Joint Optimization of Relay Strategies and Resource Allocations in Cooperative Cellular Networks

Truman Chiu-Yam Ng and Wei Yu
University of Toronto, Toronto, Ontario Canada
{ngtruman, weiyu}@comm.utoronto.ca

Jianzhong(Charlie) Zhang and Anthony Reid
Nokia Research Center, Irving, TX USA
{charlie.zhang, tony.reid}@nokia.com

***Abstract*— This paper considers a wireless cooperative cellular data network with a base station and many subscribers in which the subscribers have the ability to relay information for each other to improve the overall network performance. For a wireless network operating in a frequency-selective fading environment, the choices of relay node, relay strategy, and the allocation of power and bandwidth for each user are important design parameters. The design challenge is compounded further by the need to take user traffic demands into consideration. This paper proposes a utility maximization framework for such a network. We show that for a cellular system employing orthogonal frequency-division multiple-access (OFDMA), the optimization of physical-layer transmission strategies can be done efficiently by introducing a set of pricing variables. The proposed solution incorporates both user traffic demand and the physical channel realization in a cross-layer design that not only allocates power and bandwidth optimally for each user, but also selects the best relay node and best relay strategy (i.e. decode-and-forward vs. amplify-and-forward) for each source-destination pair.**

I. Introduction

In a wireless network with many source-destination pairs, cooperative transmission by relay nodes has been shown to improve transmission rate and diversity of the wireless network [1], [2]. The relays may facilitate transmission by first decoding the transmitted codeword, then forwarding the decoded codeword to the destination using a strategy known as "decode-and-forward" (DF). Alternatively, a relay may simply amplify its received signal and employ a so-called "amplify-and-forward" (AF) strategy.

This paper is motivated by the following questions: In a cooperative wireless network, which node should act as a relay? What relay strategy should be used? When, and in which frequency should relaying be employed? Clearly, the answers to these questions depend on the topology of the network. For example, when a relay is located closer to the source than to the destination, decode-and-forward appears to be a natural choice. On the other hand, when the relay is located closer to the destination, its received signal-to-noise ratio (SNR) may not be high enough to allow decoding, in which case amplify-and-forward is more suited. However, the optimal operation of relays is more complicated than that provided by a rule-of-thumb. This is because the choices of relay node and relay strategy also depend on the amount of transmit power available at the source and at the relay, and further on the user traffic patterns. This is particularly true in a power- and bandwidth-limited network in which each node may act as a source/destination or relay simultaneously. In this case, the partitioning of the power and bandwidth between the transmission of one's own data vs. the relaying of other user's traffic becomes crucial. The optimal power and bandwidth allocation is further coupled with the choice of relay and the choice of relay strategies.

This paper takes a *system* view of the cooperative network, and aims to jointly optimize relay strategies and physical-layer resources in a network. We focus on a cellular data network with a single base station and many subscribers in each cell, where each subscriber has the ability to relay information for each other. Further, we consider a network employing orthogonal frequency-division multiple-access (OFDMA) where each user node may simultaneously act as a source, a destination, or as a relay, but at different frequency tones. In such a network, the power allocation among the transmitting nodes in the network and across the frequency tones can greatly affect network performance. Our target application is a fixed broadband access network in which channel estimation is feasible, and where centralized resource allocation can be implemented. The cooperative strategies considered in this paper take advantage of the broadcast nature of wireless channel, and allow the destination to cooperatively "combine" signals sent by both the source and the relay.

The issue of power control for relay-assisted networks has been dealt with in [3], [4] and many others. In particular, the question of which node to use as the relay [5], [6] and which relay strategy to employ [7] has always been regarded as important but difficult. This paper makes progress by showing that the joint relay-strategy selection and resource allocation problem may be solved globally and efficiently by using a set of dual variables. Our approach accounts for user traffic demand, and it represents a *cross-layer* optimization framework for cooperative networks. A key technique in our study is the use of pricing to determine the optimal relay strategy. Pricing has been used in earlier studies of both multihop [8] and ad-hoc relay networks [9], where pricing information is used to give selfish nodes an incentive to relay. The present paper describes a centralized system where prices are used to regulate system resources in order to achieve an overall global optimal performance for the network.

The notation used in the paper is as follows. Boldface lower-case letters are used for column vectors. $\mathbf{0}$ and $\mathbf{1}$ denote the all-zero and all-one column vectors respectively. For two vectors of the same length, "$\succeq$" and "$\preceq$" denote component-wise inequalities. Lower-case letter x_i denotes the i^{th} entry of $\boldsymbol{x}$. Boldface upper-case letters denote matrices. For a matrix

$\boldsymbol{X}$, $X(i,j)$ represents the entry on the i^{th} row and j^{th} column. The vector $X(i,:)$ is the i^{th} row of $\boldsymbol{X}$, and $X(:,j)$ is the j^{th} column of $\boldsymbol{X}$. The superscripts $(\cdot)^T$, $(\cdot)^H$ denote transpose, and the Hermitian respectively.

II. Utility Maximization Framework

This paper adopts a network utility maximization (NUM) framework in which each data stream has an associated utility function. A utility function is a concave function of data rates that reflects user satisfaction. The choice of utility function depends on the user application (e.g. data, video.) The network objective is to maximize the sum utility. The NUM framework originated from the work of Kelly *et al* [10], and it has been applied to many physical-layer design problems [11]. In the context of a cooperative network where a cooperative node must spend its own resources to relay information for other nodes, maximizing sum utility serves as a common objective to create incentives for user nodes to act as relays.

A. System Model

The cooperative cellular network considered in this paper consists of a base station and K user nodes. Let $\mathcal{K} = \{1,2,...,K\}$ be the set of user nodes. Denote the base station as node $K+1$. Let $\mathcal{K}_+ = \{1,2,...,K+1\}$ be the extended set of nodes. Each of the K user nodes has both downlink (d) and uplink (u) communications with the base station, so there are $2K$ data streams in total. Let $\mathcal{M}$ be the set of all data streams, i.e. $\mathcal{M} = \mathcal{K} \times \{d,u\} = \{(1,d),(2,d),...,(K,d),(1,u),(2,u),...,(K,u)\}$. Each node in the system is equipped with a single antenna.

