Fair Scheduling and Resource Allocation for Wireless Cellular Network with Shared Relays

Yicheng Lin, *Student Member, IEEE,* and Wei Yu, *Senior Member, IEEE*

***Abstract*—This paper examines the shared relay architecture for the wireless cellular network, where instead of deploying multiple separate relays within each cell sector, a single relay with multiple antennas is placed at the cell edge and is shared by multiple sectors. The advantage of shared relaying is that the joint processing of signals at the relay enables the mitigation of intercell interference. To maximize the benefit of shared relaying, the resource allocation and the scheduling of users among adjacent cell sectors need to be optimized jointly. Based on this motivation, this paper formulates a network utility maximization problem for the shared relay system that considers the practical wireless backhaul constraint of matching the relay-to-user rate demand with the base-station-to-relay rate supply using a set of pricing variables. In addition, zero-forcing beamforming is used at the shared relay to separate users spatially; multiple users are scheduled in the frequency domain to maximize frequency reuse. A heuristic but efficient scheduling and resource allocation algorithm is proposed accordingly. System-level simulations quantify the effectiveness of the proposed approach, and show that the incorporation of the shared relay can improve the overall network performance and in particular significantly increase the throughput of cell edge users as compared to separate relaying.**

***Index Terms*—Cellular systems, shared relay, wireless backhaul, orthogonal frequency-division multiplex (OFDM), scheduling, proportional fairness, zero-forcing (ZF) beamforming.**

I. Introduction

MODERN cellular networks need to provide ubiquitous coverage and high data rates with low infrastructure deployment cost [1]. The incorporation of two-hop fixed relays, which are connected to base-stations via wireless backhaul, provides a convenient solution toward such a goal. Although fixed relays can be deployed to enhance coverage at the cell edge, they are typically not designed with intercell interference in mind. In fact, cell-edge relays from different neighboring cell sectors are often close to each other in distance, and consequently can create more intercell interference than that in a conventional cellular network. One way to tackle this intercell interference problem is to introduce the concept of the coordinated shared relay. The basic idea, which was firstly proposed in [2] and [3], is to place a multi-antenna relay at the intersection of adjacent sectors, which can be thought of as a coordinated version of multiple separate relays from different sectors.

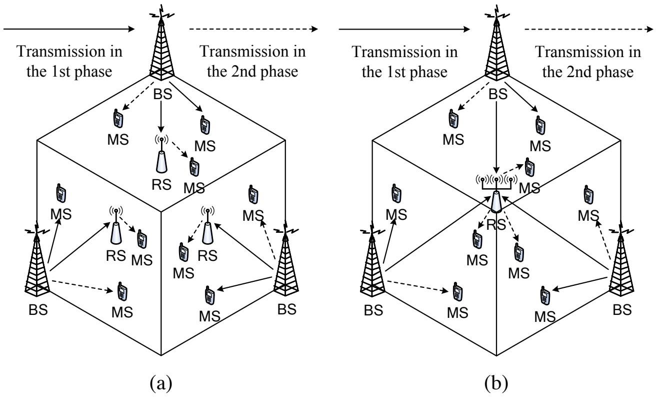

Transmission in the 1st phase
Transmission in the 2nd phase
BS
MS
RS
(a)
Transmission in the 1st phase
Transmission in the 2nd phase
BS
MS
RS
(b)

Fig. 1. Relay scenarios. (a) Separate relays at edge of each sector. (b) Coordinated relay shared by three adjacent sectors.

Fig. 1 illustrates a cellular relay network with hexagonal layout, where three downlink users are served in three adjacent cell sectors respectively. Fig. 1(a) shows a separate relay architecture. Fig. 1(b) shows a shared relay architecture, where a shared relay is deployed at the intersection of three sectors, providing coverage to three users simultaneously. The shared relay is capable of maintaining connections to multiple base-stations by resolving control messages from them, and acquiring the channel knowledge of the base-station-to-relay and relay-to-user links. In the downlink, the shared relay receives signals from all of its donor base-stations in adjacent sectors, and spatially separate these signals via receive beamforming. After receiving the data packets for each user, the shared relay then retransmits the decoded signals to multiple users via spatial multiplexing using transmit beamforming techniques. As compared to separate relaying, although the shared relay is placed further away from the base-station, its interference mitigation capability can potentially compensate the increased base-station-to-relay distance, leading to an improved overall network performance.

Shared relaying has a clear advantage for cell edge users, which would otherwise suffer from severe intercell interference. To truly quantify the benefit of shared relaying for the entire network, it is also important to evaluate the network performance from a system-level perspective. Toward this end, this paper focuses on the scheduling and resource allocation aspects of shared relaying, while adopting the following assumptions and definitions:

- Downlink transmission with a half-duplex decode-and-forward strategy [4] is assumed.
- The base-station-to-user and relay-to-user links can both be used to schedule users, and they are called the *access*

Manuscript received 1 August 2011; revised 1 May 2012. This paper has been presented in part at the 46th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, March 2012.

The authors are with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario M5S 3G4, Canada (e-mail: ylin@comm.utoronto.ca; weiyu@comm.utoronto.ca).

Digital Object Identifier 10.1109/JSAC.2012.120920.

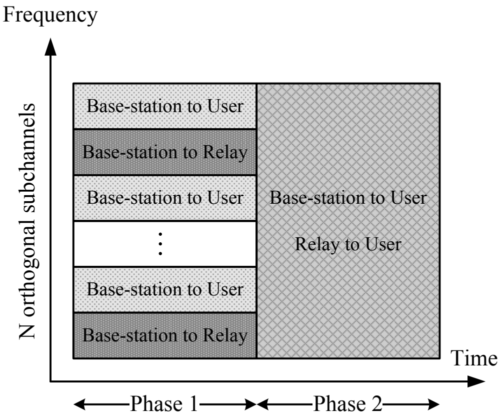


Fig. 2. Half-duplex two-phase relay frame structure

links. The base-station-to-relay links, which are called the *feeder links*, provide the wireless backhaul connection.

- Users can choose either direct transmission from the base-station or indirectly via the relay. Depending on this routing choice, the users are classified as *one-hop users* or *two-hop users*, respectively.
- The half-duplex multi-channel frame structure in Fig. 2 is adopted. In the first phase, base-station transmits to the users and to the relay on orthogonal subchannels. In the second phase, the base-station and the relay simultaneously transmit to separate sets of users on all subchannels to maximize frequency reuse. The reuse of frequencies by all serving nodes in the second phase inevitably induces more interference, which heightens the importance of scheduling and resource allocation.
- Perfect channel estimation is assumed in every time slot for both phases across subchannels.

The main contribution of this paper is as follows. We adopt a network utility maximization framework with a *proportionally fair* (PF) objective [5], and design a heuristic but efficient resource allocation and scheduling algorithm to address questions such as how the frequencies should be allocated among the different links in an OFDM system, and how the frequencies should be reused within each sector. We characterize how much performance gain can be obtained from shared relaying as compared to separate relaying for the entire network from a system-level perspective.

A. Related Work

There have only been a limited number of works in the literature on the shared relay architecture after the initial qualitative description of the concept in IEEE 802.16m [2], [3]. Most notably, [6] shows that shared relaying can approach the gains of local base-station coordination at reduced complexity. In [6], multiple-input multiple-output (MIMO) multiple-access and broadcast techniques are used at the shared relay; the time durations of the two phases are optimally adjusted. The shared relay concept is further studied in [7], where practical zero-forcing (ZF) methods are used in combination with partial or full base-station coordination for both one-way and two-way shared relaying. In [8], a joint processing scheme that improves the shared relaying strategy of [6] is proposed by letting the base-station and the relay send the same message to the corresponding user in the second phase. However, none of these works consider the impact of scheduling: [6] and [7] assume arbitrary scheduling on a single subchannel, while round robin scheduling for cell-edge users is used in [8]. In [9], a Hungarian-based scheduling scheme is proposed for shared relaying under static orthogonal subchannel segmentation among neighboring sectors. The present paper advances this line of work in addressing fairness and dynamic resource allocation issues of the shared relay architecture.

For the conventional separate relay system, resource allocation and scheduling have been studied extensively. In [10], fair resource utilization of relay nodes is considered as an integer optimization problem. The work [11] uses a cross-layer optimization framework for centralized resource allocation of OFDM-based relay networks. However, both [10] and [11] assume that the source and the relay always transmit on the same subchannel, an assumption which is removed in this paper. In [12], [13] the sum-rate maximization problem is formulated with the constraint that the receiving rate (rate supply) and transmitting rate (rate demand) of relays are approximately equal when resources are allocated optimally. Proportional fairness is considered in [14] in formulating the subchannel and rate allocation problem and the relay's rate demand and rate supply constraint is considered on a per-user basis. In [15], [16], a queue-aware resource management algorithm is proposed, and the Hungarian algorithm is used to solve the joint routing and scheduling problem. All of the above works assume that in the second phase the source and relays use orthogonal resource partition, which limits the performance gain. In this paper, a PF scheduling algorithm based on the relay frame structure in Fig. 2 is used for the shared relay system. Differing from [14], our algorithm strikes a balance between the rate demand of the access link for all scheduled users in one sector and the rate supply from the feeder link. Frequency reuse is allowed between the relay and the base-station in the second phase, offering maximum flexibility in the overall design.

