Energy-Efficient Resource Allocation in Multi-Cell OFDMA Systems with Limited Backhaul Capacity

Derrick Wing Kwan Ng, *Student Member, IEEE,* Ernest S. Lo, *Member, IEEE,* and Robert Schober, *Fellow, IEEE*

***Abstract*—We study resource allocation for energy-efficient communication in multi-cell orthogonal frequency division multiple access (OFDMA) downlink networks with cooperative base stations (BSs). We formulate the resource allocation problem for joint BS zero-forcing beamforming (ZFBF) transmission as a non-convex optimization problem which takes into account the circuit power consumption, the limited backhaul capacity, and the minimum required data rate. We transform the considered problem in fractional form into an equivalent optimization problem in subtractive form, which enables the derivation of an efficient iterative resource allocation algorithm. In each iteration, a low-complexity suboptimal semi-orthogonal user selection policy is computed. Besides, by using the concept of perturbation function, we show that in the considered systems under some general conditions, the duality gap with respect to the power optimization variables is zero despite the non-convexity of the primal problem. Thus, dual decomposition can be used in each iteration to derive an efficient closed-form power allocation solution for maximization of the energy efficiency of data transmission (bit/Joule delivered to the users). Simulation results illustrate that the proposed iterative resource allocation algorithm converges in a small number of iterations, and unveil the trade-off between energy efficiency, network capacity, and backhaul capacity: (1) In the low transmit power regime, an algorithm which achieves the maximum spectral efficiency may also achieve the maximum energy efficiency; (2) a high spectral efficiency does not necessarily result in a high energy efficiency; (3) spectral efficiency is always limited by the backhaul capacity; (4) energy efficiency increases with the backhaul capacity only until the maximum energy efficiency is achieved.**

***Index Terms*—Energy efficiency, green communication, network MIMO, limited backhaul, non-convex optimization, resource allocation.**

I. Introduction

ORTHOGONAL frequency division multiple access (OFDMA) is the preferred multiple access scheme for high speed wireless multiuser communication networks, such as 3GPP Long Term Evolution Advanced (LTE-A) [1], IEEE 802.16 Worldwide Interoperability for Microwave Access (WiMAX) [2], and IEEE 802.22 Wireless Regional Area Networks (WRANs) [3], due to its high spectral efficiency and resistance to multipath fading. OFDMA converts a wideband channel into a number of orthogonal narrowband subcarrier channels which facilitate the multiplexing of users data. In practice, the fading coefficients of different subcarriers are likely statistically independent for different users. With channel state information at the transmitter (CSIT), the maximum system capacity can be achieved by selecting the best user for each subcarrier and adapting the corresponding transmit power [4], [5].

On the other hand, cooperative communication for wireless networks has received considerable interest in both industry and academia as it provides extra degrees of freedom in resource allocation. A particularly interesting approach is base station (BS) cooperation for mitigation of strong multi-cell interference caused by aggressive/universal frequency reuse in the network. In the past decade, a number of interference mitigation techniques have been proposed in the literature, including successive interference cancellation (SIC) and interference nulling through multiple antennas, for alleviating the negative side-effects of aggressive/universal frequency reuse. Unfortunately, those techniques may be too complex for low-power battery driven mobile receiver units. On the contrary, BS cooperation, which is known as network multiple-input multiple-output (MIMO), shifts the signal processing burden to the BSs and provides a promising system performance [6]-[11]. In particular, all BSs share the channel state information (CSI) and the data of all users through backhaul communication links, which enables coordinated transmission. In [6], the sum rate of multi-cell zero-forcing beamforming (ZFBF) systems was studied for the Wyner interference model and a large number of users. In [7] and [8], the authors derived the optimal block diagonalization precoding matrix and the optimal max-min beamformer for multi-cell environments, respectively. However, the results in [6]-[8] are based on the ideal backhaul assumption such that an unlimited amount of control signals, user channel information, and precoding data can be exchanged. In practice, the backhaul capacity can be limited due to the deployment costs of the backhaul links. Besides, if a multi-carrier system is considered, the results in [6]-[8], which are valid for the single-carrier case, may no longer be applicable. Furthermore, numerous resource allocation algorithms were designed for different system configurations utilizing only the CSI in multi-cell systems but ignoring the possibility of data exchange, e.g. [9]-[11]. Yet, this kind of cooperation may not be able to fully exploit the potential performance gains achievable by BS cooperation, since the backhaul capacity is not fully utilized.

Manuscript received November 1, 2011; revised April 8 and July 31, 2012; accepted August 1, 2012. The associate editor coordinating the review of this paper and approving it for publication was S. Bhashyam.

D. W. K. Ng and R. Schober are with the Institute for Digital Communications, University of Erlangen-Nürnberg, Germany (e-mail: {kwan, schober}@LNT.de).

E. S. Lo is with the Centre Tecnològic de Telecomunicacions de Catalunya - Hong Kong (CTTC-HK) (e-mail: ernest.lo@cttc.hk).

This paper has been presented in part at the IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 2012.

