformations of the problem (e.g., change of variable) may lead to new decomposability structures. This would open the door to even more alternative decompositions, each of which has a different engineering implication to distributed and modularized network architecture [49].

To compare alternative decompositions, the following metrics all need to be considered: speed of convergence, the amount and symmetry of message passing for global communication, the distribution of local computational load, robustness to errors, failures, or network dynamics, the possibility of efficient relaxations and simple heuristics, and the ability to remain evolvable as the application needs change over time. Some of these metrics do not have any quantitative units of measurement, such as evolvability. Some do not have a universally agreed upon definition, such as the measure of *how* distributed an algorithm is. Some are difficult to analyze accurately, such as the rate of convergence. Application contexts lead to a prioritization of these possibly conflicting metrics, based on which, the "best" decomposition can be chosen from the range of alternatives.

APPENDIX
PROOF OF LEMMA 1

The concavity of $f^\star(\mathbf{x})$ follows from [11, Sec. 3.2.5]. By the definition of subgradient, we have for a given $\mathbf{y}$

$$f_0(\mathbf{x}, \mathbf{y}) \le f_0(\mathbf{x}_0, \mathbf{y}) + \mathbf{s}_{0, \mathbf{x}}^T(\mathbf{x}_0, \mathbf{y})(\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \quad (65)$$

$$f_i(\mathbf{x}, \mathbf{y}) \ge f_i(\mathbf{x}_0, \mathbf{y}) + \mathbf{s}_{i, \mathbf{x}}^T(\mathbf{x}_0, \mathbf{y})(\mathbf{x} - \mathbf{x}_0)$$

$$\forall \mathbf{x}, \ 1 \le i \le n. \quad (66)$$

Then, for a given $\mathbf{x}_0 \in \mathcal{F}$, we have that for any $(\mathbf{x}, \mathbf{y})$ satisfying $f_i(\mathbf{x}, \mathbf{y}) \leq 0, \forall i$ (see [33, Sec. 5.4.4] for a similar approach)

$$
\begin{aligned}
f^\star(\mathbf{x}_0) &= g\left(\boldsymbol{\lambda}^\star(\mathbf{x}_0)\right) \\
&\geq L\left((\mathbf{x}_0, \mathbf{y}), \boldsymbol{\lambda}(\mathbf{x}_0)\right) \\
&= f_0(\mathbf{x}_0, \mathbf{y}) - \sum_i \lambda_i^\star(\mathbf{x}_0) f_i(\mathbf{x}_0, \mathbf{y}) \\
&\geq f_0(\mathbf{x}, \mathbf{y}) - \mathbf{s}_{0,\mathbf{x}}^T(\mathbf{x}_0, \mathbf{y})(\mathbf{x} - \mathbf{x}_0) \\
&\quad - \sum_i \lambda_i^\star(\mathbf{x}_0) f_i(\mathbf{x}, \mathbf{y}) \\
&\quad + \sum_i \lambda_i^\star(\mathbf{x}_0) \mathbf{s}_{i,\mathbf{x}}^T(\mathbf{x}_0, \mathbf{y})(\mathbf{x} - \mathbf{x}_0) \\
&\geq f_0(\mathbf{x}, \mathbf{y}) - \mathbf{s}_{0,\mathbf{x}}^T(\mathbf{x}_0, \mathbf{y})(\mathbf{x} - \mathbf{x}_0) \\
&\quad + \boldsymbol{\lambda}^{\star T}(\mathbf{x}_0) \mathbf{S}_{\mathbf{x}}^\star(\mathbf{x}_0, \mathbf{y})(\mathbf{x} - \mathbf{x}_0) \tag{67}
\end{aligned}
$$

