Simultaneous Routing and Resource Allocation via Dual Decomposition

Lin Xiao, *Student Member, IEEE,* Mikael Johansson, *Member, IEEE,* and Stephen P. Boyd, *Fellow, IEEE*

***Abstract*— In wireless data networks the optimal routing of data depends on the link capacities which, in turn, are determined by the allocation of communications resources (such as transmit powers and bandwidths) to the links. The optimal performance of the network can only be achieved by simultaneous optimization of routing and resource allocation.**

In this paper, we formulate the simultaneous routing and resource allocation problem and exploit problem structure to derive efficient solution methods. We use a capacitated multi-commodity flow model to describe the data flows in the network. We assume that the capacity of a wireless link is a concave and increasing function of the communications resources allocated to the link, and the communications resources for groups of links are limited. These assumptions allow us to formulate the simultaneous routing and resource allocation problem as a convex optimization problem over the network flow variables and the communications variables. These two sets of variables are coupled only through the link capacity constraints. We exploit this separable structure by dual decomposition. The resulting solution method attains the optimal coordination of data routing in the network layer and resource allocation in the radio control layer via pricing on the link capacities.

***Index Terms*— networks, routing, communication systems, resource allocation, optimization methods.**

I. Introduction

As the demand for wireless services increases, efficient use of radio resources grows in importance. One way of improving the resource utilization in wireless data networks is to move from optimizing each networking layer in isolation to optimally coordinating the operation across the networking stack. In this paper we develop a method for joint optimization of the routing in the network layer and the resource allocation in the radio control (physical) layer.

Traditionally, routing problems for data networks have often been formulated as convex multicommodity network flow problems (*e.g.*, [1]) for which many efficient solution methods exist, *e.g.*, [2], [3], [4], [5], [6]. The optimal routing of data flows depends on the link capacities, which are usually assumed fixed. In wireless data networks, however, link capacities are not necessarily fixed but can be adjusted by the allocation of communications resources, such as transmit powers, bandwidths, or time-slot fractions, to different links. Adjusting the resource allocation changes the link capacities, influences the optimal routing of data flows, and alters the total utility of the network. Hence, the routing problem in the network layer and resource allocation problem in the radio control layer are coupled through the link capacities, and the overall optimal performance of the network can only be achieved by simultaneous optimization of routing and resource allocation.

Both optimal routing and optimal resource allocation problems have been studied in isolation: routing in data networks has a long tradition, *e.g.*, [1], [5], [6]; while optimal resource allocation problems for wireless systems have been considered more recently, *e.g.*, [7], [8], [9]. Joint optimization of routing and capacity assignment has been studied in the context of design and provisioning of computer communication networks (see, e.g., [10], [11], [12]). In this case the capacities take one of several discrete values (corresponding to, say, the number of transmission lines between two routers) and the routing is often restricted to paths, which leads to nonlinear integer programs. Related is also the joint routing and link scheduling problem studied in, *e.g.*, [13], [14]. However, these approaches do not account for the nontrivial relationship between resource allocation and the resulting capacities of the wireless links. A systematic approach for joint design across the two networking layers is needed.

In this paper, we study the simultaneous routing and resource allocation (SRRA) problem for wireless data networks within a convex optimization framework, and exploit the problem structure via dual decomposition. The resulting solution method can be interpreted as a pricing mechanism on the link capacities, which attains the optimal coordination of data routing in the network layer and resource allocation in the radio control layer. Because of our convex formulation of the SRRA problem and associated strong duality results, the dual decomposition method obtains the global optimal solution. This is in contrast to the nonlinear integer program formulation and the application of similar methods (Lagrange relaxation) in obtaining suboptimal solutions for the joint routing and capacity assignment problems in computer communication networks (see, *e.g.*, [10], [11], [12]).

This paper is organized as follows. Section II describes the

Paper approved by G. S. Kuo, the Editor for Communications Architecture of the IEEE Communications Society. Manuscript received September 19, 2002, revised June 2, 2003 and January 8, 2004. This research was sponsored in part by NSF grant No. 0140700, by AFOSR grant F49620-01-1-0365, by DARPA contracts F33615-99-C-3014 and MDA972-02-1-0004. Parts of this paper were presented at The 39th Annual Allerton Conference on Communication, Control, and Computing, Monticello, Illinois, October 2001, and at The 4th Asian Control Conference, Singapore, September 2002.

Lin Xiao is with the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305 (e-mail: lxiao@stanford.edu).

Mikael Johansson is with the Department of Signals, Sensors and Systems, Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden (e-mail: mikaelj@s3.kth.se).

Stephen P. Boyd is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 (e-mail: boyd@stanford.edu).

network topology and the multicommodity flow model for the data network. In section III, we present our model of the communication system that supports the data network, and illustrate how many classical channel models and capacity formulas fit our framework. Based on our models for the two networking layers, we formulate the SRRA problem as a convex optimization problem in section IV and demonstrate the benefits of joint optimization on a numerical example in section V. Finally in section VI, we describe how to exploit the structure of the SRRA problem via dual decomposition, and we solve the dual of the SRRA problem using a subgradient method. Section VII concludes the paper and points to several possible extensions.

II. Network flow model

We use the standard directed graph model to represent the network topology and a multicommodity flow model for the average behavior of data transmissions across the network.

A. Network topology

We represent the topology of a data network by a directed graph (we assume that the graph is always connected). In this model, a collection of nodes, labeled $n = 1, \ldots, N$, can send, receive and relay data across communication links. A communication link is represented as an ordered pair (i, j) of distinct nodes. The presence of a link (i, j) means that the network is able to send data from the start node i to the end node j. We label the links with integers $l = 1, \ldots, L$. The network topology can be represented by a *node-link incidence matrix* $A \in \mathbf{R}^{N \times L}$, whose entry A_{nl} is associated with node n and link l via

$$A_{nl} = \begin{cases} 1, & \text{if } n \text{ is the start node of link } l \\ -1, & \text{if } n \text{ is the end node of link } l \\ 0, & \text{otherwise.} \end{cases}$$

We define $\mathcal{O}(n)$ as the set of links that are outgoing from node n, and $\mathcal{I}(n)$ as the set of links that are incoming to node n.

B. Multicommodity network flows

We use a multicommodity flow model for the routing of data packets across the network. Such models are widely used in the literature of network routing and optimization (see, *e.g.*, [6], [1], [2]). In this model, each node can send (different) data to many destinations and receive data from many sources, but multicast is not considered. We assume that the data flows are lossless across links and that they satisfy flow conservation laws at each node.

We identify the flows by their destinations, *i.e.*, flows with the same destination are considered as one single commodity regardless of their sources. We assume that the destination nodes are labeled $d = 1, \ldots, D$, where $D \leq N$. For each destination d, we define a *source-sink vector* $s^{(d)} \in \mathbf{R}^N$, whose nth ($n \neq d$) entry $s_n^{(d)}$ denotes the non-negative amount of flow (data rate in bits/sec) injected into the network at node n (the source) and destined for node d (the sink). In light of the flow conservation law, the sink flow at the destination is given by $s_d^{(d)} = -\sum_{n, n \neq d} s_n^{(d)}$, where the summation is over all nodes except the destination.

