Cross-Layer Scheduling for OFDMA Amplify-and-Forward Relay Networks

Derrick Wing Kwan Ng, *Student Member, IEEE*, and Robert Schober, *Fellow, IEEE*

***Abstract*—In this paper, we consider cross-layer scheduling for the downlink of amplify-and-forward (AF) relay-assisted orthogonal frequency-division multiple-access (OFDMA) networks. The proposed cross-layer design takes into account the effects of imperfect channel-state information (CSI) at the transmitter (CSIT) in slow fading. The rate, power, and subcarrier allocation policies are optimized to maximize the system goodput (in bits per second per hertz successfully received by the users). The optimization problem is solved by using dual decomposition, resulting in a highly scalable distributed iterative resource-allocation algorithm. We also investigate the asymptotic performance of the proposed scheduler with respect to (w.r.t.) the numbers of users and relays. We find that the number of relays should grow faster than the number of users to fully exploit the *multiuser diversity* (MUD) gain. On the other hand, diversity from multiple relays can be exploited to enhance system performance when the MUD gain is saturated due to noise amplification at the AF relays. Furthermore, we introduce a feedback-reduction scheme to reduce the computational burden and the required amount of CSI feedback from the users to the relays. Simulation results confirm the derived analytical results for the growth of the system goodput and illustrate that the proposed distributed cross-layer scheduler only requires a small number of iterations to achieve practically the same performance as the optimal centralized scheduler, even if the information exchanged between the base station (BS) and the relays in each iteration is quantized, and the proposed CSI feedback reduction scheme is employed.**

***Index Terms*—Amplify-and-forward (AF) relaying, cross-layer scheduling, distributed resource allocation, dual decomposition, imperfect channel-state information (CSI), multiuser diversity (MUD).**

I. Introduction

ORTHOGONAL frequency-division multiple access (OFDMA) is a promising candidate for high-speed wireless communication networks, such as Wireless Fidelity, Worldwide Interoperability for Microwave Access (WiMAX), and future fourth-generation wireless systems [1], due to its high spectral efficiency and resistance to multipath fading. In an OFDMA system, the fading coefficients of different subcarriers are likely independent for different users, and by selecting the best user for each subcarrier and adapting the corresponding power, *multiuser diversity* (MUD) can fully be exploited. On the other hand, cooperative relaying is an attractive technique to increase the range of communication systems and to enhance the link reliability without incurring the high cost of additional base station (BS) deployment. Different relaying strategies, such as amplify-and-forward (AF), compress-and-forward, and decode-and-forward, have been proposed in the literature [2]–[4]. AF is particularly appealing as the relays only amplify and linearly process the received signal, which leads to low-complexity transceiver designs. More importantly, AF relays are transparent to the adaptive modulation techniques that are typically employed at the BS in today's wireless standards. For these reasons, AF was selected as one of the possible relaying modes in IEEE 802.16j (mobile multihop relay) [5].

Recently, there has been a growing interest in combining OFDMA/orthogonal frequency-division multiplexing (OFDM) with relaying to enhance wireless system performance. In [6] and [7], the authors study the power-allocation problem for AF OFDM systems when all transceivers are equipped with single antennas, whereas the case of multiple antennas is studied in [4]. Both works assume that perfect global channel-state information (CSI) of all links is available at the BS such that the power allocation can be done optimally. Furthermore, centralized scheduling algorithms assuming perfect global CSI at the BS have been proposed in, e.g., [8]–[10]. However, for these centralized algorithms, the computational complexity at the BS exponentially increases with the numbers of relays, users, and subcarriers, and the overhead for CSI feedback becomes significant, which limits the scalability of the system. In addition, in practice, perfect CSI at the transmitter (CSIT) is difficult to obtain for the relay-to-user links due to the mobility of the users. However, the ergodic channel capacity, which has been adopted as a performance criterion in the existing literature [4], [6]–[10], is a meaningful metric only when the channel is fast fading or when perfect CSIT is available. In slow fading without perfect CSIT, a channel outage occurs whenever the transmit data rate exceeds the channel capacity due to the imperfect CSIT. However, ergodic capacity fails to capture this effect. Therefore, for practical implementation, scalable distributed-scheduling algorithms that take into account imperfect CSIT and converge fast to the optimal solution are needed. Moreover, although it is well known that the system throughput scales with the number of users K on the order of $\mathcal{O}(\log\log K)$ [11], [12] in single-hop systems with perfect CSIT, it is unclear how the system performance scales with the number of users and relays in a relay-assisted OFDMA system with imperfect CSIT.

In this paper, we address the foregoing issues. For this purpose, we formulate the scheduling problem in AF relay-assisted OFDMA systems as an optimization problem. To make the problem tractable, we transform it into a convex optimization

Manuscript received August 24, 2009; revised November 19, 2009. First published December 31, 2009; current version published March 19, 2010. This paper was presented in part at the 70th IEEE Vehicular Technology Conference, Anchorage, AK, September 2009. The review of this paper was coordinated by Prof. J. Wu.

The authors are with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: wingn@ece.ubc.ca; rschober@ece.ubc.ca).

Digital Object Identifier 10.1109/TVT.2009.2039814

problem by introducing time-sharing variables. Using dual decomposition, the optimization problem is separated into a master problem and several subproblems. Each relay solves its own subproblem by utilizing its local CSI without any help from other relays, whereas the BS solves the master problem using a gradient method and updates the dual variables through the concept of pricing. Therefore, the computational complexity at the BS and the CSI feedback overhead are both significantly reduced compared with optimal centralized scheduling. To even further reduce the complexity and the signaling overhead, we propose an efficient CSI feedback-reduction scheme for the CSI feedback from the users to the relays. Furthermore, using tools from extreme value theory, we analyze the asymptotic growth of the system goodput if the proposed distributed scheduling algorithm is employed. Simulation results illustrate the excellent performance of the proposed algorithm and confirm the asymptotic expressions for the system goodput. In particular, our results show that large savings in computational complexity and signaling overhead are possible with the proposed CSI feedback-reduction scheme and a 3-bit quantization of the information exchanged between the BS and the relays in each iteration of the distributed algorithm at the expense of a small degradation in performance.

The remainder of this paper is organized as follows: In Section II, we introduce the OFDMA AF system model. In Section III, we define the performance metric *system goodput* and formulate the cross-layer design as an optimization problem. In Section IV, the cross-layer optimization problem is solved by dual decomposition, and extreme value theory is used to evaluate the system performance for large numbers of users and relays. Section V presents the numerical performance results for the proposed distributed algorithm. In Section VI, we conclude with a summary of the provided results.

II. Orthogonal Frequency-Division Mutiple Access Amplify-and-Forward Relay Network Model

In this section, after introducing the notation used in this paper, we present the adopted network and channel models as well as the assumptions regarding the availability of CSIT.

A. Notation

In this paper, the following conventions are adopted. $\mathcal{O}(g(x))$ denotes an *asymptotic upper bound*. Specifically, $f(x) = \mathcal{O}(g(x))$ if $\lim_{x\to\infty} |f(x)/g(x)| \leq N$ for $0 < N < \infty$. $E_X\{\cdot\}$ denotes the statistical expectation w.r.t. random variable X. $\mathcal{CN}(\mu, \sigma^2)$ denotes a complex Gaussian random variable with mean μ and variance σ^2. $1(\cdot)$ denotes an indicator function that is 1 when the event is true and 0 otherwise. $(x)^+ = \max\{0, x\}$. $Q(a, b)$ is the generalized Macrum Q-function. Finally, all logarithms, unless further specified in the subscript, are assumed to have base e.

B. Physical Layer Downlink OFDMA Model

We consider an OFDMA downlink relay-assisted packet-transmission network that consists of one BS, K mobile users, and M relays. All transceivers have a single antenna. A single cell with two ring-shaped boundary regions is studied. The region between the inner boundary and the outer boundary is divided into A sectors of equal size, as shown in Fig. 1, and the users in a given sector are assigned to a group of $R_i > 0$ relays such that $M = \sum_{i=1}^{A} R_i$. We assume that there is no direct transmission between the BS and the mobile users due to heavy blockage and long distance transmission.[1] A time-division channel allocation with two time slots is used to facilitate orthogonal transmission [2]. In the first time slot, the BS broadcasts its signal to the relay stations. Then, in the second time slot, the relay stations amplify the previously received signal and forward it to the corresponding users.

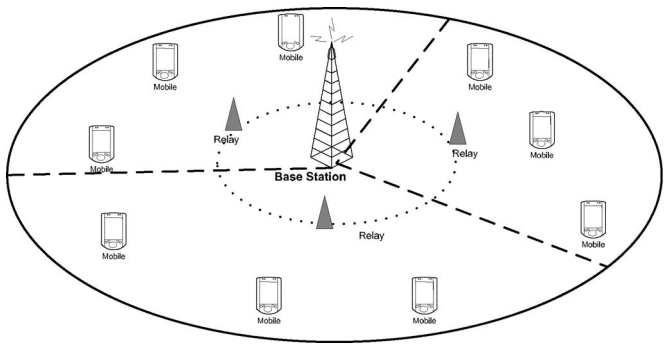

Fig. 1. Relay-assisted packet transmission system model with $K = 9$ users, $A = 3$ sectors, and $M = 3$ relays.

C. Channel Model

The channel impulse response is assumed to be time invariant (slow fading) within a frame. We consider an OFDMA system with n_F subcarriers. In the first time slot, the (frequency-domain) received symbol in subcarrier $i \in \{1, \ldots, n_F\}$ at relay $m \in \{1, \ldots, M\}$ for user $k \in \{1, \ldots, K\}$ is given by

$$Y_{SR_m,i}^{(k)} = \sqrt{P_{SR_m,i}^{(k)} l_{SR_m}} H_{SR_m,i} X_i^{(k)} + Z_{SR_m,i} \quad (1)$$

where $X_i^{(k)}$ is the symbol transmitted to user k on subcarrier i. l_{SR_m} represents the path loss between the BS and the relay m, and $Z_{SR_m,i}$ is the additive white Gaussian noise (AWGN) in subcarrier i at relay m. $P_{SR_m,i}^{(k)}$ is the transmit power for the link between the BS and the relay m in subcarrier i for user k. In practice, both the BS and the relays are placed in relatively high positions, and hence, the number of blockages or scatterers between them is limited, and a strong line of sight is expected. Hence, the channel fading coefficients between the BS and the relay m in subcarrier i, i.e., $H_{SR_m,i}$, are modeled as Rician fading with Rician factor κ, i.e., $H_{SR_m,i} \sim \mathcal{CN}(\sqrt{\kappa/(1+\kappa)}, 1/(1+\kappa))$. The received signal at relay m

[1] We assume that the resource allocation for relay-assisted users (located between the inner and outer boundaries) and nonrelay assisted users (located inside the inner boundary) is done separately. Since resource allocation for nonrelay-assisted OFDMA systems has extensively been studied in the literature, we focus on the relay-assisted case in this paper. We note that a joint resource allocation for nonrelay- and relay-assisted users would result in a better system performance, but the computational complexity of a joint optimization may be too high in practice.