B. Organization

The rest of this paper is organized as follows. The optimization framework is presented in Section II. In Sections III and IV, the resource allocation and scheduling problems are formulated, and corresponding algorithms are described for shared relaying and separate relaying, respectively. Section V contains the simulation results that quantify the benefit of shared relaying using the proposed algorithms. Finally, conclusions are drawn in Section VI.

Notation: we use upper-case bold letters (e.g., $\mathbf{I}$) for matrices, and lower-case bold letters (e.g., $\mathbf{w}$) for column vectors. The conjugate transpose and Euclidean norm of vector $\mathbf{w}$ are denoted as $\mathbf{w}^H$ and $\|\mathbf{w}\|$, respectively. Calligraphy letters (e.g., $\mathcal{K}$) are used to denote sets. The subscripts s and r refer to *source* (base-station) and *relay* respectively. In particular, s_m and r_m refer to the base-station and the separate relay in sector m, whereas r without the subscript refers to the shared relay.

II. General Optimization Framework

A. System Model

Consider a cellular network where each cell is divided into $M = 3$ sectors and users are uniformly distributed. The set of users in sector m is denoted as $\mathcal{K}^{(m)}$, which comprises of users that are associated with the source $\mathcal{K}_s^{(m)}$ (or the one-hop users), and users that are associated with the relay $\mathcal{K}_r^{(m)}$ (or the two-hop users), where $\mathcal{K}_s^{(m)} \cup \mathcal{K}_r^{(m)} = \mathcal{K}^{(m)}$. This paper aims to compare the performance of the separate relay system in Fig. 1(a) where one relay is deployed at the boresight of the cell edge in each sector, and that of the shared relay system in Fig. 1(b) where the relay is placed at the intersection of the M adjacent sectors. The base-stations and the mobile users are equipped with a single antenna each. For fair comparison, while the separated relays are assumed to have one antenna each, the shared relay is assumed to have M antennas, covering the cluster of M adjacent coordinated sectors. Only the downlink transmission in the central cluster (the central M adjacent sectors) in Fig. 1 are considered in the mathematical formulation. The out-of-cluster downlink interference is relatively weak as compared to the intra-cluster transmission, but is explicitly modeled in the simulation part.

B. Transmit and Receive Strategies

The multi-antenna shared relay can use MIMO techniques for both transmission and reception. This paper assumes linear MIMO strategies in a half-duplex decode-and-forward structure. In the first (i.e., multiple-access) phase, the shared relay uses the minimum-mean-squared-error (MMSE) receiver to spatially separate the signals from the M base-stations. In the second (i.e., broadcast) phase, zero-forcing (ZF) beamforming is used for transmission from the shared relay to the multiple selected users. Instead of using linear MIMO techniques, it is possible to further improve the link performance with nonlinear processing (e.g., with multiuser detection and dirty-paper coding). Alternative relaying strategies such as amplify-and-forward or compress-and-forward can also be used. We justify our design choice by noting that linear MIMO processing already achieve the maximum degree of freedom; decode-and-forward is sensible as the two-hop users typically have weak source-to-destination links.

In the frame structure shown in Fig. 2, the mobile users may be served directly by the base-station in both phases, or by the relay. In the case where a user is served by the relay, we assume a simple multi-commodity-flow model in which the relay decodes and re-transmits the received bits in the second phase, and the user decodes its message solely based on the signal from the relay, i.e., the user does not combine the received signals in the two phases. This assumption can again be justified by the fact that the two-hop users typically have weak direct links from the base-station.

The signal-to-interference-plus-noise ratios (SINR) for various links in the overall scheme are described as follows.

1) Base-Station-to-Relay Feeder Link: The SINR of the wireless backhaul connecting the base-station m and the shared relay on subchannel n can be expressed as

$$\gamma_{s_m,r}^{(n)} = \frac{P_{s_m}^{(n)} \left| \left(\mathbf{v}_{s_m,r}^{(n)}\right)^H \mathbf{h}_{s_m,r}^{(n)} \right|^2}{\sigma_r^2 \left\| \mathbf{v}_{s_m,r}^{(n)} \right\|^2 + \sum\limits_{j\neq m} P_{s_j}^{(n)} \left| \left(\mathbf{v}_{s_m,r}^{(n)}\right)^H \mathbf{h}_{s_j,r}^{(n)} \right|^2} \qquad (1)$$

where $P_{s_m}^{(n)}$ is the transmit power of base-station m, $\mathbf{h}_{s_m,r}^{(n)} \in \mathbb{C}^{M\times 1}$ is the channel vector from base-station m to the shared relay, $\left(\mathbf{v}_{s_m,r}^{(n)}\right)^H \in \mathbb{C}^{1\times M}$ is the corresponding receive beamformer, and σ_r^2 is the combined out-of-cluster interference and noise at the shared relay. The optimal MMSE receive beamformer at the relay is

$$\mathbf{v}_{s_m,r}^{(n)} = \left(\sigma_r^2 \mathbf{I} + \sum_{j\neq m} P_{s_j}^{(n)} \mathbf{h}_{s_j,r}^{(n)} \left(\mathbf{h}_{s_j,r}^{(n)}\right)^H \right)^{-1} \mathbf{h}_{s_m,r}^{(n)}, \qquad (2)$$

which suppresses mutual interference among M base-stations and maximizes the receiver SINR.

2) Relay-to-Two-Hop-User Access Link: The shared relay schedules up to M two-hop users in the second phase across M sectors per subchannel. Note that the shared relay is not limited to choosing exactly one user per sector, i.e., the relay may schedule multiple users in one sector, and no user in another sector. Let $\mathcal{S}_r^{(n)}$ be the selected user set, so $\left|\mathcal{S}_r^{(n)}\right| \leq M$. To eliminate the mutual interference among users in $\mathcal{S}_r^{(n)}$, ZF beamforming is used due to its simplicity. The SINR of the link between the shared relay and a scheduled two-hop user k on subchannel n is

$$\begin{aligned}\gamma_{r,k}^{(n)} &= \frac{P_{r,k}^{(n)} \left| \left(\mathbf{h}_{r,k}^{(n)}\right)^H \hat{\mathbf{w}}_{r,k}^{(n)} \right|^2}{\sigma_k^2 + \sum\limits_j P_{s_j}^{(n)} \left| h_{s_j,k}^{(n)} \right|^2 + \sum\limits_{j\in\mathcal{S}_r^{(n)}, j\neq k} P_{r,j}^{(n)} \left| \left(\mathbf{h}_{r,k}^{(n)}\right)^H \hat{\mathbf{w}}_{r,j}^{(n)} \right|^2} \\ &\overset{\text{ZF}}{=} \frac{P_{r,k}^{(n)}}{\left\| \mathbf{w}_{r,k}^{(n)} \right\|^2 \left(\sigma_k^2 + \sum\limits_j P_{s_j}^{(n)} \left| h_{s_j,k}^{(n)} \right|^2 \right)}\end{aligned} \qquad (3)$$

where $P_{r,k}^{(n)}$ is the shared relay's power allocation for user k, $\left(\mathbf{h}_{r,k}^{(n)}\right)^H \in \mathbb{C}^{1\times M}$ is the channel vector from the shared relay to user k, $\mathbf{w}_{r,k}^{(n)}$ and $\hat{\mathbf{w}}_{r,k}^{(n)} \in \mathbb{C}^{M\times 1}$ are the ZF transmit beamformer and its normalized version, $h_{s_j,k}^{(n)}$ is the channel response between base-station j and user k, and σ_k^2 is the combined out-of-cluster interference and noise at user k. The second equality in (3) comes from that

$$\left(\mathbf{h}_{r,k}^{(n)}\right)^H \hat{\mathbf{w}}_{r,k}^{(n)} = \left(\mathbf{h}_{r,k}^{(n)}\right)^H \frac{\mathbf{w}_{r,k}^{(n)}}{\left\| \mathbf{w}_{r,k}^{(n)} \right\|} = \frac{1}{\left\| \mathbf{w}_{r,k}^{(n)} \right\|} \qquad (4a)$$

$$\left(\mathbf{h}_{r,k}^{(n)}\right)^H \hat{\mathbf{w}}_{r,j}^{(n)} = 0. \qquad (4b)$$

Note that the intra-cluster interference for the two-hop users comes from the fact that the frequency is maximally reused by the base-station; the interference from the relay is eliminated by ZF beamforming.

3) Base-Station-to-One-Hop-User Access Link: The one-hop users can be scheduled by the base-stations in both phases. The SINR of the link between the base-station m and the one-hop user k on subchannel n in the ith phase is

$$\gamma_{s_m,k}^{(n,i)} = \frac{P_{s_m}^{(n)} \left| h_{s_m,k}^{(n)} \right|^2}{\sigma_k^2 + \sum\limits_{j \neq m} P_{s_j}^{(n)} \left| h_{s_j,k}^{(n)} \right|^2 + \Delta_k^{(n,i)}} \tag{5}$$

where

$$\Delta_k^{(n,i)} = \begin{cases} 0, & i = 1 \\ \sum\limits_{j \in \mathcal{S}_r^{(n)}} P_{r,j}^{(n)} \left| \left(\mathbf{h}_{r,k}^{(n)}\right)^H \hat{\mathbf{w}}_{r,j}^{(n)} \right|^2, & i = 2 \end{cases} \tag{6}$$

for the given beamforming vectors $\hat{\mathbf{w}}_{r,j}^{(n)}$ at the shared relay. Note that the inter-user interference term $\Delta_k^{(n,i)}$ only exists in the second phase when the relay is transmitting. This is due to the reuse of frequencies at the relay and at the neighboring base-stations.