Digital Object Identifier 10.1109/TWC.2012.083112.111951

Recently, an increasing interest in high data rate services such as video conferencing and online high definition video streaming has led to a tremendous demand for a better utilization of limited resources, i.e., energy and bandwidth. Multi-cell OFDMA with BS cooperation is considered as a possible solution for fulfilling this demand [12]-[14]. In [12] and [13], user assignment and BS assignment in multi-cell OFDMA systems with backhaul capacity constraints were studied, respectively. In [14], the authors proposed dynamic frequency allocation with fractional frequency reuse and equal power allocation across all cooperating BSs. In all studies [6]-[14], a substantial capacity gain and better interference management are reported compared to non-cooperative systems. Yet, the advantages of BS cooperation do not come for free. They have significant financial implications for service providers due to the high power consumption in electronic circuitries, radio frequency (RF) transmission, and data exchange via backhaul links. Thus, *energy efficiency* (bit-per-Joule) may be a better performance metric compared to *system capacity* (bit-per-second-per-Hz) in evaluating the utilization of resources in such systems [15]-[20]. In [18], power adaptation for maximizing the energy efficiency in frequency-selective channels is considered for a multi-carrier single cell system. In [19], a risk-return model was proposed for energy-efficient power allocation. In [20], the authors studied joint power allocation and user selection for OFDMA hybrid marco-cell and micro-cell systems. However, energy efficiency has not been considered for multi-cell systems with limited backhaul capacity in the literature, e.g. [6]-[20], at least not from a resource allocation point of view.

In this paper, we address the above issues and study the trade-off between energy efficiency, backhaul capacity, and network capacity. To this end, we formulate the resource allocation problem for energy-efficient communication in multi-cell OFDMA systems with limited backhaul capacity as an optimization problem. By exploiting the properties of fractional programming, the considered non-convex optimization problem[1] in fractional form is transformed into an equivalent optimization problem in subtractive form with a tractable solution, which can be found with an iterative algorithm. In each iteration, a sub-optimal low-complexity user selection policy is computed and ZFBF is performed. We show that the duality gap for the resulting power allocation problem is zero[2] when the number of subcarriers is sufficiently large, despite the non-convexity of the problem[3]. As a result, dual decomposition is used in each iteration to derive a closed-form power allocation solution for maximization of the network energy efficiency.

II. Multi-Cell OFDMA Network Model

In this section, after introducing the notation used in this paper, we present the adopted channel and signal models.

[1]Non-convex optimization is a general terminology referring to an optimization problem neither minimizing a convex function over convex sets, nor maximizing a concave function over convex sets.

[2]Note that the result of zero duality gap in multi-carrier system derived in this paper can be applied to any well defined function and the considered power allocation optimization problem is just an illustrative example.

[3]In other words, the proposed resource allocation algorithm is optimal with respect to the power allocation variables.

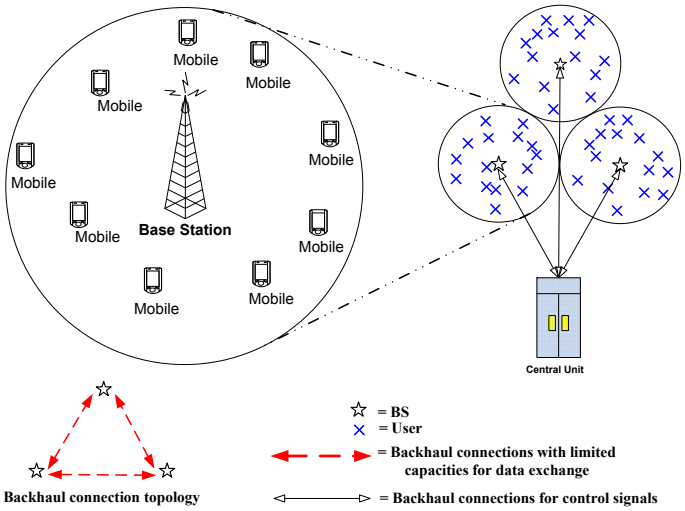


Fig. 1. A multi-cell system with $M = 3$ cells and a fully connected backhaul link topology. There are in total $K = 45$ users in the system. Each transceiver is equipped with a single antenna.

A. Notation

In this paper, the following conventions are adopted. $[x]^+ = \max\{0, x\}$. $\mathbb{C}^{N\times M}$ is the space of all $N \times M$ matrices with complex entries. $\mathbb{R}^N$ represents an N-dimensional real-valued column vector. $\|\cdot\|$ and $|\cdot|$ denote the Euclidean norm of a matrix/vector and the absolute value of a complex number, respectively. Operator $\left[\,\cdot\,\right]_{a,b}$ refers to the element in row a and column b of a matrix. $[\cdot]^\dagger$ and $[\cdot]^T$ represent the conjugate transpose and transpose operations, respectively. $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S}$. The operators " $\geq$ " and " $\leq$ " for vectors are defined element-wise. $\mathcal{E}\{\cdot\}$ denotes statistical expectation.