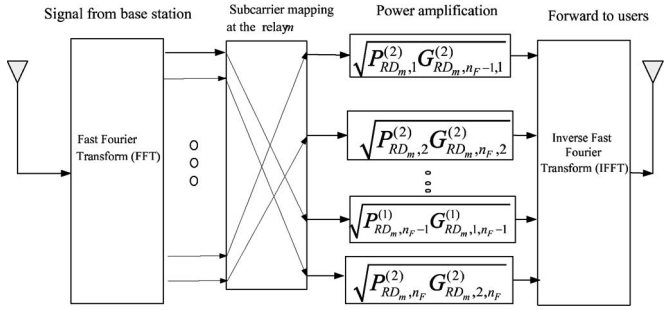

Fig. 2. Example for subcarrier mapping and power amplification at the mth relay for users 1 and 2.

on subcarrier i is mapped to subcarrier $j \in \{1, \ldots, n_F\}$ in the second time slot to optimize performance [13], [14]. Furthermore, the signal in subcarrier j is amplified by a gain factor $\sqrt{P^{(k)}_{RD_m,j} G^{(k)}_{RD_m,i,j}}$ and forwarded to the destination, as shown in Fig. 2, where $P^{(k)}_{RD_m,j}$ is the transmit power for the link between relay m and user k in subcarrier j, and $G^{(k)}_{RD_m,i,j}$ normalizes the input power of relay m in subcarrier i. The signal received at user k in subcarrier j from relay m is given by

$$Y^{(k)}_{RD_m,i,j} = \sqrt{G^{(k)}_{RD_m,i,j} P^{(k)}_{RD_m,j} l^{(k)}_{RD_m}}\, H^{(k)}_{RD_m,j} \times \left(\sqrt{P^{(k)}_{SR_m,i} l_{SR_m}}\, H_{SR_m,i} X^{(k)}_i + Z_{SR_m,i} \right) + Z^{(k)}_j \quad (2)$$

where the variables $P^{(k)}_{RD_m,j}$, $l^{(k)}_{RD_m}$, $H^{(k)}_{RD_m,j}$, and $Z^{(k)}_j$ are defined in a similar manner as the corresponding variables for the BS-to-relay links. Since the users are generally surrounded by a large number of scatterers, we model the small-scale fading coefficients between relay m and user k as Rayleigh distributed, i.e., $H^{(k)}_{RD_m,j} \sim \mathcal{CN}(0,1)$. To simplify the subsequent mathematical expressions and without loss of generality, we assume in the following a noise variance of $N_0 = 1$ at all relay and user stations. Based on this assumption and (2), the overall receive SNR of user k using subcarrier pair (i,j) in the first and second time slots through relay m can be expressed as

$$\Gamma^{(k)}_{\mathrm{eq}_m,i,j} = \frac{P^{(k)}_{RD_m,j} l^{(k)}_{RD_m} \left| H^{(k)}_{RD_m,j} \right|^2 P^{(k)}_{SR_m,i} l_{SR_m} |H_{SR_m,i}|^2}{P^{(k)}_{RD_m,j} l^{(k)}_{RD_m} \left| H^{(k)}_{RD_m,j} \right|^2 + \frac{1}{\left| G^{(k)}_{RD_m,i,j} \right|^2}}. \quad (3)$$

If the noise statistic is known at the relays, then a popular choice for $G^{(k)}_{RD_m,i,j}$ is [2]

$$\left| G^{(k)}_{RD_m,i,j} \right|^2 = \frac{1}{1 + P^{(k)}_{SR_m,i} l_{SR_m} |H_{SR_m,i}|^2} \quad (4)$$

which normalizes the instantaneous input power of relay m in subcarrier i to 1. Substituting (4) into (3) yields the final equivalent receive SNR as

$$\Gamma^{(k)}_{\mathrm{eq}_m,i,j} = \frac{P^{(k)}_{SR_m,i} l_{SR_m} |H_{SR_m,i}|^2 P^{(k)}_{RD_m,j} l^{(k)}_{RD_m} \left| H^{(k)}_{RD_m,j} \right|^2}{1 + P^{(k)}_{SR_m,i} l_{SR_m} |H_{SR_m,i}|^2 + P^{(k)}_{RD_m,j} l^{(k)}_{RD_m} \left| H^{(k)}_{RD_m,j} \right|^2}. \quad (5)$$

D. CSI

The cross-layer optimization problem presented in the next section can be solved either centrally at the BS or in a distributed fashion. For the centralized solution, the BS requires the CSI of all BS-to-relay and relay-to-user links at the beginning of each scheduling slot. In contrast, for the distributed solution, the relays only require the CSI[2] of their own BS-to-relay and relay-to-user links, whereas the BS needs no CSI. In the following, since large-scale fading is a slowly varying random process that changes on the order of seconds, we assume that the path loss coefficients l_{SR_m} and $l^{(k)}_{RD_m}$, $m \in \{1, \ldots, M\}$, $k \in \{1, \ldots, K\}$ can be estimated perfectly. For the small-scale fading, we take into account the different natures of the BS-to-relay and relay-to-user links. In particular, since both the BS and the relays are static, the BS-to-relay links are assumed to be time invariant. Thus, the BS-to-relay fading gains $H_{SR_{m,i}}$, $m \in \{1, \ldots, M\}$, $i \in \{1, \ldots, n_F\}$ can be reliably estimated at the relays with negligible estimation error. Therefore, we can assume perfect CSIT for the BS-to-relay links. On the other hand, although we also assume that the users can obtain perfect estimates of the relay-to-user fading gains $H^{(k)}_{RD_{m,j}}$, $m \in \{1, \ldots, M\}, j \in \{1, \ldots, n_F\}, k \in \{1, \ldots, K\}$, the corresponding CSI may be outdated at the relays (for the distributed solution) and at the BS (for the centralized solution) because of the mobility of the users. To capture this effect, we model the small-scale fading CSIT of the link between user k and relay m in subcarrier j as

$$\hat{H}^{(k)}_{RD_m,j} = H^{(k)}_{RD_m,j} + \Delta H^{(k)}_{RD_m,j}, \quad \Delta H^{(k)}_{RD_m,j} \sim \mathcal{CN}\left(0, \sigma_e^2\right) \quad (6)$$

where $H^{(k)}_{RD_m,j} \sim \mathcal{CN}(0,1)$ and $\Delta H^{(k)}_{RD_m,j} \sim \mathcal{CN}(0, \sigma_e^2)$ denote, respectively, the actual CSI and the CSIT error, which are mutually uncorrelated. σ_e^2 is the variance of the CSIT error.

III. Cross-Layer Design for Orthogonal Frequency-Division Multiple Access Amplify-and-Forward Relay Systems

In this section, we introduce the adopted system performance metric and formulate the related cross-layer optimization problem. Since the adopted approach is based on information theory, the buffers at the BS are assumed to be always full, and there are no empty scheduling slots due to an insufficient number of source packets at the buffers. To facilitate the formulation of the cross-layer scheduling problem as an optimization problem, we first introduce the following definitions.

Definition 1 (Subcarrier-Allocation Policy $\mathcal{S}$): Let $s^{(k)}_{m,i,j}$ be the subcarrier assignment variable of user k in using subcarrier pair (i,j) through relay m. The subcarrier-allocation policy is given by $\mathcal{S} = \{s^{(k)}_{m,i,j} \in \{0,1\}, \forall m,i,j,k\}$.

[2]We assume a frequency-division duplex system where the CSI of the relay-to-user links is obtained through feedback from the users to the relays at the beginning of each scheduling slot, whereas the CSI of the BS-to-relay links can be obtained either in the handshaking phase or from a previous transmission.

Definition 2 (Power-Allocation Policy $\mathcal{P}$): Let $P_{SR_m,i}^{(k)}$ and $P_{RD_m,j}^{(k)}$ be the transmit power of user k in using subcarrier pair (i,j) through relay m. The power-allocation policy is $\mathcal{P}=\{P_{SR_m,i}^{(k)}, P_{RD_m,j}^{(k)} \geq 0, \forall m,i,j,k\}$ such that the total power used is less than P_t.

Definition 3 (Rate-Allocation Policy $\mathcal{R}$): Let $r_{m,i,j}^{(k)}$ be the scheduled data rate of user k using subcarrier pair (i,j) through relay m in the first and second time slots. The rate-allocation policy is given by $\mathcal{R}=\{r_{m,i,j}^{(k)} \geq 0, \forall m,i,j,k\}$.

A. Instantaneous Mutual Information and System Goodput

In this section, we define the adopted system performance measure. Given perfect CSI at the receiver, the instantaneous mutual information between the BS and the user k in subcarrier pair (i,j) through relay m is given by

$$C_{m,i,j}^{(k)} = \frac{1}{2}\log_2\left(1+\Gamma_{\mathrm{eq}_m,i,j}^{(k)}\right). \tag{7}$$

In most existing cross-layer designs, the system performance is measured in terms of ergodic capacity. This is a meaningful measure when the schedulers have perfect CSIT or the channels are fast fading (ergodic realizations of CSI within the encoding frame) such that an arbitrarily small decoding error probability can be achieved as long as the channel error-correction code is strong enough. However, when the cross-layer schedulers have imperfect CSIT in slow fading, a packet outage occurs whenever the transmit data rate exceeds the instantaneous channel capacity, even when channel capacity achieving coding is applied for error protection. This is because the actual instantaneous mutual information is a random variable for both the relays and the BS. In this case, ergodic capacity fails to capture the effect of potential packet errors. Thus, using ergodic capacity as a system performance measure may not be a good choice in this situation since it fails to account for the penalty of channel outage. To model the effect of packet errors, we adopt the system goodput [15] as a performance measure. We first define the instantaneous goodput of a packet transmission for user k who is assigned to relay m as (in bits per second per hertz successfully delivered to user k)

$$\rho_m^{(k)} = \frac{1}{n_F}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F} s_{m,i,j}^{(k)} r_{m,i,j}^{(k)} \times 1\left(r_{m,i,j}^{(k)} \leq C_{m,i,j}^{(k)}\right). \tag{8}$$

Next, we define $\mathcal{U}_m$ as the set of users associated with relay m. The *average weighted system goodput* is defined as the total average bits per second per hertz successfully delivered to the K mobile stations through the M relays (averaged over multiple scheduling slots) and is given by

$$\begin{aligned}
&U_{\mathrm{goodput}}(\mathcal{P},\mathcal{R},\mathcal{S}) \\
&\quad = E\left\{\sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \rho_m^{(k)}\right\} \\
&\quad = E_{\mathbf{H}_{SR},\hat{\mathbf{H}}_{RD}} \\
&\qquad \times\left\{\sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\frac{s_{m,i,j}^{(k)} r_{m,i,j}^{(k)}}{n_F} E_{\mathbf{H}_{RD_m}}\right. \\
&\qquad \left.\times\left[1\left(r_{m,i,j}^{(k)} \leq C_{m,i,j}^{(k)} | \mathbf{H}_{SR_m}, \hat{\mathbf{H}}_{RD_m}, \mathbf{L}_m\right)\right]\right\} \\
&\quad = \frac{1}{n_F} E_{\mathbf{H}_{SR},\hat{\mathbf{H}}_{RD}} \\
&\qquad \times\left\{\sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F} s_{m,i,j}^{(k)} r_{m,i,j}^{(k)}\right. \\
&\qquad \left.\times \underbrace{\Pr\left[r_{m,i,j}^{(k)} \leq C_{m,i,j}^{(k)} | \mathbf{H}_{SR_m}, \hat{\mathbf{H}}_{RD_m}, \mathbf{L}_m\right]}_{\text{conditional goodput prob.}}\right\}
\end{aligned} \tag{9}$$

where the weights $w_k > 0$, $\sum_k w_k = K$, which are specified in the media access control (MAC) layer, allow the scheduler to give different priorities to different users and to enforce certain notions of fairness such as proportional fairness and max–min fairness [16], [17]. Matrices $\mathbf{H}_{RD}$ and $\mathbf{H}_{SR}$ contain vectors $\hat{\mathbf{H}}_{RD_m}$ and $\mathbf{H}_{SR_m}$, $m \in \{1,\ldots,M\}$, respectively, and vectors $\hat{\mathbf{H}}_{RD_m}$, $\mathbf{H}_{SR_m}$, and $\mathbf{L}_m$ contain the estimated CSIT $\hat{H}_{RD_m,j}^{(k)}$ for all links from relay m to users $k \in \mathcal{U}_m$, the actual CSIT $H_{SR_m,i}$ for the link between the BS and the relay m, and the path loss for all links involving relay m, respectively.

B. Problem Formulation for Cross-Layer Design

The cross-layer scheduling algorithm is responsible for the resource allocation for relay-assisted transmission at every scheduling slot. Based on the available CSI, the schedulers compute the power, rate, and subcarrier-allocation policies $\{\mathcal{P},\mathcal{R},\mathcal{S}\}$ to maximize the total average weighted system goodput $U_{\mathrm{goodput}}(\mathcal{P},\mathcal{R},\mathcal{S})$ for a target packet-outage probability $\varepsilon^{(k)}$ of user k. This leads to the following optimization problem.

Problem 1 (Cross-Layer Optimization Problem): The optimal power-allocation policy $\mathcal{P}^*$, the rate-allocation policy $\mathcal{R}^*$, and the subcarrier-allocation policy $\mathcal{S}^*$ are given by[3]

$$\begin{aligned}
&(\mathcal{P}^*,\mathcal{R}^*,\mathcal{S}^*) = \arg\max_{\mathcal{P},\mathcal{R},\mathcal{S}} U_{\mathrm{goodput}}(\mathcal{P},\mathcal{R},\mathcal{S}) \\
&\text{s.t.}\quad \mathrm{C1}: \Pr\left[r_{m,i,j}^{(k)} > C_{m,i,j}^{(k)} | \mathbf{H}_{SR_m}, \hat{\mathbf{H}}_{RD_m}, \mathbf{L}_m\right] \\
&\qquad\qquad \leq \varepsilon^{(k)} \qquad \forall k,i,j \\
&\qquad \mathrm{C2}: \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F} s_{m,i,j}^{(k)}\left(P_{RD_m,i}^{(k)} + P_{SR_m,j}^{(k)}\right) \leq P_t \\
&\qquad \mathrm{C3}: \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F} s_{m,i,j}^{(k)} = 1 \qquad \forall j \\
&\qquad \mathrm{C4}: \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{j=1}^{n_F} s_{m,i,j}^{(k)} = 1 \qquad \forall i \\
&\qquad \mathrm{C5}: P_{SR_m,i}^{(k)}, P_{RD_m,j}^{(k)} \geq 0 \qquad \forall m,i,j,k \\
&\qquad \mathrm{C6}: s_{m,i,j}^{(k)} \in \{0,1\} \qquad \forall m,i,j,k.
\end{aligned} \tag{10}$$

[3]Finding the optimal value of w_k to ensure target fairness is out of the scope of this paper. An iterative method for finding the optimal w_k is given in [18].