For comparison purposes, we can also formulate the SINR expressions of the links for the separate relaying in a similar fashion. The SINR from base-station m to relay m is denoted as $\gamma_{s_m,r_m}^{(n)}$; the SINR from relay m to two-hop user k is denoted as $\gamma_{r_m,k}^{(n)}$; and the SINR from base-station m to one-hop user k in phase i is denoted as $\gamma_{s_m,k}^{(n,i)\prime}$ (to differentiate from the one-hop user SINR $\gamma_{s_m,k}^{(n,i)}$ in shared relaying in (5)).

The SINRs of each link on each subchannel can be mapped to the corresponding transmission rate by

$$r = \log_2\left(1 + \frac{\gamma}{\Gamma}\right), \tag{7}$$

where $\Gamma = -\ln(5\text{BER})/1.5$ is the SNR gap corresponding to a target bit-error-rate (BER) [17]. Thus, the per-subchannel rate $r_{s_m,r}^{(n)}$, $r_{s_m,r_m}^{(n)}$, $r_{r,k}^{(n)}$, $r_{r_m,k}^{(n)}$, $r_{s_m,k}^{(n,i)}$, and $r_{s_m,k}^{(n,i)\prime}$ can be computed using (7) from their respective SINR formulae.

C. Utility Maximization Framework

The objective is to maximize the total network PF utility defined as [5]:

$$\max \sum_m \sum_{k \in \mathcal{K}^{(m)}} \ln\left(R_k(t)\right) \tag{8}$$

where $R_k(t)$ is the long-term average rate of user k up to time t, which is updated using exponential averaging [18]:

$$R_k(t) = \left(1 - \frac{1}{T}\right) R_k(t-1) + \frac{1}{T} r_k(t) \tag{9}$$

where T is the predefined averaging window size and $r_k(t)$ is user k's instantaneous transmission rate at time t, which is a function of the per-subchannel user rate $r_{r,k}^{(n)}$ and $r_{s_m,k}^{(n,i)}$ for shared relaying, or $r_{r_m,k}^{(n)}$ and $r_{s_m,k}^{(n,i)\prime}$ for separate relaying, defined in the previous section.

Proportional fairness maximization can be implemented in practice using a weighted rate sum maximization formulation (e.g. for the multi-channel system as in [19]):

$$\sum_m \sum_{k \in \mathcal{K}^{(m)}} \alpha_k r_k, \quad \text{where } \alpha_k = \frac{1}{R_k(t-1)}. \tag{10}$$

In this paper, the base-station has a fixed transmit power spectral density (PSD) constraint, and the shared relay has a fixed sum PSD constraint across all its antennas. With multiple users served by the shared relay, its transmit power $P_{r,k}^{(n)}$ can be optimally allocated among all of its scheduled users $k \in \mathcal{S}_r^{(n)}$ on each subchannel. The optimization in this paper can thus be formulated as that of deciding: (a) for base-stations, which user should be scheduled in two phases, and which subchannels should be reserved for wireless backhaul in the first phase; (b) for the shared relay, which users should be scheduled with ZF beamforming, and what the appropriate power level is for each scheduled users.

III. Resource Allocation and Scheduling for Shared Relaying

A. Problem Formulation

Users in each sector m are partitioned into the one-hop set which are associated with the source, $\mathcal{K}_s^{(m)}$, and the two-hop set which are associated with the relay, $\mathcal{K}_r^{(m)}$. The one-hop users can be served in any of the two phases, while two-hop users can only be served in the second phase. The user partitioning process is also known as routing. This paper adopts a simple intra-sector routing metric where users are partitioned based on their received signal strength from the base-stations and the relays.

With a fixed user partition, we can formulate a scheduling and resource allocation problem based on the utility maximization objective (10) for the shared relay system. Define binary indicators $\rho_{s_m,k}^{(n,i)}$ and $\rho_{r,k}^{(n)}$, such that $\rho_{s_m,k}^{(n,i)} = 1$ indicates that the base-station m schedules user k on subchannel n in the ith phase, and $\rho_{r,k}^{(n)} = 1$ indicates that the shared relay schedules user k on subchannel n. The one-hop users are served by the base-station in either or both of the two phases, but only one user is scheduled in each subchannel in every sector in each phase, i.e.,

$$\sum_{k \in \mathcal{K}_s^{(m)}} \rho_{s_m,k}^{(n,i)} \leq 1, \ \ \rho_{s_m,k}^{(n,i)} \in \{0,1\}, \ \ \forall m, n, i \in \{1,2\}. \tag{11}$$

The two-hop users are served by the relay; a maximum of M users are served at the same time in each subchannel using spatial multiplexing, i.e.,

$$\sum_{m=1}^{M} \sum_{k \in \mathcal{K}_r^{(m)}} \rho_{r,k}^{(n)} \leq M, \quad \rho_{r,k}^{(n)} \in \{0,1\}, \quad \forall n. \tag{12}$$

The user rate in (10) can now be expressed in terms of the indicator variables as

$$r_k = \begin{cases} \sum\limits_{n=1}^{N} \sum\limits_{i=1}^{2} \rho_{s_m,k}^{(n,i)} r_{s_m,k}^{(n,i)}, & k \in \mathcal{K}_s^{(m)} \\ \sum\limits_{n=1}^{N} \rho_{r,k}^{(n)} r_{r,k}^{(n)}, & k \in \mathcal{K}_r^{(m)}. \end{cases} \tag{13}$$

Note that the actual user rate is half of the above one due to the half-duplex loss; the $1/2$ factor is not included here for simplicity since it would not change the optimization result.

We assume that the data received by the relay in each time instance must be forwarded to the users at the next period, with no possibility of buffering at the relay. This assumption is valid for delay-sensitive services. At each time frame, we have the following wireless backhaul constraint: the total rate demand of the shared relay for all of its serving users in sector m in the access link, denoted as $R_{r,d}^{(m)}$, should be no larger than the total rate supply in the feeder link from the base-station m to this relay, denoted as $R_{s,r}^{(m)}$, i.e.,

$$\begin{aligned} R_{r,d}^{(m)} &\triangleq \sum_{k\in\mathcal{K}_r^{(m)}} \sum_{n=1}^{N} \rho_{r,k}^{(n)} r_{r,k}^{(n)} \\ &\le \sum_{n=1}^{N} \left(1 - \sum_{k\in\mathcal{K}_s^{(m)}} \rho_{s_m,k}^{(n,1)} \right) r_{s_m,r}^{(n)} \triangleq R_{s,r}^{(m)}, \ \ \forall m. \quad (14) \end{aligned}$$

Note that in the first phase, if a subchannel is used for the wireless backhaul transmission, then it is not used for the scheduling of one-hop users of the base-station, i.e., $\rho_{s_m,k}^{(n,1)} = 0, \forall k$.

The base-stations transmit PSD is assumed to be fixed. For the shared relay, the total allocated power for its scheduled users on any subchannel should be bounded by its PSD constraint:

$$\sum_{m=1}^{M} \sum_{\left\{k\in\mathcal{K}_r^{(m)} \middle| \rho_{r,k}^{(n)}=1\right\}} P_{r,k}^{(n)} \le P_r^{\max}, \quad \forall n. \quad (15)$$

We consider the adjacent M sectors, whose resource allocation is jointly coordinated by the shared relay. The resource allocation problem can now be reformulated from (10) as

$$\max_{\boldsymbol{\rho},\mathbf{w},\mathbf{P}} \ \sum_{m=1}^{M} \left(\sum_{k\in\mathcal{K}_s^{(m)}} \alpha_k r_k + \sum_{k\in\mathcal{K}_r^{(m)}} \alpha_k r_k \right), \quad (16)$$

subject to the constraints (11), (12), (14), and (15), where r_k is defined in (13) and α_k is the weight in (10). The maximization is over the variables of the scheduling indicators $\boldsymbol{\rho} = \left\{ \left\{ \rho_{s_m,k}^{(n,i)} \right\}_{m,k,n,i}, \left\{ \rho_{r,k}^{(n)} \right\}_{k,n} \right\}$, the relay's normalized beamforming vectors $\mathbf{w} = \left\{ \hat{\mathbf{w}}_{r,k}^{(n)} \right\}_{k,n}$ and its power allocation $\mathbf{P} = \left\{ P_{r,k}^{(n)} \right\}_{k,n}$.