The cooperative network of interest employs an OFDMA physical-layer with N tones. Let $\mathcal{N} = \{1,...,N\}$ denote the set of tones. To prevent inter-stream interference, we restrict that there is only one active data stream in each tone. However, each of the $2K$ data streams can be active in more than one tone. Moreover, we impose the condition that an active data stream can use at most one relay (where the relay node can be any one of the other $K-1$ user nodes.) We further assume that diversity combining at the destination occurs only when the source-relay, source-destination, and relay-destination links are all at the same tone. Note that uplink and downlink transmissions take place simultaneously in the network, so the base station is both a source and a destination (but in different tones.) However, the base station can never be the relay. On the other hand, all user nodes in the network can simultaneously be a source, a relay, and a destination (again in different tones.) The wireless channel is modelled as a frequency-selective fading channel with coherence bandwidth in the order of the bandwidth of a few tones. We further assume a slow fading environment, so that full channel side-information (CSI) is available at the base station.

Let $\boldsymbol{P}$ be a $(K+1)\times N$ matrix such that $P(i,n)$ denotes the power spent by node i in tone n. Because only the source and relay spend power in each tone, the column vector $P(:,n)$ has at most two non-zero entries. Similarly, let $\boldsymbol{R}$ be a $2K\times N$ matrix such that $R(m,n)$ denotes the actual rate achieved by stream m in tone n. Since only one stream can be active in each tone, the column vector $R(:,n)$ has at most one non-zero entry. Power and rate are related by the *achievable rate region*, denoted as $\boldsymbol{R} \in \mathcal{C}(\boldsymbol{P})$. The achievable rate region implicitly accounts for the best possible use of relay strategies, and design restrictions mentioned this paragraph.

By definition, $(\boldsymbol{P1})_i$, i.e. the row sum of $\boldsymbol{P}$, is the total power spent at node i, summed across all tones. Similarly, $(\boldsymbol{R1})_m$, i.e. the row sum of $\boldsymbol{R}$, gives the data rate for data stream m, summed across all tones. A separate power constraint, $\boldsymbol{p^{max}} = [p_1^{max}, p_2^{max}, ..., p_{K+1}^{max}]^T$, is imposed on each node. A separate utility function U_m is associated with each data stream m. The objective is to optimally allocate power among the frequency tones, while choosing the best relay node and strategy, in order to maximize the network sum utility. Expressed succinctly, the optimization problem is

$$\begin{aligned} \text{maximize} \quad & \sum_{m\in\mathcal{M}} U_m((\boldsymbol{R1})_m) \\ \text{subject to} \quad & \boldsymbol{P1} \preceq \boldsymbol{p^{max}}, \quad \boldsymbol{R} \in \mathcal{C}(\boldsymbol{P}) \end{aligned} \tag{1}$$

B. Cross-layer Optimization via Dual Decomposition

In general, finding the achievable rate region $\mathcal{C}(\boldsymbol{P})$ involves a search over all possible power allocations, relays, and relay strategies. So, the optimization problem (1) is a mixed integer programming problem, and the structure of $\mathcal{C}(\boldsymbol{P})$ is complicated. However, in an OFDMA system with many narrow subcarriers, $\mathcal{C}(\boldsymbol{P})$ is always convex because the time-sharing of two transmission strategies can always be implemented across the frequency tones via frequency-division multiplexing. The idea is that if two sets of rates using two different power allocations and relay strategies are achievable individually, then their linear combination is also achievable by a frequency-division multiplex of the two sets of strategies.

The key point is that this observation, made early in [12] for a spectrum balancing problem and in [11], is applicable *even as discrete relay-selection and relay-strategy-selection are involved*. This opens the door for using convex optimization techniques for solving the mixed integer programming problem (1). In particular, using the duality theory of [12], the following is true:

Proposition 1: The optimization problem (1) has zero duality gap in the limit as the number of OFDM tones goes to infinity.

The zero-duality result implies that the Lagrangian technique can be used to solve the mixed integer-programming problem efficiently. In particular, the Lagrangian method leads to a decomposition of the utility maximization problem into two smaller subproblems, each of which may be solved independently. The rest of this section develops this dual decomposition result. First, rewrite (1) as

$$\begin{aligned} \max_{\boldsymbol{P},\boldsymbol{R}} \quad & \sum_{m\in\mathcal{M}} U_m(t_m) \\ \text{s.t.} \quad & \boldsymbol{P1} \preceq \boldsymbol{p^{max}}, \quad \boldsymbol{R1} \succeq \boldsymbol{t}, \quad \boldsymbol{R} \in \mathcal{C}(\boldsymbol{P}) \end{aligned} \tag{2}$$

where $\boldsymbol{t} = [t_{(1,d)}, t_{(2,d)}, ..., t_{(K,u)}]^T$ are extra variables. The key step is to relax $\boldsymbol{R1} \succeq \boldsymbol{t}$ by first forming the Lagrangian

$$L = \sum_{m\in\mathcal{M}} \left(U_m(t_m) + \lambda_m \left(\sum_{n\in\mathcal{N}} R(m,n) - t_m \right) \right)$$

where $\boldsymbol{\lambda} = [\lambda_{(1,d)}, \lambda_{(2,d)}, ..., \lambda_{(K,u)}]^T$, with λ_m being a dual variable corresponding to stream m. The dual function

$$g(\boldsymbol{\lambda}) = \left\{ \begin{array}{ll} \max\limits_{\boldsymbol{P},\boldsymbol{R},\boldsymbol{t}} & L(\boldsymbol{P},\boldsymbol{R},\boldsymbol{t},\boldsymbol{\lambda}) \\ \text{s.t.} & \boldsymbol{P1} \preceq \boldsymbol{p^{max}}, \quad \boldsymbol{R} \in \mathcal{C}(\boldsymbol{P}) \end{array} \right. \tag{3}$$