On each link l, we let $x_l^{(d)} \geq 0$ be the amount of flow destined for node d. We call $x^{(d)} \in \mathbf{R}^L$ the *flow vector* for destination d. At each node n, components of the flow vector and the source-sink vector with the same destination satisfy the flow conservation law:

$$\sum_{l \in \mathcal{O}(n)} x_l^{(d)} - \sum_{l \in \mathcal{I}(n)} x_l^{(d)} = s_n^{(d)}, \quad d = 1, \ldots, D.$$

The flow conservation law across the whole network can be compactly written as

$$A x^{(d)} = s^{(d)}, \quad d = 1, \ldots, D \tag{1}$$

where A is the node-link incidence matrix defined in section II-A.

Finally, we impose capacity constraints on the individual links. Let c_l be the capacity of link l and $t_l = \sum_d x_l^{(d)}$ be the total amount of traffic on link l. We then require that $t_l \leq c_l$.

In summary, our network flow model imposes the following group of constraints on the network flow variables $x^{(d)}$, $s^{(d)}$ and t:

$$\begin{array}{lll} A x^{(d)} = s^{(d)}, & & d = 1, \ldots, D \\ x^{(d)} \succeq 0, & s^{(d)} \succeq_d 0, & d = 1, \ldots, D \\ t_l = \sum_d x_l^{(d)}, & & l = 1, \ldots, L \\ t_l \leq c_l, & & l = 1, \ldots, L \end{array} \tag{2}$$

where $\succeq$ means component-wise inequality, and $\succeq_d$ means component-wise inequality except for the dth component (the sink flow $s_d^{(d)}$ is always negative). We will use x to denote the collection of flow vectors $x^{(d)}$ and use s to denote the collection of source vectors $s^{(d)}$.

This model describes the average behavior of data transmissions, *i.e.*, the average data rates on the communication links, and ignores packet-level details of transmission protocols and forwarding mechanisms. The link capacity in practical communication systems should be defined appropriately, taking into account packet loss and retransmission, so the flow conservation law holds for the effective throughput or goodput (see, *e.g.*, [15]).

C. Multicommodity flow problems with fixed link capacities

In traditional multicommodity network flow problems, the capacities c_l are usually assumed fixed and one is to minimize some convex function of the network flow variables subject to the set of constraints (2). For example, one of the most common cost functions used in the communication network literature is the total delay function (see, *e.g.*, [16], [1])

$$f_{\text{delay}}(t) = \sum_l \frac{t_l}{c_l - t_l}, \tag{3}$$

which is a convex function of t. In the minimum delay routing problem, the source vectors s (*i.e.*, the load to be supported by the network) are given and one is to minimize $f_{\text{delay}}(t)$ by selecting the optimal flow variables x and t subject to the constraints (2).

There is a vast literature on convex multicommodity network flow problems and many efficient solution methods have been developed; see, *e.g.*, [2], [3], [4], [5] and references therein. In this paper, however, we are interested in the interplay between resource allocation, link capacities, and optimal routing present in wireless data networks. The dependence of link capacities on communications resources will be described next.

III. Communications model and assumptions

In this section, we derive a model of the wireless communication system that supports the data traffic. In a wireless system, the capacities of individual links depend on the media access scheme and the selection of certain critical parameters, such as transmit powers, bandwidths or time-slot fractions, allocated to individual links or groups of links. We refer to these critical communications parameters collectively as *communications variables*, and denote the vector of communications variables by r. We assume that the medium access methods and coding and modulation schemes of the communication system are fixed, but that we can optimize over the communications variables r. The communications variables are themselves limited by various resource constraints, such as limits on the total transmit power at each node or the total signal bandwidth available across the whole network.

A. A generic model for communications resource constraints

Let r_l be a vector of communications variables associated with link l. In general, the capacity c_l depends not only on r_l, but also on communications resources allocated to other links in the network (due to interferences). However, in this paper we will focus on the case where the link capacity is only a function of the local resource allocation r_l, *i.e.*, $c_l = \phi_l(r_l)$. For example, communication systems with frequency-division and time-division multiple access fit this model (see section III-B). We will use the following generic model to relate the vector of total traffic t and the vector of communications variables r:

$$\begin{aligned} & t_l \leq c_l = \phi_l(r_l), \quad l = 1, \ldots, L \\ & Fr \preceq g, \quad r \succeq 0 \end{aligned} \tag{4}$$

We make the following assumptions about this generic model:

- The functions ϕ_l are concave and monotone increasing in r_l. The concavity of ϕ_l implies that the first set of constraints are jointly convex in t and r. The monotonicity condition means that the link capacities increase with increasing resources.
- The second set of constraints are in the form of linear inequalities. They describe resource limits, such as the total available transmit power at each node, and/or the total bandwidth for group of links (examples will be given in section III-B). They also specify that the communications variables are non-negative.

This generic communications model and the network flow model in section II will allow us to formulate the SRRA problem as a convex optimization problem in section IV.

B. Examples of communications resource constraints

Capacity formulas of many important communication channel models satisfy the concavity and monotonicity assumptions of the generic model (see, *e.g.*, [17], [7], [8]). Here we will only illustrate how the Gaussian broadcast channels with FDMA (frequency-division multiple access) and TDMA (time-division multiple access) fit into this framework.

1) Gaussian broadcast channel with FDMA: In the Gaussian broadcast channel using FDMA, the transmitters at node n send information to receivers at the end nodes of its outgoing links. The outgoing links $l \in \mathcal{O}(n)$ are assigned disjoint frequency bands with bandwidths $W_l \geq 0$ and powers $P_l \geq 0$. The receivers at the end of the links are subject to independent additive white Gaussian noises with power spectral densities σ_l. The classical Shannon capacity formula relates the capacity c_l and the communications variables $r_l = (P_l, W_l)$ by

$$c_l = \phi_l(P_l, W_l) = W_l \log_2 \left(1 + \frac{P_l}{\sigma_l W_l}\right), \quad l \in \mathcal{O}(n). \tag{5}$$

It can be easily verified that ϕ_l is concave and monotone increasing in the variables (P_l, W_l). Hence, formula (5) is in the generic form of the first set of constraints in (4).

