Resource Allocation and Scheduling in Multi-Cell OFDMA Systems with Decode-and-Forward Relaying

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

***Abstract*—In this paper, we formulate resource allocation and scheduling for multi-cell orthogonal frequency division multiple access (OFDMA) systems with half-duplex decode-and-forward (DF) relaying as a joint optimization problem taking into account multi-cell interference and heterogeneous user data rate requirements. For efficient multi-cell interference mitigation, we incorporate a time slot allocation strategy into the problem formulation. We transform the resulting non-convex and combinatorial optimization problem into a standard convex problem by imposing an interference temperature constraint, which yields a lower bound for the original problem. Subsequently, the transformed optimization problem is solved by dual decomposition and a semi-distributed iterative resource allocation algorithm with closed-form power and subcarrier allocation policies is derived to maximize the average weighted system throughput (bit/s/Hz/base station). Simulation results illustrate that our proposed semi-distributed algorithm achieves practically the same performance as the centralized optimal solution of the original non-convex problem and provides a substantial performance gain compared to single-cell resource allocation and scheduling schemes.**

***Index Terms*—Heterogeneous users, decode-and-forward relay, multi-cell resource allocation, base station coordination, multiuser diversity.**

I. Introduction

Orthogonal frequency division multiple access (OFDMA) is a promising candidate for high speed wireless communication networks including IEEE 802.22 Wireless Regional Area Networks (WRAN), IEEE 802.16 Worldwide Interoperability for Microwave Access (WiMAX), and Long Term Evolution (LTE). In OFDMA, a wide-band frequency spectrum is shared by many orthogonal narrowband subcarriers and data streams from different users are multiplexed on different subcarriers according to a scheduling policy [1], [2]. In a single-cell OFDMA system, the fading coefficients of different subcarriers are likely to be independent for different users, which is known as *multiuser diversity* (MUD). Maximum system spectral efficiency can be achieved by selecting the best user for each subcarrier and adapting the corresponding power. On the other hand, a large amount of work has been devoted to cooperative relaying in wireless networks as it provides coverage extension and throughput gains [3]-[6]. Several efficient relaying protocols such as decode-and-forward (DF), amplify-and-forward (AF), and compress-and-forward (CF) have been proposed in the literature to facilitate relaying. There is no uniformly optimal relaying protocol and each protocol can outperform the others, depending on the system configuration. However, DF relaying has the advantage that conventional transmitter and receiver structures can be employed.

The next generation broadband wireless communication systems are expected to support different data rate services for real-time applications such as video games and video conferencing with certain quality of service (QoS) requirements. This translates into a heavy demand for the spectral resources. The combination of OFDMA and DF relaying provides a possible solution in meeting these demanding requirements, particularly for users at the cell edge. In [3]-[6], best effort resource allocation and scheduling for homogeneous users in DF OFDMA systems are studied for different system configurations. In practice, users are heterogeneous with different QoS requirements such as minimum required data rate, which best effort resource allocation cannot fulfill. Furthermore, [1]-[7] focus on single-cell systems and ignore the co-channel interference caused by adjacent cells. This assumption is valid for small frequency reuse factors. However, aggressive/universal frequency reuse with interference coordination techniques is a new trend in next generation communication systems since it achieves a higher system capacity [8] and existing works considering only a single cell may not be able to reveal the actual performance in a practical system.

In the past decade, a number of interference mitigation techniques have been proposed in the literature, such as successive interference cancellation and interference nulling through multiple antennas [9], [10], for alleviating the negative side-effects of aggressive/universal frequency reuse. Unfortunately, interference cancellation and multiple antenna receivers may be too complex for low-power battery driven mobile units. Recently, base station (BS) coordination, where BSs only share channel state information (CSI), has been proposed as a major technique to mitigate co-channel interference, since it shifts the signal processing burden to the BSs. In [11] and [12], the sum rate performance of a multi-cell time-division multiple access (TDMA) system with half-duplex and full-duplex AF relays is studied, respectively. In [13], the authors investigate

Manuscript received July 2, 2010; revised December 16, 2010; accepted April 2, 2011. The associate editor coordinating the review of this paper and approving it for publication was S. Bhashyam.

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

This paper has been presented in part at the IEEE Global Communications Conference (Globecom 2010). This work has been supported in part by the Natural Science and Engineering Council of Canada (NSERC) under Project STPGP 396545.

Digital Object Identifier 10.1109/TWC.2011.042211.101183

1536-1276/11$25.00 © 2011 IEEE

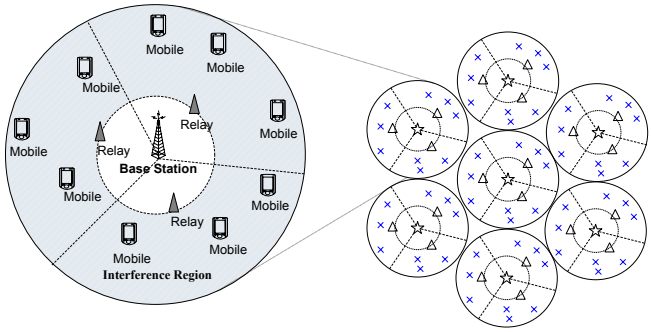

Fig. 1. A cluster with $P = 7$ cells. There are a total of $K = 63$ users in the cluster. Each cell is modeled as two concentric ring-shaped discs and contains $\frac{K}{P} = 9$ users and 3 relays. The shaded part is the region served by the relays.

the sum-rate scaling law of cooperative multi-cell downlink transmission with zero-forcing beamforming. However, the results in [11]-[13] are not suitable for resource allocation and scheduling purposes due to the adopted oversimplified (Wyner) interference model. On the other hand, resource allocation for single hop multi-cell networks with single-carrier and multi-carrier transmission is considered in [14] and [15], [16], respectively. Yet, in all these works, fairness of users is not taken into account for resource allocation which leads to the starvation of weak cell edge users. Furthermore, the results of [14]-[16] cannot be directly extended to the considered case of a multi-cell OFDMA system with DF relaying.

In this paper, we address the above issues. For this purpose, we formulate the scheduling problem in multi-cell OFDMA systems with DF relaying as an optimization problem. We incorporate an effective time slot allocation strategy into the problem formulation to mitigate the interference. To make the problem tractable, we transform it into a convex optimization problem by imposing an interference constraint and introducing time-sharing variables, which results in a lower bound for the original problem. Using dual decomposition, the optimization problem is separated into a master problem and several subproblems which can be solved by the proposed semi-distributed iterative algorithm. Each BS solves its own problem by utilizing its local CSI and exchanges partial interference information with all BSs through the concept of pricing. Therefore, the computational complexity at the BS and the CSI feedback overhead are both significantly reduced compared to optimal centralized scheduling which requires global CSI. In particular, our results show that large savings in computational complexity and signaling overhead are possible with the proposed semi-distributed algorithm at the expense of a small degradation in performance.

The rest of the paper is organized as follows. In Section II, we outline the model for the considered multi-cell OFDMA system with DF relaying. In Section III, we formulate the resource allocation and scheduling design as an optimization problem, and solve this problem by dual decomposition in Section IV. In Section V, we present numerical performance results for the proposed semi-distributed iterative algorithm for multi-cell resource allocation and scheduling. In Section VI, we conclude with a brief summary of our results.

II. System Model for Multi-Cell OFDMA with DF Relaying

In this section, after introducing the notation used in this paper, we present the adopted multi-cell network and channel models.

A. Notation

A complex Gaussian random variable with mean μ and variance σ^2 is denoted by $\mathcal{CN}(\mu, \sigma^2)$, and $\sim$ means "distributed as". $\left[x\right]_b^a = a$, if $x > a$, $\left[x\right]_b^a = x$, if $b \leq x \leq a$, and $\left[x\right]_b^a = b$, if $b < x$. $\left[x\right]^+ = \max\{0, x\}$. $\mathcal{E}\{\cdot\}$ denotes statistical expectation. $|\mathcal{A}|$ represents the cardinality of set $\mathcal{A}$.

B. System Model

We consider a multi-cell OFDMA system with half-duplex DF relaying which consists of P coordinated BSs, M relays, and K mobile users. The users belong to one of two categories, namely, *delay sensitive* users and *non-delay sensitive* users. The *delay-sensitive* users require a minimum constant data rate while *non-delay sensitive* users have no data rate constraint and can be served in a best-effort manner. We adopt an information theoretic approach for the design of resource allocation and scheduling. Therefore, the buffers at the BSs are assumed to be always full and there are no empty scheduling slots due to an insufficient number of source packets at the buffers. All transceivers are equipped with single antennas. We assume universal frequency reuse and the P coordinated BSs share a total bandwidth $\mathcal{B}$. Each cell is modeled by two concentric ring-shaped discs as shown in Figure 1. In this paper, we focus on the resource allocation and scheduling with interference coordination for heterogeneous users who need help from relays, i.e., cell edge users in the shaded region of Figure 1. We assume that there is a separated resource for those users who do not need relays[1]. In the considered model, there is no direct link between the BSs and the users due to heavy blockage and long distance transmission. BSs are connected to a centralized unit with optical fiber backhaul links to facilitate the proposed semi-distributed resource allocation and scheduling algorithm. Nevertheless, each user is only served by one relay and one BS and each relay only serves one BS. In particular, we assume that the users are associated with the relays with the strongest average channels. Furthermore, the information for the desired user is not jointly encoded in different BSs and dirty-paper coding is not considered. The transmission is organized in different time frames. The CSI of all links of a cell is assumed to be perfectly known at the BS of the same cell. All CSI is time invariant within a transmission frame, but time varying from one frame to the next. In each scheduling slot, at the beginning of each frame, scheduling and resource allocation are performed at each BS with the help of a centralized unit. In each frame, the downlink transmission between the BSs and the users via the relays consists of two

[1]The resource allocation for relay assisted users (located between the inner and the outer boundaries in Figure 1) and non-relay assisted users (located inside the inner boundary) is assumed to be done separately. We note that a joint resource allocation for non-relay assisted 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.