B. Multi-Cell System Model and Central Unit

We consider a multi-cell OFDMA network which consists of a total of M BSs and K mobile users. All transceivers are equipped with a single antenna, cf. Figure 1. We assume universal frequency reuse and the M BSs share a total bandwidth of $\mathcal{B}$ Hertz. The global CSI is assumed to be perfectly known at a *central unit* and all computations are performed in this unit. Based on the available CSI, the central unit decides the resource allocation policy and broadcasts it to all BSs via backhaul connections which are dedicated to control signals only. On the other hand, all BSs are cooperating with each other by sharing the CSI and the data symbols of all selected users via capacity limited backhaul communication links. Note that the energy consumptions incurred by exchanging CSI and other overheads such as control signals are not considered here since they are relatively insignificant compared to the resources used for data exchange.

C. OFDMA Channel Model

We consider an OFDMA system with n_F subcarriers. The channel impulse response is assumed to be time-invariant within each frame. Suppose user $k \in \{1, \ldots, K\}$ is associated with BS $m \in \{1, \ldots, M\}$. Let $w^k_{B_m}(i)$ be the precoding coefficient used by BS m in subcarrier $i \in \{1, \ldots, n_F\}$ for

user k. Then, the transmitted signal from BS m to all selected users on subcarrier i is given by

$$\sum_{k\in\mathcal{S}(i)} x_m^k(i) = \sum_{k\in\mathcal{S}(i)} w_{B_m}^k(i)\sqrt{P_{B_m}^k(i)}u^k(i), \quad (1)$$

where $x_m^k(i) = w_{B_m}^k(i)\sqrt{P_{B_m}^k(i)}u^k(i)$ is the pre-coded signal transmitted from BS m for user k on subcarrier i, $P_{B_m}^k(i)$ is the transmit power for the link between BS m and user k in subcarrier i, $u^k(i)$ is the transmitted information symbol for user k on subcarrier i, and $\mathcal{S}(i)$ is the set of users selected for using subcarrier i and the cardinality of the set is $|\mathcal{S}(i)| \leq M, \forall i$.

The signal received from the M BSs at user k on subcarrier i is given by

$$\begin{aligned} & Y^k(i) \quad (2)\\ &= \left(\sum_{c=1}^{M} H_{B_c}^k(i)w_{B_c}^k(i)\sqrt{P_{B_c}^k(i)l_{B_c}^k}\right)u^k(i)\\ &+ \underbrace{\sum_{m=1}^{M}\sum_{\substack{j\in\mathcal{S}(i)\\ j\neq k}} \sqrt{P_{B_m}^j(i)l_{B_m}^k}H_{B_m}^k(i)w_{B_m}^j(i)u^j(i) + z^k(i)}_{\text{Multiple Access Interference}}, \end{aligned}$$

where $l_{B_m}^k$ represents the path loss between BS m and user k, $z^k(i)$ is the additive white Gaussian noise (AWGN) in subcarrier i at user k with zero mean and variance σ_z^2, and $H_{B_m}^k(i)$ is the small scale fading coefficient between BS m and user k in subcarrier i.

D. Backhaul Model

In practice, the backhaul signal model depends on the specific implementation. For instance, digital subscriber lines (DSL) and optical fibers are able to deliver high data rates by using orthogonal frequency division multiplexing (OFDM) and wavelength division multiplexing (WDM), respectively. Yet, the media over which is transmitted in the backhaul in both cases are different. In order to provide a general model for the backhaul, we do not assume a particular type/medium for the backhaul. Instead, we focus on the backhaul capacity of the N_m outgoing backhaul connections of BS m, i.e., $R_{B_{m_{N_m}}}$. The value of N_m depends on the backhaul connection topology. For instance, a fully connected topology in a 3-cell system, cf. Figure 1, requires $N_m = 2$ outgoing connections for each BS. Furthermore, in order to isolate the considered problem from specific implementation assumptions, we assume that each backhaul has a fixed average power consumption of P_{BH}.

III. Resource Allocation and Scheduling

A. Instantaneous Channel Capacity

In this subsection, we define the adopted system performance measure. Given perfect CSI at the receiver, the maximum channel capacity between all the cooperating BSs and user k on subcarrier i with subcarrier bandwidth $\frac{\mathcal{B}}{n_F}$ is given by

$$C^k(i) = \frac{\mathcal{B}}{n_F}\log_2\Big(1+\Gamma^k(i)\Big), \quad (3)$$

$$\Gamma^k(i) = \frac{\left|\sum_{c=1}^{M} H_{B_c}^k(i)w_{B_c}^k(i)\sqrt{P_{B_c}^k(i)l_{B_c}^k}\right|^2}{\sigma_z^2 + I^k(i)}, \quad (4)$$

$$I^k(i) = \sum_{\substack{j\in\mathcal{S}(i)\\ j\neq k}}\Big|\sum_{m=1}^{M}\sqrt{P_{B_m}^j(i)}w_{B_m}^j(i)\sqrt{l_{B_m}^k}H_{B_m}^k(i)\Big|^2, \quad (5)$$

where $\Gamma^k(i)$ and $I^k(i)$ are the received signal-to-interference-plus-noise ratio (SINR) and the received interference power at user k on subcarrier i, respectively.