consists of application-layer variables $\boldsymbol{t}$, and physical layer variables $\boldsymbol{P}$ and $\boldsymbol{R}$. Moreover, $g(\boldsymbol{\lambda})$ can be separated into two maximization subproblems, namely a utility maximization problem, corresponding to a rate adaptation problem in the application layer

$$g_{appl}(\boldsymbol{\lambda}) = \max_{\boldsymbol{t}} \sum_{m \in \mathcal{M}} \left(U_m(t_m) - \lambda_m t_m \right) \tag{4}$$

and a joint relay-strategy selection and power and bandwidth allocation problem in the physical layer

$$g_{phy}(\boldsymbol{\lambda}) = \left\{ \begin{array}{ll} \max\limits_{\boldsymbol{P},\boldsymbol{R}} & \sum\limits_{m \in \mathcal{M}} \lambda_m \sum\limits_{n \in \mathcal{N}} R(m,n) \\ \text{s.t.} & \boldsymbol{P1} \preceq \boldsymbol{p^{max}}, \quad \boldsymbol{R} \in \mathcal{C}(\boldsymbol{P}) \end{array} \right. \tag{5}$$

Thus, the optimization framework provides a layered approach to the sum utility maximization problem. The application layer adaptively adjusts user's traffic demand based on the current channel conditions, while the physical layer adaptively allocates power and bandwidth and selects the best choice of relay and relaying scheme to obtain rates required by the upper layer. The interaction between the layers is now controlled through the use of the dual variable $\boldsymbol{\lambda}$, which coordinates the user *demand* and the physical layer *supply* of rates. Because (1) has zero duality gap, it can be solved via its dual

$$\begin{array}{ll} \text{minimize} & g(\boldsymbol{\lambda}) \\ \text{subject to} & \boldsymbol{\lambda} \succeq \boldsymbol{0} \end{array} \tag{6}$$

In particular, the update of $\boldsymbol{\lambda}$ may be done using a subgradient method [13] as follows:

Subroutine 1: Subgradient-based method for solving (6)

1) Initialize $\boldsymbol{\lambda^{(0)}}$.
2) Given $\boldsymbol{\lambda^{(l)}}$, solve (4) and (5) separately to obtain the optimal values $\boldsymbol{t^*}$, $\boldsymbol{P^*}$, and $\boldsymbol{R^*}$.
3) Set $\boldsymbol{\lambda^{(l+1)}} = \boldsymbol{\lambda^{(l)}} + (\boldsymbol{\nu^{(l)}})^T (\boldsymbol{t^*} - \boldsymbol{R^*1})$
4) Return to step 2 until convergence. □

Subroutine 1 is guaranteed to converge to the optimal dual variable, if the step sizes $\boldsymbol{\nu^{(l)}}$ are chosen following a diminishing step size rule [14]. From the optimal dual variables, the optimal primal variables can then be found easily.

C. Solutions of individual subproblems

We now describe efficient methods to solve the two subproblems, which together with the first decomposition described in the previous section, solve the overall utility maximization problem globally and efficiently.

1) Application Layer Subproblem: Note that $g_{appl}(\boldsymbol{\lambda})$ as in (4) can be solved by maximizing each of the summation terms separately. Specifically, since U_m is a concave function of t_m, so is $U_m(t_m) - \lambda_m t_m$. Therefore, t_m^* can be found by taking the derivative of $(U_m(t_m) - \lambda_m t_m)$ with respect to t_m and setting it to zero. Example 1 shows a class of utility functions that will later be used in simulation results.

Example 1: Let t be data rate, and define

$$U(t) = \left\{ \begin{array}{ll} a\left(1 - e^{-bt}\right), & \text{if } t \geq 0 \\ -\infty, & \text{if } t < 0 \end{array} \right. , \tag{7}$$

where a and b are strictly positive real numbers. a represents the upper limit of the utility, while b is chosen such that at some rate c, the utility is equal to $0.9a$. Given c, $b = \frac{\ln(0.1)}{-c}$. In the application layer subproblem, the per stream maximization is of the form $(U(t) - \lambda t)$. By calculus,

$$t^* = \max\left(0, -\frac{1}{b} \ln \frac{\lambda}{ab}\right) \tag{8}$$

2) Physical Layer Subproblem: The physical layer subproblem is the more difficult of the two. Finding $\boldsymbol{R^*}$ involves selecting the best data stream, power allocation, relay node and relay strategy in all tones. Further, the per-node power constraint introduces coupling across the tones. This section shows that by a second decomposition step, the coupling across the tones can be removed, resulting in a procedure that is *linear* in the number of tones. The main technique here is reminiscent of the weighted sum-rate problem in [15]. A Lagrangian can be formed by relaxing the constraint $\boldsymbol{P1} \preceq \boldsymbol{p^{max}}$ and introducing prices into the objective function of (5):

$$\begin{aligned} Q \quad = \quad & \sum_{m \in \mathcal{M}} \lambda_m \sum_{n \in \mathcal{N}} R(m,n) \\ & + \sum_{i \in \mathcal{K}_+} \mu_i \left(p_i^{max} - \sum_{n \in \mathcal{N}} P(i,n) \right) \end{aligned} \tag{9}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_{K+1}]^T$. The key observation is that

$$q(\boldsymbol{\mu}) = \left\{ \begin{array}{ll} \max\limits_{\boldsymbol{P},\boldsymbol{R}} & Q(\boldsymbol{P},\boldsymbol{R},\boldsymbol{\lambda},\boldsymbol{\mu}) \\ \text{s.t.} & \boldsymbol{R} \in C(\boldsymbol{P}) \end{array} \right. \tag{10}$$

can now be decoupled into N per-tone maximization subproblems: $\forall n \in \mathcal{N}$,

$$\begin{array}{ll} \max & \lambda_m R(m,n) - (\mu_{\mathcal{S}} P(\mathcal{S},n) + \mu_{\mathcal{R}} P(\mathcal{R},n)) \\ \text{s.t.} & R(:,n) \in C(P(:,n)) \end{array} \tag{11}$$

The complexity of the physical layer subproblem is now linear in N. The dual variables $\boldsymbol{\mu}$ represent the cost of power for each node. Together, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ coordinate "supply" of power and "demand" for rates.