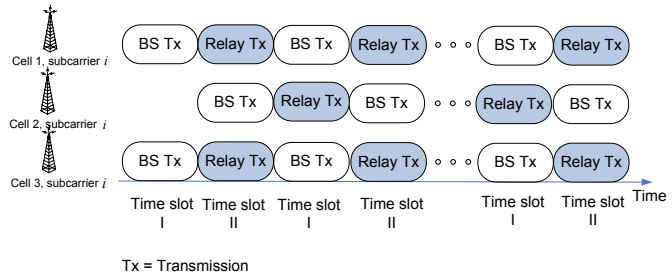

Fig. 2. Example for time slot allocation strategy on subcarrier i with 3 coordinated BSs. The shaded part represents the signals causing strong interference to cell edge users. By considering the link budget, it can be observed that the received signals from neighbouring BSs at the cell edge users are weak because of the large BS-user distances and only transmissions from the relays in neighbouring cells cause noticeable interference to the cell edge users. Thus, in this simple example, the relays in cells 1 and 3 heavily interfere the cell edge users in cells 3 and 1, respectively. On the contrary, the cell edge users in cell 2 experience only a weak interference.

phases. In the first phase, the BSs transmit the signals to the relay stations. Then, in the second phase, the relay stations decode the previously received signals and forward them to the corresponding receivers. Since BS coordination is considered, an effective strategy to mitigate the multi-cell interference is to allocate the two transmission phases in adjacent cells to different time slots (if possible). In this paper, we consider two different time slot allocation strategies on a per subcarrier basis. For example, if cell 1 uses time slot allocation strategy $t = 1$ for subcarrier i, the BS in cell 1 transmits the signal in subcarrier i in the first time slot and a relay forwards the signal to a user in the second time slot. On the other hand, the adjacent cell 2 may use time slot allocation strategy $t = 2$ for subcarrier i and the corresponding BS transmits the signal in subcarrier i in the second time slot and a relay forwards the signal in the next first time slot. Such a strategy helps to minimize the interference to cell edge users as is illustrated in Figure 2. The optimization of the time slot allocation strategy will be incorporated in to the considered resource allocation and scheduling problem. Furthermore, time division duplex (TDD) is assumed for the downlink and uplink transmission.

C. OFDMA Relay Channel Model

We consider the downlink of an OFDMA system with n_F subcarriers. The channel impulse response is assumed to be time-invariant within a frame. Suppose user k is served by BS $p \in \{1, \ldots, P\}$ and relay $m \in \{1, \ldots, M\}$. The received symbol with time slot allocation strategy $t \in \{1, 2\}$ at relay m for user k in subcarrier $i \in \{1, \ldots, n_F\}$ is given by

$$\begin{aligned} Y_{R_m,p}^{(t,k)}(i) &= \sqrt{P_{B_p,R_{m,p}}^{(t,k)}(i) l_{B_p,R_{m,p}}} H_{B_p,R_{m,p}}(i) X^{(k)}(i) \\ &\quad + I_{R_{m,p}}^{(t)}(i) + Z_{R_m}(i), \end{aligned} \tag{1}$$

where $P_{B_p,R_{m,p}}^{(t,k)}(i)$ and $X^{(k)}(i)$ are the transmit power and the transmit symbol for the link between BS p and relay m in subcarrier i in the first relaying phase with time slot allocation strategy t, respectively. $I_{R_{m,p}}^{(t)}(i)$ is the received multi-cell interference in relay m in subcarrier i. $l_{B_p,R_{m,p}}$ represents the path loss between BS p and relay m. $Z_{R_m}(i) \sim \mathcal{CN}(0, \sigma_z^2)$ is the additive white Gaussian noise (AWGN) in subcarrier i at relay m. $H_{B_p,R_{m,p}}(i)$ is the small scale fading coefficient between the BS and relay m in subcarrier i.

Relay m decodes[2] message $X^{(k)}(i)$ and forwards it to user k. Therefore, the signal received at user k in subcarrier i from relay m using time slot allocation strategy t is given by

$$\begin{aligned} Y_{U_m,p}^{(t,k)}(i) &= \sqrt{P_{R_{m,p}}^{(t,k)}(i) l_{R_{m,p}}^{(k)}} H_{R_{m,p}}^{(k)}(i) X^{(k)}(i) \\ &\quad + I^{(t,k)}(i) + Z^{(k)}(i). \end{aligned} \tag{2}$$

Variables $P_{R_{m,p}}^{(t,k)}(i)$, $l_{R_{m,p}}^{(k)}$, and $H_{R_{m,p}}^{(k)}(i)$ are defined in a similar manner as the corresponding variables for the BS-to-relay links except that the signalling direction is from relay m in cell p to user k. $Z^{(k)}(i) \sim \mathcal{CN}(0, \sigma_z^2)$ is the AWGN in subcarrier i at user k. $I^{(t,k)}(i)$ is the received multi-cell interference of user k in subcarrier i in using time slot allocation strategy t and its variance is given by

$$\begin{aligned} \sigma_{t,k}^2(i) &= \mathcal{E}\{|I^{(t,k)}(i)|^2\} \\ &= \sum_{\substack{c=1 \\ c \neq p}}^{P} \sum_{a \in \mathcal{R}_c} \sum_{j \neq k} s_{a,c}^{(t,j)}(i) P_{R_{a,c}}^{(t,j)}(i) \left(l_{R_{a,c}}^{(k)} |H_{R_{a,c}}^{(k)}(i)|^2 \right), \end{aligned} \tag{3}$$

where $l_{R_{a,c}}^{(k)}$ and $H_{R_{a,c}}^{(k)}(i)$ are the path loss and the small scale fading gain in subcarrier i between relay a in cell c and user k, respectively. $s_{a,c}^{(t,j)}(i) \in \{0, 1\}$ is the subcarrier allocation indicator. $\mathcal{R}_c$ is the set of relays which belong to BS c. Note that the amount of interference at user k is a function of the time slot allocation strategy. In (3), for modeling purposes, the interference generated by the neighbouring BSs to the cell edge users is ignored. This is because the relays in the neighbouring cells generate a much larger interference than the corresponding BSs for a typical cell size. For instance, it can be shown that for a cell with radius 2 km and BS-to-relay distances of 1 km[3], the relays in the neighbouring cells cause a 10 dB larger interference than the corresponding BSs to cell edge users.

In practice, different links in a relay network experience asymmetric fading conditions [17]. For example, users are generally surrounded by a large number of scatterers and their locations are random. Hence, a non-line-of sight (NLoS) communication link is expected between the relays and the users. Thus, we model $H_{R_{m,p}}^{(k)}(i)$ as Rayleigh distributed, i.e., $H_{R_{m,p}}^{(k)}(i) \sim \mathcal{CN}(0, 1)$. On the other hand, a strong line-of-sight (LoS) propagation channel is expected between the BS and the relays, since they are placed in relatively high positions in practice and the number of blockages between them are limited. Hence, $H_{B_p,R_{m,p}}(i)$ is modeled as Rician fading with Rician factor κ, i.e., $H_{B_p,R_{m,p}}(i) \sim \mathcal{CN}(\sqrt{\kappa/(1+\kappa)}, 1/(1+\kappa))$.

III. Resource Allocation and Scheduling Design

A. Instantaneous Channel Capacity and System Throughput

In this subsection, we define the adopted system performance measure. Given perfect CSI at the receiver, the

[2]We note that each relay requires a buffer to store the received packets for decoding and re-encoding.

[3]Please note that the interference from the BSs in neighbouring cells is included in the simulation results in Section V.

instantaneous channel capacity between BS p and relay m for user k in subcarrier i using time slot allocation strategy t is given by

$$C_{B_p,R_{m,p}}^{(t,k)}(i) = \frac{1}{2}\log_2\left(1+\Gamma_{B_p,R_{m,p}}^{(t,k)}(i)\right) \quad (4)$$

with signal-to-interference-plus-noise ratio (SINR)

$$\begin{aligned}\Gamma_{B_p,R_{m,p}}^{(t,k)}(i) &= \frac{P_{B_p,R_{m,p}}^{(t,k)}(i) l_{B_p,R_{m,p}} |H_{B_p,R_{m,p}}(i)|^2}{\sigma_{t,R_{m,p}}^2(i)+\sigma_z^2} \\ &\approx \frac{P_{B_p,R_{m,p}}^{(t,k)}(i) l_{B_p,R_{m,p}} |H_{B_p,R_{m,p}}(i)|^2}{\sigma_z^2}, \quad (5)\end{aligned}$$

where the the pre-log factor $\frac{1}{2}$ is due to the two channel uses required for transmitting one message and the approximation in (5) is because the channel capacity between the BS and the relay is limited by channel noise[4], i.e., $\sigma_{t,R_{m,p}}^2(i) = \mathcal{E}\{|I_{R_{m,p}}^{(t)}|^2\} \ll \sigma_z^2$. Similarly, the channel capacity of user k in using subcarrier i and time slot allocation strategy t via relay m in cell p is given by

$$\begin{aligned}C_{U_m,p}^{(t,k)}(i) &= \frac{1}{2}\log_2\left(1+\Gamma_{U_m,p}^{(t,k)}(i)\right) \\ &= \frac{1}{2}\log_2\left(1+\frac{P_{R_{m,p}}^{(t,k)}(i) l_{R_{m,p}}^{(t,k)} |H_{R_{m,p}}^{(k)}(i)|^2}{\sigma_{t,k}^2(i)+\sigma_z^2}\right), \quad (6)\end{aligned}$$

where $\Gamma_{U_m,p}^{(t,k)}(i)$ is the received SINR at user k in subcarrier i.

Now, we define the instantaneous throughput (bit/s/Hz successfully delivered) for user k who is assigned to relay m and BS p as

$$\begin{aligned}\rho_{m,p}^{(k)} = \frac{1}{n_F}\min\Bigg\{&\sum_{i=1}^{n_F}\sum_{t=1}^{2} s_{m,p}^{(t,k)}(i) C_{B_p,R_{m,p}}^{(t,k)}(i), \\ &\sum_{i=1}^{n_F}\sum_{t=1}^{2} s_{m,p}^{(t,k)}(i) C_{U_m,p}^{(t,k)}(i)\Bigg\} \quad (7)\end{aligned}$$

where $s_{m,p}^{(t,k)}(i) \in \{0,1\}$ is the subcarrier and time slot allocation strategy indicator. The *average weighted system throughput* is defined as the total average number of bit/s/Hz/BS successfully decoded at the K users via the M relays and P coordinated BSs and given by

$$\mathcal{U}_{TP}(\mathcal{P},\mathcal{S}) = \frac{1}{P}\sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}} w^{(k)}\rho_{m,p}^{(k)}, \quad (8)$$

where $\mathcal{P}$ and $\mathcal{S}$ are the power and subcarrier allocation policies, respectively. $\mathcal{U}_{m,p}$ is the set of users served by relay m and BS p. $w^{(k)}$ is a positive constant, which is specified in the media access control (MAC) layer and allows the scheduler to give different priorities to different users and to enforce certain notions of fairness such as proportional fairness and max-min fairness [18].

[4]Note that although the multi-cell interference at the relays is ignored in the system model, it is taken into account for the simulation results shown in Section V. It can be verified by simulation and link budget calculations that the multi-cell interference received at the relays is negligible compared to the noise variance for typical cell sizes, relay locations, and transmit powers.

B. Problem Formulation for Resource Allocation and Scheduling Design

In a practical multi-cell system, users located at the cell edge suffer from strong multi-cell interference and weak desired signal strengths which results in poor SINR and poor QoS. Simply increasing the transmission power in one cell does not necessarily improve the overall system performance as interference to other cells increases concurrently. On the other hand, users are heterogeneous with different minimum data rate requirements regardless of their current channel conditions. Therefore, a practical multi-cell scheduler should be able to coordinate the amount of interference created by each cell and fulfill the different data rate requirements of the users even if the corresponding channels are weak. This leads to the following optimization problem.