where $g(\boldsymbol{\lambda}(\mathbf{x}_0)) \triangleq \sup_{\mathbf{y}} L((\mathbf{x}_0, \mathbf{y}), \boldsymbol{\lambda}(\mathbf{x}_0))$ is the dual function and strong duality has been used in the first equality, $L((\mathbf{x}_0, \mathbf{y}), \boldsymbol{\lambda}(\mathbf{x}_0))$ is the Lagrangian, $\boldsymbol{\lambda}^\star(\mathbf{x}_0)$ is the optimal Lagrangian multiplier, the definition of subgradients has been used in the second inequality, and $\lambda_i^\star(\mathbf{y}_0) \geq 0$ and $f_i(\mathbf{x}, \mathbf{y}) \leq 0$ have been used in the last inequality. Now, simply particularize the previous inequality for $\mathbf{y}^\star(\mathbf{x})$, to write

$$
\begin{aligned}
f^\star(\mathbf{x}_0) \geq f^\star(\mathbf{x}) - \big(\mathbf{s}_{0,\mathbf{x}}(\mathbf{x}_0, \mathbf{y}^\star(\mathbf{x})) \\
- \mathbf{S}_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}^\star(\mathbf{x})) \boldsymbol{\lambda}^\star(\mathbf{x}_0)\big)^T (\mathbf{x} - \mathbf{x}_0), \quad \forall \mathbf{x}. \tag{68}
\end{aligned}
$$

Since the mapping $\mathbf{y}^\star(\mathbf{x})$ is continuous, for $\mathbf{x}$ sufficiently close to $\mathbf{x}_0$, $\mathbf{y}^\star(\mathbf{x})$ can be made arbitrarily close to $\mathbf{y}^\star(\mathbf{x}_0)$ and

$$
\begin{aligned}
f^\star(\mathbf{x}_0) \geq f^\star(\mathbf{x}) - \big(\mathbf{s}_{0,\mathbf{x}}(\mathbf{x}_0, \mathbf{y}^\star(\mathbf{x}_0)) \\
- \mathbf{S}_{\mathbf{x}}(\mathbf{x}_0, \mathbf{y}^\star(\mathbf{x}_0)) \boldsymbol{\lambda}^\star(\mathbf{x}_0)\big)^T (\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \approx \mathbf{x}_0. \tag{69}
\end{aligned}
$$

However, since $f^\star(\mathbf{x})$ is a concave function and the previous inequality holds for a nonempty set around $\mathbf{x}_0$, it must also hold for all $\mathbf{x}$ (even those not in $\mathcal{F}$ for which $f^\star(\mathbf{x}) = -\infty$) and the subgradient is then readily identified. ∎

ACKNOWLEDGMENT

REFERENCES

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

[2] R. Srikant, *The Mathematics of Internet Congestion Control*. New York: Birkhauser, 2004.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–213, Jan. 2007.

[10] L. Chen, S. H. Low, and J. C. Doyle, "Joint congestion control and media access control design for ad hoc wireless networks," presented at the IEEE INFOCOM, Miami, FL, Mar. 13–27, 2005.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[12] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[13] J. L. R. Ford and D. R. Fulkerson, *Flow in Networks*. Princeton, NJ: Princeton Univ. Press, 1962.

[14] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[15] L. S. Lasdon, *Optimization Theory for Large Systems*. New York: Macmillian, 1970.

[16] D. P. Palomar, M. Bengtsson, and B. Ottersten, "Minimum BER linear transceivers for MIMO channels via primal decomposition," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2866–2882, Aug. 2005.

[17] D. P. Palomar, "Convex primal decomposition for multicarrier linear MIMO transceivers," *IEEE Trans. Signal Process.*, vol. 53, no. 12, Dec. 2005.

[18] D. P. Palomar and M. Chiang, "Choices of distributed algorithms for wireless network resource allocation," presented at the IEEE Globecom, St. Louis, MO, Nov. 2005.

[19] B. Johansson and M. Johansson, "Primal and dual approaches to distributed cross-layer optimization," presented at the 16th IFAC World Congr., Prague, Czech Republic, 2005.

[20] M. Andrews, "Joint optimization of scheduling and congestion control in communication networks," presented at the CISS, 2006.