A critical requirement for the decomposition of the physical layer subproblems into N per-tone subproblems is the convexity structure of the problem, namely $\mathcal{C}(\boldsymbol{P})$ can always be made a convex region if time- or frequency-sharing can be implemented. Therefore, the physical layer subproblem also has zero duality gap, and can be solved optimally via the dual problem

$$\begin{array}{ll} \text{minimize} & q(\boldsymbol{\mu}) \\ \text{subject to} & \boldsymbol{\mu} \succeq \boldsymbol{0} \end{array} \tag{12}$$

Again, a subgradient approach with appropriate step sizes may be used to solve the dual problem.

Subroutine 2: Subgradient-based method for solving (12)

1) Initialize $\boldsymbol{\mu^{(0)}}$.

2) Given $\boldsymbol{\mu}^{(l)}$, solve the N per-tone maximization problems (11) separately to obtain $\boldsymbol{P}^*$ and $\boldsymbol{R}^*$.
3) Set $\boldsymbol{\mu}^{(l+1)} = \boldsymbol{\mu}^{(l)} + (\boldsymbol{\epsilon}^{(l)})^T (\boldsymbol{P}^*\mathbf{1} - \boldsymbol{p}^{max})$
4) Return to step 2 until convergence. □

Now, it remains to solve the per-tone optimization problem (11). The optimizing variables are

- Data stream m ($m \in \mathcal{M}$)
- Relaying scheme = {DC, DF, AF}[1]
- Choice of relay node $\mathcal{R}$ ($\mathcal{R} \in \mathcal{K}, \mathcal{R} \neq \mathcal{S}$ or $\mathcal{D}$)
- Bit rate $R(m,n)$

As the search variables are discrete, the per-tone maximization problem above can be solved by simply searching over a discrete set. Size of the set is the product of the number of bits, data streams, relay nodes and relay strategies. Such a search is often feasible for a practical network.

III. Optimal Relay-Strategy Selection

A main point of the previous section is that the joint relay operation and power allocation problem across frequency tones can be solved globally and efficiently. This hinges upon an efficient solution to the per-tone problem. This section provides a solution to the per-tone power allocation problem for each relay strategy. The main idea is to express the required power at the source and at the relay as a function of the given rate for each relay strategy. This is done for each stream m (which uniquely determines $\mathcal{S}$ and $\mathcal{D}$) with a possible participation of a relay $\mathcal{R}$ at each tone.

This paper focuses on two-time-slot implementation of AF and DF, and imposes the practical constraint that a relay node cannot send and receive at the same time in the same tone [16]. In both relaying schemes, during the first time-slot, only $\mathcal{S}$ transmits, which simplifies expression with insignificant performance loss. The difference between AF and DF is in the operation of $\mathcal{R}$. In AF, $\mathcal{R}$ amplifies the signal it receives in the first time-slot, and sends it out in the second time-slot. In DF, $\mathcal{R}$ attempts to decode its received signal in the first time-slot. If decoding is unsuccessful, $\mathcal{R}$ will remain silent in the second time-slot. Otherwise, $\mathcal{R}$ re-encodes the decoded data and then transmits it in the second time-slot.

Achievable rates for two-time-slot cheap relay channel as a function of transmit powers have been previously derived in [1]. What is required here is, however, the optimal transmit power as a function of rates. With a participation of a relay, an entire range of power allocations is possible at $\mathcal{S}$ and $\mathcal{R}$ for each fixed bit rate. The main idea is to show that by using an extra optimization step that accounts for the pricing structure of the power availability, the optimal powers can be readily found.

As mentioned earlier, perfect knowledge of channel gain and noise variance is assumed. Power spent at node i and data rate of stream m in tone n are denoted as $P(i,n)$ and $R(m,n)$ respectively. However, since transmission takes place in two time-slots, both actual power and actual data rate should be halved. Throughout this section, $P^*(\cdot,n)$ denotes the optimal transmit power (at either $\mathcal{S}$ or $\mathcal{R}$) at a data rate $R(m,n)$.

[1]DC stands for "direct channel", i.e. no relay.

Subscripts 1 and 2 are used to denote the first and the second time-slots, respectively. We denote $x_{\mathcal{S}1}$ as the symbols sent by $\mathcal{S}$, $y_{\mathcal{D}1}$ and $y_{\mathcal{D}2}$ as the received symbol at $\mathcal{D}$, and $y_{\mathcal{R}1}$ as the received symbol at $\mathcal{R}$. The complex channel gains from $\mathcal{S}$ to $\mathcal{D}$, $\mathcal{S}$ to $\mathcal{R}$, and $\mathcal{R}$ to $\mathcal{D}$ are denoted as $h_{\mathcal{SD}}$, $h_{\mathcal{SR}}$, and $h_{\mathcal{RD}}$, respectively. The channel gains are assumed to be identical in both time-slots. Moreover, $n_{\mathcal{D}1}$, $n_{\mathcal{D}2}$, and $n_{\mathcal{R}1}$ are circularly symmetric complex Gaussian noises $\mathcal{CN}(0, N_oW)$.