Problem 1 (Optimization Problem Formulation): The optimal power allocation policy, $\mathcal{P}^*$, and subcarrier allocation policy, $\mathcal{S}^*$, are given by

$$\begin{aligned}&(\mathcal{P}^*,\mathcal{S}^*) = \arg\max_{\mathcal{P},\mathcal{S}} \quad \mathcal{U}_{TP}(\mathcal{P},\mathcal{S}) \\ \text{s.t. } &\text{C1: } \sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F} P_{B_p,R_{m,p}}^{(t,k)}(i) s_{m,p}^{(t,k)}(i) \le P_{B_T},\ \forall p \\ &\text{C2: } \sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F} P_{R_{m,p}}^{(t,k)}(i) s_{m,p}^{(t,k)}(i) \le P_{R_T}, \quad \forall m,p \\ &\text{C3: } \rho_{m,p}^{(k)} \ge R^{(k)}, \quad \forall k \in \mathcal{D}_{m,p} \\ &\text{C4: } \sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2} s_{m,p}^{(t,k)}(i) = 1, \quad \forall p,i \\ &\text{C5: } s_{m,p}^{(t,k)}(i) \in \{0,1\},\ \forall m,p,k,i,t \\ &\text{C6: } P_{B_p,R_{m,p}}^{(t,k)}(i), P_{R_{m,p}}^{(t,k)}(i) \ge 0, \quad \forall m,p,i,k,t, \quad (9)\end{aligned}$$

where $\mathcal{D}_{m,p}$ is the set of *delay sensitive* users who are served by relay m and BS p. Here, C1 (C2) represents the individual power constraint for each BS (relay) with maximum transmit power P_{B_T} (P_{R_T}). C3 enforces the minimum required data rate $R^{(k)}$ for *delay sensitive* users which are chosen by the application layer. C6 is the positive power constraint. Constraints C4 and C5 are imposed to guarantee that each subcarrier is only used by one user in each cell for any two time slots. In other words, intra-cell interference does not exist in the system. Also, C4 ensures that each subcarrier can be transmitted with one time slot allocation strategy only. Note that if the equality in C4 is replaced by an inequality, i.e, $\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2} s_{m,p}^{(t,k)}(i) \le 1$, the same solution as for the considered equality constraint is obtained, as can be verified by the Karush-Kuhn-Tucker (KKT) conditions.

IV. Semi-Distributed Resource Allocation Algorithm

In this section, the considered problem is transformed into a convex optimization problem and solved via dual decomposition. A novel semi-distributed iterative resource allocation algorithm with closed-form power and subcarrier allocation is derived to maximize the average weighted system throughput.

A. Transformation of Optimization Problem

The considered problem is a mixed combinatorial and non-convex optimization problem. The combinatorial nature comes from the integer constraint for the subcarrier and time slot allocation strategy while the non-convexity is caused by the multi-cell interference in (6). In general, a brute force approach is needed to obtain the global optimal solution. In a multi-cell system with P coordinated cells, K users, n_F subcarriers, and two time slot allocation strategies in each cell, there are $P^{K^{2^{n_F}}}$ possible subcarrier assignments which limits the scalability in practical systems. In order to make the problem tractable, we perform a three-step transformation to simplify the problem.

The first step in solving the considered problem is to handle the multi-cell interference. To this end, we introduce an additional constraint C7 to the original problem which is given by

$$\text{C7: } \sigma^2_{t,k}(i) \leq I, \quad \forall k, i, m, p, t. \tag{10}$$

C7 can be interpreted as the maximum multi-cell interference temperature [19] (tolerable interference level) in each subcarrier. By varying[5] the value of I, the schedulers are able to control the amount of interference in each subcarrier to improve the system performance. Furthermore, by substituting $\sigma^2_{t,k}(i)$ in (6) by I, the multi-cell interference can be decoupled from the objective function, which facilitates the design of an efficient resource allocation algorithm. Secondly, we handle the combinatorial constraint in C5 by introducing the following lemma.

Lemma 1 (Optimality of the Time-Sharing relaxation): If a new optimization problem is formed by relaxing the subcarrier selection variable $s^{(t,k)}_{m,p}(i)$ in (9) to be a real value between zero and one instead of a Boolean, i.e., $0 \leq s^{(t,k)}_{m,p}(i) \leq 1$, then the relaxed problem has the same solution as the original optimization problem in (9).

Proof: The above lemma can be proved by using a similar approach as in [20], [21].

For facilitating the time sharing on each subcarrier, we introduce two new variables and define them as $\tilde{P}^{(t,k)}_{B_p,R_{m,p}}(i) = P^{(t,k)}_{B_p,R_{m,p}}(i)s^{(t,k)}_{m,p}(i)$ and $\tilde{P}^{(t,k)}_{R_{m,p}}(i) = P^{(t,k)}_{R_{m,p}}(i)s^{(t,k)}_{m,p}(i)$. These two variables are the actual transmit power of BS p and relay m on subcarrier i for user k in using time slot allocation strategy t under the time-sharing assumption. Then, we can also rewrite constraint C7 in the time-sharing form

$$\text{C7: } \widetilde{\sigma}^2_{t,k}(i) \leq I \quad \forall k, i, m, p, t, \tag{11}$$

where $\widetilde{\sigma}^2_{t,k}(i) = \sigma^2_{t,k}(i)|_{\tilde{P}^{(t,k)}_{R_{m,p}}(i)=P^{(t,k)}_{R_{m,p}}(i)s^{(t,k)}_{m,p}(i)}$. By combing the above two steps, the channel capacities of the first and second hop for user k through relay m in subcarrier i using time slot allocation strategy t in cell p are given by

$$\widetilde{C}^{(t,k)}_{B_p,R_{m,p}}(i) = \frac{1}{2}\log_2\left(1+\widetilde{\Gamma}^{(t,k)}_{B_p,R_{m,p}}(i)\right) \text{ and} \tag{12}$$

$$\widetilde{C}^{(t,k)}_{U_m,p}(i) = \frac{1}{2}\log_2\left(1+\frac{\tilde{P}^{(t,k)}_{R_{m,p}}(i)l^{(t,k)}_{R_{m,p}}|H^{(k)}_{R_{m,p}}(i)|^2}{(I+\sigma^2_z)s^{(t,k)}_{m,p}(i)}\right), \tag{13}$$

[5]The maximum multi-cell interference temperature variable I is not an optimization variable in the proposed framework. However, a suitable value of I can be found via simulation in an off-line manner.

respectively, where $\widetilde{\Gamma}^{(t,k)}_{B_p,R_{m,p}}(i) = \Gamma^{(t,k)}_{B_p,R_{m,p}}(i)|_{P^{(t,k)}_{Bp,R_{m,p}}(i)=\tilde{P}^{(t,k)}_{B_p,R_{m,p}}(i)/s^{(t,k)}_{m,p}(i)}$ is the equivalent SINR for the link between BS p and relay m on subcarrier i under the time-sharing condition. Note that the actual channel capacity in (6) is larger than the scheduled capacity in (13), i.e., $C^{(t,k)}_{U_m,p}(i) \geq \widetilde{C}^{(t,k)}_{U_m,p}(i)$. In fact, $\widetilde{C}^{(t,k)}_{U_m,p}(i)$ can be viewed as the worst-case capacity for resource allocation in the second hop since it is a lower bound for the actual capacity. Finally, the max-min formulation in the objective function can be handled by introducing the extra auxiliary variables $z^{(k)}_{m,p}$, $m \in \{1,\ldots,M\}, p \in \{1,\ldots,P\}, k \in \{1,\ldots,K\}$ and transforming the resource allocation and scheduling optimization problem into its epigraph form[6] [22]:

Problem 2 (Transformed Optimization Problem):

$$\begin{aligned}
&\max_{\mathcal{P},\mathcal{S},z^{(k)}_{m,p}} \sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}} w^{(k)} z^{(k)}_{m,p} \\
&\text{s.t.} \quad \text{C4, C6, C7} \\
&\text{C1: } \sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F}\tilde{P}^{(t,k)}_{B_p,R_{m,p}}(i) \leq P_{B_T}, \ \forall p, \\
&\text{C2: } \sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F}\tilde{P}^{(t,k)}_{R_{m,p}}(i) \leq P_{R_T}, \quad \forall m, p, \\
&\text{C3: } \tilde{\rho}^{(k)}_{m,p} \geq R^{(k)}, \quad \forall k \in \mathcal{D}_{m,p} \\
&\text{C5: } 0 \leq s^{(t,k)}_{m,p}(i) \leq 1, \quad \forall m, p, k, i, t \\
&\text{C8: } \sum_{i=1}^{n_F}\sum_{t=1}^{2} s^{(t,k)}_{m,p}(i)\widetilde{C}^{(t,k)}_{B_p,R_{m,p}}(i) \geq z^{(k)}_{m,p}, \forall p, m, k, \\
&\text{C9: } \sum_{i=1}^{n_F}\sum_{t=1}^{2} s^{(t,k)}_{m,p}(i)\widetilde{C}^{(t,k)}_{U_m,p}(i) \geq z^{(k)}_{m,p}, \forall p, m, k,
\end{aligned} \tag{14}$$

where $\tilde{\rho}^{(k)}_{m,p} = \rho^{(k)}_{m,p}\big|_{C^{(t,k)}_{U_m,p}(i)=\widetilde{C}^{(t,k)}_{U_m,p}(i)}$. Note that the constant term $\frac{1}{n_F\times P}$ is removed from the transformed objective function for simplicity of notation as it does not affect the values of the arguments which maximize the objective function. The extra constraints C8 and C9 represent the hypograph [22] of the original optimization problem in (9). Now, the problem is jointly concave with respect to the optimization variables since the Hessian matrix of the objective function in (14) is negative semi-definite and the inequality constraints are convex. Therefore, the transformed problem is a convex optimization problem and the local optimal solution is identical to the global optimal solution, since the duality gap is equal to zero under some mild conditions [22]. More importantly, it is guaranteed that the global optimal solution can be obtained in polynomial time. In the next section, the considered optimization problem will be solved in the dual domain.