Here, C1 represents a per-subcarrier packet outage-probability constraint for user k. We note that other outage-probability constraints, such as chunk-based outage constraints [19], can be included in the problem formulation. However, the resulting optimization problem does not lend itself to an efficient distributed solution. Therefore, in this paper, a per-subcarrier outage constraint is imposed to obtain a tractable and scalable scheduling and resource-allocation algorithm. Since the outage probability is a good approximation for the frame error rate if strong error-correction coding is applied for error protection in slow fading [20], [21], C1 may be considered as a quality-of-service constraint. C2 is a joint power constraint for the BS and the relays with total maximum power P_t. Although, in a practical system, the BS and the relays have different power supplies, a joint power optimization provides useful insight into the power usage of the whole system rather than the per-hop required power. Moreover, for separate power constraints for the BS and the relays [4], [22], a distributed implementation of the resource allocation with provable convergence to the globally optimal solution does not seem possible. Thus, to obtain first-order system design insight and a highly scalable resource-allocation algorithm for relay-assisted OFDMA networks, a joint power constraint is imposed in this paper. Constraints C3, C4, and C6 are imposed to guarantee that each subcarrier will be used at most once in each time slot.

IV. Cross-Layer Optimization Solution

In this section, the cross-layer optimization problem is solved by dual decomposition, and a practical distributed scheduling algorithm is derived to alleviate the computational complexity burden at the BS. In addition, to further reduce the signaling overhead and the required signal processing, an efficient feedback-reduction scheme is proposed for the CSI feedback from the users to the relays. Furthermore, tools from extreme value theory are applied to analyze the system performance for large numbers of users and relays.

A. Transformation of the Optimization Problem

For the derivation of an efficient distributed scheduling algorithm, it is convenient to incorporate the outage-probability constraint in C1 in (10) into the objective function. This is possible if the constraint in C1 is fulfilled with equality for the optimal solution, which is the case for the low outage probabilities typically required in practical applications (e.g., $\varepsilon^{(k)} \leq 0.1$). Thus, in the following, we replace the "$\leq$" sign in C1 with a "$=$" sign, and the resulting optimization problem may be viewed as a more constrained version of the original problem [see (10)]. We are now ready to introduce the following lemma.

Lemma 1 (Equivalent Data Rate Constraint):

$$\Pr\left[r_{m,i,j}^{(k)} > C_{m,i,j}^{(k)} | \mathbf{H}_{SR_m}, \hat{\mathbf{H}}_{RD_m}, \mathbf{L}_m\right] = \varepsilon^{(k)} \Rightarrow r_{m,i,j}^{(k)} = \frac{1}{2}\log_2\left(1+\Lambda_{\mathrm{eq}_m,i,j}^{(k)}\right) \quad (11)$$

with equivalent receive SNR

$$\Lambda_{\mathrm{eq}_m,i,j}^{(k)} = \frac{P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 P_{RD_m,j}^{(k)} l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right)}{1 + P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 + P_{RD_m,j}^{(k)} l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right)} \quad (12)$$

where $F_{RD_m,j}^{-1(k)}(\varepsilon^{(k)})$ denotes the inverse cumulative distribution function (cdf) of a noncentral chi-square random variable with 2 degrees of freedom and noncentrality parameter $|\hat{H}_{RD_m,j}^{(k)}|^2/\sigma_e^2$. The inverse function of the noncentral chi-square cdf can be evaluated directly[4] or be stored in a lookup table in a practical implementation. Hence, by substituting (11) into the original objective function [see (9)], a new objective function that incorporates outage can be obtained.

Proof: See Appendix A. ■

The cross-layer scheduling problem is now a mixed combinatorial and nonconvex optimization problem, which is $\mathcal{NP}$-hard. The combinatorial nature comes from the integer constraint for subcarrier allocation, whereas the nonconvexity is caused by the power-allocation variables in the objective function. A sufficient condition for a function to be jointly concave w.r.t. its variables is that its Hessian is negative semi-definite [23]. Unfortunately, this is not the case for the considered problem. However, if the SNR of each link is high[5] enough, then the equivalent receive SNR in (12) can be approximated as

$$\Lambda_{\mathrm{eq}_m,i,j}^{(k)} \approx \frac{P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 P_{RD_m,j}^{(k)} l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right)}{P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 + P_{RD_m,j}^{(k)} l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right)} \quad (13)$$

which leads to a jointly concave objective function w.r.t. the power-allocation variables (cf. Appendix B). The next step is to handle the combinatorial subcarrier assignment constraint. The traditional brute force approach can be used to obtain a global optimal solution but results in exponential complexity w.r.t. the numbers of users and relays. To strike a balance between complexity and optimality, we follow the approach in [25] and relax $s_{m,i,j}^{(k)}$ in constraint C6 to be a real value between zero and one instead of a Boolean. Then, $s_{m,i,j}^{(k)}$ can be interpreted as a time-sharing factor for the K users to utilize the subcarrier pair (i,j) through relay m. Although the relaxation of the subcarrier-allocation constraint is generally suboptimal, the authors in [26] analytically show that the duality gap due to the relaxation becomes zero when the number of subcarriers goes to infinity. Furthermore, since the power constraint is instantaneous, as far as the optimization problem is concerned, the average weighted system goodput maximization is equivalent to the maximization of the instantaneous weighted goodput for each set of channel gains, although both criteria are different in general. Thus,

[4]The inverse of the noncentral chi-square cdf is commonly implemented as in-built function in software such as MATLAB.

[5]We note that the high SNR assumption seems reasonable for next-generation wireless systems, which are being designed to provide high spectral efficiencies [24]. Furthermore, the cross-layer scheduler will primarily select users that experience high SNRs.

by following a similar approach as in [27], the cross-layer scheduling optimization problem can be transformed into the following convex optimization problem.

Problem 2 (Transformed Cross-layer Optimization Problem):

$$\begin{aligned}
&\arg\max_{\mathcal{P},\mathcal{R},\mathcal{S}} \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F} \frac{\left(1-\varepsilon^{(k)}\right) s_{m,i,j}^{(k)}}{2} \\
&\quad \times \log_2\left(1+\frac{\Lambda_{\mathrm{eq}_m,i,j}^{(k)}}{s_{m,i,j}^{(k)}}\right) \\
&\text{s.t.} \quad \mathrm{C2}: \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\left(\tilde{P}_{RD_m,j}^{(k)}+\tilde{P}_{SR_m,i}^{(k)}\right) \le P_t \\
&\qquad \mathrm{C3}, \quad \mathrm{C4} \\
&\qquad \mathrm{C5}: \tilde{P}_{SR_m,i}^{(k)}, \tilde{P}_{RD_m,j}^{(k)} > 0 \qquad \forall m,i,j,k \\
&\qquad \mathrm{C6}: 0 \le s_{m,i,j}^{(k)} \le 1 \qquad \forall m,i,j,k
\end{aligned} \tag{14}$$

where $\tilde{P}_{SR_m,i}^{(k)} = P_{SR_m,i}^{(k)} s_{m,i,j}^{(k)}$ and $\tilde{P}_{RD_m,j}^{(k)} = P_{RD_m,j}^{(k)} s_{m,i,j}^{(k)}$ are auxiliary power variables. Problem 2 is now jointly concave w.r.t. the optimization variables (cf. Appendix B). Under some mild conditions, as illustrated in [23], it can be shown that strong duality holds, and the duality gap is equal to zero. Therefore, centralized numerical methods, such as the interior-point method and the ellipsoid method, can be used to solve Problem 2, and convergence to the optimal solution in polynomial time is guaranteed. On the other hand, centralized algorithms such as those in [8], [9], and [28] are alternative approaches to solve the foregoing optimization problem. However, all centralized methods require a large amount of CSI feedback to the BS, and the computational complexity exponentially increases w.r.t. the number of relays and users, which may overload the BS. Therefore, a distributed solution is preferable to alleviate the CSI overhead and the computational complexity at the BS. Hence, in the following sections, a distributed optimal solution for *Problem 2* will be derived based on dual decomposition.

B. Dual Problem Formulation

In this section, we formulate the dual for the considered cross-layer scheduling optimization problem. For this purpose, we first need the Lagrangian function of the primal problem. Upon rearranging the terms, the Lagrangian can be written as

$$\begin{aligned}
&\mathcal{L}(\lambda,\boldsymbol{\gamma},\boldsymbol{\beta},\mathcal{P},\mathcal{R},\mathcal{S}) \\
&\quad = \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F} \frac{\left(1-\varepsilon^{(k)}\right) s_{m,i,j}^{(k)}}{2} \\
&\qquad \times \log_2\left(1+\frac{\Lambda_{\mathrm{eq}_m,i,j}^{(k)}}{s_{m,i,j}^{(k)}}\right) \\
&\qquad - \lambda \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\left(\tilde{P}_{RD_m,j}^{(k)}+\tilde{P}_{SR_m,i}^{(k)}\right) \\
&\qquad - \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\beta_i s_{m,i,j}^{(k)} \\
&\qquad - \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\gamma_j s_{m,i,j}^{(k)} \\
&\qquad + \lambda P_t + \sum_{j=1}^{n_F}\gamma_j + \sum_{i=1}^{n_F}\beta_i
\end{aligned} \tag{15}$$

where λ is the Lagrange multiplier corresponding to the joint power constraint, and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are Lagrange multiplier vectors associated with the subcarrier usage constraints with elements $\gamma_j, j \in \{1,\ldots,n_F\}$, and $\beta_i, i \in \{1,\ldots,n_F\}$, respectively.

Thus, the dual problem is given by

$$\min_{\lambda,\boldsymbol{\gamma},\boldsymbol{\beta}\ge 0} \max_{\mathcal{P},\mathcal{R},\mathcal{S}} \mathcal{L}(\lambda,\boldsymbol{\gamma},\boldsymbol{\beta},\mathcal{P},\mathcal{R},\mathcal{S}). \tag{16}$$

In the following sections, we solve the preceding dual problem in (16) by decomposing it into two parts: The first part is a subproblem to be solved by each relay station, and the second part is the master dual problem to be solved by the BS.

C. Distributed Solution—Subproblem for Each Relay Station

By dual decomposition, the dual problem in (16) can be decomposed into a master problem and serval subproblems. The dual problem can iteratively be solved, where in each iteration, each relay solves one local subproblem with no assistance from the other relays and passes its local solution to the BS, which solves the master problem. The subproblem to be solved by relay m is given by

$$\max_{\mathcal{P},\mathcal{R},\mathcal{S}} \quad \mathcal{L}_m(\lambda,\boldsymbol{\gamma},\boldsymbol{\beta},\mathcal{P},\mathcal{R},\mathcal{S}) \tag{17}$$

with

$$\begin{aligned}
&\mathcal{L}_m(\lambda,\boldsymbol{\gamma},\boldsymbol{\beta},\mathcal{P},\mathcal{R},\mathcal{S}) \\
&\quad = \sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F} \frac{\left(1-\varepsilon^{(k)}\right) s_{m,i,j}^{(k)}}{2} \log_2\left(1+\frac{\Lambda_{\mathrm{eq}_m,i,j}^{(k)}}{s_{m,i,j}^{(k)}}\right) \\
&\qquad - \sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\beta_i s_{m,i,j}^{(k)} - \sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\gamma_j s_{m,i,j}^{(k)} \\
&\qquad - \lambda \sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\left(\tilde{P}_{RD_m,j}^{(k)}+\tilde{P}_{SR_m,i}^{(k)}\right)
\end{aligned} \tag{18}$$

where the Lagrange multipliers λ, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are provided by the BS. Using standard optimization techniques and the Karush–Kuhn–Tucker condition, the optimal power allocation for subcarrier pair (i,j) is obtained as

$$\begin{aligned}
\tilde{P}_{SR_m,i}^{(k)*} &= s_{m,i,j}^{(k)} P_{SR_m,i}^{(k)*} \\
&= s_{m,i,j}^{(k)} \left(w_k\left(1-\varepsilon^{(k)}\right)/\lambda - \Phi_{m,i,j}^{(k)}\right)^{+} \\
&\quad / \left(1+\Omega_{m,i,j}^{(k)}\right)
\end{aligned} \tag{19}$$

$$\begin{aligned}\tilde{P}_{RD_m,j}^{(k)*} &= s_{m,i,j}^{(k)} P_{RD_m,j}^{(k)*} \\ &= s_{m,i,j}^{(k)} \left(w_k \left(1-\varepsilon^{(k)}\right)/\lambda - \Phi_{m,i,j}^{(k)} \right)^+ \\ &\quad / \left(1+\frac{1}{\Omega_{m,i,j}^{(k)}}\right) \end{aligned} \tag{20}$$

where

$$\Phi_{m,i,j}^{(k)} = \frac{\left(\sqrt{l_{SR_m}|H_{SR_m,i}|^2} + \sqrt{l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right)}\right)^2}{l_{SR_m}|H_{SR_m,i}|^2 l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right)}$$

$$\Omega_{m,i,j}^{(k)} = \sqrt{\frac{l_{SR_m}|H_{SR_m,i}|^2}{l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right)}}.$$