B. Resource Allocation and Scheduling

The first step for solving the problem is to write the Lagrangian of (16) with respect to the wireless backhaul constraints (14) for all M sectors:

$$\begin{aligned} g(\boldsymbol{\rho},\mathbf{w},\mathbf{P},\boldsymbol{\lambda}) = &\sum_{m=1}^{M} \left(\sum_{k\in\mathcal{K}_s^{(m)}} \alpha_k r_k + \sum_{k\in\mathcal{K}_r^{(m)}} \alpha_k r_k \right) \\ &+ \sum_{m=1}^{M} \lambda^{(m)} \left(R_{s,r}^{(m)} - R_{r,d}^{(m)} \right) \quad (17) \end{aligned}$$

where $\lambda^{(m)}$ is the dual variable representing the wireless backhaul price in sector m, which coordinates the rate demand $R_{r,d}^{(m)}$ and the rate supply $R_{s,r}^{(m)}$ of the shared relay as computed in (14). Plugging (13) and (14) into (17), we have

$$\begin{aligned} &g(\boldsymbol{\rho},\mathbf{w},\mathbf{P},\boldsymbol{\lambda}) \\ &= \sum_{n=1}^{N} \sum_{m=1}^{M} \left\{ \sum_{k\in\mathcal{K}_s^{(m)}} \left[\rho_{s_m,k}^{(n,1)} A_k^{(m,n)} + \rho_{s_m,k}^{(n,2)} B_k^{(m,n)} \right] \right. \\ &\quad \left. + \sum_{k\in\mathcal{K}_r^{(m)}} \rho_{r,k}^{(n)} C_k^{(m,n)} \right\} + \sum_{m=1}^{M} \lambda^{(m)} \sum_{n=1}^{N} r_{s_m,r}^{(n)} \quad (18) \end{aligned}$$

where $A_k^{(m,n)}$, $B_k^{(m,n)}$, and $C_k^{(m,n)}$ are:

$$A_k^{(m,n)} = \alpha_k r_{s_m,k}^{(n,1)} - \lambda^{(m)} r_{s_m,r}^{(n)} \quad (19a)$$

$$B_k^{(m,n)} = \alpha_k r_{s_m,k}^{(n,2)} \quad (19b)$$

$$C_k^{(m,n)} = \left(\alpha_k - \lambda^{(m)} \right) r_{r,k}^{(n)}. \quad (19c)$$

The Lagrangian function (18) can now be decoupled into per-subchannel maximization subproblems where the scheduling indicators are set to be the users with the maximum positive value of $A_k^{(m,n)}$, $B_k^{(m,n)}$ and $C_k^{(m,n)}$.

The term $B_k^{(m,n)}$ is independent of $\lambda^{(m)}$ (unlike $A_k^{(m,n)}$ and $C_k^{(m,n)}$). Consequently the user scheduling at the base-station in the second phase is straightforward, while the scheduling at the base-station in the first phase and that at the relay require iterative search of the backhaul price $\lambda^{(m)}$. However, because of the frequency reuse between the base-station and the relay in the second phase, $B_k^{(m,n)}$ depends on the ZF beamformers implicitly included in $C_k^{(m,n)}$, which have impact on the scheduling at the base-station. The overall scheduling rule for the base-station and the relay is outlined below:

Algorithm 1: User scheduling and resource allocation for the shared relay system in each subchannel n for given $\lambda^{(m)}$'s. Return $\rho_{s_m,k}^{(n,1)}$, $\rho_{s_m,k}^{(n,2)}$, $\rho_{r,k}^{(n)}$, $\hat{\mathbf{w}}_{r,k}^{(n)}$, and $P_{r,k}^{(n)}$.

(a) *Base-station m in the first phase:* Select the user $\hat{k} = \arg\max_{k\in\mathcal{K}_s^{(m)}} \left\{ \alpha_k r_{s_m,k}^{(n,1)} \right\}$. If $\alpha_{\hat{k}} r_{s_m,\hat{k}}^{(n,1)} > \lambda^{(m)} r_{s_m,r}^{(n)}$, set $\rho_{s_m,\hat{k}}^{(n,1)} = 1$ and $\rho_{s_m,k}^{(n,1)} = 0$ for $k \neq \hat{k}$; otherwise this subchannel is used for relay feeder link and $\rho_{s_m,k}^{(n,1)} = 0$ for all k.

(b) *Shared Relay in the second phase:* User scheduling at the shared relay is jointly considered with ZF beamforming and power allocation on every subchannel. First, the relay schedules up to M users in all M adjacent sectors with the maximum positive values of $\left(\alpha_k - \lambda^{(m)} \right) r_{r,k}^{(n)}$. (If fewer than M users satisfy $\alpha_k \ge \lambda^{(m)}$, then fewer than M users are scheduled.) For the scheduled users, ZF transmit beamforming is employed at the relay. Finally, the transmit powers of the scheduled users are optimized to further increase the weighted sum rate. This step is explained in more detail below.

(c) *Base-station m in the second phase:* Select the user $\hat{k} = \arg\max_{k\in\mathcal{K}_s^{(m)}} \left\{ \alpha_k r_{s_m,k}^{(n,2)} \right\}$. Set $\rho_{s_m,\hat{k}}^{(n,2)} = 1$ and $\rho_{s_m,k}^{(n,2)} = 0$ for $k \neq \hat{k}$.

Part (a) of *Algorithm 1* resolves the competition of resources between the one-hop users and the relay feeder link in the first phase for a fixed $\lambda^{(m)}$. Parts (b) and (c) determine how users are scheduled in the second phase. This paper assumes that the transmit PSD of the single-antenna base-stations are fixed, while the shared relay can allocate power across the beamforming vectors. Thus the relay-to-user rate $r_{r,k}^{(n)}$ contains a fixed level of interference from the base-stations, whereas the base-station-to-user rate $r_{s_m,k}^{(n,2)}$ is a function of the relay power allocation $P_{r,k}^{(n)}$ and beamforming vector $\hat{\mathbf{w}}_{r,k}^{(n)}$ according to (5). Consequently, part (b) of *Algorithm 1* needs to be executed before part (c).

Part (b) of *Algorithm 1* involves selecting the users and finding the ZF beamformers and power allocations to maximize the weighted sum rate with weights $\alpha_k - \lambda^{(m)}$. This is a conventional multiuser MIMO problem. Although the global optimum solution for such a problem is not easy to find, many practical but suboptimal solutions exist [20], [21]. This paper adapts the heuristic approach of [21]; the algorithmic details are presented below.

Let $\hat{\mathcal{K}}_r^{(m)} = \left\{k \in \mathcal{K}_r^{(m)} \middle| \alpha_k > \lambda^{(m)}\right\}$ be the candidate set of users for relay scheduling in sector m, as the selected users should have positive values for $C_k^{(m,n)}$. The shared relay's scheduled user in sector m on subchannel n is denoted as $\mathcal{S}_r^{(m,n)} = \left\{k \in \hat{\mathcal{K}}_r^{(m)} \middle| \rho_{r,k}^{(n)} = 1\right\}$. Let $\mathcal{S}_r^{(n)} = \bigcup_m \mathcal{S}_r^{(m,n)}$. The algorithm consists of three steps.

1) User Subset Selection: The basic idea is to ensure semiorthogonality among the selected users [21]. The per-user weighted rate can be approximated, and users can be selected in each step in a greedy manner on each subchannel. The relay selects up to M users. Let $\hat{k}(l)$ be the selected user in the lth step. A sketch of the scheduling process in subchannel n is as follows:

i) Start with $l = 1$. For each sector m, initialize $\hat{\mathcal{K}}_r^{(m)} = \left\{k \in \mathcal{K}_r^{(m)} \middle| \alpha_k > \lambda^{(m)}\right\}$ and $\mathcal{S}_r^{(m,n)} = \emptyset$.

ii) For each user $k \in \bigcup_m \hat{\mathcal{K}}_r^{(m)}$, find the orthogonal component of its channel vector $\mathbf{h}_{r,k}^{(n)}$ projected to the subspace spanned by $\left\{\mathbf{g}_{r,k(1)}^{(n)}, \ldots, \mathbf{g}_{r,k(l-1)}^{(n)}\right\}$ as $\mathbf{g}_{r,k}^{(n)} = \left(\mathbf{I} - \sum_{l'=1}^{l-1} \frac{\mathbf{g}_{r,k(l')}^{(n)}\left(\mathbf{g}_{r,k(l')}^{(n)}\right)^H}{\left\|\mathbf{g}_{r,k(l')}^{(n)}\right\|^2}\right)\mathbf{h}_{r,k}^{(n)}$. If $l = 1$, $\mathbf{g}_{r,k}^{(n)} = \mathbf{h}_{r,k}^{(n)}$.

iii) Use $\mathbf{g}_{r,k}^{(n)}$ to approximate the per-user rate. The suboptimal user selection is: $k(l) = \arg\max_{1\le m\le M} \max_{k\in\hat{\mathcal{K}}_r^{(m)}} \left(\alpha_k - \lambda^{(m)}\right)\log_2\left(1 + \frac{P_r^{\max}}{M}\frac{\left\|\mathbf{g}_{r,k}^{(n)}\right\|^2}{\Gamma\left(\sigma_k^2 + \sum_j P_{s_j}^{(n)}\left|h_{s_j,k}^{(n)}\right|^2\right)}\right)$.

iv) If $k(l) \in \hat{\mathcal{K}}_r^{(m)}$, update $\mathcal{S}_r^{(m,n)} = \mathcal{S}_r^{(m,n)} \cup \{k(l)\}$, update the candidate user set in sector m as $\hat{\mathcal{K}}_r^{(m)} = \left\{k \in \hat{\mathcal{K}}_r^{(m)}, k \neq k(l) \middle| \frac{\left|\left(\mathbf{h}_{r,k}^{(n)}\right)^H \mathbf{g}_{r,k(l)}^{(n)}\right|}{\left\|\mathbf{h}_{r,k}^{(n)}\right\|\left\|\mathbf{g}_{r,k(l)}^{(n)}\right\|} < \delta\right\}$, and $l = l + 1$.

v) The user selection ends when $\bigcup_m \hat{\mathcal{K}}_r^{(m)} = \emptyset$ or $\left|\bigcup_m \mathcal{S}_r^{(m,n)}\right| = M$. Finally output $\mathcal{S}_r^{(n)} = \bigcup_m \mathcal{S}_r^{(m,n)}$, and set $\rho_{r,k}^{(n)} = 1$ for all $k \in \mathcal{S}_r^{(n)}$ and other $\rho_{r,k}^{(n)} = 0$.