B. Dual Problem Formulation

In this subsection, the transformed resource allocation and scheduling optimization problem is solved by *Lagrange dual decomposition*. For this purpose, we first need the Lagrangian

[6]The epigraph form is a useful tool from optimization theory. It represents a set of points (i.e., a graph) above or below the considered function [22].

function of the primal problem. Upon rearranging terms, the Lagrangian is given by

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta},\mathcal{P},\mathcal{S},z_{m,p}^{(k)})\\
&=\sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}(w^{(k)}-(\mu_{m,p}^{(k)}+\nu_{m,p}^{(k)}))z_{m,p}^{(k)}\\
&+\sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F}s_{m,p}^{(t,k)}(i)\times\Big((\eta^{(k)}+\mu_{m,p}^{(k)})\\
&\times\widetilde{C}_{B_p,R_{m,p}}^{(t,k)}(i)+(\delta^{(k)}+\nu_{m,p}^{(k)})\widetilde{C}_{U_m,p}^{(t,k)}(i)\Big)\\
&-\sum_{p=1}^{P}\lambda_p\Big(\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F}\tilde{P}_{B_p,R_{m,p}}^{(t,k)}(i)-P_{B_T}\Big)\\
&-\sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\gamma_{m,p}\Big(\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F}\tilde{P}_{R_{m,p}}^{(t,k)}(i)\Big)\\
&-\sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\sum_{i=1}^{n_F}\theta_{m,p}^{(t,k)}(i)(\widetilde{\sigma}_{t,k}^{2}(i)-I)\\
&-\sum_{p=1}^{P}\sum_{i=1}^{n_F}\beta_p(i)\Big(\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}s_{m,p}^{(t,k)}(i)-1\Big)\\
&-\sum_{k\in\mathcal{D}_{m,p}}R^{(k)}(\eta^{(k)}+\delta^{(k)})+\sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\gamma_{m,p}P_{R_T},\qquad(15)
\end{aligned}$$

where $\boldsymbol{\lambda}$ is the Lagrange multiplier vector associated with the individual BS power constraints with elements λ_p. $\boldsymbol{\gamma}$ is the Lagrange multiplier vector corresponding to the individual relay power constraints with elements $\gamma_{m,p}$. $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are the Lagrange multiplier vectors corresponding to the data rate constraints in the two time slots with elements $\delta^{(k)}$ and $\eta^{(k)}$, respectively. $\delta^{(k)}=0$ and $\eta^{(k)}=0$ for *non-delay sensitive* users, i.e., $k\notin\mathcal{D}_{m,p},\forall m,p$. Lagrange multiplier vector $\boldsymbol{\beta}$ is connected with the subcarrier usage constraints and has elements $\beta_p(i)$, $i\in\{1,\ldots,n_F\}$. $\boldsymbol{\mu}$ are $\boldsymbol{\nu}$ are the Lagrange multiplier vectors for constraints C8 and C9 in (14) with elements $\mu_{m,p}^{(k)}$ and $\nu_{m,p}^{(k)}$, respectively. $\boldsymbol{\theta}$ is the Lagrange multiplier vector for the maximum received interference temperature constraints in each subcarrier with elements $\theta_{m,p}^{(t,k)}(i)$. The boundary constraints C5 and C6 are absorbed into the KKT conditions when deriving the optimal solution in Section IV-C. Thus, the dual problem is given by

$$\min_{\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta}\geq 0}\ \max_{\mathcal{P},\mathcal{S},z_{m,p}^{(k)}}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta},\mathcal{P},\mathcal{S},z_{m,p}^{(k)}).\qquad(16)$$

In general, the above dual problem can be unbounded if $z_{m,p}^{(k)}\to\infty$. Consider the parts of the dual function in the inner maximization which are related to $z_{m,p}^{(k)}$:

$$\begin{aligned}
&\max_{z_{m,p}^{(k)}}\sum_{p=1}^{P}\sum_{m\in\mathcal{R}_p}\sum_{k\in\mathcal{U}_{m,p}}\Big(w^{(k)}-(\mu_{m,p}^{(k)}+\nu_{m,p}^{(k)})\Big)z_{m,p}^{(k)}\\
&=\begin{cases}0 & \text{if } \mu_{m,p}^{(k)}+\nu_{m,p}^{(k)}=w^{(k)}\\ \infty & \text{otherwise}\end{cases}.\qquad(17)
\end{aligned}$$

In order to have a bounded dual function, the Lagrange multipliers $\mu_{m,p}^{(k)}$ and $\nu_{m,p}^{(k)}$ must satisfy $\mu_{m,p}^{(k)}+\nu_{m,p}^{(k)}=w^{(k)}$. Thus, the dual problem is simplified to

$$\min_{\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta}\geq 0}\ \max_{\mathcal{P},\mathcal{S}}\ \widetilde{\mathcal{L}}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta},\mathcal{P},\mathcal{S}),\qquad(18)$$

where $\widetilde{\mathcal{L}}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta},\mathcal{P},\mathcal{S})=\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta},\mathcal{P},\mathcal{S},z_{m,p}^{(k)})|_{\nu_{m,p}^{(k)}=w^{(k)}-\mu_{m,p}^{(k)}}$. Note that the auxiliary variables $z_{m,p}^{(k)}$ vanish when we set $\nu_{m,p}^{(k)}=w^{(k)}-\mu_{m,p}^{(k)}$ in (15).

C. Semi-Distributed Solution - Subproblem

By dual decomposition, the dual problem is decomposed into a master problem and $P\times M\times n_F\times 2$ subproblems which have identical structure. The dual problem can be solved iteratively where in each iteration each BS solves one local subproblem[7] by utilizing the local CSI and exchanges some information with other BSs to jointly solve the master problem. The subproblem to be solved by BS p is given by

$$\max_{\mathcal{P},\mathcal{S}}\ \widetilde{\mathcal{L}}_{m,p,i}(\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta},\mathcal{P},\mathcal{S}),\qquad(19)$$

where $\widetilde{\mathcal{L}}_{m,p,i}\ (\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\eta},\mathcal{P},\mathcal{S})=$

$$\begin{aligned}
&\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}s_{m,p}^{(t,k)}(i)\Big((\eta^{(k)}+\mu_{m,p}^{(k)})\widetilde{C}_{B_p,R_{m,p}}^{(t,k)}(i)+\\
&(\delta^{(k)}+\nu_{m,p}^{(k)})\widetilde{C}_{U_m,p}^{(t,k)}(i)\Big)+\gamma_{m,p}P_{R_T}\\
&-\beta_p(i)\Big[\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}s_{m,p}^{(t,k)}(i)-1\Big]-\sum_{k\in\mathcal{D}}R^{(k)}(\eta^{(k)}+\delta^{(k)})\\
&-\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\theta_{m,p}^{(t,k)}(i)s_{m,p}^{(t,k)}(i)\left(\widetilde{\sigma}_{t,k}^{2}(i)-I\right)\\
&-\lambda_p\Big(\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\tilde{P}_{B_p,R_{m,p}}^{(t,k)}(i)-P_{B_T}\Big)\\
&-\gamma_{m,p}\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\tilde{P}_{R_{m,p}}^{(t,k)}(i)\\
&-\sum_{g\neq p}^{P}\sum_{q\in\mathcal{R}_g}\sum_{b\in\mathcal{U}_{q,g}}\sum_{t=1}^{2}\theta_{q,g}^{(t,b)}(i)s_{p,q}^{(t,b)}(i)\\
&\times\Big(\tilde{P}_{R_{m,p}}^{(t,k)}(i)l_{R_{m,p}}^{(b)}|H_{R_{m,p}}^{(b)}(i)|^2-I\Big)\qquad(20)
\end{aligned}$$

for a given set of Lagrange multipliers. Let $\tilde{P}_{B_p,R_{m,p}}^{(t,k)*}(i)$, $\tilde{P}_{R_{m,p}}^{(t,k)*}(i)$, and $s_{m,p}^{(t,k)*}(i)$ denote the optimal solution of the subproblem. Then, the KKT conditions reveal

[7]In deed, the subproblem solved at the BS can be further decomposed into $|\mathcal{R}_p|$ smaller subproblems to be solved by the relays. Nevertheless, the resulting extra computational burden may overload the relays and the speed of convergence of the iterative resource allocation algorithm decreases.

that

$$\frac{\partial\widetilde{\mathcal{L}}_{m,p,i}(\ldots)}{\partial\tilde{P}^{(t,k)*}_{B_p,R_{m,p}}(i)} = -\lambda_p + \left(\frac{(\mu^{(k)}_{m,p}+\eta^{(k)})}{2\ln(2)}\right) \times \left(\frac{s^{(t,k)*}_{m,p}(i)|H_{B_p,R_{m,p}}(i)|^2 l_{B_p,R_{m,p}}}{\sigma_z^2 s^{(t,k)*}_{m,p}(i)+\tilde{P}^{(t,k)*}_{B_p,R_{m,p}}(i)|H_{B_p,R_{m,p}}(i)|^2 l_{B_p,R_{m,p}}}\right) \begin{cases} =0, & \tilde{P}^{(t,k)*}_{B_p,R_{m,p}}(i)>0 \\ <0, & \text{otherwise,} \end{cases} \quad (21)$$

$$\frac{\partial\widetilde{\mathcal{L}}_{m,p,i}(\ldots)}{\partial\tilde{P}^{(t,k)*}_{R_{m,p}}(i)} = -\gamma_{m,p} + \Big(\frac{(\nu^{(k)}_{m,p}+\delta^{(k)})s^{(t,k)*}_{m,p}(i)}{2\ln(2)}\Big) \times \left(\frac{l^{(t,k)}_{R_{m,p}}|H^{(k)}_{R_{m,p}}(i)|^2}{s^{(t,k)*}_{m,p}(i)(\sigma_z^2+I)+\tilde{P}^{(t,k)*}_{R_{m,p}}(i)l^{(t,k)}_{R_{m,p}}|H^{(k)}_{R_{m,p}}(i)|^2}\right) - \sum_{g\neq p}^{P}\sum_{q\in\mathcal{R}_g}\sum_{b\in\mathcal{U}_{q,g}}\theta^{(t,b)}_{q,g}(i)\Big(s^{(t,b)}_{q,g}(i)l^{(b)}_{R_{m,p}}|H^{(b)}_{R_{m,p}}(i)|^2\Big) \begin{cases} =0, & \tilde{P}^{(t,k)*}_{R_{m,p}}(i)>0 \\ <0, & \text{otherwise} \end{cases}, \quad (22)$$

and

$$\frac{\partial\widetilde{\mathcal{L}}_{m,p,i}(\ldots)}{\partial s^{(t,k)*}_{m,p}(i)} = W^{(t,k)}_{m,p}(i) - \beta_p(i)\begin{cases} >0, & s^{(t,k)*}_{m,p}(i)=1 \\ =0, & 0<s^{(t,k)*}_{m,p}(i)<1 \\ <0, & s^{(t,k)*}_{m,p}(i)=0 \end{cases}, \quad (23)$$

where

$$W^{(t,k)}_{m,p}(i) = \frac{(\mu^{(k)}_{m,p}+\eta^{(k)})}{2}\Big(\log_2\Big(1+\Gamma^{(t,k)*}_{B_p,R_{m,p}}(i)\Big) - \frac{\Gamma^{(t,k)*}_{B_p,R_{m,p}}(i)}{\ln 2(\sigma_z^2+\Gamma^{(t,k)*}_{B_p,R_{m,p}}(i))}\Big) + \frac{(\nu^{(k)}_{m,p}+\delta^{(k)})}{2} \times \left(\log_2\Big(1+\tilde{\Gamma}^{(t,k)*}_{U_m,p}(i)\Big) - \frac{\tilde{\Gamma}^{(t,k)*}_{U_m,p}(i)}{\ln 2(\sigma_z^2+I+\tilde{\Gamma}^{(t,k)*}_{U_m,p}(i))}\right), \quad (24)$$

$\Gamma^{(t,k)*}_{B_p,R_{m,p}}(i) = \Gamma^{(t,k)}_{B_p,R_{m,p}}(i)\big|_{\tilde{P}^{(t,k)}_{B_p,R_{m,p}}(i)=\tilde{P}^{(t,k)*}_{B_p,R_{m,p}}(i)}$, and $\tilde{\Gamma}^{(t,k)*}_{U_m,p}(i) = \frac{P^{(t,k)*}_{R_{m,p}}(i)l^{(t,k)}_{R_{m,p}}|H^{(k)}_{R_{m,p}}(i)|^2}{(I+\sigma_z^2)}$. Note that there was a term $\sum_{k\in\mathcal{U}_{m,p}}\sum_{t=1}^{2}\theta^{(t,k)}_{m,p}(i)\left(\widetilde{\sigma}^2_{t,k}(i)-I\right)$ in (24), which vanishes for the optimal solution as suggested by the complementary slackness condition [22].