It can be observed that variables w_k and $\varepsilon^{(k)}$ (provided by the MAC layer) affect the power allocation by changing the water level $w_k(1-\varepsilon^{(k)})/\lambda$ of user k. On the other hand, the optimal rate allocation is given by

$$r_{m,i,j}^{(k)*} = \frac{1}{2}\log_2\left(1+\Lambda_{\mathrm{eq}_m,i,j}^{(k)*}\right) \tag{21}$$

where $\Lambda_{\mathrm{eq}_m,i,j}^{(k)*}$ is obtained by substituting $P_{RD_m,j}^{(k)*}$ and $P_{SR_m,i}^{(k)*}$ into $\Lambda_{\mathrm{eq}_m,i,j}^{(k)}$. To obtain the optimal subcarrier allocation, we take the derivative of the subproblem w.r.t. $s_{m,i,p}^{(t,k)*}$ and substitute the optimal powers in (19) and (20) into the derivative, which yields

$$\left.\frac{\partial \mathcal{L}_m}{\partial s_{m,i,j}^{(k)*}}\right|_{\substack{\tilde{P}_{SR_m,i}^{(k)}=\tilde{P}_{SR_m,i}^{(k)*} \\ \tilde{P}_{RD_m,j}^{(k)}=\tilde{P}_{RD_m,j}^{(k)*}}} \tag{22}$$

$$\Rightarrow \underbrace{\log_2\left(1+\Lambda_{\mathrm{eq}_m,i,j}^{(k)*}\right) - \frac{\Lambda_{\mathrm{eq}_m,i,j}^{(k)*}}{1+\Lambda_{\mathrm{eq}_m,i,j}^{(k)*}}}_{A_{m,i,j}^{(k)}} - \frac{2(\gamma_j+\beta_i)}{(1-\varepsilon^{(k)})w_k}. \tag{23}$$

Thus, the subcarrier pair selection determined by relay station m is given by

$$s_{m,i,j}^{(k)*} = \begin{cases} 1, & \text{if } A_{m,i,j}^{(k)} \geq \frac{2(\gamma_j+\beta_i)}{w_k\left(1-\varepsilon^{(k)}\right)} \\ 0, & \text{otherwise} \end{cases}. \tag{24}$$

The dual variables β_i and γ_j act as the global price in using subcarrier pair (i,j). Only the user who has a sufficiently large weight w_k and good channel conditions in subcarrier pair (i,j) is able to pay the price and be selected by the scheduler. We observe from (19)–(21) and (24) that relay m, $m \in \{1,\ldots,M\}$ only requires the CSI of its own BS-to-relay link, the imperfect CSI of the relay-to-user links of the users in its own sector, and the dual variables λ, γ_j, $j \in \{1,\ldots,n_F\}$, and β_i, $i \in \{1,\ldots,n_F\}$ supplied by the BS.

D. Solution of the Master Dual Problem at the BS

To solve the master problem at the BS, each relay calculates the local resource usages and passes this information, i.e., $r_{m,i,j}^{(k)*}$, $s_{m,i,j}^{*(k)}$, $P_{SR_m,i}^{(k)*}$, and $P_{RD_m,j}^{(k)*}$, to the BS. Since the dual function is differentiable, the gradient method can be used to solve the minimization of the master problem in (16). The solution is given by

$$\begin{aligned}&\gamma_j(t+1) \\ &\quad = \left[\gamma_j(t) - \xi_1(t)\left(1 - \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F} s_{m,i,j}^{(k)}\right)\right]^+ \qquad \forall j \\ &\beta_i(t+1) \\ &\quad = \left[\beta_i(t) - \xi_2(t)\left(1 - \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{j=1}^{n_F} s_{m,i,j}^{(k)}\right)\right]^+ \qquad \forall i \\ &\lambda(t+1) \\ &\quad = \left[\lambda(t) - \xi_3(t)\times\left(P_{\mathrm{t}} - \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\tilde{P}_{RD_m,j}^{(k)} + \tilde{P}_{SR_m,i}^{(k)}\right)\right]^+ \end{aligned} \tag{25}$$

where t is the iteration index, and $\xi_1(t)$, $\xi_2(t)$, and $\xi_3(t)$ are the positive step sizes. Convergence to the optimal solution is guaranteed if the chosen step sizes satisfy[6] the infinite travel condition [23], [29]

$$\sum_{t=1}^{\infty}\xi_i(t) = \infty, \qquad i \in \{1,2,3\}. \tag{26}$$

The gradient update in (25) can be interpreted as the pricing-adjustment rule of the demand and supply model in economics. If the demand of the system resource exceeds the maximum supply, then the gradient method will raise the price in the next update; otherwise, it will reduce the price until at least some users can afford it. By combining the gradient update equations at the BS and the subcarrier-selection criterion in (24) at the relays, all selected subcarrier pairs will be occupied by one user only eventually. The overall algorithm is equivalent to a centralized approach that finds the subcarrier pair that maximizes the overall SNR in both links. In other words, for each subcarrier in the relay-to-user links, the algorithm is trying to find the best user to maximize the system goodput, and a similar selection is performed in the BS-to-relay links where the best relay is selected.

We note that there is no intersector interference in the considered system since the resource-allocation algorithm is applied to the whole cell, and all sectors are competing for resources. Equation (24) shows that, for the optimal solution, there is no time sharing between the assigned subcarrier pairs, and thus, intersector interference does not exist. In this paper, sectoring is only used to limit the number of users assigned to a relay and, thus, to limit the computational complexity per relay.

[6]We note that, for simplicity, for the results shown in Section V, constant step sizes were adopted. Constant step sizes are easier to optimize than variable step sizes and guarantee convergence to a close-to-optimal solution.

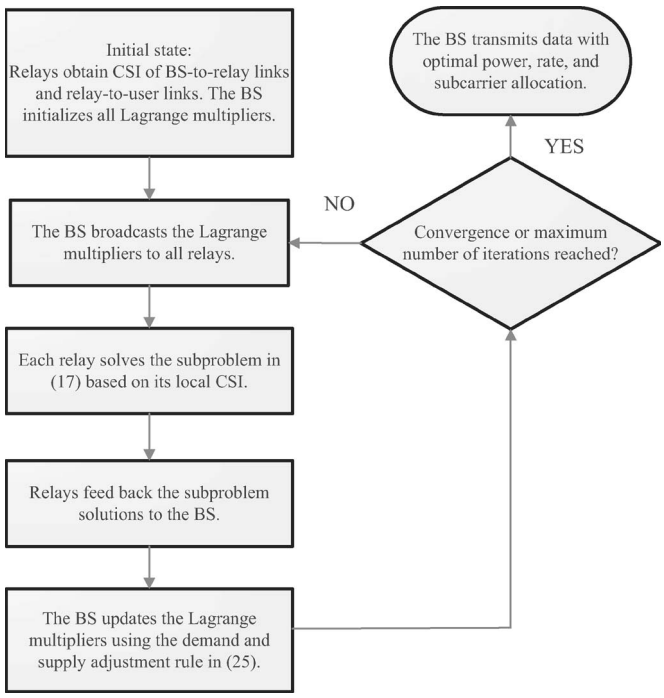

Fig. 3. Flow chart of the proposed distributed scheduling algorithm.

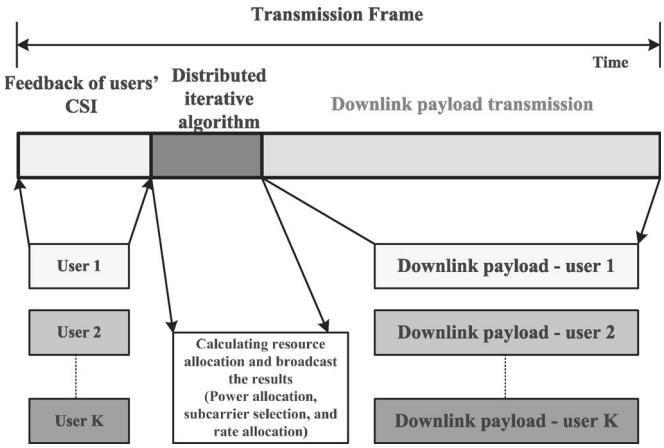

Fig. 4. Timing diagram for the proposed distributed iterative cross-layer scheduling algorithm.

We observe from (25) that the relays have to forward their respective power and subcarrier allocation policies to the BS. However, the BS does not require any form of CSI for optimal resource allocation. Thus, the feedback overhead of the resulting distributed algorithm is significantly lower than that of centralized algorithms, particularly if the number of users is large. The overall distributed algorithm is illustrated in Fig. 3, and the corresponding timing diagram for cross-layer scheduling is given in Fig. 4. The centralized brute force approach at the BS requires $\mathcal{O}(K^{n_F^2})$ operations, i.e., complexity exponentially increases with the number of subcarriers. In contrast, the complexity of the distributed scheduler at a given relay is only $\mathcal{O}(n_F^2 K/M)$. Thus, the distributed algorithm reduces the computational burden of the system considerably. Nevertheless, the complexity of the distributed algorithm still grows linearly with the number of users, and CSI feedback from all users to all relays is required. Therefore, to further reduce the complexity, in the next section, a CSI feedback reduction scheme is introduced.

E. Feedback Reduction Scheme for Relay-to-User Links

In this section, a feedback-reduction scheme is introduced such that only a subset of the users are required to feed back their CSI to the corresponding relays. As a result, each relay only needs to process a small set of user, which reduces the computational load of the system. The basic idea is similar to the SNR-based selective MUD scheme in [30]. However, our proposed scheme is based on the subcarrier selection in (24) rather than the SNR. In particular, from (24), we observe that the subcarrier selection criterion in each relay is based on the global resource usage (as indicated by the dual variables) and the channel quality of the users. For a reasonably large number of users K, it is very unlikely that a user with low $A_{m,i,j}^{(k)}$ is selected to use any subcarrier pairs. Therefore, CSI feedback from these users and allocation of computational resources to them are wasteful and should be avoided. In the proposed feedback-reduction scheme, user k only feeds back its CSI of subcarrier j to relay m when the feedback condition

$$l_{RD_m}^{(k)} F_{RD_m,j}^{-1(k)}\left(\varepsilon^{(k)}\right) \geq \left(2^{\frac{2\Theta_{\text{th}}}{w_k\left(1-\varepsilon^{(k)}\right)}+1}-1\right)\frac{2n_F}{P_t} \quad (27)$$

is fulfilled, where Θ_{th} is a threshold, which can be used to trade CSI feedback for performance. Equation (27) is derived in Appendix C. Note that (27) involves only local CSI, which is available at user k.

F. Asymptotic Analysis of AF Relay OFDMA System

In this section, we shall analyze the order growth of the average system goodput w.r.t. the numbers of users K and relays M. To obtain a tractable result, we focus on the study of proportional-fair (PF) schedulers with long-term fairness consideration. PF schedulers are popular because they allow the striking of a balance between system capacity and fairness among users and have been implemented in third-generation cellular systems for delay-tolerant applications. In [31], it has been shown that for long-term fairness, the PF scheduler symmetrizes the channel gain distribution of all users by adjusting the weighting w_k such that the path loss of the users is disregarded, and the user selection is based on the instantaneous independent identically distributed (i.i.d.) small-scale fading channel gain only. Thus, each user is selected by the scheduler with the same probability. Furthermore, we assume that all selected users have the same outage probability requirement $\varepsilon^{(k)} = \varepsilon$, and the distance between the BS and the relays is the same for simplicity.

The analysis is divided into three different scenarios. Case I illustrates the system goodput for arbitrary large K and M. In Case II, we consider an arbitrary large number of users K and a growing number of relays M such that the ratio of these two is given by $\lim_{K,M\to\infty}(K/M) \to \infty$. The physical meaning of this scenario is that there are already many users in the system, and a service provider may be interested in the gain achievable by adding more relays. In Case III, we study an arbitrary large

number of relays M and a growing number of users K for a ratio that is given by $\lim_{K,M\to\infty}(M/K) \to \infty$. This scenario corresponds to the case where there are many relays in the system in the first place, and the number of users increases. The results are summarized in the following theorem.