Note that δ is a small positive constant which force semiorthogonality between the lth selected user and the previously selected users [21]. The complexity of the user selection is upper bounded by $M\left|\bigcup_m \hat{\mathcal{K}}_r^{(m)}\right|$ times the sum complexity of one matrix multiplication, one 2-norm calculation, and one inner product calculation. The overall complexity is much lower than that required for an exhaustive search over all possible user sets [21].

2) ZF Beamforming: With the user scheduling fixed, the ZF beamformer for the shared relay can be computed. For all scheduled users in the set $\mathcal{S}_r^{(n)}$, stack their channel vectors $\left(\mathbf{h}_{r,k}^{(n)}\right)^H$ to form $\mathbf{H}_{\mathcal{S}_r^{(n)}} = \left[\cdots \mathbf{h}_{r,k}^{(n)} \cdots\right]^H$. The beamforming matrix that gives zero inter-user interference for the shared relay is $\mathbf{W}_{\mathcal{S}_r^{(n)}} = \mathbf{H}_{\mathcal{S}_r^{(n)}}^{\dagger} = \mathbf{H}_{\mathcal{S}_r^{(n)}}^H\left[\mathbf{H}_{\mathcal{S}_r^{(n)}}\mathbf{H}_{\mathcal{S}_r^{(n)}}^H\right]^{-1} = \left[\cdots \mathbf{w}_{r,k}^{(n)} \cdots\right]$, where $\mathbf{w}_{r,k}^{(n)} \in \mathbb{C}^{M\times 1}$. Finally, the normalized transmit beamformer for the scheduled user $k \in \mathcal{S}_r^{(n)}$ is $\hat{\mathbf{w}}_{r,k}^{(n)} = \mathbf{w}_{r,k}^{(n)} / \left\|\mathbf{w}_{r,k}^{(n)}\right\|$.

3) Power Allocation: Given the user scheduling and beamforming, and since $r_{r,k}^{(n)} = \log_2\left(1 + \gamma_{r,k}^{(n)}/\Gamma\right)$ where $\gamma_{r,k}^{(n)}$ is given in (3), the optimal power allocation can be calculated by water-filling, but modified by the weights. For a user $k \in \mathcal{S}_r^{(m,n)}$, the optimal transmit power $P_{r,k}^{(n)}$ is

$$P_{r,k}^{(n)} = \left[\frac{\alpha_k - \lambda^{(m)}}{\mu\ln(2)} - \left\|\mathbf{w}_{r,k}^{(n)}\right\|^2\Gamma\left(\sigma_k^2 + \sum_j P_{s_j}^{(n)}\left|h_{s_j,k}^{(n)}\right|^2\right)\right]_+ \tag{20}$$

where $[x]_+ = \max(x, 0)$ and the water level μ is chosen to satisfy the power constraint (15).

C. Update of the Lagrangian Dual Variables

Now we need to find the Lagrangian price $\boldsymbol{\lambda} = \left\{\lambda^{(m)}\right\}_{m=1}^M$ for the dual function

$$q(\boldsymbol{\lambda}) = \max_{\boldsymbol{\rho},\mathbf{w},\mathbf{P}} g(\boldsymbol{\rho}, \mathbf{w}, \mathbf{P}, \boldsymbol{\lambda}) \tag{21}$$

such that the wireless backhaul constraints (14) for all the sectors that share the relay are satisfied, and the dual objective (21) is minimized.

Intuitively, $\lambda^{(m)}$ is the pricing variable balancing the base-station-to-relay rate supply $R_{s,r}^{(m)}$ and the relay-to-user rate demand $R_{r,d}^{(m)}$. Note that $\lambda^{(m)}$ is upper bounded by $\lambda_{\max}^{(m)} = \max_{k\in\mathcal{K}_r^{(m)}} \alpha_k$, in which case $\alpha_k - \lambda_{\max}^{(m)} \le 0$ holds for all $k \in \mathcal{K}_r^{(m)}$, and consequently the relay does not schedule any users and $R_{r,d}^{(m)} = 0$. Meanwhile $\lambda^{(m)}$ is lower bounded by $\lambda_{\min}^{(m)} = 0$, in which case no subchannel is used for relay's feeder link transmission in the first phase, and $R_{s,r}^{(m)} = 0$. The standard method for updating $\boldsymbol{\lambda}$ is the subgradient approach [22] with an appropriate step size:

Algorithm 2: Search of $\left\{\lambda^{(m)}\right\}_{m=1}^M$ for shared relaying. Return $\left\{\lambda^{(m)}\right\}_{m=1}^M$.

i) Initialize $\lambda^{(m)} = 0, \forall m$.

ii) Use $\left\{\lambda^{(m)}\right\}$ to solve the parts (a) and (b) of *Algorithm 1* to obtain the indicators $\rho_{s_m,k}^{(n,i)}$, $\rho_{r,k}^{(n)}$ and relay's beamforming vector $\mathbf{w}$ and power allocation $\mathbf{P}$.

iii) Update $R_{r,d}^{(m)}$ and $R_{s,r}^{(m)}$ according to (14).

iv) Update $\lambda^{(m)}$ using the subgradient projection method: $\lambda^{(m)} = \left[\lambda^{(m)} - s^{(m)}\left(R_{s,r}^{(m)} - R_{r,d}^{(m)}\right)\right]_{\lambda_{\min}^{(m)}}^{\lambda_{\max}^{(m)}}$ for all sector m, where $[x]_b^a = \min\left(\max\left(x, b\right), a\right)$. Go back to step ii) until convergence.

The price search converges when the backhaul constraints (14) are satisfied, and the gap between $R_{s,r}^{(m)}$ and $R_{r,d}^{(m)}$ is minimized for all m. The subgradient method is guaranteed to converge only if the maximization problem in the dual objective (21) is solved exactly. Although approximations are used in the evaluation of (21) in this paper, we observe that the subgradient update still works quite well.

The subgradient method, however, can be quite slow, especially since the step size $s^{(m)}$ needs to get progressively smaller with iteration, and each iteration step requires the evaluation of parts (a) and (b) of *Algorithm 1*. To accelerate the search, we propose a faster $\boldsymbol{\lambda}$-search method as follows. In part (a) of *Algorithm 1*, whether a subchannel is used for the feeder link or the user scheduling in the first phase is determined by comparing $\alpha_{\hat{k}} r_{s_m,\hat{k}}^{(n,1)}$ and $\lambda^{(m)} r_{s_m,r}^{(n)}$, where $\hat{k}$ is the selected user. Each channel realization then gives a threshold $\lambda_{\text{th}}^{(m,n)} = \alpha_{\hat{k}} r_{s_m,\hat{k}}^{(n,1)} / r_{s_m,r}^{(n)}$ for each subchannel. For each sector m, we can sort all $\left\{\lambda_{\text{th}}^{(m,n)}\right\}_{n=1}^{N}$ from the smallest to the largest, and denote the thresholds as: $\lambda_{\text{th},1}^{(m)} < \cdots < \lambda_{\text{th},i_m}^{(m)} < \cdots < \lambda_{\text{th},N}^{(m)}$, with $\lambda_{\text{th},0}^{(m)} = 0$. If $\lambda_{\text{th},i_m}^{(m)} < \lambda^{(m)} < \lambda_{\text{th},i_m+1}^{(m)}$, the set of subchannels $\left\{n \middle| \lambda_{\text{th}}^{(m,n)} \leq \lambda_{\text{th},i_m}^{(m)}\right\}$ are all used for wireless backhaul of the feeder link, while the set of subchannels $\left\{n \middle| \lambda_{\text{th}}^{(m,n)} \geq \lambda_{\text{th},i_m+1}^{(m)}\right\}$ are all used for one-hop user scheduling in the first phase in sector m. Since $R_{s,r}^{(m)}$ is monotonically increasing with $\lambda^{(m)}$, starting from $\lambda^{(m)} = \lambda_{\text{th},0}^{(m)}$, i.e., $i_m = 0$, $\forall m$, we can use the following fast discrete $\boldsymbol{\lambda}$-search to accelerate the subgradient search in step iv) of *Algorithm 2*:

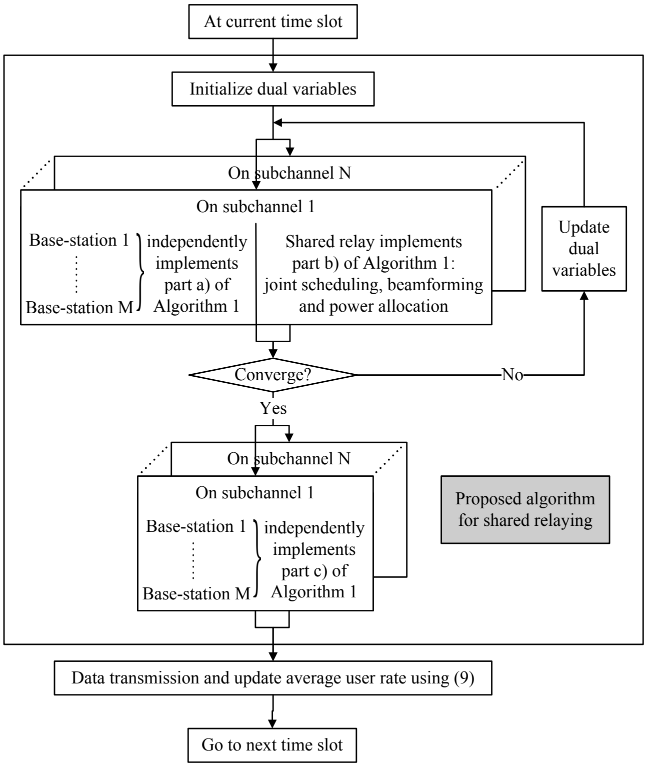
iv′) If $R_{s,r}^{(m)} < R_{r,d}^{(m)}$ and $i_m < N$, update $\lambda^{(m)} = \lambda_{\text{th},i_m+1}^{(m)}$, $i_m = i_m + 1$ and go back to step ii).