By setting the derivative in (21) to zero, we obtain the optimal transmit power allocation at BS p for user k in subcarrier i to relay m for using time slot allocation strategy t, which can be expressed as

$$\tilde{P}^{(t,k)*}_{B_p,R_{m,p}}(i) = s^{(t,k)}_{m,p}(i)P^{(t,k)*}_{B_p,R_{m,p}}(i) = s^{(t,k)}_{m,p}(i)\left[\frac{(\mu^{(k)}_{m,p}+\eta^{(k)})}{\lambda_p 2\ln(2)} - \frac{\sigma_z^2}{|H_{B_p,R_{m,p}}(i)|^2 l_{B_p,R_{m,p}}}\right]^{P_{B_T}}_0, \quad (25)$$

where $\eta^{(k)}=0$, for *non-delay sensitive* users, i.e., $k\notin\mathcal{D}_{m,p},\forall m,p$. λ_p is chosen such that it satisfies the individual BS power constraint. Power allocation (26) can be interpreted as a *multi-level* water-filling scheme as the water levels of different users can be different. Specifically, the water-levels, $\frac{(\mu^{(k)}_{m,p}+\eta^{(k)})}{\lambda_p 2\ln(2)}$, of *delay-sensitive* users, are generally higher than those of *non-delay sensitive* users, in order to satisfy constraint C3 in (9). Similarly, the optimal transmit power allocation at relay m of cell p for user k in subcarrier i for time slot allocation strategy t is obtained by setting the derivative in (22) to zero, and is given by

$$\tilde{P}^{(t,k)*}_{R_{m,p}}(i) = s^{(t,k)}_{m,p}(i)P^{(t,k)*}_{R_{m,p}}(i) = s^{(t,k)}_{m,p}(i)\left[\frac{(\nu^{(k)}_{m,p}+\delta^{(k)})}{2\ln(2)\,(\gamma_{m,p}+\Omega^{(t,k)}(i))} - \frac{\sigma_z^2+I}{l^{(t,k)}_{R_{m,p}}|H^{(k)}_{R_{m,p}}(i)|^2}\right]^{P_{R_T}}_0, \quad (26)$$

where

$$\Omega^{(t,k)}(i) = \sum_{g\neq p}^{P}\sum_{q\in\mathcal{R}_g}\sum_{b\in\mathcal{U}_{q,g}}\theta^{(t,b)}_{q,g}(i)\Big(s^{(t,b)}_{q,g}(i)l^{(b)}_{R_{m,p}}|H^{(b)}_{R_{m,p}}(i)|^2\Big) \quad (27)$$

represents the interference to the other users created by this power allocation. A large value of $\Omega^{(t,k)}(i)$ results in a lower water-level in the power allocation to reduce the impact on the other users. On the other hand, subcarrier i at BS p is allocated to user k through relay m in using time slot allocation strategy t if

$$s^{(t,k)*}_{m,p}(i) = \begin{cases} 1 & \text{if } W^{(t,k)}_{m,p}(i) = \max\limits_{j,d,z} W^{(z,j)}_{d,p}(i) \\ 0 & \text{otherwise} \end{cases}, \quad (28)$$

where $W^{(z,j)}_{d,p}(i)$ is defined in (24). The derived subcarrier allocation solution has two important implications. First, although time-sharing is assumed for solving the optimization problem, the optimal solution indicates that the maximum system performance is achieved when there is no time-sharing in any subcarrier of any cell. In other words, a subcarrier is only assigned to one user with one time slot allocation strategy in each cell and intra-cell interference is completely avoided. Second, the optimal solution of the original optimization problem is identical to the solution of the problem with time-sharing relaxation, which agrees with Lemma 1.

D. Solution of the Master Dual Problem

The Lagrange dual function is differentiable and, hence, the gradient method can be used to solve the minimization of the master problem in (16) at each BS. The solution is given by

$$\mu^{(k)}_{m,p}[n+1] = \Big[\mu^{(k)}_{m,p}[n] - \xi_1[n] \times \Big(\sum_{i=1}^{n_F}\sum_{t=1}^{2} s^{(t,k)}_{m,p}(i)\times(\widetilde{C}^{(t,k)}_{B_p,R_{m,p}}(i) - \widetilde{C}^{(t,k)}_{U_m,p}(i))\Big)\Big]^{+}_{\mathbb{U}^{(t,k)}_{m,p}}, \forall k,m,p \quad (29)$$

$$\gamma_{m,p}[n+1] = \Bigg[\gamma_{m,p}[n] - \xi_2[n] \times \Bigg(P_{R_T} - \sum_{k\in\mathcal{U}_{m,p}} \sum_{i=1}^{n_F} \sum_{t=1}^{2} \tilde{P}_{R_{m,p}}^{(t,k)}(i)\Bigg)\Bigg]^{+}, \forall m,p \tag{30}$$

$$\lambda_p[n+1] = \Bigg[\lambda_p[n] - \xi_3[n] \times \Bigg(P_{B_T} - \sum_{m\in\mathcal{R}_p} \sum_{k\in\mathcal{U}_{m,p}} \sum_{i=1}^{n_F} \sum_{t=1}^{2} \tilde{P}_{B_p,R_{m,p}}^{(t,k)}(i)\Bigg)\Bigg]^{+}, \forall p \tag{31}$$

$$\eta_k[n+1] = \Bigg[\eta_k[n] - \xi_4[n] \times \Big(\sum_{i=1}^{n_F} \sum_{t=1}^{2} s_{m,p}^{(t,k)}(i)\widetilde{C}_{B_p,R_{m,p}}^{(t,k)}(i) - R^{(k)}\Big)\Bigg]^{+}, \forall k \in \mathcal{D}_{m,p} \tag{32}$$

$$\delta^{(k)}[n+1] = \Bigg[\delta^{(k)}[n] - \xi_5[n] \times \Big(\sum_{i=1}^{n_F} \sum_{t=1}^{2} s_{m,p}^{(t,k)}(i)\widetilde{C}_{U_m,p}^{(t,k)}(i) - R^{(k)}\Big)\Bigg]^{+}, \forall k \in \mathcal{D}_{m,p} \tag{33}$$

$$\theta_{m,p}^{(t,k)}(i)[n+1] = \Bigg[\theta_{m,p}^{(t,k)}(i)[n] - \xi_6[n] \times \Big(\widetilde{\sigma}_{t,k}^2(i) - I\Big)\Bigg]^{+}, \forall m,p,k,i, \tag{34}$$

where $\mathbb{U}_{m,p}^{(t,k)}$ in (29) denotes the projection operator on the feasible set $\mathbb{U}_{m,p}^{(t,k)} = \{\mu_{m,p}^{(k)} | 0 \le \mu_{m,p}^{(k)} \le w^{(k)}\}$. The projection operator can be simply implemented by a clipping function $\Big[\mu_{m,p}^{(k)}[n+1]\Big]_0^{w^{(k)}}$ such that $0 \le \mu_{m,p}^{(k)} \le w^{(k)}$ always holds. Index $n \ge 0$ is the iteration index and $\xi_u[n]$, $u \in \{1,\ldots,6\}$ are positive step sizes. Updating $\beta_p(i)$ is not necessary since it has the same value for each user and relay served by the same BS and it does not affect the subcarrier allocation in (28). $\nu_{m,p}^{(k)}$ can be obtained from $\nu_{m,p}^{(k)} = w^{(k)} - \mu_{m,p}^{(k)}$. Since the transformed problem is convex in nature, it is guaranteed that the algorithm converges to the optimal solution if the chosen step sizes satisfy the infinite travel condition [22], [23]

$$\sum_{n=1}^{\infty} \xi_u[n] = \infty, \quad u \in \{1,\ldots,6\}. \tag{35}$$

In summary, the master problem adjusts the water-levels of (25) and (26) through the gradient update equations (31) and (32) until all individual power constraints of the BSs and the relays are satisfied, respectively. On the other hand, the Lagrange multipliers in update equations (33) and (34) act as extra weightings and water-levels in (25)-(28), respectively, forcing the scheduler to assign more subcarriers and power to *delay-sensitive* users in order to satisfy the data rate requirements. Furthermore, (35) limits the maximum amount of interference received in each subcarrier. Finally, (29) reduces the difference between the capacity of user k in the first and second time slots, which corresponds to the selection of the minimum capacity in (7).

Algorithm 1 Semi-Distributed Iterative Resource Allocation Algorithm

1: Initialize L_{max}, $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, $\boldsymbol{\eta}$, and set iteration index $n = 0$
2: Initialize $\mathcal{P}_n$ and compute $\mathcal{S}_n$ according to (28) for $n = 0$
3: **repeat** {Outer Loop}
4: **for** $i = 1$ to n_F **do**
5: **repeat** {Inner Loop}
6: **for** $p = 1$ to P **do**
7: **for** $t = 1$ to 2 **do**
8: Calculate $P_{B_p,R_{m,p}}^{(t,k)*}(i)$ and $P_{R_{m,p}}^{(t,k)*}(i), \forall k \in \mathcal{U}_{m,p}$, according to (25) and (26), respectively.
9: **end for**
10: **end for**
11: All BSs and relays transmit pilot signals on the assigned subcarrier.
12: Active users feed back $\theta_{q,g}^{(t,b)}(i)(s_{q,g}^{(t,b)}(i) l_{R_{m,p}}^{(b)} |H_{R_{m,p}}^{(b)}(i)|^2)$ to their BS and all BSs exchange the values with the help of a centralized unit through an optical backhaul. The centralized unit calculates $\Omega^{(t,k)}(i)$ in (26) for the active users and distributes the solution to each BS.
13: Update $\mathcal{S}_n(i)$ for each BS according to (28) while assuming the subcarrier allocation in other cells remains unchanged.
14: **until** $\mathcal{P}_n(i)$ and $\mathcal{S}_n(i)$ converge
15: **end for**
16: BS updates $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, $\boldsymbol{\eta}$ according to (29)-(35) and set $n = n + 1$
17: **until** convergence or $n = L_{max}$