The communications variables are constrained by total resource limits:

$$\sum_{l \in \mathcal{O}(n)} P_l \leq P_{\text{tot}}^{(n)}, \qquad \sum_{l \in \mathcal{O}(n)} W_l \leq W_{\text{tot}}^{(n)}.$$

If we denote the vector of all communications variables by $r = (P_1, \ldots, P_L, W_1, \ldots, W_L)^T$, then these resource limits can be expressed in the generic form $Fr \preceq g$ in (4) by letting

$$\begin{aligned} F &= \begin{bmatrix} A_+ & 0 \\ 0 & A_+ \end{bmatrix} \\ g &= \left(P_{\text{tot}}^{(1)}, \ldots, P_{\text{tot}}^{(n)}, W_{\text{tot}}^{(1)}, \ldots, W_{\text{tot}}^{(n)}\right)^T \end{aligned}$$

where the matrix A_+ has the same size as the incidence matrix A, and its elements are given by $(A_+)_{nl} = \max\{0, A_{nl}\}$, which only identify the outgoing links at each node.

2) Gaussian broadcast channel with TDMA: In the TDMA case, each link is assigned a time-slot fraction τ_l, and the average capacity of each link is a linear (hence concave) function of τ_l:

$$c_l = \phi_l(\tau_l) = \tau_l W_{\text{tot}}^{(n)} \log_2 \left(1 + \frac{P_{\text{tot}}^{(n)}}{\sigma_l W_{\text{tot}}^{(n)}}\right), \quad l \in \mathcal{O}(n).$$

Here the communications variables are the time-slot fractions τ_l, and they satisfy

$$\begin{aligned} \sum_{l \in \mathcal{O}(n)} \tau_l &\leq 1, \quad n = 1, \ldots, N, \\ \tau_l &\geq 0, \quad l = 1, \ldots, L. \end{aligned}$$

In terms of the generic form (4), we have $r = (\tau_1, \ldots, \tau_L)^T$, $F = A_+$ and $g = \mathbf{1}$, where $\mathbf{1}$ is the vector of all ones (here its dimension is N, the number of nodes in the network).

C. Communications resource allocation problem

Many resource allocation problems in wireless systems can be written in the form of maximizing a weighted sum of communication rates (or capacities) of multiple users. For example, varying the weights and solving the associated resource allocation problem allows us to trace the capacity region of multi-user communication systems (see, *e.g.*, [17], [7], [8]). We can formulate the following resource allocation problem based on the generic model (4):

$$\begin{array}{ll} \text{maximize} & \sum_l w_l c_l = \sum_l w_l \phi_l(r_l) \\ \text{subject to} & Fr \preceq g, \quad r \succeq 0 \end{array} \tag{6}$$

where w_l are non-negative scalar weights. This includes, for example, the problem of allocating both powers and bandwidths in the FDMA model from section III-B.1 to maximize the total communication rate. Since the functions ϕ_l are assumed to be concave, this is a convex optimization problem and the globally optimal solution can be found using a variety of methods. In addition, many specialized algorithms have been developed for problem (6) that exploits its structure. For example, if there is only one total resource limit, then it can be solved by the classical water-filling algorithm (see, *e.g.*, [17, §10.4], [18, page 245]). Actually, water-filling is the one-dimensional version of the more general dual decomposition method which will be described in section VI.

IV. The SRRA problem

A model for the wireless data network can be obtained by combining the network flow model and the communications model described in the previous two sections. This model reflects how the link capacities depend on the allocation of communications resources and how the overall optimal performance of the network can only be achieved by simultaneous optimization of routing and resource allocation. In this section, we formulate the SRRA problem as a convex optimization problem and describe some useful examples.

A. A generic convex optimization formulation

Consider the operation of a wireless data network described by the network flow model (2) and the communications model (4), and suppose that the objective is to minimize a convex cost function $f(x, s, t, r)$ (or maximize a concave utility function). We have the following generic formulation of the SRRA problem:

$$\begin{array}{lll} \text{minimize} & f(x,s,t,r) & \\ \text{subject to} & Ax^{(d)} = s^{(d)}, & d = 1, \ldots, D \\ & x^{(d)} \succeq 0, \quad s^{(d)} \succeq_d 0, & d = 1, \ldots, D \\ & t_l = \sum_d x_l^{(d)}, & l = 1, \ldots, L \\ & t_l \le \phi_l(r_l), & l = 1, \ldots, L \\ & Fr \preceq g, \quad r \succeq 0. & \end{array} \tag{7}$$

Here the optimization variables are the network flow variables x, s, t and the communications variables r. Since the constraints in (7) define a convex set and the objective function is convex, the SRRA problem is a convex optimization problem. This implies that it can be solved globally and efficiently by recently developed interior-point methods (see, *e.g.*, [19], [18]). Moreover, in the above model, the matrices A and F are sparse and highly structured, which can be exploited to develop far more efficient algorithms.

B. Examples of SRRA problem

The SRRA problem is very general and includes many important design problems for wireless data networks. We conclude this section by describing three of these in some detail.

1) Minimum power SRRA: Given a set of (fixed) source-sink vectors $s^{(d)}$ to be supported by the network, it is natural to try to find the joint routing and resource allocation that minimizes the total transmit power used by the network. This problem is readily formulated as

$$\begin{array}{ll} \text{minimize} & w^T r \\ \text{subject to} & \text{constraints in (7)} \end{array}$$

where

$$w_i = \begin{cases} 1, & \text{if } r_i \text{ is a power variable} \\ 0, & \text{otherwise} \end{cases}$$

Many variations, such as minimizing the maximum power used by any node in the network, or minimizing a weighted sum (hereby accounting for the relative costs of draining the different power sources) can be handled similarly. Another useful problem, which can be treated analogously, is to minimize the total bandwidth needed to support the desired traffic.

2) Minimax link utilization SRRA: It may be desirable to generalize the minimum delay routing problem mentioned in section II-C within the SRRA framework. However, the total delay function f_{delay} in (3) is not jointly convex in t and r when c_l is substituted by the capacity function $\phi_l(r_l)$. Another cost function with similar qualitative properties is the maximum link utilization [1]

$$f_{\text{maxu}}(t,r) = \max_l \frac{t_l}{\phi_l(r_l)}.$$

This function is quasi-convex. To see this, first notice that the functions $f_l(t_l, r_l) = t_l/\phi_l(r_l)$ are quasiconvex because the sub-level sets

$$\begin{array}{rl} & \{(t_l, r_l) \mid f_l(t_l, r_l) \le \alpha, \ t_l \ge 0, \ r_l \ge 0\} \\ = & \{(t_l, r_l) \mid t_l \le \alpha \phi_l(r_l), \ t_l \ge 0, \ r_l \ge 0\} \end{array}$$

are convex; the function $f_{\text{maxu}}(t, r)$ is the non-negative weighted maximum of the quasiconvex functions $f_l(t_l, r_l)$ and hence is quasiconvex (see [18, Chapter 3]).

We can formulate the minimax link utilization SRRA problem as

$$\begin{array}{ll} \text{minimize} & \max_l \dfrac{t_l}{\phi_l(r_l)} \\ \text{subject to} & \text{constraints in (7)} \end{array}$$

where the source-sink vectors $s^{(d)}$ are fixed. This quasiconvex optimization problem can be solved efficiently through a sequence of convex feasibility problems (see, *e.g.*, [18, §4.2]).