The discrete search can accelerate the update process in the first few iterations in a coarse manner. It can be followed by re-running step ii) to step iv) of the original *Algorithm 2* using the subgradient projection method to refine the search result.

D. Complexity and Implementation Issues

The overall algorithm for shared relaying is depicted in Fig. 3. The overall per-iteration complexity of the algorithm is linear in the total number of users in the M sectors. This can be seen as follows: Part (a) and Part (c) of *Algorithm 1* require a linear search over all candidate users. In Part (b), the beamforming and water-filling steps have trivial complexities, and the user selection step has a complexity that is linear in the total number of users as mentioned earlier. Finally, as seen in the simulation part of the paper, the algorithm takes about 20 iterations to converge. Thus, the overall complexity of the algorithm is quite low.

Distributed implementation is desirable for scalability reason, which we now discuss briefly. First, given the backhaul price $\lambda^{(m)}$ and the rates $r_{s_m,r}^{(n)}$ and $r_{s_m,k}^{(n,1)}$, part (a) of *Algorithm 1* can be done independently at each base-station m. Afterwards, $R_{s,r}^{(m)}$ can be computed independently at each base-station. Next, with intra-cluster interference from the base-stations fixed, the combined interference and noise term for two-hop user k can be measured, and the shared relay can independently implement part (b) of *Algorithm 1* given the knowledge of $\left\{\lambda^{(m)}\right\}_{m=1}^{M}$. Then, the shared relay feeds back $R_{r,d}^{(m)}$ to the base-stations; the base-station updates the price $\lambda^{(m)}$ locally in each sector (since only local rate supply $R_{s,r}^{(m)}$ and rate demand $R_{r,d}^{(m)}$ are required). After the update, each base-station can then broadcast new $\lambda^{(m)}$ to the shared relay for the next iteration. When convergence is reached, since the relay interference term $\Delta_k^{(n,i)}$ for the one-hop users in the second phase in (5) is fixed, each base-station can implement part (c) of *Algorithm 1* independently.



At current time slot
Initialize dual variables
On subchannel N
On subchannel 1
Base-station 1
Base-station M
independently implements part a) of Algorithm 1
Shared relay implements part b) of Algorithm 1: joint scheduling, beamforming and power allocation
Update dual variables
Converge?
No
Yes
On subchannel N
On subchannel 1
Base-station 1
Base-station M
independently implements part c) of Algorithm 1
Proposed algorithm for shared relaying
Data transmission and update average user rate using (9)
Go to next time slot

Fig. 3. Block diagram of the resource allocation and scheduling process of the shared relaying scheme

Note that the above implementation requires the exchange of $\lambda^{(m)}$ and $R_{r,d}^{(m)}$ between the shared relay and the base-stations in each iteration. As these are scalar quantities, the overall amount of information exchange is manageable. Moreover, since the backhaul prices are correlated in time due to the temporal correlation of the wireless channel and the partial update of the user weights, the relays do not need to search the prices from scratch each time, i.e., the backhaul prices used in the previous transmission can be used as the starting point of the current iteration to accelerate the search. Note also that the estimation and feedback of channel state information (CSI) are needed for both one-hop and two-hop users. In particular, for the two-hop users the amount of CSI feedback required is equivalent to that of an M-antenna multiuser MIMO system.

IV. Resource Allocation and Scheduling for Separate Relaying

For comparison purposes, this section briefly describes the algorithm for separate relaying in Fig. 1(a), where each sector is deployed with a separate relay. The problem formulation for separate relaying is similar to that of shared relaying, except that we need to replace $\rho_{r,k}^{(n)}$, $r_{s_m,r}^{(n)}$, $r_{r,k}^{(n)}$, and $r_{s_m,k}^{(n,i)}$ with $\rho_{r_m,k}^{(n)}$, $r_{s_m,r_m}^{(n)}$, $r_{r_m,k}^{(n)}$, and $r_{s_m,k}^{(n,i)\prime}$, respectively. The binary scheduling indicator constraints are modified as

$$\sum_{k\in\mathcal{K}_s^{(m)}} \rho_{s_m,k}^{(n,i)} \le 1, \qquad \sum_{k\in\mathcal{K}_r^{(m)}} \rho_{r_m,k}^{(n)} \le 1 \tag{22a}$$

$$\rho_{s_m,k}^{(n,i)},\ \rho_{r_m,k}^{(n)} \in \{0,1\}, \quad \forall m,n,i \in \{1,2\} \tag{22b}$$

such that each relay schedules one user in its sector.

The advantage of share relaying over separate relaying can come from multiple sources, e.g., the shared relay's ability to mitigate interference via beamforming, or its flexibility in allocating power across the antennas. To determine the relative contributions of these two factors, we considered the following three types of separate relaying for comparison purposes:

1) Single-antenna separate relaying with fixed transmit PSD;
2) Single-antenna separate relaying with relay power control under individual PSD constraints;
3) Multi-antenna separate relaying, with M antennas per relay but with fixed PSD. In this case, each relay has the ability to mitigate interference from and to the neighboring $M-1$ sectors via MMSE receive and ZF transmit beamforming, respectively.

The base-station uses the same scheduling rule as in part (a) and part (c) of *Algorithm 1*. In part (b) of *Algorithm 1*, each relay m schedules one user with the maximum positive value of the modified weighted rate $\left(\alpha_k - \lambda^{(m)}\right) r_{r_m,k}^{(n)}$. For Type 1 separate relaying, the scheduling of individual relay in each sector is independent. For Type 2 separate relaying, the optimization of power spectrum across the relays can be iterated with the relay's user scheduling until convergence [23]. For Type 3 separate relaying, the relay estimates the user rate $r_{r_m,k}^{(n)}$ based on the norm of the channel vectors for scheduling (similar to the user subset selection process at the shared relay in Section III, but without the orthogonal projection and the semi-orthogonality check since each separate relay schedules independently), and forms a ZF beamformer to null its interference to the scheduled users in other sectors.

The update of the Lagrangian prices for separate relaying can be easily modified as follows. For Type 1 separate relaying with independent scheduling in each sector, the rate supply $R_{s,r}^{(m)}$ and demand $R_{r,d}^{(m)}$ have monotone increasing and decreasing relationships with the backhaul price $\lambda^{(m)}$ respectively. So it is possible to use bisection search independently in each sector to find the proper $\lambda^{(m)}$ that satisfies the constraint (14), with the same upper and lower bounds $\lambda_{\max}^{(m)}$ and $\lambda_{\min}^{(m)}$ as in the shared relaying scheme. For both Type 2 and Type 3 separate relaying, the resource allocation and scheduling of adjacent M sectors within the cluster are interdependent, and the subgradient method used in *Algorithm 2* can be adopted to search for the backhaul prices.

V. Performance Evaluation

A. Simulation Parameters

We evaluate and compare the performances of the proposed scheduling and resource allocation schemes for the separate and the shared relay systems in a sectorized cellular networks with $M = 3$ coordinated sectors as shown in Fig. 1. The cell radius is 1km, and total bandwidth of 10MHz is divided into 64 orthogonal subchannels. One tier of explicitly modeled out-of-cluster interference is included in the simulation. The separate relays are placed at a distance of $2/3$ of the cell radius from the base-station [24], and the shared relay is placed at the intersection of adjacent sectors. Users are placed uniformly but at fixed locations. The number of users per sector is denoted as $K = \left|\mathcal{K}^{(m)}\right|$. The base-station's total transmit power P_s is set to be 46dBm over the 10MHz bandwidth, and the corresponding PSD is -24dBm/Hz. The separate relay's total power is denoted as P_r, and the comparable shared relay's power is $3P_r$ across all of its antennas (as a shared relay can be thought of as a combination of three separate relays). The path loss of the access link is $L = 128.1 + 37.6\log_{10}(d)$ dB (d is in km), with a 8-dB lognormal shadowing and Rayleigh fading. The path loss of the feeder link is $L = 128.1 + 28.8\log_{10}(d)$ dB, with a 4-dB lognormal shadowing and Rician fading with 10-dB Rician factor. The relay receives from its donor base-station with a directional antenna pattern in the feeder link ($\theta_{3dB} = 20^{\text{o}}$ [15]), and transmits to users with an omni-directional antenna. (Note that the use of directional receiving is crucial for separate relaying as the system performance with omni-directional receiving at the relay would be deficient due to the intercell interference in the feeder links.) The noise PSD is -174 dBm/Hz, and the target BER for the SNR gap is 10^{-3}. For PF scheduling, the update window size is assumed to be $T = 5$. We simulate 100 snapshots for every scenario, each with independent channel realization and user distribution.