The *weighted system capacity* is defined as the total number of bits successfully delivered to the K mobile users and is given by

$$U(\mathcal{P},\mathcal{W},\mathcal{S}) = \sum_{m=1}^{M}\sum_{k\in\mathcal{A}_m}\alpha_k\sum_{i=1}^{n_F}s^k(i)C^k(i), \quad (6)$$

where $\mathcal{P}$, $\mathcal{W}$, and $\mathcal{S}$ are the power, precoding coefficient, and subcarrier allocation policies, respectively. $\mathcal{A}_m$ is the user admission set of BS m and each user can only be admitted to one BS. $s^k(i) \in \{0,1\}$ is the subcarrier allocation indicator. Note that since the resources in the network are limited in general, not every user in the admission set $\mathcal{A}_m$ can be allocated subcarriers in all time instances. $0 < \alpha_k \leq 1$ is a positive constant provided by the upper layers, which allows the resource allocator to give different priorities to different users and to enforce certain notions of fairness such as proportional fairness and max-min fairness [21], [22]. On the other hand, for designing a resource allocation algorithm for energy-efficient communication, the total power consumption should be included in the optimization objective function. Thus, we model the power dissipation in the system as the sum of two static terms and one dynamic term as follows [23]:

$$\begin{aligned} U_{TP}(\mathcal{P},\mathcal{W},\mathcal{S}) = \quad & P_C\times M + \delta\times P_{BH} \quad (7)\\ & + \sum_{m=1}^{M}\sum_{k\in\mathcal{A}_m}\sum_{i=1}^{n_F}\varepsilon P_{B_m}^k(i)|w_{B_m}^k(i)|^2 s^k(i), \end{aligned}$$

where P_C is the constant signal processing power[4] required at each BS which includes the power dissipations in the transmit filter, mixer, frequency synthesizer, and digital-to-analog converter, etc. $P_C \times M$ represents the total signal processing power consumed by the M BSs. The second term in (7) denotes the total power dissipation in the backhaul links where δ is an integer variable which indicates the number of backhaul links in the system. For instance, the topology considered in Figure 1 requires $\delta = \sum_{m=1}^{M} N_m = 6$ backhaul connections. The last term in (7) represents the total power consumption in the power amplifiers of the M BSs. $\varepsilon \geq 1$ is

[4]Note that by taking into account the circuit power consumption in the equation, we have a more accurate system model for revealing system energy efficiency. This is because the circuit power consumption will increase substantially in next generation communication systems due to their sophisticated transceivers designs [23].

a constant which accounts for the inefficiency of the power amplifier. For example, if $\varepsilon = 5$, it means that for every 10 Watts of power radiated in the RF, 50 Watts are consumed in the power amplifier and the power efficiency is $\frac{1}{\varepsilon} = \frac{1}{5} = 20\%$.

Hence, the *weighted energy efficiency* of the considered system is defined as the total average number of bits/Joule

$$U_{eff}(\mathcal{P}, \mathcal{W}, \mathcal{S}) \quad = \quad \frac{U(\mathcal{P}, \mathcal{W}, \mathcal{S})}{U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S})}. \tag{8}$$

B. Optimization Problem Formulation

The optimal power allocation policy, $\mathcal{P}^*$, precoding policy, $\mathcal{W}^*$, and subcarrier allocation policy, $\mathcal{S}^*$, can be obtained by solving

$$\begin{aligned}
&\max_{\mathcal{P},\mathcal{W},\mathcal{S}} \; U_{eff}(\mathcal{P}, \mathcal{W}, \mathcal{S}) \\
\text{s.t. } &\text{C1: } \sum_{k\in\mathcal{A}_m} \sum_{i=1}^{n_F} |w_{B_m}^k(i)|^2 P_{B_m}^k(i) s^k(i) \le P_{T_m}, \quad \forall m, \\
&\text{C2: } \sum_{m=1}^{M} \sum_{k\in\mathcal{A}_m} \sum_{i=1}^{n_F} s^k(i) C^k(i) \ge R_{\min}, \\
&\text{C3: } \sum_{k\in\mathcal{A}_m} \sum_{i=1}^{n_F} s^k(i) C^k(i) \le R_{\max_m}, \quad \forall m, \\
&\text{C4: } \sum_{k=1}^{K} s^k(i) \le M, \quad \forall i, \quad \text{C5: } s^k(i) = \{0,1\}, \; \forall i, k, \\
&\text{C6: } P_{B_m}^k(i) \ge 0, \; \forall i, k, m,
\end{aligned} \tag{9}$$