3) Maximum utility SRRA: Let $U_n^{(d)}(\cdot)$ be a concave and strictly increasing utility function, and let $U_n^{(d)}(s_n^{(d)})$ $(n \neq d)$ represent the utility of node n for sending data at rate $s_n^{(d)}$ to destination d. Then the maximum utility SRRA problem can be formulated as

$$\begin{array}{ll} \text{maximize} & \sum_{d} \sum_{n, n\neq d} U_n^{(d)}(s_n^{(d)}) \\ \text{subject to} & \text{constraints in (7).} \end{array} \qquad (8)$$

We will give an numerical example of this problem in section V.

It is worthwhile to point out that our SRRA formulation allows a possible extension of the work on optimization based congestion control in computer networks (see, *e.g.*, [20], [21], [22]) to joint flow and power control in wireless networks. In this cases, it is natural to keep the routes between the source-destination pairs fixed and only optimize over source rates on the different routes and resource allocation on the communication links.

We label all the routes by integers $i = 1, \ldots, R$, and let s_i be the data rate sent through route i. In place of the node-link incidence matrix A, we use the *link-route incidence matrix* $B \in \mathbf{R}^{L\times R}$, whose entries B_{li} is defined as

$$B_{li} = \begin{cases} 1 & \text{if route } i \text{ pass through link } l \\ 0 & \text{otherwise.} \end{cases}$$

With this definition, the total traffic vector on the links is given by $t = Bs$. The joint rate and resource allocation problem can be formulated as

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{R} U_i(s_i) \\ \text{subject to} & Bs = t, \quad s \succeq 0 \\ & t_l \leq \phi_l(r_l), \quad l = 1, \ldots, L \\ & Fr \preceq g, \quad r \succeq 0. \end{array} \qquad (9)$$

where U_i is a concave and strictly increasing utility function of the source rate s_i. Similar to the approaches in [20], [21], the dual decomposition method in section VI can be applied to problem (9) to develop distributed joint flow and power control algorithms.

V. A NUMERICAL EXAMPLE

To demonstrate the benefits of simultaneous routing and resource allocation, we consider the wireless network shown in figure 1. The network is randomly generated by drawing node positions from a uniform distribution on the unit square $[0, 1] \times [0, 1]$ and allowing two nodes to communicate if their distance is smaller than the threshold 0.25. The network has $N = 50$ nodes and $L = 340$ links (the 170 double-directed links shown in figure 1). We randomly choose 5 source and destination nodes, labeled $S_1, \ldots, S_5$ in figure 1.

In this example, we assume that the bandwidth allocation is fixed (each link is assigned unit bandwidth) and that there is no interference among links (using FDMA). We are free to adjust the transmit powers P_l allocated to each link, but we impose a total power constraint for the outgoing links of each node

$$\sum_{l\in\mathcal{O}(n)} P_l \leq P_{\text{tot}}^{(n)} = 100, \quad n = 1, \ldots, N.$$

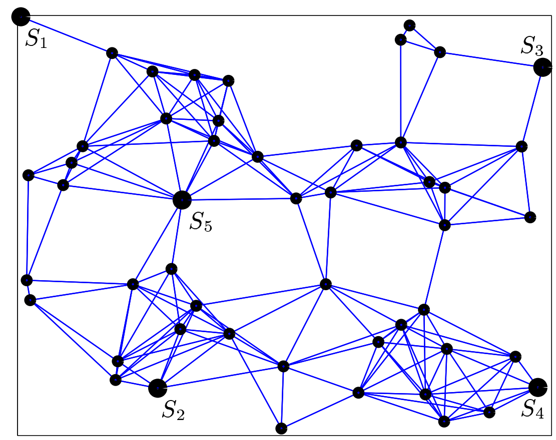


Fig. 1. Topology of a randomly generated wireless network with 50 nodes and 170 bi-directional links.

Let y_l be the distance between the two end nodes of link l. We use a inverse-square path-loss model: the power at the receiver is given by $(y_0/y_l)^2 P_l$, where $y_0 = \min_l y_l$ is a reference distance. The additive Gaussian noise powers σ_l at the receivers are generated randomly, with a uniform distribution on the interval $[0.01, 0.1]$. We use the link capacity formula (*cf.* equation (5) with unit bandwidth)

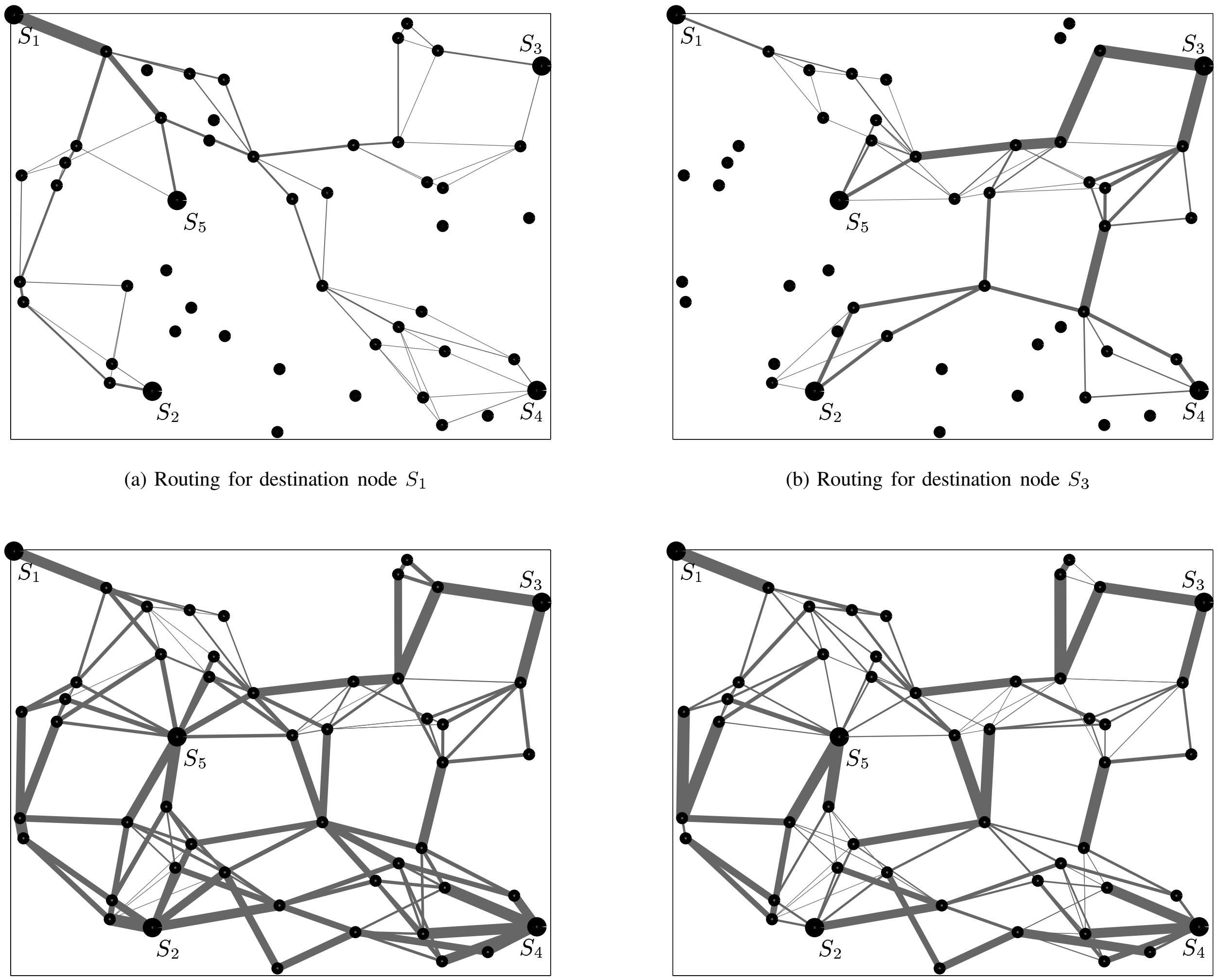
$$\phi_l(P_l) = \log\left(1 + \left(\frac{y_0}{y_l}\right)^2 \frac{P_l}{\sigma_l}\right).$$