In addition to the proposed price-based algorithm, we also evaluate a backhaul-unaware scheme as a reference (ref) for both separate and shared relaying, in which the relay uses conventional PF metric to schedule two-hop users, without considering the resource allocation in the first phase. After the relay resource allocation step, the base-station simply assigns sufficient number of subchannels for the feeder link to meet the rate demand of the relays. The no-relay cellular system is also evaluated for comparison purpose.

B. Simulation Results

Table I lists the performances of both the separate and the shared relaying schemes under different scenarios. In addition to the comparison between single-antenna separate relaying and 3-antenna shared relaying, we also evaluate separate relaying with 3 antennas per relay and the comparable shared relaying scheme with 9 antennas for comparison purpose. To evaluate the effect of power optimization, single-antenna separate relaying is evaluated with or without power control; 3-antenna shared relaying is also evaluated with or without the water-filling step (20). The performance metrics in the table include the sum rate, the 5% rate, which corresponds to the cell edge performance, and the utility as defined in equation (8). Note that the utility is negative as we take logarithmic of user rates in Mbps, which are typically below 1. Table I clearly shows that the shared relaying schemes have much better performance than the separate relaying schemes with equivalent number of antennas. In addition, the proposed

TABLE I
COMPARISON OF DIFFERENT SCHEMES IN TERMS OF PER-SECTOR PERFORMANCE METRIC, WITH $P_r = \frac{1}{3}P_s$

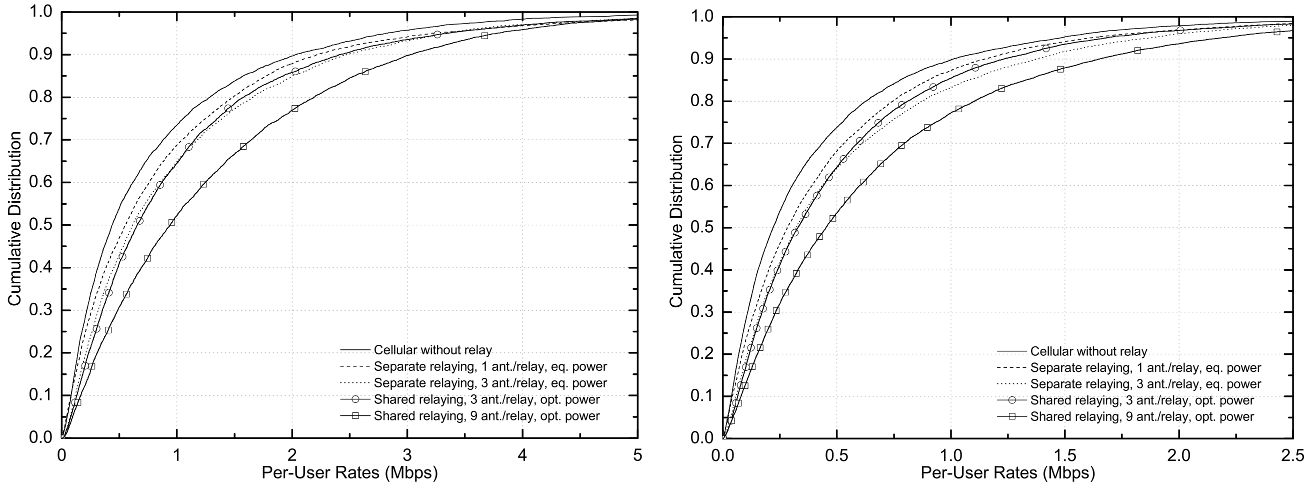
(a) $K = 40$

		Separate relaying				Shared relaying			
	No relay	1 ant. (ref) eq. power	1 ant. eq. power	1 ant. opt. power	3 ant. eq. power	3 ant. (ref) opt. power	3 ant. eq. power	3 ant. opt. power	9 ant. opt. power
Sum rate[a]	16.33	18.74	19.23	18.99	22.11	19.76	21.09	21.10	27.51
5% rate[a]	0.0203	0.0118	0.0230	0.0261	0.0337	0.0258	0.0382	0.0398	0.0438
Utility[b]	-63.98	-67.82	-56.74	-56.12	-48.83	-55.83	-48.98	-48.27	-37.55

(b) $K = 20$

		Separate relaying				Shared relaying			
	No relay	1 ant. (ref) eq. power	1 ant. eq. power	1 ant. opt. power	3 ant. eq. power	3 ant. (ref) opt. power	3 ant. eq. power	3 ant. opt. power	9 ant. opt. power
Sum rate[a]	16.12	18.19	18.97	18.98	20.69	19.51	20.56	20.65	26.43
5% rate[a]	0.0469	0.0239	0.0419	0.0492	0.0725	0.0535	0.0752	0.0784	0.0949
Utility[b]	-17.77	-20.91	-15.25	-14.33	-11.20	-14.46	-10.68	-10.30	-4.49

[a] measured in Mbps.
[b] computed with (8) as a function of per-user rates in Mbps.



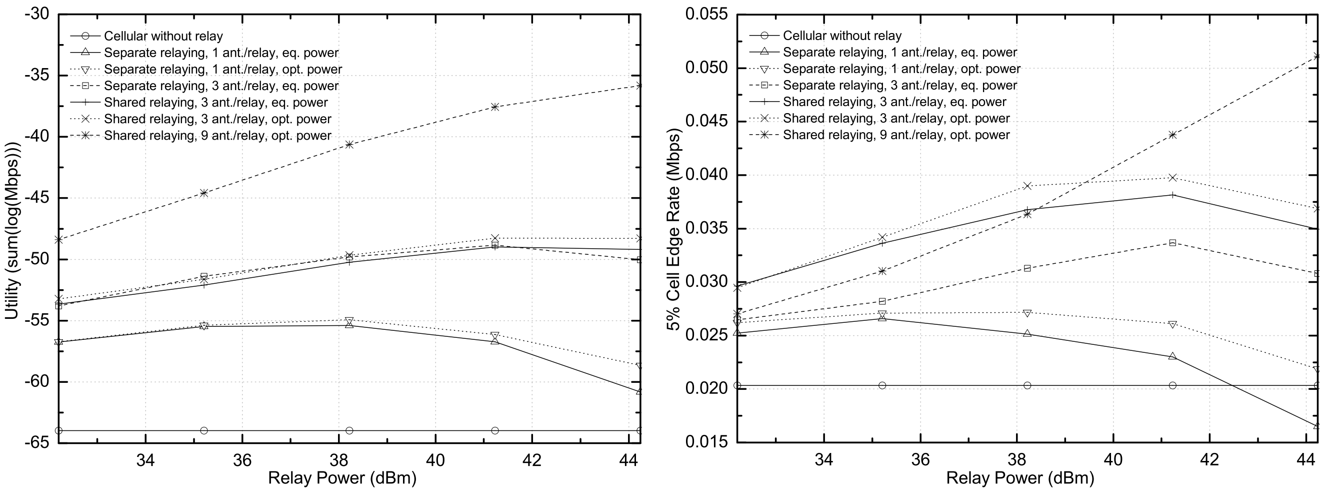
(a)

(b)

Fig. 4. CDF comparison of the separate and the shared relaying schemes, with $P_r = \frac{1}{3}P_s$ and (a) $K = 20$, (b) $K = 40$

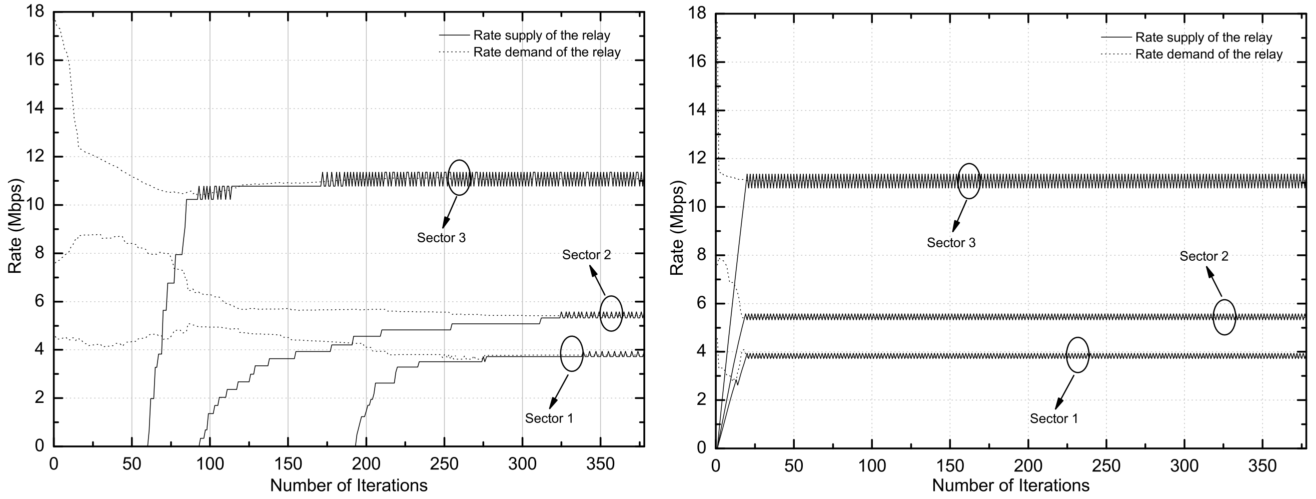
resource allocation and scheduling algorithms are effective in improving the system performance as compared to the backhaul-unaware ones (ref). Interestingly, separate relaying with 3 antennas each, which can cancel interference from/to the other 2 sectors, achieves about the same utility as that of shared relaying with 3 antennas (but has much smaller utility than the equivalent shared relaying scheme with 9 antennas). It is also observed that power optimization alone has only a positive but small improvement on the performance (except that power control on separate relaying slightly reduces sum rate when $K = 40$). This demonstrates that the advantage of shared relaying comes mostly from its interference cancelation ability rather than its ability to allocate power optimally across the antennas. Overall, shared relaying with 3 antennas achieves a $50{\sim}80\%$ increase in cell-edge rate over single-antenna separate relaying, and the improvement is about 30% when both separate and shared relays triple their antenna number. This shows that increasing the number of antennas in separate and shared relays diminishes the benefit for cell-edge rate. This is due to the enhanced ability of separate relays to cancel interference when more antennas are added. The throughput cumulative distribution functions (CDF) of shared relaying as compared to separate relaying are shown in Fig. 4. The CDF plot illustrates the same trends.