A. Direct Channel (DC)

The DC channel is modelled as

$$y_{\mathcal{D}} = \sqrt{P(\mathcal{S},n)}h_{\mathcal{SD}}x_{\mathcal{S}} + n_{\mathcal{D}} \tag{13}$$

The achievable rate (in b/s/Hz) is well-known:

$$R(m,n) \leq \log_2\left(1 + \frac{P(\mathcal{S},n)|h_{\mathcal{SD}}|^2}{\Gamma N_oW}\right), \tag{14}$$

where Γ is the gap to capacity. For discrete bit-loading,

$$P^*(\mathcal{S},n) = (2^{R(m,n)} - 1)\frac{\Gamma N_oW}{|h_{\mathcal{SD}}|^2}. \tag{15}$$

B. Decode-and-forward (DF)

The channel equations of DF relay channel are:

$$y_{\mathcal{D}1} = \sqrt{P(\mathcal{S},n)}h_{\mathcal{SD}}x_{\mathcal{S}1} + n_{\mathcal{D}1} \tag{16}$$
$$y_{\mathcal{R}1} = \sqrt{P(\mathcal{S},n)}h_{\mathcal{SR}}x_{\mathcal{S}1} + n_{\mathcal{R}1} \tag{17}$$
$$y_{\mathcal{D}2} = \sqrt{P(\mathcal{R},n)}h_{\mathcal{RD}}x_{\mathcal{S}1} + n_{\mathcal{D}2} \tag{18}$$

For successful decoding of $x_{\mathcal{S}1}$ at the relay, we need:

$$R(m,n) \leq \log_2\left(1 + \frac{P(\mathcal{S},n)|h_{\mathcal{SR}}|^2}{\Gamma N_oW}\right), \tag{19}$$

or equivalently,

$$P(\mathcal{S},n) \geq (2^{R(m,n)} - 1)\frac{\Gamma N_oW}{|h_{\mathcal{SR}}|^2}, \tag{20}$$

For successful decoding at $\mathcal{D}$, we need

$$R(m,n) \leq \log_2\left(1 + \frac{P(\mathcal{S},n)|h_{\mathcal{SD}}|^2 + P(\mathcal{R},n)|h_{\mathcal{RD}}|^2}{\Gamma N_oW}\right)$$

A maximum-ratio combining formula is used to derive the above equation. Rearranging the above gives

$$P^*(\mathcal{R},n) = \frac{\left(2^{R(m,n)} - 1\right)\Gamma N_oW - P(\mathcal{S},n)|h_{\mathcal{SD}}|^2}{|h_{\mathcal{RD}}|^2} \tag{21}$$

If $P^*(\mathcal{R},n) \leq 0$, then DF is not a suitable relay scheme.

Now, it remains to optimize $P(\mathcal{S},n)$, which is not immediate since at a fixed rate $R(m,n)$, decreasing $P(\mathcal{S},n)$ would increase $P^*(\mathcal{R},n)$, which may actually decrease the utility. Recall that the objective of the per-tone optimization problem as expressed in (11) is

$$\max_{P(\mathcal{S},n)} \lambda_m R(m,n) - \mu_{\mathcal{S}}P(\mathcal{S},n) - \mu_{\mathcal{R}}P^*(\mathcal{R},n) \tag{22}$$

From (21), $P^*(\mathcal{R},n)$ may be obtained as a function of $P(\mathcal{S},n)$, so that the entire objective is a function of $P(\mathcal{S},n)$ only. Since the optimization problem is now unconstrained, it

can be solved by looking at the first order derivative of the objective function (called f below):

$$\frac{df}{dP(\mathcal{S},n)} = -\mu_{\mathcal{S}} + \mu_{\mathcal{R}} \frac{|h_{\mathcal{SD}}|^2}{|h_{\mathcal{RD}}|^2} \tag{23}$$

Note that the objective is a linear function of $P(\mathcal{S},n)$, so the derivative is a constant. This implies that if $\frac{df}{dP(\mathcal{S},n)} > 0$, then $P(\mathcal{S},n)$ should be set to infinity. From (21), this means DF is unnecessary. On the other hand, if $\frac{df}{dP(\mathcal{S},n)} \leq 0$, then $P^*(\mathcal{S},n)$ should be set to the minimum as expressed in (20). Note that by (21), such a $P^*(\mathcal{S},n)$ guarantees that $P^*(\mathcal{R},n)$ is positive.

The DF mode power allocation procedure is summarized below:

Subroutine 3: Optimal power allocation for a fixed $R(m,n)$ in the DF relay mode:

1) If $|h_{\mathcal{SR}}| <= |h_{\mathcal{SD}}|$, set $P^*(\mathcal{S},n) = P^*(\mathcal{R},n) = \infty$,
2) else if $-\mu_{\mathcal{S}} + \mu_{\mathcal{R}} \frac{|h_{\mathcal{SD}}|^2}{|h_{\mathcal{RD}}|^2} > 0$, then set $P^*(\mathcal{S},n) = P^*(\mathcal{R},n) = \infty$,
3) else set $P^*(\mathcal{S},n)$ and $P^*(\mathcal{R},n)$ according to (20) and (21), respectively, with equality.
4) Divide $R(m,n)$, $P^*(\mathcal{S},n)$, and $P^*(\mathcal{R},n)$ by 2. □

C. Amplify-and-forward (AF)

The channel equations of AF relay channel are:

$$y_{\mathcal{D}1} = \sqrt{P(\mathcal{S},n)} h_{\mathcal{SD}} x_{\mathcal{S}1} + n_{\mathcal{D}1}, \tag{24}$$

$$y_{\mathcal{R}1} = \sqrt{P(\mathcal{S},n)} h_{\mathcal{SR}} x_{\mathcal{S}1} + n_{\mathcal{R}1}, \tag{25}$$

$$y_{\mathcal{D}2} = \beta y_{\mathcal{R}1} h_{\mathcal{RD}} + n_{\mathcal{D}2}, \tag{26}$$

where

$$\beta = \sqrt{\frac{P(\mathcal{R},n)}{P(\mathcal{S},n)|h_{\mathcal{SR}}|^2 + N_o W}} \tag{27}$$

is the power amplification factor at $\mathcal{R}$.