Theorem 1 (Asymptotic System Goodput for PF Scheduler): In high SNR, the asymptotic system goodput for the PF scheduler can be generalized into the following three cases:

$$U_{\text{goodput}}(\mathcal{P},\mathcal{R},\mathcal{S}) = \begin{cases} \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F} \frac{s_{m,i,j}^{(k)}}{2n_F} \\ \quad \times\mathcal{O}\left(\log_2\left(\frac{\nu_{m,j}^{(k)}\delta_{m,i}^{(k)}}{\nu_{m,j}^{(k)}+\delta_{m,i}^{(k)}}\right)\right), & \text{Case I} \\ \mathcal{O}\left(\log_2\left(\frac{\log M}{\kappa+1}\right)\right), & \text{Case II} \\ \mathcal{O}\left(\log_2\left(\left(1-\sigma_e^2\right)\log K\right)\right), \text{ for } 0\le\sigma_e^2<1 & \text{Case III} \end{cases} \tag{28}$$

where $\delta_{m,i}^{(k)} = P_{SR_m,i}^{(k)} l_{SR_m}(\log M/(\kappa+1))$, $\nu_{m,j}^{(k)} = P_{RD_m,j}^{(k)} l_{RD_m}(1-\sigma_e^2)\log K$, and κ is the Rician factor of the Rician fading in the BS-to-relay links.

Proof: Please refer to Appendix D. ■

For a better illustration, we preserve some terms in Cases II and III, which do not grow with either K or M. The results of Cases II and III have a simple max-flow min-cut interpretation. They illustrate that the maximum asymptotic growth of system goodput in an AF system is limited by its bottleneck link. Unlike the results of traditional multiuser systems, a large number of users do not necessarily lead to a MUD gain due to noise amplification in the AF relays. In Case II, the term $\kappa+1$ acts as a growth deduction factor because the line-of-sight path reduces the channel fluctuations required to exploit diversity. The analytical expression for Case II illustrates the diversity gain achievable by increasing the number of relays if the number of users is large. In particular, since the goodput depends double logarithmically on the number of relays, the relative improvement in system goodput reduces with growing M. In Case III, the term $(1-\sigma_e^2)$ acts as a penalty to the growth of goodput due to imperfect CSIT. The asymptotic expression in Case III shows that, to fully exploit the traditional MUD gain $\mathcal{O}(\log_2((1-\sigma_e^2)\log K))$, the number of relays should grow faster than the number of users, which corresponds to an impractical scenario.

V. Results

In this section, we evaluate the system performance using simulations. A single cell with two ring-shaped boundary regions is considered. The outer boundary and the inner boundary have radii of 1 km and 500 m, respectively. The K users are uniformly distributed between the inner and outer boundaries. The M relay stations are equally distributed at the inner boundary, and the cell is divided into A sectors of equal sizes. The number of subcarriers is $n_F = 128$, and the Third-Generation Partnership Project path loss model is used [32]. The small-scale fading coefficients of the BS-to-relay links are modeled as i.i.d. Rician random variables with Rician factor $\kappa = 6$ dB, whereas the small-scale fading coefficients of the relay-to-user links are i.i.d. Rayleigh random variables. The target packet outage probability is set to $\varepsilon^{(k)} = 0.01\ \forall k \in \{1,\ldots,K\}$ for illustration. The average weighted system goodput is obtained by counting the number of packets successfully decoded by all users averaged over both macroscopic and microscopic fading.

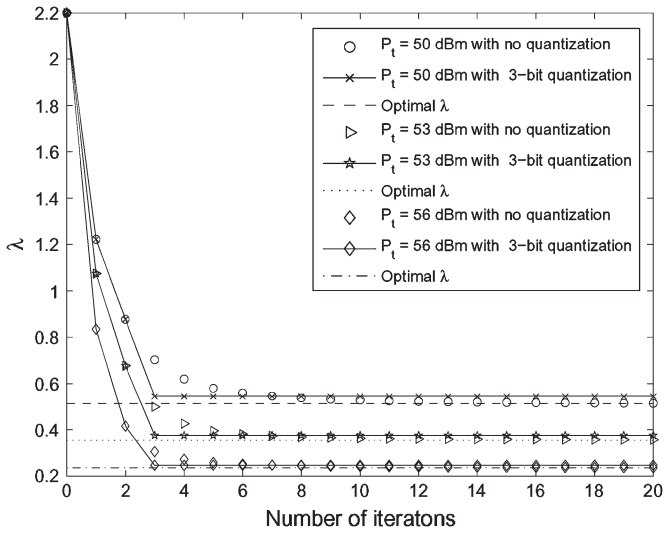

Fig. 5. Dual variable λ versus number of iterations with $K = 15$ users, $A = 3$ sectors, $M = 3$ relays, packet outage probability $\varepsilon^{(k)} = 0.01$, and $\sigma_e^2 = 0.01$.

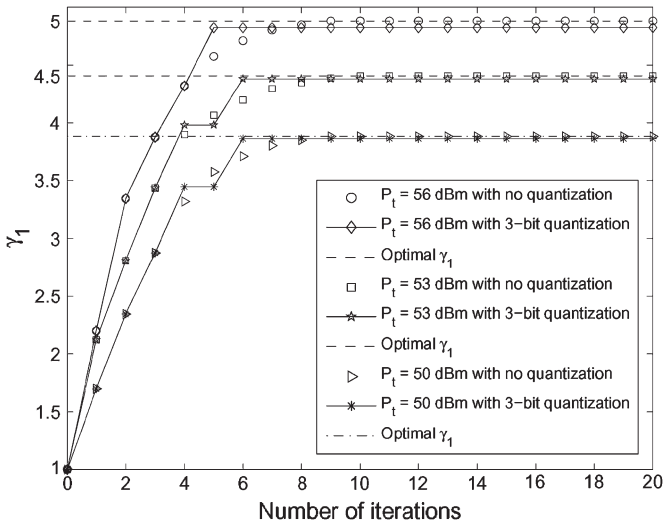

Fig. 6. Dual variable γ_1 versus number of iterations with $K = 15$ users, $A = 3$ sectors, $M = 3$ relays, packet outage probability $\varepsilon^{(k)} = 0.01$, and $\sigma_e^2 = 0.01$.

A. Convergence of Distributed Algorithm and Signaling Overhead Reduction

Figs. 5 and 6 illustrate the evolution of the Lagrange multipliers λ and γ_1 of the distributed algorithm over time for different maximum transmit powers P_t, $K = 15$ users, $M = 3$ relays, $A = 3$ sectors, and CSIT error variance $\sigma_e^2 = 0.01$. Positive constant step sizes $\xi_1(t)$, $\xi_2(t)$, and $\xi_3(t)$, which were optimized for fast convergence, were adopted. The results in Figs. 5 and 6 were averaged over 1000 independent adaptation processes. For comparison, Figs. 5 and 6 also contain results

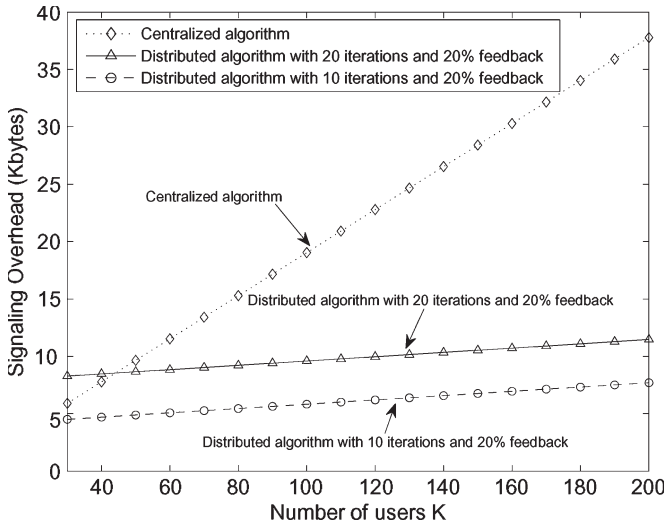

Fig. 7. Signaling overhead versus number of users for $n_F = 128$, $A = 3$ sectors, and $M = 3$.

TABLE I
QUANTIZATION TABLE FOR THE FEEDBACK VARIABLES

Feedback Variables	Number of bits
$H^{(k)}_{RD_m,j}, H_{SR_m,i}$	6
$\sum_{k\in\mathcal{U}_m}\sum_{j=1}^{n_F} s^{(k)}_{m,i,j}$	3
$\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F} s^{(k)}_{m,i,j}$	3
$\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\left(\tilde{P}^{(k)}_{SR_m,i}+\tilde{P}^{(k)}_{RD_m,j}\right)$	3
$r^{(k)*}_{m,i,j}, P^{(k)*}_{SR_m,i}$	3
$\lambda, \beta_i, \gamma_j$	3

for the realistic case where the information (dual variables and solution of subproblems) exchanged between the BS and the relays in each iteration is quantized to 3 bit. Thereby, the quantizer was designed offline using the Lloyd–Max algorithm. Figs. 5 and 6 show that the distributed algorithm converges fast and typically achieves 90%–95% of the optimal value within ten iterations. Thereby, quantization does not negatively affect the speed of convergence but causes a small deviation from the optimal value in the steady state, as expected.

Fig. 7 depicts the signaling overhead[7] versus the number of users for both the centralized scheme and the proposed distributed algorithm with feedback reduction. The number of bits used for quantization of the variables in this simulation is listed in Table I. Basically, we quantize the channel information of the relay-to-user links and the BS-to-relay links with 6 bits, the dual variables and the intermediate resource allocation results[8] with 3 bits, and the final resource allocation result with 6 bits. The threshold Θ_{th} defined in Section IV-E was chosen such that the amount of CSI feedback from the users to the relays is only 20% of the full feedback. For the centralized scheme, all channel gains fed back to the BS were quantized with 6 bits.

It can be observed from Fig. 7 that the proposed distributed iterative algorithm (together with the feedback reduction scheme) results in a significant decrease in signaling overhead compared with the centralized scheduling algorithm, particularly when the number of users in the system is large. However, even for a comparatively small number of users (e.g., $K = 30$), the amount of overhead for the distributed iterative algorithm is still less than that of the centralized algorithm if the distributed algorithm is limited to ten iterations. We will show in the next section that ten iterations are typically enough to achieve a close-to-optimal performance.

Remark 1: For demonstrating the time scale of the proposed distributed iterative algorithm, a fixed WiMAX system is a good example since we assume that both the BS and the relays are in fixed positions. In the fixed WiMAX system, there are 195 OFDM symbols per frame, and each frame has a length of 5 ms for a 10-MHz wideband system [33]. Suppose that there are $n_F = 128$ subcarriers. If each subcarrier is modulated with 64 quadrature-amplitude modulation, then the data rate is $((1 \times 128 \times 6 \times 195)/5\text{ ms}) \times (5/6) = 24.96$ Mbit/s if a code rate of 5/6 is used. As illustrated in Fig. 7, 10 kB of overhead is needed for $K = 100$ users and 20 iterations. Therefore, in the considered case, the information exchange required for the distributed algorithm takes $((10 \times 1024 \times 8)/24.96\text{ Mbit/s}) \approx 3$ ms. Furthermore, for an OFDMA system with a central carrier frequency of 2.5 GHz, the coherence time of the relay-to-user links is roughly 10 ms and 43 ms for users with a mobility of 45 and 10 km/h [34], respectively. Therefore, the scheduling and resource-allocation results obtained with the distributed algorithm are still valid[9] after 20 iterations.

B. Average System Goodput Versus Transmit Power and CSIT Error Variance σ_e^2

Fig. 8 depicts the average system goodput versus the transmit power for $K = 15$ users, $M = 3$ relays, $A = 3$ sectors, and CSIT error variance $\sigma_e^2 = 0.01$. In particular, the results for the proposed distributed and centralized scheduling algorithms are shown with maximum system goodput and PF scheduling as design goals. The maximum goodput scheduling can be obtained by using weights $w_k = 1\ \forall k$, whereas PF scheduling is performed by adapting the weights of each user according to [31]. For the centralized scheduler, the BS is assumed to have the CSI of each link in the network to perform subcarrier allocation based on an exhaustive search and optimal power allocation based on a standard water-filling procedure as well as rate adaption. As can be observed, even with only ten iterations, the proposed distributed algorithm closely approaches the performance of the optimal centralized scheduling algorithm for both maximum goodput and PF scheduling. Furthermore, Fig. 8 shows that the proposed CSI feedback-reduction scheme and quantization of the information exchanged by the BS and the

[7]In this paper, "overhead" refers to the amount of feedback required for the distributed iterative algorithm and centralized algorithm, respectively.

[8]Each relay station feeds back the intermediate resource allocation results $\sum_{k\in\mathcal{U}_m}\sum_{j=1}^{n_F} s^{(k)}_{m,i,j}$, $\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F} s^{(k)}_{m,i,j}$, and $\sum_{k\in\mathcal{U}_m}\sum_{i=1}^{n_F}\sum_{j=1}^{n_F}(\tilde{P}^{(k)}_{SR_m,i}+\tilde{P}^{(k)}_{RD_m,j})$ to solve the master problem at the BS.