We consider the problem of joint optimization of routing and power allocation to maximize the total utility of the network, where all source-destination pairs (chosen from the five nodes $S_1, \ldots, S_5$) have the logarithmic utility function

$$U_n^{(d)}(s_n^{(d)}) = \log s_n^{(d)}, \quad n \neq d, \quad n, d \in \{S_1, \ldots, S_5\}.$$

This maximum-utility SRRA problem has total 2060 variables, of which 340 are power variables and 1720 are network flow variables (there are 5 destinations, each with 340 flow routing variables and 4 source variables).

While this problem can be solved by general interior-point methods, we solved it by implementing the dual decomposition method that will be described in section VI, which exploits the layered structure of the wireless network. Figures 2(a) and 2(b) show the optimal data routing for the destinations S_1 and S_3 respectively (others are omitted since they have similar patterns), and figure 2(c) shows the aggregate flow, *i.e.*, the total traffic on all links. Figure 2(d) shows the optimal power allocation over the links across the network. In all these figures, the thickness of the link drawn is roughly proportional to the associated flow amount or power allocation. Table I shows the source and sink flows which achieve the maximum total utility 17.27. The diagonals are the negative total flows (sinks) at the five diestinations. To compare with the SRRA approach, we also solved a maximum-utility routing problem with uniform power allocation where all the nodes distribute their total powers evenly to their outgoing links. The result of routing under uniform power allocation is shown in table II, with the maximum total utility 12.77. We see that the SRRA formulation gives a 35% improvement of performance.



(a) Routing for destination node S_1

(b) Routing for destination node S_3

(c) Aggregate flow for all destinations

(d) Optimal power allocation over the links

Fig. 2. Optimal routing and power allocation solutions. The thickness of the link drawn is proportional to the associated flow amount or power allocation.

TABLE I
OPTIMAL SOURCE-SINK FLOW SOLUTIONS $s_n^{(d)}$ FOR THE SRRA PROBLEM; THE TOTAL UTILITY IS 17.27.

n	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
1	-3.88	1.11	0.92	1.12	1.13
2	1.03	-16.05	2.93	6.98	6.97
3	0.84	2.69	-9.43	2.69	2.77
4	0.96	4.80	2.46	-18.23	4.80
5	1.05	7.45	3.12	7.44	-15.67

TABLE II
OPTIMAL SOURCE-SINK FLOW SOLUTIONS $s_n^{(d)}$ WITH UNIFORM POWER ALLOCATION; THE TOTAL UTILITY IS 12.77.

n	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
1	-2.26	1.03	0.88	1.01	1.37
2	0.56	-13.95	1.73	9.59	5.92
3	0.54	2.07	-6.61	1.97	4.14
4	0.54	6.70	1.55	-16.34	4.20
5	0.62	4.15	2.45	3.77	-15.63

VI. THE DUAL DECOMPOSITION METHOD

We now turn our attention to the development of efficient solution methods for the SRRA problem (7). Our approach is based on exploiting problem structure via the dual decomposition method (see, *e.g.*, [23, §6]).

A. Formulation of the dual SRRA problem

Although the dual decomposition method applies to the generic SRRA formulation (7), we will illustrate it on the maximum-utility problem (8), which is rewritten here:

$$
\begin{array}{lll}
\text{maximize} & \sum_d \sum_{n, n\neq d} U_n^{(d)}(s_n^{(d)}) & \\
\text{subject to} & A^{(d)}x^{(d)} = s^{(d)}, & d=1,\ldots,D \\
& x^{(d)} \succeq 0, \quad s^{(d)} \succeq_d 0, & d=1,\ldots,D \\
& t_l = \sum_d x_l^{(d)}, & l=1,\ldots,L \\
& t_l \leq \phi_l(r_l), & l=1,\ldots,L \\
& Fr \preceq g, \quad r \succeq 0. &
\end{array}
\qquad (10)
$$

Due to the rich structure of this problem, there are many ways to formulate the dual problem depending on for which constraints the Lagrange multipliers are introduced. We will focus on the layered structure: the network flow variables x, s, t and the communications variables r (that appear in different layers of the network) are only coupled through the capacity constraints $t_l \leq \phi_l(r_l)$.

We form the dual problem by introducing Lagrange multipliers $p \in \mathbf{R}^L$ only for the L coupling constraints $t_l \leq \phi_l(r_l)$. This results in the *partial Lagrangian*

$$\begin{aligned} L(x,s,t,r,p) &= \sum_d \sum_{n, n\neq d} U_n^{(d)}(s_n^{(d)}) - \sum_l p_l(t_l - \phi_l(r_l)) \\ &= \left(\sum_d \sum_{n, n\neq d} U_n^{(d)}(s_n^{(d)}) - \sum_l p_l t_l \right) \\ &\quad + \sum_l p_l \phi_l(r_l). \end{aligned}$$

The dual function, *i.e.*, the objective function of the dual problem, is defined as

$$V(p) = \sup_{x,s,t,r} \left\{ L(x,s,t,r,p) \;\middle|\; \begin{array}{l} Ax^{(d)} = s^{(d)},\ x^{(d)} \succeq 0 \\ s^{(d)} \succeq_d 0,\ d = 1, \ldots, D \\ t_l = \sum_d x_l^{(d)},\ l = 1, \ldots, L \\ Fr \preceq g,\ r \succeq 0 \end{array} \right\} \tag{11}$$

One immediate observation is that the dual function can be evaluated separately in the network flow variables x, s, t and the communications variables r, *i.e.*,

$$V(p) = V_{\mathrm{net}}(p) + V_{\mathrm{comm}}(p),$$

where

$$\begin{aligned} V_{\mathrm{net}}(p) &= \sup_{x,s,t} \left\{ \sum_d \sum_{n, n\neq d} U_n^{(d)}(s_n^{(d)}) - \sum_l p_l t_l \right. \\ &\qquad \left. \;\middle|\; \begin{array}{l} Ax^{(d)} = s^{(d)},\ x^{(d)} \succeq 0, \\ s^{(d)} \succeq_d 0,\ d = 1, \ldots, D \\ t_l = \sum_d x_l^{(d)},\ l = 1, \ldots, L \end{array} \right\} \quad (12) \\ V_{\mathrm{comm}}(p) &= \sup_r \left\{ \sum_l p_l \phi_l(r_l) \;\middle|\; Fr \preceq g,\ r \succeq 0 \right\}. \quad (13) \end{aligned}$$

Moreover, as we will see shortly, $V_{\mathrm{net}}(p)$ and $V_{\mathrm{comm}}(p)$ can be evaluated very efficiently.