E. Semi-Distributed Iterative Algorithm for Practical Implementation

Theoretically, equations (25)-(35) provide a complete solution for the considered multi-cell resource allocation problem. However, (26) involves non-causal knowledge of the resource allocation policies in other cells due to multi-cell interference, which is a hurdle for practical implementation. In this section, we present a semi-distributed and iterative algorithm (Algorithm 1) to bridge the gap between theory and practice. In Algorithm 1, L_{max} is the maximum number of iterations, and $\mathcal{P}_n$ and $\mathcal{S}_n$ are the power allocation and subcarrier allocation policy in the nth iteration, respectively. $\mathcal{P}_n(i)$ and $\mathcal{S}_n(i)$ are the power allocation and subcarrier allocation policy vectors in subcarrier i for all BSs and relays in the nth iteration, respectively. The overall iterative algorithm is implemented by two nested loops. The inner loop, i.e., line 5 to line 14, is solving the maximization in (18) for a given set of Lagrange multipliers for subcarrier i. In line 8, we keep the subcarrier allocation in subcarrier i in other cells fixed and optimize $P_{B_p,R_{m,p}}^{(t,k)}(i)$ and $P_{R_{m,p}}^{(t,k)}(i)$. Then, we use the updated variables in cell p and optimize the power allocation variables in subcarrier i for cell $p+1$, and so on. The same logic is also applied in line 13. In other words, a multi-variable function is optimized over each variable iteratively in the inner loop which is known as the coordinate ascent method. Convergence to the optimal solution for a given set of Lagrange multipliers is ensured since the optimization problem is jointly concave with respect to the optimization variables [24]. On the other hand, the outer loop, i.e., line 3 to line 17, solves the minimization of the master problem by updating the Lagrange multipliers. Table I illustrates the required information exchange between

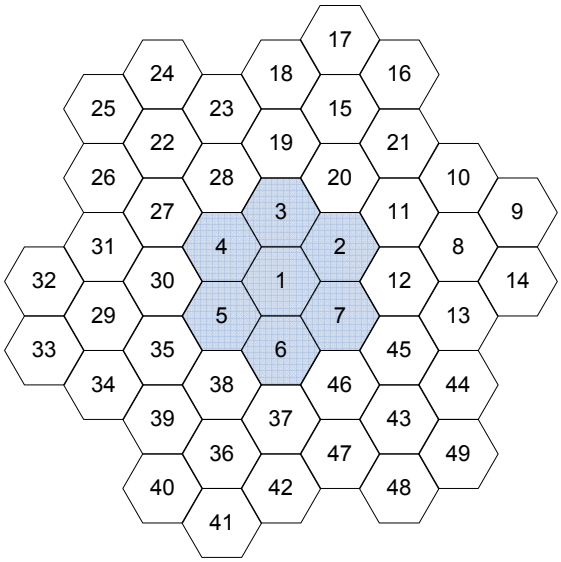

Fig. 3. A multi-cell networks with 49 cells which share the same bandwidth $\mathcal{B}$. The shaded central part is the cluster of $P = 7$ coordinated cells.

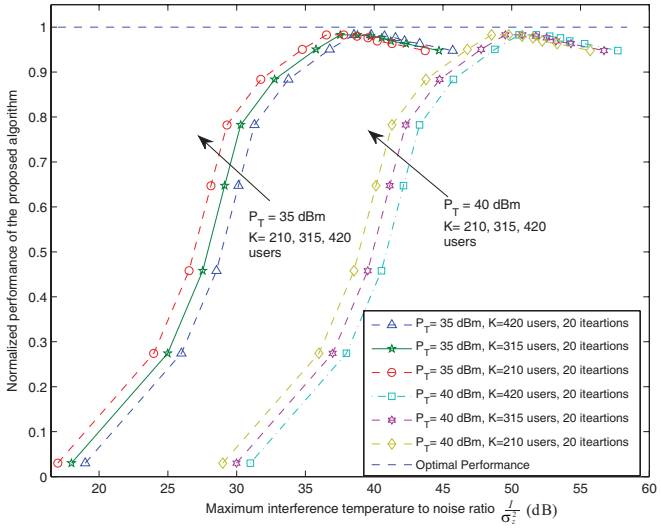

Fig. 4. The normalized performance of the proposed algorithm versus the maximum interference-temperature-to-noise ratio $\frac{I}{\sigma_z^2}$ for different values of P_T and different numbers of users K with $P = 7$ coordinated cells and $M = 21$ relays. The y-axis is normalized by the performance of the optimal centralized algorithm.

the different entities of the network for solving the master problem and the subproblems with the proposed semi-distributed iterative algorithm.

V. Results and Discussions

In this section, we evaluate the system performance using simulations. A multi-cell system with 49 cells is considered where the central cluster of $P = 7$ cells are coordinated as shown in Figure 3. Each coordinated cell is modeled as two concentric ring-shaped discs where the outer boundary has a radius of 2 km and the inner boundary a radius of 1 km, cf. Figure 1. There are $M = 21$ relay stations in the cluster and each cell has $N = \frac{M}{P} = \frac{21}{7} = 3$ relays which are equally distributed at the inner cell boundary in each cell for assisting the transmission. There are K/P active cell edge users uniformly distributed in the outer ring of each cell. The number of subcarriers is $n_F = 128$ with carrier center frequency 2.5 GHz, bandwidth $\mathcal{B} = 5$ MHz, and $w^{(k)} = 1, \forall k$. Each subcarrier has a bandwidth of 39 kHz and a noise variance of $N_0 = -128$ dBm. The 3GPP path loss model is used [25]. The small scale fading coefficients of the BS-to-relay links are generated as independent and identically distributed (i.i.d.) Rician random variables with Rician factor $\kappa = 6$ dB, while the small scale fading coefficients of the relay-to-user links are i.i.d. Rayleigh fading. For simplicity, log-normal shadowing is ignored in the simulations, since its impact on the performance of cell edge users is small compared to the impact of path loss. Since not all BSs are coordinated, we model the uncoordinated multi-cell interference as part of the noise variance σ_z^2 [16]. We assume that the maximum transmit power per cell is P_T and each transmission device, i.e., BS or relay, has a maximum transmit power $P_{R_T} = P_{B_T} = \frac{P_T}{N+1}$. The average weighted system throughput is obtained by counting the number of packets which are successfully decoded by the users averaged over both macroscopic (path loss) and microscopic (multipath) fading. Unless further specified, the number of iterations is defined as the product of the number of inner loops and the number of outer loops in Algorithm 1.

A. System Throughput versus Maximum Interference Temperature I

In this section, we focus on the impact of the value of I on the system performance. As can be seen from Section IV-A, the interference temperature I, which is the key for transforming the original problem into a convex optimization problem, plays an important role in the proposed semi-distributed algorithm. The value of I puts a limit on the maximum transmit power from other coordinated cells by controlling the amount of interference temperature[8]. Figure 4 shows the normalized performance of the proposed algorithm versus the value of I for different P_T and different numbers of users K in $P = 7$ coordinated BSs. The y-axis is normalized by the optimal performance obtained by solving the original non-convex problem (9) using a centralized algorithm[9], such that it demonstrates the achievable fraction of the optimal performance. The x-axis is the interference temperature-to-noise ratio, i.e., $\frac{I}{\sigma_z^2}$. We assume that there are always 7 *delay sensitive* users with data rate requirement $R^{(k)} = 0.1$ bit/s/Hz in the coordinated cells[10], $k \in \mathcal{D}_{m,p}$, while the remaining users are *non-delay sensitive*. It can be seen that for a wide range of $\frac{I}{\sigma_z^2}$, we can achieve more than 95% of the optimal performance and enjoy the convexity of the

[8]In practice, the values of I for implementing the proposed algorithm can be found in an off-line manner.

[9]For the centralized algorithm, the centralized unit is assumed to have the CSI of all links in the network (including the interference links) to perform subcarrier allocation and optimal power allocation by following a similar approach as in [21]. Note, however, that the proposed semi-distributed algorithm is guaranteed to be solved in polynomial time with linear complexity in each iteration while the original problem in (9) has an exponential complexity in M, P, t, and n_F.

[10]The target cell-edge performance for 4G is around 0.1 bit/s/Hz/user [26].

TABLE I
QUANTIZATION TABLE FOR THE FEEDBACK/FEEDFORWARD VARIABLES AND LAGRANGE MULTIPLIERS .

Variables	Number of bits
$\tilde{P}^{*(t,k)}_{R_{m,p}}(i)$	6
$\theta^{(t,b)}_{q,g}(i)s^{(t,b)}_{q,g}(i)l^{(b)}_{R_{m,p}}\|H^{(b)}_{R_{m,p}}(i)\|^2$	6
$\lambda_p, \gamma_{m,p}, \mu^{(k)}_{m,p}, \eta_k, \delta_k$	3
$\Omega^{(t,k)}(i)$	6

transformed problem. Furthermore, the choice of I is not sensitive to the number of users for a given value of P_T. On the other hand, as expected, the optimal value of I is proportional to P_T since the amount of multi-cell interference increases with P_T; a higher value of I is needed to reflect the actual interference temperature.

B. Convergence of the Semi-Distributed Algorithm and Signaling Overhead

Figures 5 and 6 illustrate the evolution of the Lagrange multipliers of the semi-distributed algorithm over time for different transmit powers P_T. The x-axis represents the number of outer loop iterations and the number of inner loop iterations is set to 2. The results in Figures 5 and 6 are averaged over 10000 independent adaptation processes where each adaptation process involves different realizations for the path loss and the multipath fading. For comparison, the figures also contain results for the realistic case where the CSI feedback, the Lagrange multipliers, and the information exchanged between the BSs and the relays in each iteration are quantized with finite bit resoultion[11]. The number of bits used for quantization of the variables in this simulation are listed in Table I. The results show that the semi-distributed algorithm converges fast and typically achieves 90-95% of the optimal value within 10 outer-loop iterations. As expected, the quantization does not affect the speed of convergence significantly but causes a small deviation from the optimal value in the steady state.

Figure 7 depicts the signaling overhead versus the number of users for both the centralized scheme and the proposed semi-distributed algorithm. For the semi-distributed algorithm, we quantize the involved variables as shown in Table I. For the centralized scheme, all CSI and interference levels fed back to the centralized unit are quantized with 6 bits. It can be observed from Figure 7 that the proposed semi-distributed iterative algorithm results in a significant reduction in signaling overhead compared to the centralized resource allocation algorithm, especially when the number of users in the coordinated cells is large. However, even for a comparatively small number of users in the cluster (e.g. $K = 50$), the amount of overhead for the semi-distributed iterative algorithm is still less than that of the centralized algorithm if the proposed algorithm is limited to 10 iterations (product of inner and outer loop iterations). We will show in the next subsection that 10 iterations are typically enough to achieve a close-to-optimal performance.

[11]The quantizer was designed off-line using the Lloyd-Max algorithm.