[9]The decoding time of signaling overhead at both the BS and the relays in each iteration is negligible when compared with the coherence time in the slow fading channel.

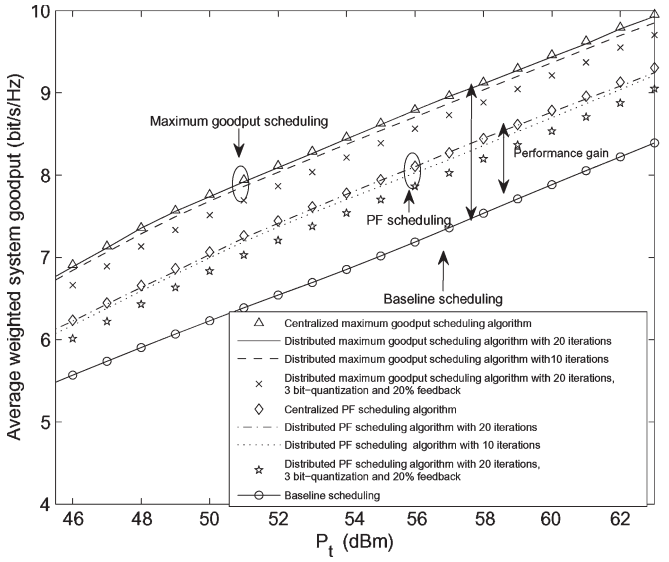

Fig. 8. Average weighted system goodput versus total transmit power for different scheduling algorithms with $K = 15$ users, $A = 3$ sectors, $M = 3$ relays, packet outage probability $\varepsilon^{(k)} = 0.01$, and $\sigma_e^2 = 0.01$.

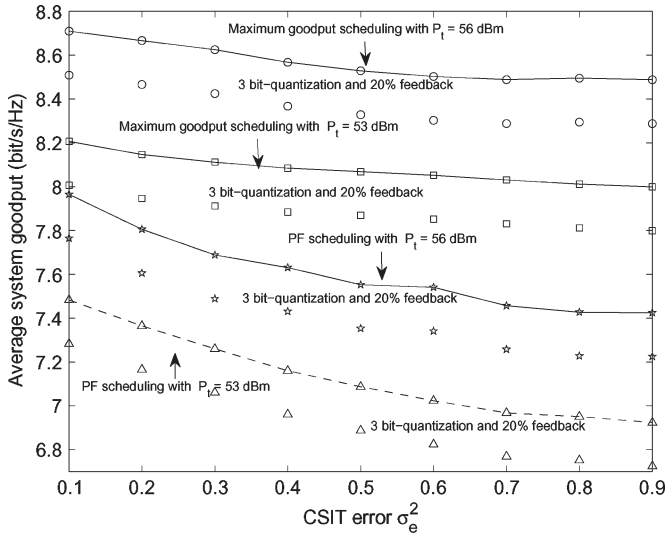

Fig. 9. Average weighted system goodput versus CSIT error variance σ_e^2 with $K = 15$ users, $A = 3$ sectors, $M = 3$ relays, and packet outage probability $\varepsilon^{(k)} = 0.01$.

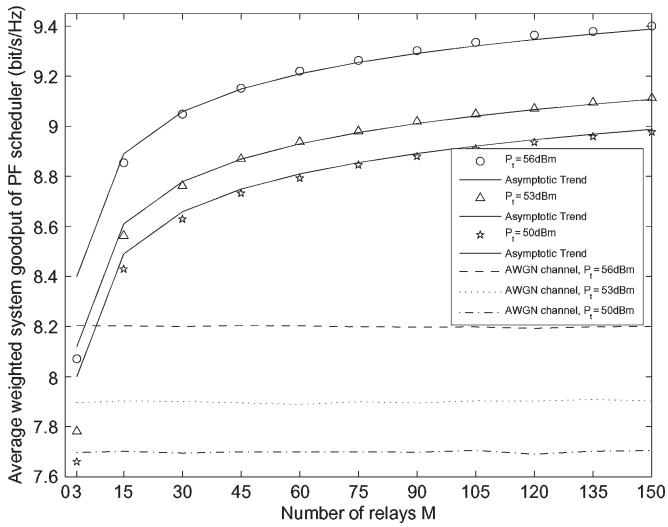

Fig. 10. Average weighted system goodput versus number of relays with $K = 300$ users $A = 3$ sectors, packet outage probability $\varepsilon^{(k)} = 0.01$, and $\sigma_e^2 = 0.01$.

relays in each iteration cause only a small loss in performance while significantly reducing the required signaling overhead and computational complexity. Similarly, the threshold Θ_{th} defined in Section IV-E was chosen such that the amount of CSI feedback from the users to the relays is only 20% of the full feedback. In other words, the CSI feedback and the signal processing at the relay are reduced by 80%.

For comparison, Fig. 8 also contains the goodput for a baseline round-robin scheduler in which subcarrier mapping is not performed, and the optimal power is allocated in a centralized manner to each subcarrier link. The performance of the baseline scheme is always worse than that of the two proposed schedulers. This is because the proposed schedulers can fully utilize the CSI of both links to perform resource allocation, while the baseline scheduler can only guarantee the outage requirement without taking further advantage of the CSI. We note that, as expected, the maximum goodput scheduler achieves a higher average goodput than the PF scheduler since the latter only considers the instantaneous small-scale fading and discards the path-loss information of the users in the long run. However, the superior average performance of the maximum goodput scheduler comes at the expense of starving users with weak channels, since only users with good channel conditions are allocated nonzero power. On the contrary, the PF scheduler maintains fairness among users such that each user has the same channel-access probability by sacrificing performance.

Fig. 9 depicts the average system goodput versus the CSIT error variance σ_e^2 of the proposed schedulers for $K = 15$ users, $M = 3$ relays, $A = 3$ sectors, 20 iterations, and different transmit powers. For comparison, we also show the goodput for the case of 3-bit quantization and CSI feedback reduction with 20% feedback load. It can be observed that as the CSIT error variance increases, the system performance decreases since the proposed schedulers have to be less aggressive in the resource allocation to satisfy the outage probability requirements of each user. Unlike traditional cross-layer schedulers, the proposed schedulers still achieve a reasonable performance for $\sigma_e^2 \to 1$. This is because the proposed schedulers consider the CSI of both the BS-to-relay and relay-to-user links. Although the scheduler has no accurate information of the small-scale fading in the relay-to-user links when $\sigma_e^2 \to 1$, power and rate adaptation can still be performed based on the path-loss information of users and the CSI of the BS-to-relay links. On the other hand, the performance loss due to quantization and feedback reduction remains roughly constant over the entire range of CSIT error variances, which suggests that the proposed schedulers are robust to CSI errors.

C. Asymptotic System Goodput Performance of PF Scheduler w.r.t. K and M

In this section, we focus on the asymptotic performance of the PF scheduler w.r.t. the numbers of users K and relays M for $A = 3$ sectors and CSIT error variance $\sigma_e^2 = 0.01$. To

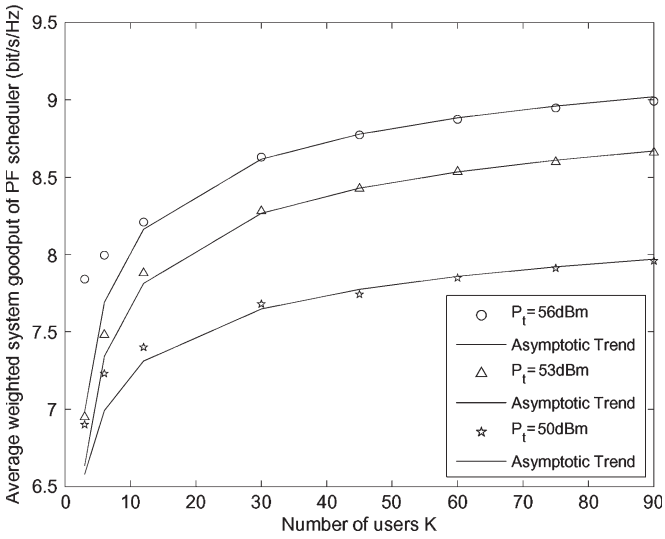

Fig. 11. Average weighted system goodput versus number of users with $M = 300$ relays, $A = 3$ sectors, packet outage probability $\varepsilon^{(k)} = 0.01$, and $\sigma_e^2 = 0.01$.

confirm the order growth of the system goodput in different scenarios, we do not perform quantization or feedback reduction in this section. Fig. 10 illustrates the average system goodput versus the number of relays for $K = 300$ users and different transmit powers. For comparison, we also plot the average system goodput for nonfading BS-to-relay links (pure AWGN channel) with the same total transmit power. The average system goodput grows with order $\mathcal{O}(\log_2(\log_e M/(\kappa + 1)))$, which matches the predicted asymptotic trend closely. This result suggests that when the traditional MUD gain is saturated due to noise amplification in the AF relays, another form of diversity gain from the multiple relays can be exploited to enhance system performance. Fig. 11 illustrates the average system goodput as a function of the number of users K for $M = 300$ relays for different total transmit powers. It can be seen that the average system goodput follows the order growth of $\mathcal{O}(\log_2((1 - \sigma_e^2) \log K))$ closely. This result suggests that, to fully exploit the MUD gain in the considered system, the number of relay stations should grow faster than the number of users to compensate for the noise amplification in the AF process at the relay.

VI. Conclusion

In this paper, taking into account imperfect CSIT, the cross-layer design of scheduling and resource allocation for AF relay-assisted OFDMA downlink transmission has been formulated as a mixed combinatorial and convex optimization problem. Based on dual decomposition of the primal problem, a highly scalable distributed resource-allocation algorithm is derived, which requires only local CSIT at the relays. Furthermore, an efficient CSI feedback reduction scheme is proposed, which allows a significant reduction of both CSI feedback from the users to the relays and the computational complexity at the relays. The asymptotic order growth of the average system goodput in terms of the number of users and relays is derived to obtain useful system design insights. The asymptotic analysis reveals that, to fully exploit the traditional MUD gain, the number of relays should grow faster than the number of users, which is impractical. On the other hand, diversity from multiple relays can be obtained when the MUD gain is saturated due to noise amplification in the AF relays. Our simulation results show that the performance of the distributed algorithm approaches that of the optimal centralized scheduler in a small number of iterations, even if the proposed CSI feedback-reduction scheme is employed, and the information exchanged between the BS and the relays in each iteration is quantized to 3 bits, which confirms the practicality of the proposed scheduler.

Appendix A
Proof of Lemma 1

We assume that subcarrier pair (i, j) is used for transmission in the first and second time slots through relay m for user k. Let $\mathbf{\Upsilon_m} = \{\mathbf{H}_{SR_m}, \hat{\mathbf{H}}_{RD_m}, \mathbf{L}_m\}$ for notational convenience. Then, the outage probability in C1 is given by (29), shown at the bottom of the page, where $F_{RD_m,j}^{(k)}(\cdot)$ denotes the cdf

$$
\begin{aligned}
&\Pr\left[r_{m,i,j}^{(k)} > \frac{1}{2}\log_2\left(1 + \frac{P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 P_{RD_m,j}^{(k)} l_{RD_m}^{(k)} \left|H_{RD_m,j}^{(k)}\right|^2}{1 + P_{SR_m,i}^{(k)} l_{SR_m} \left|H_{SR_m,i}^{(k)}\right|^2 + P_{RD_m,j}^{(k)} l_{RD_m}^{(k)} \left|H_{RD_m,j}^{(k)}\right|^2}\right) | \mathbf{\Upsilon}_m\right] \\
&= \Pr\left[\left|H_{RD_m,j}^{(k)}\right|^2 \le \frac{z\left(1 + P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2\right)}{\left(P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 - z\right) P_{RD_m,j}^{(k)} l_{RD_m}^{(k)}} | \mathbf{\Upsilon}_m\right], \qquad \text{where } z = 2^{2r_{m,i,j}^{(k)}} - 1 \\
&= F_{RD_m,j}^{(k)}\left(\frac{z\left(1 + P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2\right)}{\left(P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 - z\right) P_{RD_m,j}^{(k)} l_{RD_m}^{(k)}}\right) \\
&= 1 - Q\left(\sqrt{\frac{\left|\hat{H}_{RD_m,j}^{(k)}\right|^2}{\sigma_e^2}}, \sqrt{\frac{z\left(1 + P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2\right)}{\left(P_{SR_m,i}^{(k)} l_{SR_m} |H_{SR_m,i}|^2 - z\right) P_{RD_m,j}^{(k)} l_{RD_m}^{(k)} \sigma_e^2}}\right)
\end{aligned} \tag{29}
$$

of a noncentral chi-square random variable with 2 degrees of freedom and noncentrality parameter $|\hat{H}^{(k)}_{RD_m,j}|^2/\sigma_e^2$. Note that $P^{(k)}_{SR_m,i} l_{SR_m} |H_{SR_m,i}|^2 > z$ since $r^{(k)}_{m,i,j}$ will not exceed the channel capacity of the BS-to-relay links as the corresponding perfect CSI is available at the scheduler. Using the foregoing result, the target outage probability in constraint C1 in (10) is equivalent to

$$\begin{aligned}
\text{C1} \Rightarrow & F^{(k)}_{RD_m,j}\left(\frac{z\left(1+P^{(k)}_{SR_m,i} l_{SR_m}|H_{SR_m,i}|^2\right)}{\left(P^{(k)}_{SR_m,i} l_{SR_m}|H_{SR_m,i}|^2 - z\right) P^{(k)}_{RD_m,j} l^{(k)}_{RD_m}}\right) \\
& = \varepsilon^{(k)} \\
\Rightarrow & \frac{z\left(1+P^{(k)}_{SR_m,i} l_{SR_m}|H_{SR_m,i}|^2\right)}{\left(P^{(k)}_{SR_m,i} l^{(k)}_{SR_m}|H_{SR_m,i}|^2 - z\right) P^{(k)}_{RD_m,j} l^{(k)}_{RD_m}} \\
& = F^{-1(k)}_{RD_m,j}\left(\varepsilon^{(k)}\right) \\
\Rightarrow & r^{(k)}_{m,i,j} = \log_2\left(1+\Lambda^{(k)}_{\text{eq}_m,i,j}\right)
\end{aligned} \tag{30}$$

where $\Lambda^{(k)}_{\text{eq}_m,i,j}$ is defined in (12).