To analyze the power requirement for AF, recognize that in order for the destination $\mathcal{D}$ to decode the signal $x_{\mathcal{S}1}$, which is sent across two time-slots, we must have

$$R(m,n) \leq \log_2(1 + \text{SNR}_{\text{AF}}) \tag{28}$$

where SNR_{AF} is

$$\frac{1}{\Gamma}\left[\frac{P(\mathcal{S},n)|h_{\mathcal{SD}}|^2}{N_o W} + \frac{\frac{P(\mathcal{R},n)P(\mathcal{S},n)|h_{\mathcal{RD}}|^2|h_{\mathcal{SR}}|^2}{P(\mathcal{S},n)|h_{\mathcal{SR}}|^2+N_o W}}{N_o W\left(1+\frac{P(\mathcal{R},n)|h_{\mathcal{RD}}|^2}{P(\mathcal{S},n)|h_{\mathcal{SR}}|^2+N_o W}\right)}\right]$$

Rearranging the above, we get

$$P(\mathcal{R},n) = \frac{(c_1 P(\mathcal{S},n) + c_2)(c_3 P(\mathcal{S},n) + c_4)}{c_5 P(\mathcal{S},n) + c_6}, \tag{29}$$

where

$$c_1 = |h_{\mathcal{SD}}|^2, \quad c_2 = -(2^{R(m,n)} - 1)\Gamma N_o W, \quad c_3 = |h_{\mathcal{SR}}|^2,$$
$$c_4 = N_o W, \quad c_5 = |h_{\mathcal{RD}}|^2(-|h_{\mathcal{SD}}|^2 - |h_{\mathcal{SR}}|^2),$$
$$c_6 = (2^{R(m,n)} - 1)\Gamma N_o W |h_{\mathcal{RD}}|^2.$$

Now, observe that $(c_3 P(\mathcal{S},n) + c_4) > 0$ always. Thus, to ensure $P^*(\mathcal{R},n) > 0$, the terms $(c_1 P(\mathcal{S},n) + c_2)$ and $(c_5 P(\mathcal{S},n) + c_6)$ must either be both greater than zero or both less than zero. It is not hard to see that a valid solution is obtained only when both terms are less than zero, leading to a feasible region for $P(\mathcal{S},n)$ as

$$P_{\min}(\mathcal{S},n) < P(\mathcal{S},n) < P_{\max}(\mathcal{S},n), \tag{30}$$

where

$$P_{\min}(\mathcal{S},n) = \frac{(2^{R(m,n)} - 1)\Gamma N_o W}{|h_{\mathcal{SD}}|^2 + |h_{\mathcal{SR}}|^2} \tag{31}$$

$$P_{\max}(\mathcal{S},n) = \frac{(2^{R(m,n)} - 1)\Gamma N_o W}{|h_{\mathcal{SD}}|^2}. \tag{32}$$

Now, it remains to choose the optimal power allocation for a fixed $R(m,n)$. Similar to the analysis in the DF mode, the per-tone objective is as expressed in (11):

$$\max_{P(\mathcal{S},n)} \lambda_m R(m,n) - \mu_{\mathcal{S}} P(\mathcal{S},n) - \mu_{\mathcal{R}} P^*(\mathcal{R},n) \tag{33}$$

Let's call the objective function above f. First, we show that f is a concave function of $P(\mathcal{S},n)$. Compute

$$\frac{df}{dP(\mathcal{S},n)} = -\frac{\mu_{\mathcal{S}}}{2} - \frac{\mu_{\mathcal{R}}}{2}\frac{dP^*(\mathcal{R},n)}{dP(\mathcal{S},n)}, \tag{34}$$

$$\frac{d^2 f}{dP(\mathcal{S},n)^2} = -\frac{\mu_{\mathcal{R}}}{2}\frac{d^2 P^*(\mathcal{R},n)}{dP(\mathcal{S},n)^2} \tag{35}$$

It is not difficult to verify by algebra that for $P(\mathcal{S},n)$ within the feasible region (30), $\frac{dP^*(\mathcal{R},n)}{dP(\mathcal{S},n)} < 0$ and $\frac{d^2 P^*(\mathcal{R},n)}{dP(\mathcal{S},n)^2} > 0$. Substituting the result into (34) and (35) tells us that $\frac{d^2 f}{dP(\mathcal{S},n)^2} < 0$ within the feasible region of $P(\mathcal{S},n)$, thus proving concavity.

The concavity of f and the observation that f is continuous ensure that there is a unique optimal value of f within the feasible region for $P(\mathcal{S},n)$. The optimal value can be found by solving for the root of $\frac{df}{dP(\mathcal{S},n)}$. This can be done with a root-finding method such as the Newton's method. The AF mode power allocation procedure is summarized below:

Subroutine 4: The best power allocation for a fixed $R(m,n)$ in the AF relay mode:

1) Solve the equation $\frac{df}{dP(\mathcal{S},n)} = 0$ to obtain $P^*(\mathcal{S},n)$, using either bisection or a Newton's method, within the feasibility range (30).
2) Set $P^*(\mathcal{R},n)$ according to of (29).
3) Divide $R(m,n)$, $P^*(\mathcal{S},n)$, and $P^*(\mathcal{R},n)$ by 2. □

D. Summary of the Algorithm

The subroutines presented in this paper are interconnected hierarchically. The original sum utility problem can be solved optimally in the dual domain using Subroutine 1, which requires the solutions to both the application-layer subproblem (which is trivial) and the physical-layer subproblem (which requires Subroutine 2.) An important step of Subroutine 2 is the solution to N per-tone maximization problems. Each per-tone problem can be solved efficiently by searching over a discrete set. The discrete search involves the expressions of the required power at the source and at the relay as a function of the bit rate for different relaying schemes. Subroutines 3 and 4 describe the associated procedures for DF and AF relaying respectively. Because of the convexity of $\mathcal{C}(\boldsymbol{P})$ and the fact

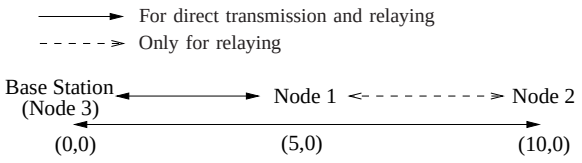

Fig. 1. Topology of a cooperative network with 2 user nodes.

that the utility maximization problem has zero duality gap, we have:

Theorem 1: The algorithm summarized in the preceding paragraph always converges to the global optimal value of (1). This is true whenever $\mathcal{C}(\boldsymbol{P})$ is convex, which is obtained in the limit as the number of OFDM tones goes to infinity.