where C1 is the individual power constraint of BS m. The value of P_{T_m} in C1 puts a limit on the amount of interference generated to the non-cooperative cells in the downlink. C2 specifies the minimum system data rate requirement $R_{\min}$. Note that although variable $R_{\min}$ in C2 is not an optimization variable in this paper, a balance between energy efficiency and aggregate system capacity can be struck by varying $R_{\min}$. In particular, $R_{\min}$ can be used to provide a guaranteed quality of service to the system and to avoid solutions that are highly energy-efficient but have a low spectral efficiency. Furthermore, we note that even with the system level minimum data rate requirement, the considered network MIMO case is different from the single-cell MIMO case. This is because the performance of the network MIMO system may be limited by the backhaul capacity and the antennas in the system are not co-located. In C3, $R_{\max_m} = \min\{R_{B_{m_1}}, R_{B_{m_2}}, \ldots, R_{B_{m_{N_m}}}\}$. The operator $\min\{\cdot\}$ in $R_{\max_m}$ accounts for the fact that the system capacity contributed by a BS is limited by the bottleneck backhaul capacity of that BS. Indeed, C3 is a generalized constraint on the backhaul capacities which is applicable to different topologies such as the *star connection topology* [24, Chapter 1] and the *fully connected topology*[5] [24, Chapter 4]. Besides, C3 puts a limit on the maximum data rate of each BS due to the limited backhaul capacity. If $R_{B_{\max_m}} \to \infty \, \forall m$, then C3 is always satisfied automatically, i.e., the backhaul capacity is much larger than the wireless link capacity. C4 is the subcarrier reuse constraint. C4 and C5 are imposed to guarantee that each subcarrier can be shared by M users, but each user can only use a subcarrier once. In other words, selected users are not allowed to multiplex different messages on the same subcarrier, since a sophisticated receiver would be required at the users, such as an SIC receiver, to recover more than one messages.

[5]In a star connection topology, the network has a hub to convey messages. In other words, all the message exchanges between BSs have to first pass through this hub. In a fully connected topology, all BSs are connected to each other. As a result, a fully connected network does not require the usage of switches/hubs.

IV. Solution of the Optimization Problem

The objective function in (9) is a ratio of two functions which generally results in a non-convex function. We note that there is no standard approach for solving non-convex optimization problems. However, in order to derive an efficient resource allocation algorithm for the considered problem, we introduce the following transformation.

A. Transformation of the Objective Function

The objective function in (9) can be classified as nonlinear fractional program [25]. For the sake of notational simplicity, we define $\mathcal{F}$ as the set of feasible solutions of the optimization problem in (9). Without loss of generality, we define the maximum energy efficiency q^* of the considered system as

$$\begin{aligned}
q^* &= \frac{U(\mathcal{P}^*, \mathcal{W}^*, \mathcal{S}^*)}{U_{TP}(\mathcal{P}^*, \mathcal{W}^*, \mathcal{S}^*)} \\
&= \max_{\mathcal{P},\mathcal{W},\mathcal{S}} \frac{U(\mathcal{P}, \mathcal{W}, \mathcal{S})}{U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S})}, \quad \forall \{\mathcal{P}, \mathcal{W}, \mathcal{S}\} \in \mathcal{F}.
\end{aligned} \tag{10}$$

We are now ready to introduce the following Theorem.

Theorem 1: The maximum energy efficiency q^* is achieved *if and only if*

$$\begin{aligned}
&\max_{\mathcal{P},\mathcal{W},\mathcal{S}} \quad U(\mathcal{P}, \mathcal{W}, \mathcal{S}) - q^* U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S}) \\
&= \quad U(\mathcal{P}^*, \mathcal{W}^*, \mathcal{S}^*) - q^* U_{TP}(\mathcal{P}^*, \mathcal{W}^*, \mathcal{S}^*) = 0,
\end{aligned} \tag{11}$$

for $U(\mathcal{P}, \mathcal{W}, \mathcal{S}) \ge 0$ and $U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S}) > 0$.

Proof: Please refer to [25] or [26] for a proof of Theorem 1.

By Theorem 1, for any optimization problem with an objective function in fractional form, there exists an equivalent[6] objective function in subtractive form, e.g. $U(\mathcal{P}, \mathcal{W}, \mathcal{S})$ $-q^* U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S})$, in the considered case. As a result, we can focus on the equivalent objective function in the rest of the paper.

Remark 1: The problem formulation above focuses on energy efficiency maximization which can be interpreted as a generalized problem formulation for aggregate weighted network throughput maximization. Indeed, the value of q^* in (11) can be interpreted as the penalty to the energy efficiency due to exceedingly high power consumption. If we force $q^* = 0$, i.e., there is no penalty in using exceedingly high power, then the transformed objective function $U(\mathcal{P}, \mathcal{W}, \mathcal{S}) - q^* U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S})$ will become the weighted network aggregate throughput[7].

[6]Here, "equivalent" means that both problem formulations lead to the same resource allocation policies.

[7]Note that the optimal value of q^* for energy efficiency maximization has to be found via optimization.

TABLE I
ITERATIVE RESOURCE ALLOCATION ALGORITHM.