Appendix B
Proof of the Concavity of Problem 2

We first consider the concavity of a function $f(P^{(k)}_{SR_m,i}, P^{(k)}_{RD_m,j}) = 1 + \Lambda^{(k)}_{\text{eq}_m,i,j}$ w.r.t. the power allocation variables $P^{(k)}_{SR_m,i}$ and $P^{(k)}_{RD_m,i}$. Let the Hessian matrix of $f(P^{(k)}_{SR_m,i}, P^{(k)}_{RD_m,j})$ be $\mathbf{H}(f(P^{(k)}_{SR_m,i}, P^{(k)}_{RD_m,j}))$. It can be shown that the eigenvalues of $\mathbf{H}(f(P^{(k)}_{SR_m,i}, P^{(k)}_{RD_m,j}))$ are given by

$$\begin{aligned}
1^{\text{st}}\text{ eigenvalue} &= \frac{-2\left(P^{2(k)}_{SR_m,i} + P^{2(k)}_{RD_m,j}\right)\varsigma^{(k)}_{m,i,j}}{\left(P^{(k)}_{SR_m,i} l_{SR_m}|H_{SR_m,i}|^2\right)^3 + \varrho^{(k)}_{m,i,j} + 3\Xi} \\
2^{\text{nd}}\text{ eigenvalue} &= 0
\end{aligned} \tag{31}$$

where $\varsigma^{(k)}_{m,i,j} = (l_{SR_m}|H_{SR_m,i}|^2 l^{(k)}_{RD_m} |F^{-1(k)}_{RD_m,j}(\varepsilon^{(k)}))^2$, $\varrho^{(k)}_{m,i,j} = (P^{(k)}_{RD_m,i} l^{(k)}_{RD_m} F^{-1(k)}_{RD_m,j}(\varepsilon^{(k)}))^3$, and $\Xi = P^{(k)}_{SR_m,i} P^{(k)}_{RD_m,i} l_{SR_m} |H_{SR_m,i}|^2 l^{(k)}_{RD_m} F^{-1(k)}_{RD_m,j} (P^{(k)}_{SR_m,i} l_{SR_m} |H_{SR_m,i}|^2 + P^{(k)}_{RD_m,i} l^{(k)}_{RD_m} F^{-1(k)}_{RD_m,j})$. Since the Hessian matrix only has nonpositive eigenvalues, it is a negative semi-definite matrix, and thus, $1 + \Lambda^{(k)}_{\text{eq}_m,i,j}$ is a concave function w.r.t. the power-allocation variables. On the other hand, $\log_2(\cdot)$ is a nondecreasing concave function, and thus, $\log_2(1 + \Lambda^{(k)}_{\text{eq}_m,i,j})$ is a concave function as well.

Furthermore, the transformation from $\log_2(1 + \Lambda^{(k)}_{\text{eq}_m,i,j})$ to $s^{(k)}_{m,i,j}\log_2(1 + (\Lambda^{(k)}_{\text{eq}_m,i,j}/s^{(k)}_{m,i,j}))$ is a perspective operation [23], and thus, concavity w.r.t. $s^{(k)}_{m,i,j}$ is preserved. Note also that for the function $g(s^{(k)}_{m,i,j}) = s^{(k)}_{m,i,j}\log_2(1 + (\Lambda^{(k)}_{\text{eq}_m,i,j}/s^{(k)}_{m,i,j}))$, it can be shown by L'Hôspital's rule that $g(0) = 0$, and thus, the case of $s^{(k)}_{m,i,j} = 0$ does not affect the concavity of the objective function.

In summary, the foregoing considerations show that the objective function in Problem 2 is jointly concave in the optimization variables.

Appendix C
Derivation of Feedback Condition

Let Θ_{th} be the threshold representing the global subcarrier usage and replacing $(\gamma_j + \beta_i)$ in (24). Furthermore, assume that the SNR is high such that $\Lambda^{(k)}_{\text{eq}_m,i,j}/(1 + \Lambda^{(k)}_{\text{eq}_m,i,j}) \approx 1$ in (23). With these assumptions and $s^{(k)}_{m,i,j} = 1$, based on (23) and (24), we can establish the following upper bound:

$$\begin{aligned}
&\log_2\left(1 + P^{(k)}_{RD_m,j} l^{(k)}_{RD_m} F^{-1(k)}_{RD_m,j}\left(\varepsilon^{(k)}\right)\right) - 1 \\
&\quad \geq \log_2\left(1 + \Lambda^{(k)}_{\text{eq}_m,i,j}\right) - \frac{\Lambda^{(k)}_{\text{eq}_m,i,j}}{1 + \Lambda^{(k)}_{\text{eq}_m,i,j}} \\
&\quad \geq \frac{2\Theta_{\text{th}}}{w_k\left(1 - \varepsilon^{(k)}\right)}.
\end{aligned} \tag{32}$$

Interestingly, the first line in (32) does not involve the CSI of the BS-to-relay links and only requires CSI information available to user k. In addition, note that the upper bound in (32) becomes tight if the relay-to-user channel is weak compared with the BS-to-relay channel. Based on (32), the proposed user feedback criterion can be obtained as

$$\begin{aligned}
l^{(k)}_{RD_m} F^{-1(k)}_{RD_m,j}\left(\varepsilon^{(k)}\right) &\geq \left(2^{\frac{2\Theta_{\text{th}}}{w_k\left(1-\varepsilon^{(k)}\right)}+1} - 1\right) P^{(k)}_{RD_m,j} \\
&\approx \left(2^{\frac{2\Theta_{\text{th}}}{w_k\left(1-\varepsilon^{(k)}\right)}+1} - 1\right) \frac{2n_F}{P_t}
\end{aligned} \tag{33}$$

where the simplification in the last line is necessary since the users do not know the final power allocation before the schedulers perform the resource allocation.

Appendix D
Asymptotic Analysis

The asymptotic analysis is divided into three parts. We first introduce the following lemma from extreme value theory.[10]

Lemma 2 (Converging to Gumbel Distribution [35]): Let $\{\zeta_1, \zeta_2, \zeta_3, \ldots, \zeta_Z\}$ be a sequence of Z positive i.i.d. random variables with probability density function (pdf) $f(\zeta)$ and cdf $F(\zeta)$, which is twice differentiable for all ζ. Define $\zeta_{\max} = \max\{\zeta_1, \zeta_2, \zeta_3, \ldots, \zeta_Z\}$ as the maximum among the Z random variables. If the growth function $g(\zeta)$ satisfies $\lim_{\zeta\to\infty} g(\zeta) = \lim_{\zeta\to\infty}((1 - F(\zeta))/f(\zeta)) = c$, where c is a constant, then $\zeta_{\max} - l_Z$ converges in distribution to the Gumbel distribution with cdf $\Psi(\zeta) = \exp(-e^{-\zeta})$, $\zeta \in \mathbb{R}$, where l_Z is given by $F(l_Z) = 1 - (1/Z)$. This result suggests that $\zeta_{\max}$ grows like $\mathcal{O}(l_Z)$ in the limiting case of $Z \to \infty$.

[10]Since we are interested in the asymptotic performance of the PF scheduler with long-term fairness, the selection of the relay-to-user links will be based on the i.i.d. small-scale fading coefficients only [31], and thus, the extreme value theory for i.i.d. random variables is applicable.

In the second part, we use the foregoing lemma to derive the order of growth of Rician fading and Rayleigh fading channel coefficients. Suppose that the random variables $\{H_1, H_2, \ldots, H_M\}$ are i.i.d. Rician random variables with Rician factor κ. Let $a = \kappa/(1+\kappa)$, $v = 1/(\kappa+1)$, and let $\zeta_i = |H_i|^2$. Then, the set $\{\zeta_1, \zeta_2, \zeta_3, \ldots, \zeta_M\}$ represents the magnitude square of the Rician random variables, and the tail of the cdf and pdf of ζ for $\zeta \to \infty$ can be approximated by [36]

$$1 - F(\zeta) \approx \frac{(a\zeta)^{\frac{-1}{4}}}{2}\sqrt{\frac{v}{\pi}}\exp\left(\frac{-(\sqrt{\zeta}-\sqrt{a})^2}{v}\right)$$
$$f(\zeta) \approx \frac{(a\zeta)^{\frac{-1}{4}}}{2\sqrt{v\pi}}\exp\left(\frac{-(\sqrt{\zeta}-\sqrt{a})^2}{v}\right). \quad (34)$$

Therefore, the growth function is given by

$$\lim_{\zeta\to\infty} g(\zeta) = \lim_{\zeta\to\infty}\frac{1-F(\zeta)}{f(\zeta)} = \frac{1}{v} \quad (35)$$

which satisfies the sufficient condition in Lemma 2, and therefore, extreme value theory can be applied. From [37] and [38], we have the following expression:

$$\log\left[-\log F^M\left(l_M + \zeta g(l_M)\right)\right]$$
$$= -\zeta + \frac{\zeta^2}{2!}g'(l_M) + \frac{\zeta^3}{3!}\left[g(l_M)g''(l_M) - 2g'^2(l_M)\right]\ldots$$
$$+\cdots+\frac{e^{-\zeta}+\cdots}{2M}+\frac{5e^{-2\zeta+\cdots}}{2M}+\cdots-\frac{e^{-3\zeta}}{8M^3}+\cdots+\cdots \quad (36)$$

where g' and g'' represent the first and second derivatives of function $g(\cdot)$, respectively. l_M is given by $F(l_M) = 1 - (1/M)$. Then, solving (36) for l_M, we obtain

$$l_M = (\sqrt{v\log M} + \sqrt{a})^2 + \mathcal{O}(\log\log M). \quad (37)$$

Following [39] and substituting $\zeta = \pm\log\log M$ in (36), we can show that

$$\Pr\left\{(\sqrt{v\log M}+\sqrt{a})^2 - \log\log M \le \zeta_{\max} - \mathcal{O}(\log\log M)\right.$$
$$\left.\le (\sqrt{v\log M}+\sqrt{a})^2 + \log\log M\right\} \ge 1 - \mathcal{O}\left(\frac{1}{\log M}\right). \quad (38)$$

Therefore, the growth of $\zeta_{\max}$ is given by $\mathcal{O}(\log M/(\kappa+1))$ for sufficiently large M. Here, $\kappa+1$ acts as a deduction factor to the growth since most of the energy is concentrated in the line-of-sight path, and thus, the channel fluctuation is small for exploiting diversity.

Furthermore, we consider $F^{-1(k)}_{RD_m,j}(\varepsilon)$, $k \in \{1, 2, \ldots, K\}$, as defined in Section IV-A, with $\varepsilon^{(k)} = \varepsilon$. By a similar framework as previously and [40], it can be shown that for CSIT error variance $\sigma_e^2 \in [0, 1)$, $F^{-1(k)}_{RD_m,j}(\varepsilon)$ grows with its noncentrality parameter on the order of $\mathcal{O}((1-\sigma_e^2)|\hat{H}^{(k)}_{RD_m,j}|^2)$, and the growth of $\max_{1\le k\le K}|\hat{H}^{(k)}_{RD_m,j}|^2$ is given by

$$\Pr\left\{-\log\log K \le \max_{1\le k\le K}\left|\hat{H}^{(k)}_{RD_m,j}\right|^2 - \log K \le \log\log K\right\}$$
$$\ge 1 - \mathcal{O}\left(\frac{1}{\log K}\right). \quad (39)$$

Hence, the growth of $\max_{1\le k\le K} F^{-1(k)}_{RD_m,j}(\varepsilon)$ is given by $\mathcal{O}((1-\sigma_e^2)\log K)$ for sufficiently large K, and the term $(1-\sigma_e^2)$ acts as a penalty on the MUD gain due to imperfect CSIT for $\sigma_e^2 \in [0, 1)$. In the final part, we combine the foregoing results to prove the growth of system goodput in different situations.