IV. Simulations

This section illustrates the proposed joint resource allocation and relay-strategy selection procedure by providing simulation results for the following two networks. In both cases, the total system bandwidth is set to 80MHz, with the number of OFDM tones $N = 256$. All links in the networks are assumed to have small-scale i.i.d. Rayleigh fading and a large-scale path loss, with path loss exponent of 4. Utility functions are chosen as described in Example 1 with parameters $a = 10$ and $c = 125$Mbps for downlinks and $a = 1$ and $c = 12.5$Mbps for uplinks.

A. Network with 2 User Nodes

Consider a network with a base station and $K = 2$ user nodes, for a total of 3 nodes. The base station (node 3) is fixed at (0,0) and node 2 is fixed at (10,0). The location of node 1 changes in different parts of this simulation. The power constraints of all nodes are the same such that $\frac{p_i^{max}}{N_o W} = 23$dB, corresponding to a medium SNR environment. It is clear that relaying is most beneficial to the second user. Moreover, simulation shows that if one restricts relaying for the downlink of the second user $(2, d)$ only, a negligible decrease in the maximum system utility is observed. This is because our choice of utility functions favor downlink transmission. Therefore, we show results for the relaying of $(2, d)$ only.

1) Fix Node 1 at $(5, 0)$: As a first example, node 1 is fixed at $(5, 0)$. Fig. 1 shows the locations of nodes and describes whether a link can be used for direct transmission, relaying transmission, or both. It is found that by allowing relaying, the maximum sum utility is 34.4% closer to the upper limit (which is 22, calculated by summing the value of the parameter a for all data streams.) This quantifies the merit of relaying.

Table I compares the achievable rates for various data streams. Each user node spends part of its power to transmit its own uplink data, and the rest of its power to act as relay. The optimization technique proposed in this paper allows each user node to find the optimal power and bandwidth division between the two roles. Table I suggests that node 1 sacrifices its own uplink data rate in return for higher data rates for node 2. In fact, node 1 spends 47.6% of its power in the relay mode.

TABLE I

Rates for various data streams in the 2-user network.

Stream	No Relay	Allow Relay	Percentage Change
$(1, d)$	130.0Mbps	115.9Mbps	-10.8%
$(2, d)$	50.8Mbps	88.8Mbps	74.8%
$(1, u)$	27.0Mbps	19.4Mbps	-28.2%
$(2, u)$	16.2Mbps	15.8Mbps	-2.5%

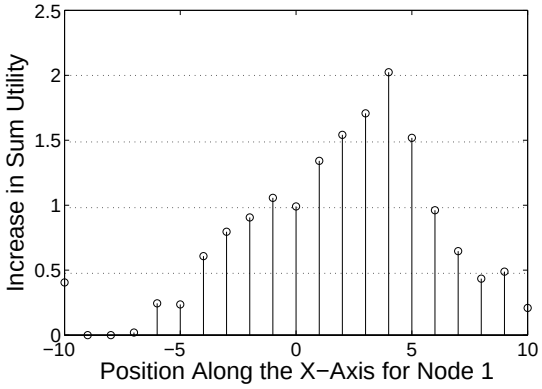

Fig. 2. The effectiveness of relaying for various positions of node 1.

2) Effectiveness of Relaying vs User Location: Fig. 2 shows the increase in sum utility for various position of node 1. Increase in performance is most significant when node 1 is at $(4, 0)$, and gradually decreases in both directions.

3) Relaying Scheme vs User Location: Fig. 3 shows the dominant relaying scheme(s) for stream $(2, d)$ for different positions of node 1. The results of Fig. 3 nicely follow the rule-of-thumb that when $\mathcal{R}$ can decode the received data, DF is the preferred relaying scheme. As $\mathcal{R}$ moves further away from $\mathcal{S}$, AF relaying scheme is preferred. However, the simple rule-of-thumb, without taking into consideration other factors, does not tell us where the change of DF to AF occur, whereas Fig. 3 shows exactly where the transition occurs.

B. Network with 4 User Nodes

The second set of simulations illustrate the optimal relay strategy for a larger network with a base station and $K = 4$ user nodes. Fig. 4 shows the locations of different nodes and describes possible relaying links. Note that in this example, both relays 1 and 2 can potentially help downlink transmissions for nodes 3 and 4. The proposed optimization procedure selects the best relay in accordance with the realization of the channel and the availability of power and bandwidth.

The power constraints of all nodes are such that $\frac{p_i^{max}}{N_o W} = 20$dB, $(i \in \mathcal{K}_+)$. This corresponds to a medium SNR environment. Using the proposed optimization procedure, it is found that by allowing relaying, the maximized sum utility increases from 34.38 to 37.20, which is 29.3% closer to the upper limit of 44. The result again quantifies the merit of cooperative relaying. Note that in this simulation, each source-destination pair selects both the best relay and the best relay strategy in each frequency tone. This is done in a globally optimal way.