Algorithm 1 Iterative Resource Allocation Algorithm

1: Initialize the maximum number of iterations L_{max} and the maximum tolerance ϵ
2: Set maximum energy efficiency $q = 0$ and iteration index $n = 0$
3: **repeat** {Main Loop}
4: Solve the inner loop problem in (12) for a given q and obtain resource allocation policies $\{\mathcal{P}', \mathcal{W}', \mathcal{S}'\}$
5: **if** $U(\mathcal{P}', \mathcal{W}', \mathcal{S}') - qU_{TP}(\mathcal{P}', \mathcal{W}', \mathcal{S}') < \epsilon$ **then**
6: Convergence = **true**
7: **return** $\{\mathcal{P}^*, \mathcal{W}^*, \mathcal{S}^*\} = \{\mathcal{P}', \mathcal{W}', \mathcal{S}'\}$ and $q^* = \frac{U(\mathcal{P}', \mathcal{W}', \mathcal{S}')}{U_{TP}(\mathcal{P}', \mathcal{W}', \mathcal{S}')}$
8: **else**
9: Set $q = \frac{U(\mathcal{P}', \mathcal{W}', \mathcal{S}')}{U_{TP}(\mathcal{P}', \mathcal{W}', \mathcal{S}')}$ and $n = n + 1$
10: Convergence = **false**
11: **end if**
12: **until** Convergence = **true** or $n = L_{max}$

B. Iterative Algorithm for Energy Efficiency Maximization

In this section, we propose an iterative algorithm (known as the Dinkelbach method [25]) for solving (9) with an equivalent objective function. The proposed algorithm is summarized in Table I and the convergence to the optimal energy efficiency is guaranteed if we are able to solve the inner problem (12) in each iteration.

Proof: Please refer to [25] or [26] for a proof of convergence.

As shown in Table I, in each iteration of the main loop, we solve the following optimization problem for a given parameter q:

$$\max_{\mathcal{P},\mathcal{W},\mathcal{S}} \quad U(\mathcal{P}, \mathcal{W}, \mathcal{S}) - qU_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S}) \quad \text{s.t. C1, C2, C3, C4, C5, C6.} \tag{12}$$

Solution of the Main Loop Problem: The transformed problem is a mixed combinatorial and non-convex optimization problem. The non-convex nature comes from the power allocation variables and precoding coefficients. The multiuser interference appears in the denominator of the SINR expression in (4) which couples the power allocation variables. On the other hand, the combinatorial nature comes from the integer constraint for subcarrier allocation. To obtain an optimal solution, an exhaustive search is needed with complexity $\sum_{g=1}^{M} \binom{K}{g}^{n_F}$, which is computationally infeasible for $K \gg M$. In order to derive an efficient resource allocation algorithm, we solve the above problem in three steps. In the first step, we employ a sub-optimal low-complexity user selection scheme. Then, in the second step, we calculate the ZFBF coefficients for a given selected user set $\mathcal{S}$. In the final step, we optimize the transmit power at each BS for energy efficiency maximization. We note that by fixing resource allocation policies $\{\mathcal{W}, \mathcal{S}\}$, Algorithm 1 in Table I converges to the maximum energy efficiency with respect to the power allocation variables. However, it is a suboptimal solution from a joint optimization point of view.

TABLE II
SEMI-ORTHOGONAL USER SELECTION ALGORITHM.

Algorithm 2 Semi-Orthogonal User Selection Algorithm

1: Initialize auxiliary user set $\mathcal{T}_t = \{1, \ldots, K\}$, orthogonality parameter η, vector subspace $\Phi = \{\vec{\phi}_{(1)}, \ldots, \vec{\phi}_{(t)}\}$, iteration index $t = 1$, and $\mathcal{S}_\perp(i) = \emptyset$, where $\vec{\phi}_{(t)} \in \mathbb{C}^{1\times M}$.
2: Update $\mathcal{S}_\perp(i) \rightarrow \mathcal{S}_\perp(i) \bigcup \pi(t), \pi(t) = \arg\max_{a\in\mathcal{T}_t} \|\vec{H}^a_{BS}(i)\|^2$, $\vec{\phi}_{(1)} = \vec{H}^{\pi(t)}_{BS}(i)$, $\mathcal{T}_{t+1} = \mathcal{T}_t / \{\pi(t)\}$.
3: **repeat**
4: For each user $k \in \mathcal{T}_t$, calculate a vector $\vec{H}^k_\perp(i) \in \mathbb{C}^{1\times M}$ which is orthogonal to Φ as
$$\vec{H}^k_\perp(i) = \vec{H}^k_{BS}(i) - \sum_{r=1}^{t-1} \frac{\vec{H}^k_{BS}(i)\vec{\phi}^\dagger_{(r)}}{\|\vec{\phi}_{(r)}\|^2}\vec{\phi}_{(r)}.$$
5: Update $\mathcal{S}_\perp(i) \rightarrow \mathcal{S}_\perp(i) \bigcup \pi(t), \pi(t) = \arg\max_{a\in\mathcal{T}_t} \|\vec{H}^a_\perp(i)\|^2$, $\vec{\phi}_{(t)} = \vec{H}^{\pi(t)}_\perp(i)$.
6: **if** $|\mathcal{S}_\perp(i)| \le M$, **then**
7: Calculate $\mathcal{T}_{t+1}$ as
$$\mathcal{T}_{t+1} = \left\{ k \in \mathcal{T}_t, k \neq \pi(t), \frac{|\vec{H}^k_{BS}(i)\vec{\phi}^\dagger_{(t)}|}{\|\vec{H}^k_{BS}(i)\|\|\vec{\phi}_{(t)}\|} < \eta \times \alpha_k \right\},$$
$t = t + 1$.
8: **end if**
9: **until** $\mathcal{T}_t = \emptyset$ or $|\mathcal{S}_\perp(i)| = M$