[21] L. Chen, S. H. Low, M. Chiang, and J. C. Doyle, "Jointly optimal congestion control, routing, and scheduling for wireless ad hoc networks," presented at the IEEE INFOCOM, 2006.

[22] A. Eryilmaz and R. Srikant, "Joint congestion control, routing and MAC for stability and fairness in wireless networks," *IEEE J. Sel. Areas Comm.*, vol. 24, no. 8, Aug. 2006.

[23] X. Lin and N. Shroff, "The impact of imperfect scheduling on crosslayer rate control in multihop wireless networks," presented at the IEEE INFOCOM, 2005.

[24] A. L. Stolyar, "Maximizing queuing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Syst.*, vol. 50, no. 4, pp. 401–457, 2005.

[25] X. Wang and K. Kar, "Cross-layer rate control for end-to-end proportional fairness in wireless networks with random access," presented at the ACM Mobihoc, 2005.

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

[27] D. S. Lun, N. Ratnakar, M. Medard, R. Koetter, D. R. Karger, T. Ho, and E. Ahmed, "Minimum-cost multicast over coded packet networks," *IEEE Trans. Inf. Theory*, 2005.

[28] Y. Wu and S. Y. Kung, "Distributed utility maximization for network coding based multicasting: A shortest path approach," *IEEE J. Sel. Areas. Commun.*, Aug. 2006.

[29] Y. Wu, M. Chiang, and S. Y. Kung, "Distributed utility maximization for network coding based multicasting: A critical cut approach," presented at the IEEE NetCod Workshop, Apr. 2006.

[30] S. Bhadra, S. Shakkottai, and P. Gupta, "Min-cost selfish multicast with network coding," presented at the IEEE NetCod Workshop, Apr. 2006.

[31] L. Chen, T. Ho, S. H. Low, M. Chiang, and J. C. Doyle, "Optimization based rate control for multicasting with network coding," presented at the IEEE INFOCOM, Anchorage, AK, May 2007.

[32] Y. Xi and E. M. Yeh, "Distributed algorithms for minimum cost multicast with network coding," presented at the IEEE NetCod Workshop, Apr. 2006.

[33] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[34] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*. Berlin, Germany: Springer-Verlag, 1985.

[35] C. W. Tan, D. P. Palomar, and M. Chiang, "Distributed optimization of coupled systems with applications to network utility maximization," in *Proc. IEEE Int. Conf. Acoustics Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 14–19, 2006, vol. V, pp. 981–984.

[36] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA: Athena Scientific, 2003.

[37] M. Mehyar, D. Spanos, and S. H. Low, "Optimization flow control with estimation error," presented at the IEEE INFOCOM, Hong Kong, China, Mar. 2004.

[38] J. Zhang, D. Zheng, and M. Chiang, "The impact of stochastic noisy feedback on distributed network utility maximization," presented at the IEEE INFOCOM, Anchorage, AK, May 2007.

[39] 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.

[40] K. Chandrayana and S. Kalyanaraman, "Uncooperative congestion control," presented at the ACM SIGMETRICS, New York, Jun. 2004.

[41] M. Chiang, S. Zhang, and P. Hande, "Distributed rate allocation for inelastic flows: Optimization frameworksoptimaltiy condition, and optimal algorithms," presented at the IEEE INFOCOM, Miami, FL, Mar. 13–17, 2005.

[42] I. C. Paschalidis and J. N. Tsitsiklis, "Congestion-dependent pricing of networks services," *IEEE/ACM Trans. Netw.*, vol. 8, no. 2, pp. 171–184, Apr. 2000.

[43] I. C. Paschalidis and Y. Liu, "Pricing in multiservice loss networks: Static pricing, asymptotic optimality, and demand substitution effects," *IEEE/ACM Trans. Netw.*, vol. 10, no. 3, pp. 425–438, Jun. 2002.

[44] M. H. D. Katabi and C. Rohrs, "Internet congestion control for high bandwidth-delay product networks," presented at the ACM Sigcomm, Aug. 2002.