The Lagrange dual problem associated with the primal problem (10) is given by

$$\begin{array}{ll} \text{minimize} & V(p) = V_{\mathrm{net}}(p) + V_{\mathrm{comm}}(p) \\ \text{subject to} & p \succeq 0. \end{array} \tag{14}$$

Since the dual function V is always convex (see, *e.g.*, [18], [23]), this is a convex optimization problem. We assume that Slater's condition for constraint qualification (see, *e.g.*, [18, §5.2] and [23, §3.3]) is satisfied for the SRRA problem, *i.e.*, there exist a feasible solution x, s, t, r such that the capacity constraints (the only nonlinear constraints) hold with strict inequality:

$$t_l < \phi_l(r_l), \qquad l = 1, \ldots, L.$$

(This is almost always true in practice.) With this assumption, we conclude that strong duality holds, *i.e.*, the optimal values of the dual problem (14) and the primal problem (10) are equal (see, *e.g.*, [18], [23]). This allows us to solve the primal via the dual.

Because the objective function in (10) is not strictly concave in the variables x and t, the dual function is usually only piecewise differentiable. Hence, the dual problem (14) is a non-differentiable convex optimization problem. Effective methods to solve non-differentiable convex problems include the subgradient method and cutting-plane methods, which we will describe in some detail in section VI-C.

As another consequence of the non-strict concavity of the primal objective function, extra care must be taken to recover optimal primal solutions in the dual decomposition method (see, *e.g.*, [23, §6]). One simple effective approach is to add a strictly concave regularization term to the primal objective function. For example, we added a small quadratic term of x in solving the numerical example of section V. More sophisticated approaches include augmented Lagrangian methods and proximal bundle methods (see, *e.g.*, [24], [25], [23]).

B. Evaluating the dual function and its subgradient

To solve the dual problem (14), we need to be able to evaluate the dual function $V(p)$ and compute its subgradient for any given dual variable $p \succeq 0$.

To evaluate the dual function $V(p)$, we compute $V_{\mathrm{net}}(p)$ and $V_{\mathrm{comm}}(p)$ separately and add them together. By their definitions (12) and (13), we can find $V_{\mathrm{net}}(p)$ by solving the problem

$$\begin{array}{ll} \text{maximize} & \sum_d \sum_{n, n\neq d} U_n^{(d)}(s_n^{(d)}) - \sum_l p_l t_l \\ \text{subject to} & Ax^{(d)} = s^{(d)}, \quad d = 1, \ldots, D \\ & x^{(d)} \succeq 0, \quad s^{(d)} \succeq_d 0, \quad d = 1, \ldots, D \\ & t_l = \sum_d x_l^{(d)}, \quad l = 1, \ldots, L \end{array} \tag{15}$$

and find $V_{\mathrm{comm}}(p)$ by solving the problem

$$\begin{array}{ll} \text{maximize} & \sum_l p_l \phi_l(r_l) \\ \text{subject to} & F \preceq g, \quad r \succeq 0. \end{array} \tag{16}$$

We call (15) the *network flow subproblem* and (16) the *resource allocation subproblem*. They are parametrized by the dual variable p. Both subproblems are convex optimization problems with special structure that allows them to be solved very efficiently. In particular, the network flow subproblem (15) is naturally decomposed into D single-commodity flow problems.

A subgradient of a non-differentiable convex function V at p is a vector $h \in \mathbf{R}^L$ such that

$$V(q) \geq V(p) + h^T(q - p) \tag{17}$$

for all q (see, *e.g.*, [26]). Given a dual variable $p \succeq 0$, let $x^\star(p), s^\star(p), t^\star(p)$ be an optimal solution to the network flow subproblem (15) and $r^\star(p)$ be an optimal solution to the resource allocation subproblem (16). From the definition of the dual function in (11), we find that a subgradient $h \in \mathbf{R}^L$ of V at p is given by

$$h_l = \phi_l(r_l^\star(p)) - t_l^\star(p), \quad l = 1, \ldots, L. \tag{18}$$

Note that h_l can be interpreted as the excess capacity on link l, *i.e.*, the difference between the capacity provided by the communication system and the proposed traffic by the routing.

Interpreting the dual variable p_l as the price for the capacity of link l (in dollars per unit flow), the dual decomposition method has an interesting economic interpretation. Given the prices p_l, the network layer solves the uncapacitated network flow problem (15), trying to maximize the total utility function discounted by $\sum_l p_l t_l$, the total cost of the link capacities used; the radio control layer solves the resource allocation problem (16), trying to maximize $\sum_l p_l \phi_l(r_l)$, the total revenue from capacities that it supports. The operation of the two layers are coordinated by the master dual problem (14) through the vector of prices p. The subgradient method for solving the master dual problem described in section VI-C can be interpreted as specific rules for updating the prices in order to arrive the optimal coordination.

C. Solving the dual problem by subgradient method

With the ability of evaluating the dual function and its subgradients, we now discuss how to solve the master dual problem (14) by the subgradient method.