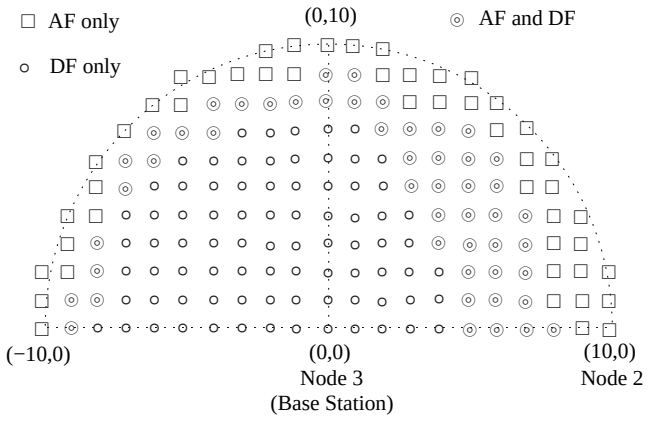

Fig. 3. Dominating relay scheme(s) for stream $(2, d)$ for various positions of node 1. ("AF and DF" indicates that both AF and DF are possible at that location, depending on resource constraints and channel realization at a tone.)

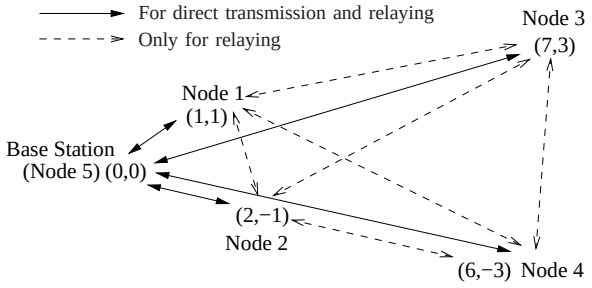

Fig. 4. Topology of a cooperative network with 4 user nodes.

Table II shows the rates for various data streams, for both when relay modes are allowed and when they are not. Again, as downlink is preferred by the virtue of the choice of the utility function, both $(3, d)$ and $(4, d)$ streams benefit tremendously from relaying. In fact, as shown in Table III, over 90% of power is used for relaying in nodes 1 and 2. These results illustrate that in a system with optimal allocation of bandwidth and power that truly maximizes the sum utility, nodes 1 and 2 would sacrifice its own data rates for the benefit of nodes 3 and 4.

V. Conclusion

This paper proposes a utility maximization framework that is capable of selecting the best relay, the best relay-strategy, and the best power, bandwidth and rate allocation in a cellular network with relays. By using a dual optimization technique for OFDMA systems, we show that the seemingly difficult joint system optimization problem can be solved efficiently and globally under a pricing structure. The proposed resource allocation scheme realizes the cooperative gain of a relay network by taking into account both physical-layer resource availability and the application-layer user traffic demands in a cross-layer approach.

TABLE II

Rates for various data streams in the 4-user network.

Stream	No Relay	Allow Relay	Percentage Change
$(1, d)$	152.8Mbps	148.4Mbps	-2.9%
$(2, d)$	135Mbps	129.4Mbps	-4.1%
$(3, d)$	46.6Mbps	71.3Mbps	53.0%
$(4, d)$	54.1Mbps	80.5Mbps	48.8%
$(1, u)$	18.8Mbps	16.6Mbps	-11.7%
$(2, u)$	18.8Mbps	16.3Mbps	-13.3%
$(3, u)$	11.9Mbps	13.8Mbps	16.0%
$(4, u)$	14.1Mbps	13.4Mbps	-5.0%

TABLE III

Percentage of power spent as relay in the 4-user network.

Node	Percentage of power spent as relay
1	94.9%
2	92.2%
3	0%
4	0%

References

[1] R. U. Nabar, H. Bölcskei, and F. W. Kneubühler, "Fading relay channels: Performance limits and space-time signal design," *IEEE J. Select. Areas Commun.*, Aug. 2004.

[2] J. N. Laneman, D. N. C. Tse, and G.W.Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[3] A. Host-Madsen and J. Zhang, "Capacity bounds and power allocation for wireless relay channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 2020–2040, June 2005.

[4] M. Chen, S. Serbetli, and A. Yener, "Distributed power allocation for parallel relay networks," in *IEEE Global Telecommun. Conf. (GLOBECOM'05)*, vol. 3, Nov. 2005, pp. 1177–1181.

[5] V. Sreng, H. Yanikomeroglu, and D. Falconer, "Relay selection strategies in cellular networks with peer-to-peer relaying," in *IEEE Veh. Tech. Conf. (VTC'03-Fall)*, Oct. 2003, pp. 1949–1953.

[6] X. Cai, Y. Yao, and G. B. Giannakis, "Achievable rates in low-power relay links over fading channels," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 184–194, Jan. 2005.

[7] M. Yu and J. Li, "Is amplify-and-forward practically better than decode-and-forward or vice versa?" in *IEEE Inter. Conf. Acoustics, Speech, and Signal Processing, (ICASSP '05)*, vol. 3, Mar. 2005, pp. 365–368.

[8] K. Chen, Z. Yang, C. Wagener, and K. Nahrstedt, "Market models and pricing mechanisms in a multihop wireless hotspot network," in *The Second Annual Inter. Conf. on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous 2005)*, July 2005, pp. 73–82.

[9] O. Ileri, S.-C. Mau, and N. Mandayam, "Pricing for enabling forwarding in self-configuring ad hoc networks," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, pp. 151–162, Jan. 2005.

[10] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of Operations Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[11] G. Song and Y. G. Li, "Cross-layer optimization for OFDM wireless networks – Part I: Theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[12] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, 2006, to appear.

[13] N. Z. Shor, *Minimization Methods for Non-differentiable Functions*. Springer, 1985.

[14] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," in lecture notes of EE392o, Stanford University, Autumn Quarter 2003-2004.

[15] R. Cendrillon, W. Yu, M. Moonen, J. Verlinden, and T. Bostoen, "Optimal spectrum balancing for digital subscriber lines," *IEEE Trans. Commun.*, 2006, to appear.

[16] M. A. Khojastepour, A. Sabharwal, and B. Aazhang, "On the capacity of 'cheap' relay networks," in *Proc. 37th Annu. Conf. Information Sciences and Systems (CISS)*, Baltimore, MD, Mar. 2003, pp. 12–14.