1) Case 1 (Asymptotic System Goodput for General Growth of Numbers of Users K and Relays M): From (38) and (39), we observe that $\max_{1\le k\le K} F^{-1(k)}_{RD_m,j}(\varepsilon)$ grows with order $\mathcal{O}((1-\sigma_e^2)\log K)$, and the maximum magnitude square among M Rician random variables grows with $\mathcal{O}(\log M/\kappa)$ for large K and M, respectively. Therefore, by considering only the first-order growing terms, the growth of the average system goodput can be written as

$$U_{\text{goodput}}(\mathcal{P}, \mathcal{R}, \mathcal{S})$$
$$= \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F}\frac{s^{(k)}_{m,i,j}}{2n_F} \times \mathcal{O}\left(\log_2\left(\frac{\nu^{(k)}_{m,j}\delta^{(k)}_{m,i}}{\nu^{(k)}_{m,j}+\delta^{(k)}_{m,i}}\right)\right) \quad (40)$$

where $\delta^{(k)}_{m,i} = P^{(k)}_{SR_m,i} l_{SR_m}(\log M/(\kappa+1))$, and $\nu^{(k)}_{m,j} = P^{(k)}_{RD_m,j} l_{RD_m}(1-\sigma_e^2)\log K$.

2) Case 2 (Asymptotic System Goodput for a Large Number of Users K and a Growing Number of Relays M): In this case, we assume that the number of users K is always larger than the number of relays M, and K grows with M such that $\lim_{K,M\to\infty}(K/M) \to \infty$. From (39), there exists a $K_0 > 0$ such that for $K > K_0$, the growth of the maximum of the inverse noncentral chi-square cdf among K users is bounded by $[(1-\sigma_e^2)(\log K - \log\log K)] \le \max_{1\le k\le K} F^{-1(k)}_{RD_m,j}(\varepsilon) \le [(1-\sigma_e^2)(\log K + \log\log K)]$. As a result, we can consider the case for large M and $K > K_0$. By only considering the growing terms and using the fact that the maximum magnitude square among M Rician random variables grows with $\mathcal{O}(\log M/\kappa)$ for sufficiently large M and a large growing number of relays M, the growth of the average system goodput is given by

$$U_{\text{goodput}}(\mathcal{P}, \mathcal{R}, \mathcal{S})$$
$$\overset{(a)}{\approx} \sum_{m=1}^{M}\sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F}$$
$$\times\left\{\frac{s^{(k)}_{m,i,j}}{2n_F}\times\mathcal{O}\left(\log_2\left(\frac{P^{(k)}_{SR_m,i} l_{SR_m}\log M}{\kappa+1}\right)\right)\right\}$$
$$\overset{(b)}{=} \mathcal{O}\left(\log_2\left(\frac{\log M}{\kappa+1}\right)\right)\Bigg\} \quad (41)$$

where (a) is due to the assumption that $\lim_{K,M\to\infty}(K/M) \to \infty$, and (b) is because the channel coefficients of the BS-to-relay links are identical distributed.

3) Case 3 (Asymptotic System Goodput for a Large Number of Relays M and a Growing Number of Users K): In this

case, we assume that the number of relays M is always larger than the number of users K, and M grows with K such that $\lim_{K,M\to\infty}(M/K) \to \infty$. By using the same arguments as in Case II and (38), for a large growing number of users K, we obtain for the growth of the average system goodput

$$\begin{aligned} & U_{\text{goodput}}(\mathcal{P},\mathcal{R},\mathcal{S}) \\ & \overset{(a)}{\approx} \sum_{m=1}^{M} \sum_{k\in\mathcal{U}_m} w_k \sum_{i=1}^{n_F}\sum_{j=1}^{n_F} \\ & \quad \times \left\{ \frac{s_{m,i,j}^{(k)}}{2n_F} \times \mathcal{O}\left(\log_2\left(P_{R_{m,j}}^{(k)} l_{RD_m}\left(1-\sigma_e^2\right)\log K\right)\right)\right\} \\ & \overset{(b)}{=} \mathcal{O}\left(\log_2\left(\left(1-\sigma_e^2\right)\log K\right)\right) \end{aligned} \tag{42}$$

where (a) is due to the assumption that $\lim_{K,M\to\infty}(M/K) \to \infty$, and (b) is because the PF scheduler selects the users according to the small-scale fading coefficient of the relay-to-user links, which are identically distributed.

REFERENCES

[1] A. Doufexi and S. Armour, "Design considerations and physical layer performance results for a 4G OFDMA system employing dynamic subcarrier allocation," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun.*, Sep. 2005, vol. 1, pp. 357–361.

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

[3] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3037–3063, Sep. 2005.

[4] I. Hammerstrom and A. Wittneben, "Power allocation schemes for amplify-and-forward MIMO-OFDM relay links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 2798–2802, Aug. 2007.

[5] S. W. Peters and R. W. Heath, "The future of WiMAX: Multihop relaying with IEEE 802.16j," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 104–111, Jan. 2009.

[6] G. D. Yu, Z. Y. Zhang, Y. Chen, S. Chen, and P. L. Qiu, "Power allocation for non-regenerative OFDM relaying channels," in *Proc. IEEE Int. Conf. Wireless Commun., Netw. Mobile Comput.*, Sep. 2005, pp. 185–188.

[7] Y. Li, W. Wang, J. Kong, W. Hong, X. Zhang, and M. Peng, "Power allocation and subcarrier pairing in OFDM-based relaying networks," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 2602–2606.

[8] L. Huang, R. Mengtain, W. Lan, X. Yisheng, and E. Schulz, "Resource allocation for OFDMA based relay enhanced cellular networks," in *Proc. IEEE 65th Veh. Technol. Conf.*, Apr. 2007, pp. 3160–3164.

[9] M. K. Awad and S. Xuemin, "OFDMA based two-hop cooperative relay network resources allocation," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 4414–4418.

[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] G. Song and Y. Li, "Asymptotic throughput analysis for channel-aware scheduling," *IEEE Trans. Commun.*, vol. 54, no. 10, pp. 1827–1834, Oct. 2006.

[12] S. Sanayei and A. Nosratinia, "Opportunistic downlink transmission with limited feedback," *IEEE Trans. Inf. Theory*, vol. 53, no. 11, pp. 4363–4372, Nov. 2007.

[13] M. Herdin, "A chunk based OFDM amplify-and-forward relaying scheme for 4G mobile radio systems," in *Proc. IEEE Int. Commun. Conf.*, Jun. 2006, pp. 4507–4512.

[14] C. R. N. Athaudage, M. Saito, and J. Evans, "Performance analysis of dual-hop OFDM relay systems with subcarrier mapping," in *Proc. IEEE Int. Commun. Conf.*, May 2008, pp. 4419–4423.

[15] Z. K. M. Ho, V. K. N. Lau, and R. S. K. Cheng, "Closed loop cross layer scheduling for goodput maximization in frequency selective environment with no CSIT," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2007, pp. 299–303.

[16] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks—Part II: Algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[17] I. C. Wong and B. L. Evans, "Optimal OFDMA resource allocation with linear complexity to maximize ergodic weighted sum capacity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2007, pp. 601–604.

[18] Y. Ma, "Rate-maximization scheduling for downlink OFDMA with long term rate proportional fairness," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 3480–3484.

[19] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems—Part I: Chunk allocation," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2734–2744, Sep. 2009.

[20] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1468–1489, Jul. 1999.

[21] R. Narasimhan, "Finite-SNR diversity–multiplexing tradeoff for correlated Rayleigh and Rician MIMO channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3965–3979, Sep. 2006.

[22] N. Yi, Y. Ma, and R. Tafazolli, "Bit and power loading for OFDM with an amplify-and-forward cooperative relay," in *Proc. IEEE Pers., Indoor, Mobile Radio Commun.*, Sep. 2008, pp. 1–5.

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

[24] "Air Interface for Fixed and Mobile Broadband Wireless Access Systems," IEEE 802.16e Working Group, Tech. Rep., Dec. 2005. [Online]. Available: http://standards.ieee.org/getieee802/download/802.16e-2005.pdf

[25] C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[26] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1321, Jul. 2006.

[27] J. A. Bazerque and G. B. Giannakis, "Distributed scheduling and resource allocation for cognitive OFDMA radios," *Mob. Netw. Appl.*, vol. 13, no. 5, pp. 452–462, Oct. 2008. [Online]. Available: http://www.springerlink.com/content/x84742jx50086q6l/

[28] T. C. Y. 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.

[29] S. Boyd, L. Xiao, and A. Mutapcic, *Subgradient Methods*. Stanford, CA: Stanford Univ. Press, Autumn 2003.

[30] D. Gesbert and M. S. Alouini, "How much feedback is multi-user diversity really worth?" in *Proc. IEEE Int. Conf. Commun.*, Jun. 2004, vol. 1, pp. 234–238.

[31] G. Caire, R. R. Muller, and R. Knopp, "Hard fairness versus proportional fairness in wireless communications: The single-cell case," *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1366–1385, Apr. 2007.

[32] "Spatial Channel Model For Multiple Input Multiple Output (MIMO) Simulations," 3GPP TR 25.996 V7.0.0 (2007-06), Tech. Rep.

[33] A. Kumar, *Mobile Broadcasting With WiMAX Principles, Technology & Applications,* 1st ed. Boston, MA: Focal, 2008.

[34] J. Andrews, A. Ghosha, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking,* 1st ed. Englewood Cliffs, NJ: Prentice-Hall, 2007.

[35] H. A. David, *Order Statistics,* 1st ed. New York: Wiley, 1970.

[36] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.

[37] N. T. Uzgoren, "The asymptotic development of the distribution of the extreme values of a sample," in *Studies in Mathematics and Mechanics Presented to Richard von Mises*. New York: Academic, 1954, pp. 346–353.

[38] Q. Zhou and H. Dai, "Asymptotic analysis in MIMO MRT/MRC systems," *Eur. J. Wireless Commun. Netw.*, vol. 2006, no. 2, pp. 1–8, Apr. 2006.

[39] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[40] V. K. N. Lau, W. K. Ng, and D. S. W. Hui, "Asymptotic tradeoff between cross-layer goodput gain and outage diversity in OFDMA systems with slow fading and delayed CSIT," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2732–2739, Jul. 2008.

Derrick Wing Kwan Ng (S'06) received the Bachelor (with First-Class Honors) and Master of Philosophy (M.Phil.) degrees in electronic engineering from the Hong Kong University of Science and Technology (HKUST), Kowloon, Hong Kong, in 2006 and 2008, respectively. He is currently working toward the Ph.D. degree with the University of British Columbia (UBC), Vancouver, BC, Canada.

His research interests include cross-layer optimization for wireless communication systems, resource allocation in multiple-input–multiple-output and orthogonal frequency division multiplexing wireless systems, and communication theory.

Mr. Ng received the Best Paper Award from the 2008 IEEE Third International Conference on Communications and Networking in China. He was also the recipient of the 2009 Four-Year Doctoral Fellowship from UBC, the Sumida and Ichiro Yawata Foundation Scholarship in 2008, and the R&D Excellence scholarship from the Center for Wireless Information Technology from HKUST in 2006.

Robert Schober (M'01–SM'08–F'10) was born in Neuendettelsau, Germany, in 1971. He received the Diplom (Univ.) and Ph.D. degrees in electrical engineering from the University of Erlangen-Nuermberg, Nuermberg, Germany, in 1997 and 2000, respectively.

From May 2001 to April 2002, he was a Postdoctoral Fellow with the University of Toronto, Toronto, ON, Canada, where he was sponsored by the German Academic Exchange Service. Since May 2002, he has been with the University of British Columbia (UBC), Vancouver, BC, Canada, where he is currently a Full Professor and a Canada Research Chair (Tier II) in wireless communications. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing.

Dr. Schober is the Area Editor for modulation and signal design for the IEEE TRANSACTIONS ON COMMUNICATIONS. He received the 2002 Heinz MaierVLeibnitz Award of the German Science Foundation, the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, and the 2008 Charles McDowell Award for Excellence in Research from UBC. In addition, he received best paper awards from the German Information Technology Society, the European Association for Signal, Speech, and Image Processing, the 2006 IEEE International Conference on Ultra-Wideband, the International Zurich Seminar on Broadband Communications, and the 2000 European Wireless Conference.