In subgradient method, we start with an initial point $p^{(1)}$. At each iteration step $k = 1, 2, 3, \ldots$, we compute the dual function $V(p^{(k)})$ and a subgradient $h^{(k)}$ (see section VI-B), then update the dual variable by

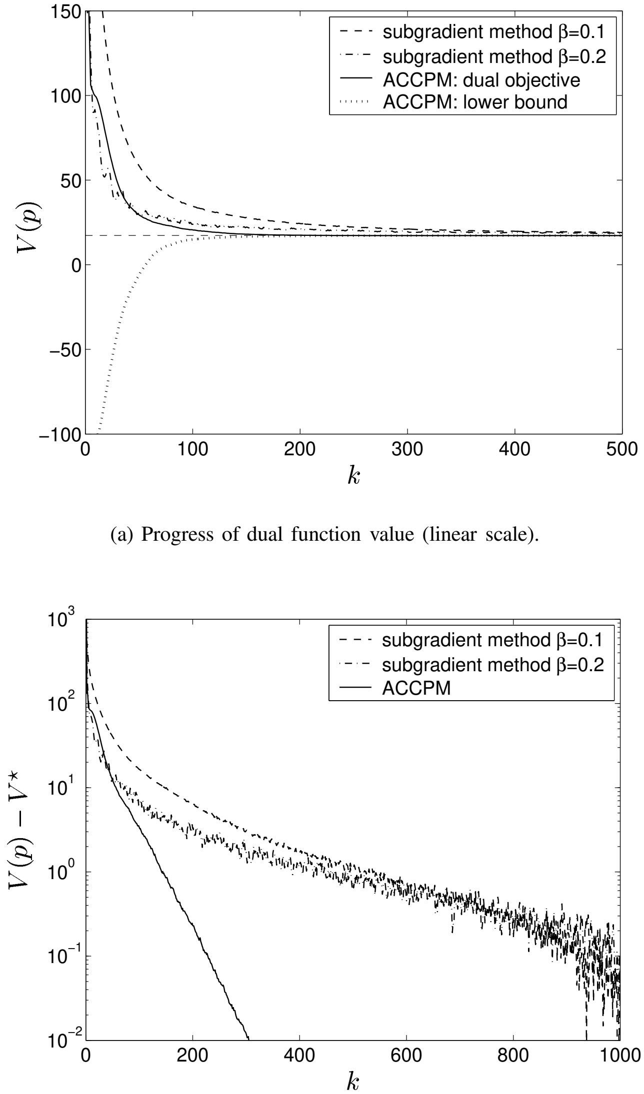
$$p^{(k+1)} = \left[p^{(k)} - \alpha_k h^{(k)}\right]_+ \tag{19}$$

Here, $[\cdot]_+$ denotes projection on the non-negative orthant and α_k is a positive scalar stepsize. There are many ways to select the stepsize in subgradient methods. One simple convergence condition (see [26, §2]) requires that the stepsize sequence satisfies

$$\alpha_k \to 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty,$$

An extensive account of subgradient methods, as well as many acceleration techniques to improve their convergence properties can be found in, *e.g.*, [26], [23]. In the numerical example of section V, we used the simple stepsize rule $\alpha_k = \beta/k$ where β is a positive constant. Figure 3 shows the dual objective function versus the number of iterations for $\beta = 0.1$ and $\beta = 0.2$.

With the economic interpretation at the end of section VI-B, the subgradient method (19) simply follows the laws of supply and demand at each link: if the link is under-utilized (*i.e.*, we have positive excess capacity $h_l = \phi_l(r_l) - t_l > 0$) the price is decreased, otherwise it is increased. Notice that the subgradient information h_l can be obtained locally at link l based on its own traffic t_l and available capacity $\phi_l(r_l)$. This property facilitates distributed implementation of subgradient methods: each link can update its own link price (dual variable) independently, without a central coordinator.



(a) Progress of dual function value (linear scale).

(b) Progress of dual optimality gap (semi-logarithmic scale).

Fig. 3. Convergence of the subgradient method and ACCPM.

In [27] we discuss another method, the analytic center cutting-plane method (ACCPM), for solving the dual problem and compare its performance with the subgradient method. The progress of the dual objective function as well as the lower bound obtained by ACCPM for the numerical example are shown in figure 3. Although ACCPM converges faster (fewer iterations) than the subgradient method, it is computationally more demanding (per iteration) and does not allow for a distributed implementation.

D. Hierarchical dual decomposition

The dual decomposition method can be applied hierarchically to exploit the structure of the SRRA problem at several different levels, illustrated in figure 4. At the first level, we decompose the SRRA problem into a network flow problem (15) and a resource allocation problem (16) and coordinate them by the master dual problem (14). At the second level, the

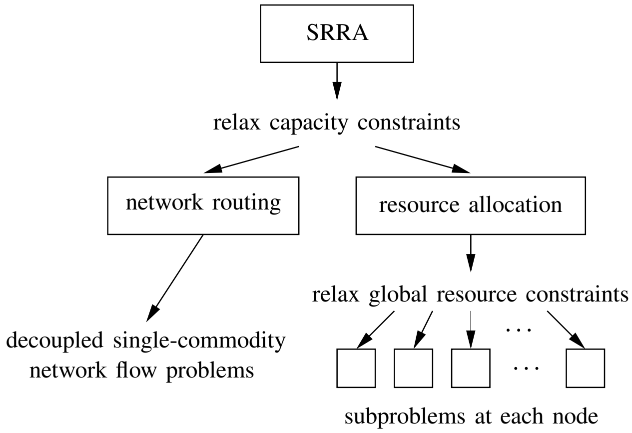


Fig. 4. Hierarchical dual decomposition for the SRRA problem.

network flow problem is naturally decomposed into single-commodity network flow problems for each destination; the resource allocation problem can be further decomposed into subproblems at each node, involving only local communications variables of the outgoing links. These local resource allocation subproblems are possibly coordinated through the price for globally shared resources (*e.g.*, bandwidths), as in the classical water-filling solution. In this paper, we have focused on the first level — vertical decomposition of two networking layers.

VII. Conclusions

We have considered the problem of simultaneous optimization of routing and resource allocation (SRRA) in wireless data networks. Our model captures the interplay between the resource allocation problem and the routing problem in two different networking layers, and our solution introduces a pricing mechanism on the capacities of communication links to optimally coordinate the operation of the two layers.

We have concentrated on a theoretical model that describes the average behavior of the network and disregards many detailed aspects such as packet loss and retransmissions, time-varying fading of wireless channels, and topology changes in the network. The model appears to be very useful for network provisioning, planning, and high-level management of the network. While much need to be done to extend this work to joint real-time power control and dynamic routing in wireless networks, we believe that the model and methodology presented in this paper opens the door toward this direction. One promising approach is to investigate the possibility of combining distributed algorithms for the master dual problem (*e.g.*, subgradient methods) with those for the routing and resource allocation subproblems under the hierarchical dual decomposition framework.

It is important to notice that we assume fixed technology (*e.g.*, coding and modulation schemes) and find the system parameters that attain the optimal performance within the specified infrastructure. Although the SRRA formulation often gives significant performance improvements over classical (non-coordinated) approaches, it does not address the information-theoretic question about the ultimate capacity of a wireless network (see, *e.g.*, [17], [28], [29]). In addition, this paper does not directly address some important practical issues in wireless data networks, such as quality of service. An extension of the SRRA formulation in this direction seems very attractive and needs further investigation.

Finally, the communications model in this paper does not include some important wireless systems, *e.g.*, systems using code-division multiple access (CDMA), and random access protocols such as carrier sense multiple access with collision avoidance (CSMA/CA). In recent work, we have extended the SRRA framework to CDMA wireless networks [30], and joint link scheduling, routing and power allocation in wireless networks [31].

Acknowledgment

The authors are grateful to Haitham Hindi and Andrea Goldsmith for helpful discussions. We also thank the anonymous reviewers for several comments and suggestions that helped improve the quality of this paper.