Step 1 (Semi-Orthogonal User Selection [27]): We propose an efficient user selection algorithm. Without loss of generality, we define a row vector $\vec{H}^k_{BS}(i) = \left[H^k_{B_1}(i)\sqrt{l^k_{B_1}} \; H^k_{B_2}(i)\sqrt{l^k_{B_2}} \; \ldots H^k_{B_M}(i)\sqrt{l^k_{B_M}} \right]$ is a super-channel vector between all BSs and user k with elements $H^k_{B_m}(i)\sqrt{l^k_{B_m}}$, $k \in \{1, \ldots, K\}$, $m \in \{1, \ldots, M\}$, representing the channel coefficient between BS m and user k on subcarrier i. Let $\mathcal{S}_\perp(i)$ be a semi-orthogonal user set for subcarrier i. Then, the adopted semi-orthogonal user selection procedure for each subcarrier is summarized in Table II. Φ is a vector subspace which is defined as $\Phi = \{\vec{\phi}_{(1)}, \ldots, \vec{\phi}_{(t)}\}$ where $\vec{\phi}_{(t)}$ are the vectors which span the the subspace. $\mathcal{T}_t$ is an auxiliary user set for executing the algorithm. η in line 7 in Table II represents a threshold for measuring orthogonality. Note that a user with a higher value of α_k (priority) has a higher chance of being selected. On the other hand, as $\eta \to 0$, the selected users in the set are increasingly orthogonal to each other. In other words, users associated with set $\mathcal{S}_\perp(i)$ cause less interference to other users in the set. Note that with the proposed user selection scheme, the search space for each subcarrier decreases from $\sum_{a=1}^{M} \binom{K}{a}$ to $2KM$ and $\frac{2KM}{\sum_{a=1}^{M}\binom{K}{a}} \ll 1$ for $K \gg M$. Note that although the proposed algorithm is suboptimal, it has been shown in [27] that the proposed scheme performs well in combination with zero-forcing beamforming.

Step 2 (Zero-Forcing Beamformimg): A multi-cell network with full BS cooperation can be interpreted as a MIMO broadcast channel. It can be shown that dirty paper coding (DPC) is optimal in achieving the multiuser broadcast capac-

ity region. However, DPC requires a very high complexity which is considered impractical. On the contrary, although ZFBF is a suboptimal precoding scheme, it is considered a practical solution, due to its linear complexity and promising performance. Besides, it can be shown that the proposed semi-orthogonal user selection algorithm together with ZFBF can achieve the same asymptotic sum capacity performance as DPC [27]. Therefore, we focus on ZFBF in the rest of the paper.

If ZFBF is used for transmission, the capacity equation in (3) can be rewritten as

$$\begin{aligned} C^k(i) &= \frac{\mathcal{B}}{n_F}\log_2\Big(1+\Gamma^k(i)\Big) \quad \text{where} \\ \Gamma^k(i) &= \frac{\Big|\sum_{c=1}^{M}\sqrt{l^k_{B_c}}H^k_{B_c}(i)w^k_{B_c}(i)\Big|^2 P^k_{B_m}(i)}{\sigma_z^2}. \end{aligned} \quad (13)$$

$P^k_{B_m}(i) = P^k_{B_1}(i) = P^k_{B_2}(i) = \ldots = P^k_{B_M}(i)$ is an imposed constraint together with ZFBF[8]. There are two reasons for imposing this constraint. First, it allows us to separate the power allocation variables from the precoding coefficients. Second, it simplifies the design of power control[9]. Let us consider the above scenario and ZFBF precoding with the above assumption, then the total transmit power of the M BSs to user k on subcarrier i is given by $\sum_{c=1}^{M} P^k_{B_c}(i)|w^k_{B_c}(i)|^2 = P^k_{B_m}(i)$, since $\sum_{c=1}^{M}|w^k_{B_c}(i)|^2 = 1$ [29]. In other words, the precoding coefficients are decoupled from the power allocation variables and the precoding coefficients do not increase the total power consumption. Besides, we can directly control the total amount of transmit power from the M BSs to user k on subcarrier i via optimizing $P^k_{B_m}(i)$. Without loss of generality, we assume that user 1 to user k are selected for using subcarrier i, i.e., $\{1,\ldots,k\} \in \mathcal{S}_\perp(i)$. Let us define a super channel matrix $\mathbf{H}_B(i) \in \mathbb{C}^{|\mathcal{S}_\perp(i)|\times M}$ such that

$$\mathbf{H}_B^T(i) = \Big[\big(\vec{H}^1_{BS}(i)\big)^T\big(\vec{H}^2_{BS}(i)\big)^T\ldots\big(\vec{H}^k_{BS}(i)\big)^T\Big]. \quad (14)$$