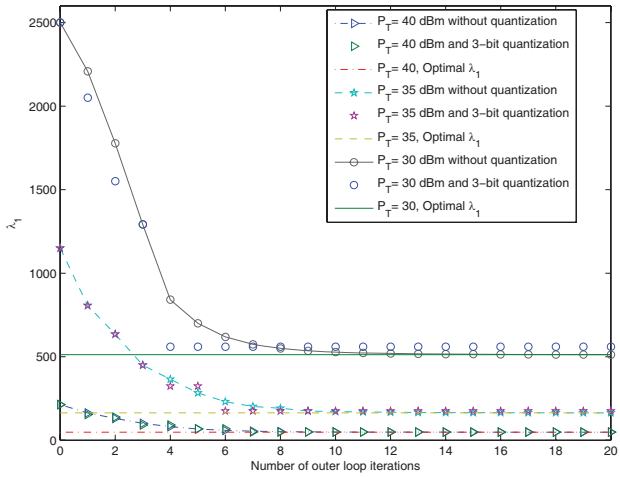

Fig. 5. Lagrange multiplier λ_1 versus number of outer loop iterations with $K = 210$ users and $M = 21$ relays for different transmit power levels. There are 7 delay-sensitive users with data rate requirement $R^{(k)} = 0.1$ bit/s/Hz.

Remark 1: For calculation of the total amount of signalling overhead of the proposed algorithm shown in Figure 7, we first define $Q[\cdot]_x$ as a x bit quantizer used to quantize the input variable and $\Delta^{(t,b)}_{q,g,m,p}(i) = \theta^{(t,b)}_{q,g}(i)s^{(t,b)}_{q,g}(i)l^{(b)}_{R_{m,p}}|H^{(b)}_{R_{m,p}}(i)|^2$. For the proposed semi-distributed iterative algorithm, the amount of feedback required can be calculated as

$$\underbrace{K \times n_F \times Q\left[|H^{(k)}_{R_{m,p}}(i)|^2\right]_6}_{A} + \text{Number of iterations} \times \Bigg\{ \underbrace{Q\left[\Delta^{(t,b)}_{q,g,m,p}(i)\right]_6 \times n_F \times 7}_{B} \underbrace{Q\left[\Omega^{(t,k)}(i)\right]_6 \times n_F \times 7}_{C} + \underbrace{2 \times \left(Q\left[\Delta^{(t,b)}_{q,g,m,p}(i)\right]_6 \times n_F \times 7\right)}_{D} + \underbrace{Q\left[P^{*(t,k)}_{R_{m,p}}(i)\right]_6 \times 7 \times n_F}_{E} \Bigg\}, \quad (36)$$

where variable A is the CSI feedback of the relay-to-user links from the relays to their home BS in the 7 cells; B represents the information passing from the 7 BSs to the centralized unit via the optical backhaul; C corresponds to the feedforward information from the centralized unit to the 7 BSs; D is the feedback information from the users to their home BS in the 7 cells; E is the feedforward power allocation information from the home BS to its relays in the 7 cells.

Remark 2: To illustrate the time scale of the proposed distributed iterative algorithm, we adopt the following assumptions. Suppose we use part of the uplink data channel for information exchange purposes in a TDD system and assume baseline scheme 1 is used in the uplink for transmission. In baseline scheme 1, each BS performs its own resource allocation and completely ignores the multi-cell interference. For baseline scheme 1 and the considered model parameters, the estimated average capacity of the uplink channel in each cell is

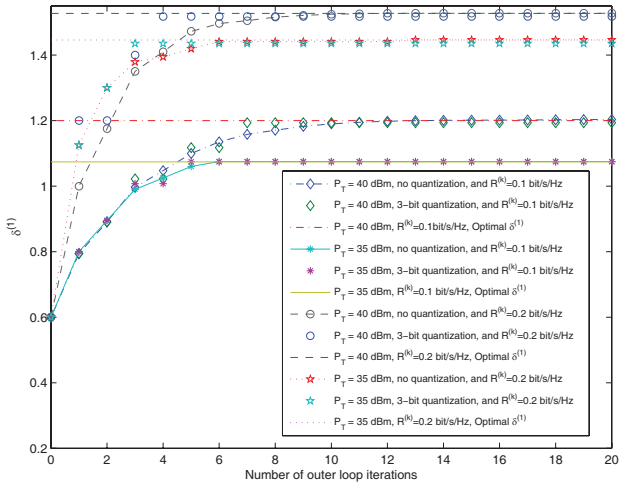

Fig. 6. Lagrange multiplier $\delta^{(1)}$ versus number of outer loop iterations with $K = 210$ users and $M = 21$ relays for different transmit power levels and data rate requirements for 7 *delay sensitive* users.

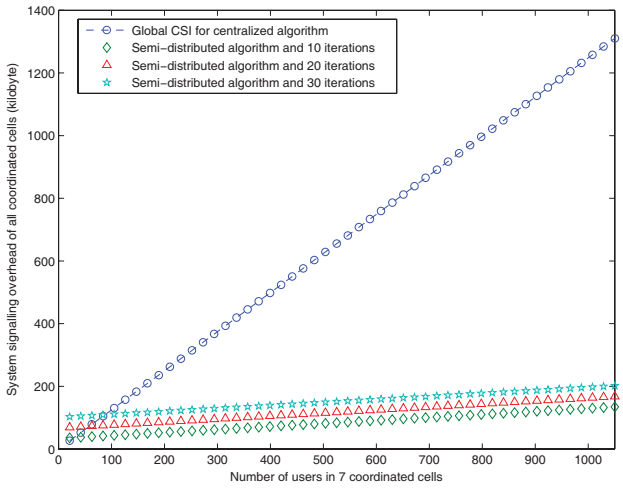

Fig. 7. Signaling overhead versus number of users for $n_F = 128$, $P = 7$ coordinated cells, $M = 21$ relays, and 7 *delay-sensitive users* with data rate requirement $R^{(k)} = 0.1$ bit/s/Hz.

around 0.4 bit/s/Hz/cell[12], as can be observed in Figure 8. The total amount of information exchange for 20 and 10 iterations is 672000 bits and 403200 bits for 7 cells with a total of $K = 175$ users, respectively, cf. Figure 7. Assuming all the information is conveyed by the bottleneck wireless links, the execution time for 20 and 10 iterations of the proposed algorithm are upper bounded by $672000 \text{ bits}/(0.4 \text{ bit/s/Hz/cell} \times 5 \text{ MHz} \times 7 \text{ cells}) = 48$ ms and $403200 \text{ bits}/(0.4 \text{ bit/s/Hz/cell} \times 5 \text{ MHz} \times 7 \text{ cells}) = 28$ ms, respectively. 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 200 ms for pedestrian users [27]. Therefore, the scheduling and resource allocation results obtained with the distributed algorithm are still valid after 10-20 iterations.

[12]The wireless link capacity is the bottleneck in the feedback path of the proposed algorithm.

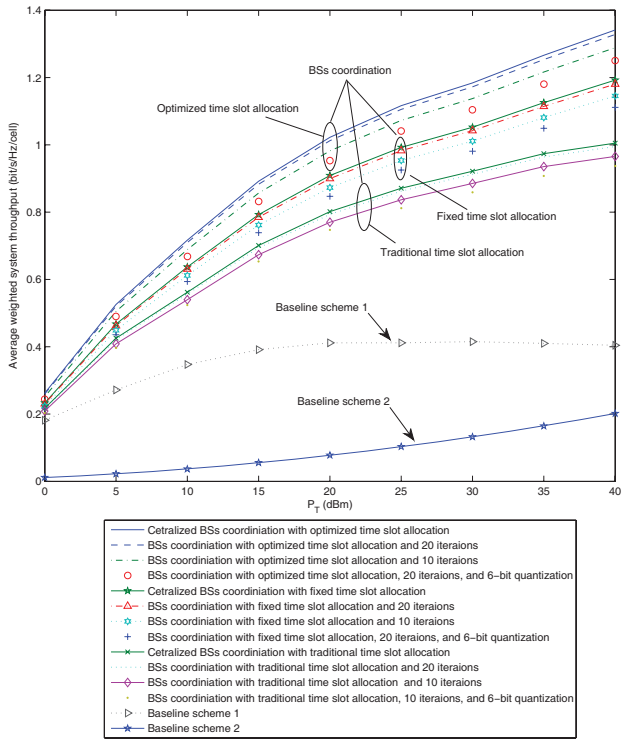

Fig. 8. Average weighted system throughput versus total transmit power per cell for different resource allocation and scheduling algorithms. $K = 210$ users and $P = 7$ coordinated BSs.

C. System Throughput versus Transmit Power

Figure 8 illustrates the average weighted system throughput versus the total transmit power in each cell P_T. There are $K = 210$ users in the cluster and there are 7 *delay sensitive* users with data rate requirement $R^{(k)} = 0.1$ bit/s/Hz, $k \in \mathcal{D}_{m,p}$, while the remaining users are *non-delay sensitive* and served by best effort. We are interested in studying the performance of the proposed semi-distributed algorithm under different system configurations. The value of I in the proposed algorithm in each simulation point is chosen such that we always achieve more than 95% of the performance of the optimal resource allocation and scheduling algorithm. We first demonstrate the performance gain achieved by the time slot allocation strategy for the proposed semi-distributed algorithm with different system configurations, namely, BS coordination with optimized time slot allocation strategy (presented in the main text), BS coordination with fixed time slot allocation strategy, and BS coordination with traditional time slot allocation strategy. For the BS coordination with fixed time slot allocation strategy, the cells with odd index p employ $t = 1$ for all subcarriers and the remaining cells use $t = 2$ for all channel realizations. The traditional time slot allocation strategy is realized by assigning $t = 1$ to all BSs and subcarriers. As can be observed, the optimized time slot allocation strategy provides a significant power gain in the high transmit power regime (multi-cell interference limited environment) compared to the fixed and traditional time slot allocation strategies. In all cases, even with only 20 iterations, the proposed semi-distributed algorithm closely approaches the performance of the optimal centralized scheduling algorithm. On the other hand, the performance loss due to quantization is small which

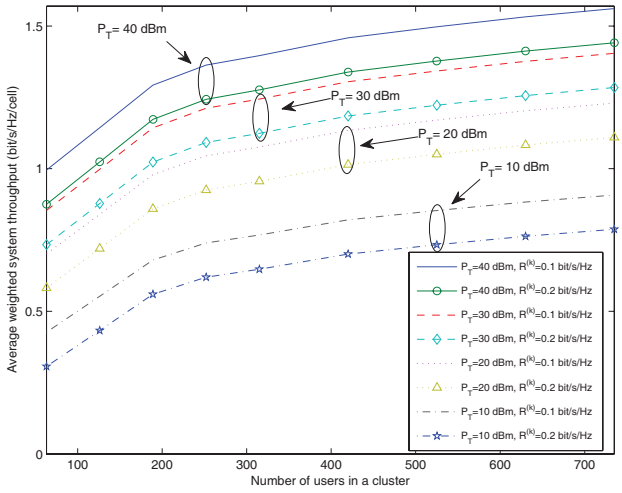

Fig. 9. Average weighted system throughput versus number of users in a cluster with different data rate requirements and transmit power levels for the proposed semi-distributed resource allocation and scheduling algorithms.