[45] N. Dukkipati, M. Kobayashi, Z. S. Rui, and N. McKeown, "Processor sharing flows in the internet," presented at the 13th Int. Workshop Quality of Service (IWQoS), Passau, Germany, Jun. 2005.

[46] X. Lin and N. B. Shroff, "Utility maximization for communication networks with multipath routing," *IEEE Trans. Autom. Control*, vol. 51, no. 5, pp. 766–781, May 2006.

[47] J. Wang, L. Li, S. H. Low, and J. C. Doyle, "Cross-layer optimization in TCP/IP networks," *IEEE/ACM Trans. Netw.*, vol. 13, pp. 582–595, Jun. 2005.

[48] J. Rexford, "Route optimization in IP networks," in *Handbook of Optimization in Telecommunications*. Norwell, MA: Kluwer, 2005.

[49] D. Palomar and M. Chiang, "A tutorial on decomposition method and distributed network resource allocation," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

Daniel P. Palomar (S'99–M'03) received the Electrical Engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

He is an Assistant Professor in the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong. He has held several visiting research appointments, namely, at King's College London (KCL), London, U.K., during 1998; the Technical University of Catalonia (UPC), Barcelona, from January 1999 to December 2003; Stanford University, Stanford, CA, from April to November 2001; Telecommunications Technological Center of Catalonia (CTTC), Barcelona, from January to December 2002; Royal Institute of Technology (KTH), Stockholm, Sweden, from August to November 2003; University of Rome "La Sapienza," Rome, Italy, from November 2003 to February 2004; Princeton University, Princeton, NJ, from March 2004 to July 2006.

Dr. Palomar is an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) 2008 Special Issue on "Game Theory in Communication Systems," and the Lead Guest Editor of the 2007 Special Issue on "Optimization of MIMO Transceivers for Realistic Communication Networks." He received a 2004/2006 Fulbright Research Fellowship; the 2004 Young Author Best Paper Award from the IEEE Signal Processing Society; the 2006 Best Student Paper Award at ICASSP 2006 (co-recipient); the 2002/2003 best Ph.D. prize in Information Technologies and Communications from the Technical University of Catalonia (UPC); the 2002/2003 Rosina Ribalta first prize for the Best Doctoral Thesis in Information Technologies and Communications by the Epson Foundation; and the 2004 prize for the best Doctoral Thesis in Advanced Mobile Communications by the Vodafone Foundation and COIT.

Mung Chiang (S'00–M'03) received the B.S. (Hon.) degree in electrical engineering and in 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 and an affiliated faculty of Applied and Computational Mathematics and of Computer Science at Princeton University, Princeton, NJ. He conducts research in the areas of nonlinear optimization of communication systems, theoretical foundation of network architectures, algorithms for broadband access networks, and stochastic analysis of communications and networking.

Dr. Chiang received the CAREER Award from the National Science Foundation, the Young Investigator Award from the Office of Naval Research, the Howard B. Wentz Junior Faculty Award from Princeton University, the School of Engineering Terman Award from Stanford University, the New Technology Introduction Award from SBC Communications, and he was a Hertz Foundation Fellow and Stanford Graduate Fellow. For his work on broadband access networks and Internet traffic engineering, he was selected for the TR35 Young Technologist Award in 2007, a list of the top 35 innovators in the world under the age of 35. His work on geometric programming was selected by the Mathematical Programming Society as one of the top three papers by authors under the age of 35 in the area of continuous optimization during 2004–2007. His work on layering as optimization decomposition became a Fast Breaking Paper in Computer Science by ISI citation. He also coauthored papers that were IEEE Infocom best paper finalists and IEEE Globecom best student papers. He has served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, a Lead Guest Editor for the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, a Guest Editor for IEEE/ACM TRANSACTIONS ON NETWORKING and the IEEE TRANSACTIONS ON INFORMATION THEORY, a Program Co-Chair of the 38th Conference on Information Sciences and Systems, and a coeditor of the new Springer book series on "Optimization and Control of Communication Systems."