References

[1] D. P. Bertsekas and R. G. Gallager. *Data Networks*. Prentice Hall, 1992.
[2] D. P. Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific, 1998.
[3] A. Ouorou, P. Mahey, and J.-Ph. Vial. A survey of algorithms for convex multicommodity flow problems. *Management Science*, vol. 46, pp. 126-147, January 2000.
[4] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.
[5] R. G. Gallager. A minimum delay routing algorithm using distributed computation. *IEEE Trans. Commun.*, vol. 25, pp. 73–85, January 1977.
[6] T. E. Stern. A class of decentralized routing algorithms using relaxation. *IEEE Trans. Commun.*, vol. 25, pp. 1092–1102, October 1977.
[7] L. Li and A. J. Goldsmith. Capacity and optimal resource allocation for fading broadcast channels — part I: ergodic capacity and part II: outage capacity. *IEEE Trans. Inform. Theory*, vol. 47, pp. 1083–1127, March 2001.
[8] D. N. C. Tse and S. V. Hanly. Multiaccess fading channels — part I: Polymatroid structure, optimal resource allocation and throughput capacities. *IEEE Trans. Inform. Theory*, vol. 44, pp. 2796–2815, November 1998.
[9] D. N. Tse. Optimal power allocation over parallel gaussian broadcast channels. In *Proc. of International Symp. Inform. Theory*, Ulm, Germany, June 1997, page 27.
[10] B. Gavish and I. Neuman. A system for routing and capacity assignment in computer communication networks. *IEEE Trans. Commun.*, vol. 37, pp. 360–366, 1989.
[11] H.-H. Yen and F. Y.-S. Lin. Near-optimal delay constrained routing in virtual circuit networks. In *Proc. IEEE INFOCOM 2001*, vol. 2, pp. 750–756, Anchorage, AK, April 2001.
[12] M. Prytz. *On Optimization in Design of Telecommunications Networks with Multicast and Unicast Traffic*. PhD thesis, Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden, March 2002.
[13] B. Hajek and Sasaki. Link scheduling in polynomial time. *IEEE Trans. Inform. Theory*, vol. 34, pp. 910–917, 1988.
[14] L. Tassiulas and A. Ephremides. Jointly optimal routing and scheduling in packet radio networks. *IEEE Trans. Inform. Theory*, vol. 38, pp. 165–168, 1992.
[15] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz. A comparison of mechanisms for improving TCP performance over wireless links. *IEEE/ACM Trans. Networking*, vol. 5, pp. 756–769, December 1997.
[16] L. Kleinrock. *Communication Nets: Stochastic Message Flow and Delay*. McGraw-Hill, New York, 1964.
[17] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. Available at: http://www.stanford.edu/people/boyd/cvxbook.html

[19] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics. SIAM, 1994.

[20] F. P. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, vol. 49, pp. 237–252, March 1998.

[21] S. H. Low and D. E. Lapsley. Optimization flow control — I: Basic algorithm and convergence. *IEEE/ACM Trans. Networking*, vol. 7, pp. 861–874, December 1999.

[22] S. H. Low, F. Paganini, and J. C. Doyle. Internet congestion control. *IEEE Control Syst. Mag.*, vol. 22, pp. 28–43, February 2002.

[23] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, 1999.

[24] R. T. Rockafellar. Augmented lagrangians and applications of the proximal point algorithms in convex programming. *Mathematics of Operations Research*, vol. 1, pp. 97–116, 1976.

[25] K. C. Kiwiel. Approximations in bundle methods and decomposition of convex programs. *Journal of Optimization Theory and Applications*, vol. 84, pp. 529–548, 1995.

[26] N. Z. Shor. *Minimization Methods for Non-differentiable Functions*. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1985.

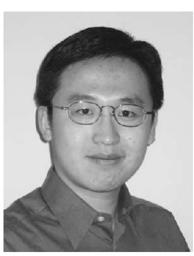
[27] L. Xiao, M. Johansson, and S. Boyd. Simultaneous routing and resource allocation via dual decomposition. In *Proc. of The 4th Asian Control Conf.*, pp. 29-34, Singapore, September 2002.

[28] A. Ephremides and B. Hajek. Information theory and communication networks: An unconsummated union. *IEEE Trans. Inform. Theory*, vol. 44, pp. 2416–2434, 1998.

[29] L.-L. Xie and P. R. Kumar. A network information theory for wireless communication: Scaling laws and optimal operation. IEEE Trans. Inform. Theory, Volume: 50 , Issue: 5 , Pages: 748 - 767, April 2002.

[30] M. Johansson, L. Xiao, and S. Boyd. Simultaneous routing and resource allocation in CDMA wireless data networks. In *Proc. IEEE Intern. Conf. Commun.*, vol. 1, pp. 51-55, Anchorage, Alaska, May 2003.

[31] M. Johansson and L. Xiao. Cross-layer optimization of wireless networks using nonlinear column generation. Technical report IR-S3-REG-0302, Royal Institute of Technology, Sweden. Also submitted for journal publication.



Lin Xiao is currently a Ph.D. candidate in the Department of Aeronautics and Astronautics at Stanford University, with a minor in Electrical Engineering. He received his M.S. degree in EE from Stanford in 2002, and B.E. and M.E. from Beijing University of Aeronautics and Astronautics, in 1994 and 1997 respectively. His current research interests include convex optimization with applications in networking and communication systems, control over networks, and distributed computation.

Mikael Johansson received the Ph.D. degree in Automatic Control from Lund University, Sweden, in 1999. During 1999-2002, he was a consulting professor and postdoctoral scholar at Stanford University and UC Berkeley. He is currently Assistant Professor at the Department of Signals, Sensors and Systems at the Royal Institute of Technology (KTH), Stockholm, Sweden. His research interests include wireless systems, data networks, optimization and hybrid and embedded control.

Stephen P. Boyd is the Samsung Professor of Engineering, Professor of Electrical Engineering, and Director of the Information Systems Laboratory at Stanford University. His current interests include computer-aided control system design, and convex programming applications in control, signal processing, and circuit design.

He received an AB degree in Mathematics, summa cum laude, from Harvard University in 1980, and a PhD in EECS from U. C. Berkeley in 1985. He is the author of Linear Controller Design: Limits of Performance (with Craig Barratt, 1991), Linear Matrix Inequalities in System and Control Theory (with L. El Ghaoui, E. Feron, and V. Balakrishnan, 1994), and Convex Optimization (with Lieven Vandenberghe, 2003). He received an ONR Young Investigator Award, a Presidential Young Investigator Award, the 1992 AACC Donald P. Eckman Award. He has received the Perrin Award for Outstanding Undergraduate Teaching in the School of Engineering, and an ASSU Graduate Teaching Award. In 2003, he received the AACC Ragazzini Education award. He is a fellow of the IEEE, and a Distinguished Lecturer of the IEEE Control Systems Society.