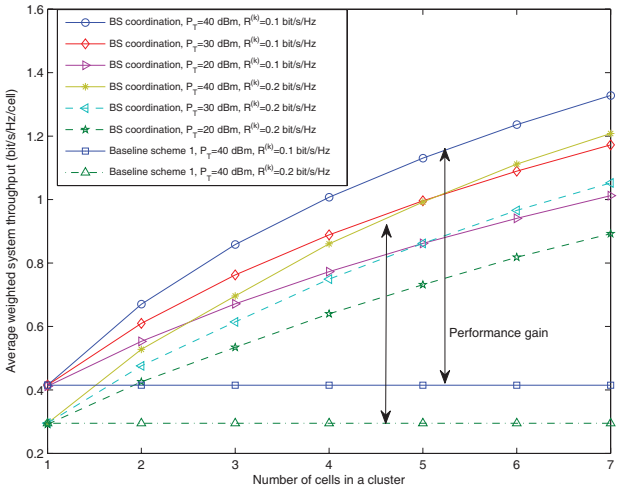

Fig. 10. Average weighted system throughput versus number of coordinated cells in a cluster with different data rate requirements and transmit power levels for different resource allocation algorithms. The double sided arrows represent the performance gain achieved by the proposed semi-distributed resource allocation and scheduling algorithm.

suggests that the proposed algorithm can be implemented in practice with finite bit resolution.

For comparison, Figure 8 also contains the performance of two baselines schemes. As mentioned before, in baseline scheme 1, each BS performs its own resource allocation for the DF relaying system and completely ignores the multi-cell interference. For baseline scheme 2, interference from coordinated cells is completely avoided by setting the frequency reuse factor to $\frac{1}{P}$. In both baseline schemes, each BS is assumed to have the CSI of its own cell. Then, the optimal power allocation and subcarrier allocation for these two schemes can be calculated by equation (9)[13]. In the low transmit power regime, i.e., $P_T < 5$ dBm, baseline scheme 1 achieves a similar performance as the proposed semi-distributed resource allocation algorithm. This is because noise is the dominating factor for system performance for low transmit powers and the effect of interference coordination is less significant. However, as the total transmit power increases, the operating point of the cluster is shifting from noise limited to interference limited. The performance of the proposed semi-distributed algorithm scales with the transmit power and achieves a substantial performance gain compared to baseline scheme 1, since the throughput of the latter is saturated due to strong multi-cell interference. On the other hand, although baseline scheme 2 is able to scale with the transmit power, the proposed algorithm achieves a much higher spectral efficiency. This is due to the fact that baseline scheme 2 sacrifices spectrum efficiency to avoid multi-cell interference.

D. System Throughput versus Number of Users

Figure 9 depicts the average weighted system throughput versus the number of users with different transmit power and user data rate requirements. The number of iterations for the proposed algorithm is 10. It can be observed that the average system throughput increases with the number of users for the proposed resource allocation and scheduling algorithm. This is because as the number of active users increase, the scheduler has a higher chance to select some *non-delay sensitive users* who have both strong channel conditions to their home cell and weak channels to the other cells. This effect can be interpreted as multi-user diversity (MUD) in multi-cell systems. However, when the data rate requirements become more stringent, the scheduler loses degrees of freedom in the resource allocation and scheduling since it needs to serve the *delay sensitive* users despite their potentially weak channel qualities, which diminishes the MUD gain.

[13]To solve the single-cell optimization problem, we need to set the multi-cell interference level to zero and reduce the number of coordinated cells to 1.

E. System Throughput versus Number of Coordinated Cells

Figure 10 shows the average weighted system throughput versus the number of coordinated cells (the size of a cluster) with different transmit power levels and data rate requirements. There are $K = 30$ users in each cell and each cell has one *delay sensitive* user with certain data rate requirements, while the remaining users are *non-delay sensitive*. The number of iterations for the proposed algorithm is set to 20, for a better illustration of the performance gain achieved by BS coordination. The average weighted system throughput increases with the number of cells in a cluster for the proposed resource allocation and scheduling algorithm. This is because additional cells in a cluster provide additional degrees of freedom in the spatial dimension, which can effectively mitigate the multi-cell interference. However, if the data rate requirements become more stringent, less resources in the BSs can be coordinated which diminishes the ability to mitigate multi-cell interference. On the other hand, the system performance of baseline scheme 1 does not scale with the size of the cluster, as the multi-cell interference is completely ignored for resource allocation and scheduling. Note that when the number of coordinated cells in the cluster is equal to one, the proposed algorithm is equivalent to baseline scheme 1.

VI. Conclusion

In this paper, we have formulated the resource allocation and scheduling design for multi-cell OFDMA systems with DF relaying as a mixed non-convex and combinatorial optimization problem, in which multi-cell interference and heterogenous user data rate requirements are taken into consideration. For improved interference mitigation, we have incorporated an effective time slot allocation strategy into the problem formulation. By imposing an additional interference temperature constraint and relaxing the subcarrier allocation constraints, the considered problem has been transformed into a convex problem. An iterative semi-distributed resource allocation algorithm with closed-form power and subcarrier allocation policies has been derived via dual decomposition and requires only local CSI at each BS. Simulation results have shown that the proposed algorithm approaches the optimal performance in a small number of iterations even if the information exchanged between the different nodes of the network is quantized, which confirms the practicality of the proposed scheme.

References

[1] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, pp. 127–134, Dec. 2005.

[2] G. Song, Y. Li, and L. J. Cimini, "Joint channel-and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Trans. Commun.*, vol. 57, pp. 2109–2121, July 2009.

[3] Y. Cui, V. K. N. 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, pp. 5096–5102, Oct. 2009.

[4] H.-X. Li, H. Yu, H.-W. Luo, J. Guo, and C. Li, "Dynamic subchannel and power allocation in OFDMA-based DF cooperative relay networks," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2008, pp. 1–5.

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

[6] L. Weng and R. D. Murch, "Cooperation strategies and resource allocations in multiuser OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 2331–2342, June 2009.

[7] T. Wang, A. Cano, G. B. Giannakis, and J. N. Laneman, "High-performance cooperative demodulation with decode-and-forward relays," *IEEE Trans. Commun.*, vol. 55, pp. 1427–1438, July 2007.

[8] "Downlink inter-cell interference co-ordination/avoidance–evaluation of frequency reuse," 3GPP TSG-RAN WG1 contribution R1-061374, tech. rep., 2006.

[9] J. G. Andrews, "Interference cancellation for cellular systems: a contemporary overview," *IEEE Wireless Commun. Mag.*, vol. 12, pp. 19–29, Apr. 2005.

[10] J. G. Andrews, W. Choi, and R. W. Heath, "Overcoming interference in spatial multiplexing MIMO cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, pp. 95–104, Dec. 2007.

[11] O. Simeone, O. Somekh, Y. Bar-Ness, and U. Spagnolini, "Uplink throughput of TDMA cellular systems with multicell processing and amplify-and-forward cooperation between mobiles," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 2942–2951, Aug. 2007.

[12] O. Somekh, O. Simeone, H. V. Poor, and S. Shamai, "Cellular systems with full-duplex amplify-and-forward relaying and cooperative base-stations," in *Proc. IEEE Intern. Sympos. on Inform. Theory*, June 2007, pp. 16–20.

[13] O. Somekh, O. Simeone, Y. Bar-Ness, A. M. Haimovich, and S. Shamai, "Cooperative multicell zero-forcing beamforming in cellular downlink channels," *IEEE Trans. Inf. Theory*, vol. 55, pp. 3206–3219, July 2009.

[14] A. Gjendemsj, D. Gesbert, G. E. Oien, and S. G. Kiani, "Binary power control for sum rate maximization over multiple interfering links," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 3164–3173, Aug. 2008.

[15] Z. Liang, Y. H. Chew, and C. C. Ko, "A linear programming solution to subcarrier, bit and power allocation for multicell OFDMA systems," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Apr. 2008, pp. 1273–1278.

[16] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allocation in downlink multicell OFDMA networks," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 2835–2848, July 2009.

[17] H. A. Suraweera, R. H. Y. Louie, Y. Li, G. K. Karagiannidis, and B. Vucetic, "Two hop amplify-and-forward transmission in mixed Rayleigh and Rician fading channels," *IEEE Commun. Lett.*, vol. 13, pp. 227–229, Apr. 2009.

[18] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks–part II: algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 625–634, Mar. 2005.

[19] "Report of the Spectrum Ffficiency Working," FCC Spectrum Policy Task Force, tech. rep., Nov 2002. Available: http://www.fcc.gov/sptf/reports.html.

[20] 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, pp. 1747–1758, Oct. 1999.

[21] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, pp. 1310–1321, July 2006.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[23] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," notes for EE392o Stanford University Autumn, 2003-2004.

[24] D. P. Bertsekas, *Nonlinear Programming*, 2nd edition. Athena Scientific, 1999.

[25] "Spatial channel model for multiple input multiple output (MIMO) simulations," 3GPP TR 25.996 V7.0.0 (2007-06), tech. rep.

[26] M. Rumney, *LTE and the Evolution to 4G Wireless: Design and Measurement Challenges*, 1st edition. Agilent Technologies, 2009.

[27] J. G. Andrews, A. Ghosha, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*, 1st edition. Prentice Hall PTR, 2007.

Derrick Wing Kwan Ng (S'06) received the bachelor degree with First class honor and Master of Philosophy (M. Phil.) degree in electronic engineering from the Hong Kong University of Science and Technology (HKUST) in 2006 and 2008, respectively. He is currently working toward the Ph.D. degree in the University of British Columbia (UBC). His research interests include cross-layer optimization for wireless communication systems, resource allocation in OFDMA wireless system and communication theory. He received the Best Paper Award at the IEEE Third International Conference on Communications and Networking in China 2008. He was awarded the IEEE Student Travel Grants for attending the IEEE WCNC 2010 and the IEEE ICC 2011. He was also the recipient of the 2009 Four Year Doctoral Fellowship from the UBC, Sumida & Ichiro Yawata Foundation Scholarship in 2008 and R&D Excellence scholarship from the Center for Wireless Information Technology in the HKUST in 2006.

Robert Schober (M'01, SM'08, F'10) was born in Neuendettelsau, Germany, in 1971. He received the Diplom (Univ.) and the Ph.D. degrees in electrical engineering from the University of Erlangen-Nuermberg in 1997 and 2000, respectively. From May 2001 to April 2002 he was a Postdoctoral Fellow at the University of Toronto, Canada, sponsored by the German Academic Exchange Service (DAAD). Since May 2002 he has been with the University of British Columbia (UBC), Vancouver, Canada, where he is now a Full Professor and 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 received the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation (DFG), 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 (ITG), the European Association for Signal, Speech and Image Processing (EURASIP), IEEE ICUWB 2006, the International Zurich Seminar on Broadband Communications, and European Wireless 2000. Dr. Schober is also the Area Editor for Modulation and Signal Design for the IEEE TRANSACTIONS ON COMMUNICATIONS.