It is instructive to analyze the performance of shared relaying vs. separate relaying as functions of the relay power as in Fig. 5. It is shown that for single-antenna separate relaying, utility first increases with relay power, then decreases significantly due to the increased intra-cluster interference. Power control at the separate relays can alleviate interference but cannot eliminate it. In contrast, the utilities of both shared relaying with 3 antennas and separate relaying with 3 antennas per relay encounter no such significant decrease at high relay power. Similar trends can be observed for the cell edge rate (although at very high relay power, shared relaying with 3 antennas and separate relaying with 3 antennas per relay also see a decreasing edge rate due to the rising interference from out-of-cluster relays). Finally, the utility and 5% rate for shared relaying with 9 antennas are always increasing functions of the relay power due to its enhanced spatial diversity. It is again confirmed that power optimization for both separate and shared relaying brings marginal improvement in utility, while



(a) (b)

Fig. 5. Per-sector performance in terms of (a) utility, (b) 5% cell edge rate, as a function of P_r, with $K = 40$



(a) (b)

Fig. 6. Convergence of rate supply and rate demand of the shared relay (a) with subgradient method (b) with discrete $\boldsymbol{\lambda}$-search and subgradient method. $K = 40$, $P_r = \frac{1}{3}P_s$, $s = 10^{-5}$.

separate relaying with 3 antennas per relay achieves nearly the same utility as that of shared relaying with 3 antennas in total.

With the proposed *Algorithm 2*, the relay rate supply and rate demand in each sector should match. This convergence behavior is confirmed in the experiment shown in Fig. 6, where the step size is fixed at $s = 10^{-5}$. The oscillation of the shared relay's rate supply is due to the fact that a small change of the $\lambda^{(m)}$ around the threshold $\lambda_{\text{th}}^{(m,n)}$ can cause subchannel n in the first phase to switch between the feeder and access links, which leads to a fluctuation of the rate supply around the optimum. Although this means time sharing is needed to achieve the optimum, in practice one can simply proceed as long as the backhaul constraints (14) are satisfied. The simulation shows that the proposed discrete $\boldsymbol{\lambda}$-search converges to the same values as the conventional subgradient update but with a much faster speed. The number of iterations is reduced from about 350 to 20 in the experiment, where each iteration involves the broadcast of prices from the base-station to the relay, and the feedback of the total access link rates from the relay to the base-station in each sector. As stated previously, the iterations number can be made een smaller in practice by exploiting the temporal correlation of the prices.

VI. Conclusion

This paper illustrates the benefit of shared relaying from a system-level perspective. To realize the full potential of shared relaying, practical scheduling and resource allocation algorithms under the network utility maximization framework are proposed. The proposed algorithm uses a set of backhaul prices to balance the rate supply and demand at the relay, and to coordinate the backhaul-aware user scheduling at the base-station and the joint scheduling, beamforming, and power allocation at the relay. This paper also proposes a fast method for the iterative update of the prices. System-level simulations illustrate that shared relaying is effective in mitigating intercell interference and in improving the overall system performance in terms of both utility and rates.

References

[1] R. Pabst, B. Walke, D. Schultz, P. Herhold, H. Yanikomeroglu, S. Mukherjee, H. Viswanathan, M. Lott, W. Zirwas, M. Dohler *et al.*, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Commun. Magazine*, vol. 42, no. 9, pp. 80–89, Sept. 2004.

[2] M. Liu, X. Chang, Y. Lu, and H. Si, "A novel network structure based on multi cell coordinated relay," IEEE C802.16m-08/029, Jan. 2008.

[3] Y. Song, H. Yang, J. Liu, L. Cai, D. Li, X. Zhu, K. Wu, and H. Liu, "Relay station shared by multiple base stations for inter-cell interference mitigation," IEEE C802.16m-08/1436r1, Nov. 2008.

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

[5] F. Kelly, "Charging and rate control for elastic traffic," *European Trans. Telecom.*, vol. 8, no. 1, pp. 33–37, 1997.

[6] S. Peters, A. Panah, K. Truong, and R. Heath, "Relay architectures for 3GPP LTE-advanced," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, pp. 1–14, 2009.

[7] A. Panah, K. Truong, S. Peters, and R. Heath, "Interference management schemes for the shared relay concept," *EURASIP J. Advances in Signal Process.*, vol. 2011, pp. 1–14, 2011.

[8] J. Kim, J. Hwang, and Y. Han, "Joint processing in multi-cell coordinated shared relay network," in *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Sept. 2010, pp. 702–706.

[9] J. Lee and H. Yanikomeroglu, "A novel architecture for multi-hop wimax systems: Shared relay segmentation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2010, pp. 1–6.

[10] G. Li and H. Liu, "Resource allocation for OFDMA relay networks with fairness constraints," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 11, pp. 2061–2069, Nov. 2006.

[11] T. Ng and W. Yu, "Joint optimization of relay strategies and resource allocations in cooperative cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 2, pp. 328–339, Feb. 2007.

[12] W. Nam, W. Chang, S. Chung, and Y. Lee, "Transmit optimization for relay-based cellular OFDMA systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, June 2007, pp. 5714–5719.

[13] B. Kim and J. Lee, "Joint opportunistic subchannel and power scheduling for relay-based OFDMA networks with scheduling at relay stations," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2138–2148, June 2010.

[14] Y. Cui, V. Lau, and R. Wang, "Distributive subband allocation, power and rate control for relay-assisted OFDMA cellular system with imperfect system state knowledge," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5096–5102, Oct. 2009.

[15] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, and Y. Kim, "Fairness-aware radio resource management in downlink OFDMA cellular relay networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1628–1639, May 2010.

[16] M. Salem, A. Adinoyi, H. Yanikomeroglu, and D. Falconer, "Opportunities and challenges in OFDMA-based cellular relay networks: A radio resource management perspective," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2496–2510, June 2010.

[17] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans Commun.*, vol. 47, no. 6, pp. 884–895, June 1999.

[18] E. Chaponniere, P. Black, J. Holtzman, and D. Tse, "Transmitter directed, multiple receiver system using path diversity to equitably maximize throughput," U.S. patent, July 1999.

[19] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett.*, vol. 9, no. 3, pp. 210–212, Mar. 2005.

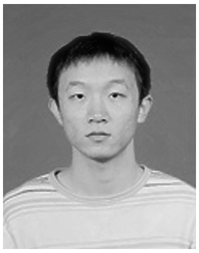
[20] G. Dimic and N. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.

[21] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

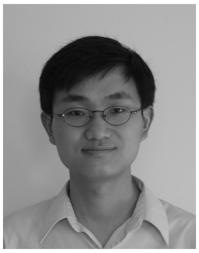
[22] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," Stanford Univ. Lecture Notes, Oct. 2003.

[23] W. Yu, T. Kwon, and C. Shin, "Joint scheduling and dynamic power spectrum optimization for wireless multicell networks," in *Proc. Information Sciences and Systems (CISS)*, Mar. 2010, pp. 1–6.

[24] Y. Liu, R. Hoshyar, X. Yang, and R. Tafazolli, "Integrated radio resource allocation for multihop cellular networks with fixed relay stations," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 11, pp. 2137–2146, Nov. 2006.



Yicheng Lin (S09) received the B.E. and M.E. degrees in Communication Engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007 and 2010, respectively. He is currently working towards the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada. His research interests include wireless communications and signal processing.



Wei Yu (S97-M02-SM08) received the B.A.Sc. degree in Computer Engineering and Mathematics from the University of Waterloo, Waterloo, Ontario, Canada in 1997 and M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, in 1998 and 2002, respectively. Since 2002, he has been with the Electrical and Computer Engineering Department at the University of Toronto, Toronto, Ontario, Canada, where he is now a Professor and holds a Canada Research Chair in Information Theory and Digital Communications. His main research interests include multiuser information theory, optimization, wireless communications and broadband access networks.

Prof. Wei Yu currently serves as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION THEORY. He served as an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS (2009-2011), as an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2004-2007), and as a Guest Editor for several special issues of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and EURASIP JOURNAL ON APPLIED SIGNAL PROCESSING. He is member of the Signal Processing for Communications and Networking Technical Committee of the IEEE Signal Processing Society. He received the IEEE Signal Processing Society Best Paper Award in 2008, the McCharles Prize for Early Career Research Distinction in 2008, the Early Career Teaching Award from the Faculty of Applied Science and Engineering, University of Toronto in 2007, and the Early Researcher Award from Ontario in 2006.