Then, the corresponding ZFBF super matrix $\mathbf{B}(i) \in \mathbb{C}^{M\times|\mathcal{S}_\perp(i)|}$ can be calculated in the centralized unit and is given by

$$\mathbf{B}(i) = \mathbf{H}_B^\dagger(i)\Big(\mathbf{H}_B(i)\mathbf{H}_B^\dagger(i)\Big)^{-1}\mathbf{D}(i), \quad (15)$$

where $\mathbf{D}(i) \in \mathbb{C}^{|\mathcal{S}_\perp(i)|\times|\mathcal{S}_\perp(i)|}$ is a diagonal matrix with diagonal elements $\gamma^k(i) = 1/\sqrt{\Big[\Big(\mathbf{H}_B(i)\mathbf{H}_B^\dagger(i)\Big)^{-1}\Big]_{k,k}} = \Big|\sum_{c=1}^{M}\sqrt{l^k_{B_c}}H^k_{B_c}(i)w^k_{B_c}(i)\Big|$. Note that $\gamma^k(i)$ represents the equivalent channel gain between all BSs and user k on subcarrier i for ZFBF transmission. Hence, the ZFBF coefficient $w^k_{B_c}(i)$ is given by

$$w^k_{B_c}(i) = \Big[\mathbf{B}(i)\Big]_{c,k} \quad \forall k \in \mathcal{S}_\perp(i) \quad (16)$$

and the central unit delivers the relevant ZFBF coefficients to each BS via additional backhaul connections which are dedicated to control signals.

Dual Problem: The final step in solving the main loop problem is to optimize the power allocation. For a given set of selected users and ZFBF transmission, the problem in (12) is still non-convex due to constraint C3. In general, a non-zero duality gap exists if we solve (12) by solving its dual. However, we will demonstrate that the duality gap is always zero when the number of subcarriers is sufficiently large. This result is summarized in the following theorem.

Theorem 2: Let P and D denote the optimal values of the primal and the dual problem in (12), respectively. For a given selected user set and ZFBF transmission, if the number of subcarriers is sufficiently large[10], then strong duality holds and the duality gap is zero, i.e., P = D.

Proof: Please refer to Appendix for a proof of Theorem 2.

By Theorem 2, we solve the main loop problem in (12) by solving its dual. For this purpose, we first need the Lagrangian function of the primal problem. Upon rearranging terms, the Lagrangian can be written as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\theta,\mathcal{P}) \\ =\ & \sum_{m=1}^{M}\sum_{i=1}^{n_F}\sum_{k\in\mathcal{A}_m\cap\mathcal{S}_\perp(i)}(\alpha_k+\theta-\beta_m)C^k(i) - \theta R_{\min} \\ -\ & \sum_{m=1}^{M}\lambda_m\Big(\sum_{i=1}^{n_F}\sum_{k\in\mathcal{S}_\perp(i)}|w^k_{B_m}(i)|^2P^k_{B_m}(i) - P_{T_m}\Big) \\ -\ & q\Big(\sum_{m=1}^{M}\sum_{k\in\mathcal{A}_m\cap\mathcal{S}_\perp(i)}\sum_{i=1}^{n_F}\varepsilon P^k_{B_m}(i)|w^k_{B_m}(i)|^2 + \delta P_{BH}\Big) \\ +\ & \sum_{m=1}^{M}\beta_m R_{\max_m} - q(P_C\times M), \end{aligned} \quad (17)$$

where $\theta \geq 0$ is the Lagrange multiplier corresponding to the required minimum capacity constraint C2. $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ are the Lagrange multiplier vectors associated with individual power constraint C1 and maximum backhaul capacity constraint C3 with elements $\lambda_m \geq 0$ and $\beta_m \geq 0$, $m \in \{1,\ldots,M\}$, respectively. Boundary constraint C6 will be absorbed into the Karush-Kuhn-Tucker (KKT) conditions when deriving the optimal power allocation in the following. Thus, the dual problem of (12), for a given selected user set and ZFBF transmission, is given by

$$\mathrm{D} = \min_{\boldsymbol{\lambda},\boldsymbol{\beta},\theta\geq 0}\ \max_{\mathcal{P}}\quad \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\beta},\theta,\mathcal{P}). \quad (18)$$

[8]Indeed, the transmit power from BS m to user k on subcarrier i is $P^k_{B_m}(i)|w^k_{B_m}(i)|^2$ instead of $P^k_{B_m}(i)$. So even if we enforce $P^k_{B_m}(i) = P^k_{B_1}(i) = P^k_{B_2}(i) = \ldots = P^k_{B_M}(i)$, the actual transmit powers to user k from M base stations (BSs) are not identical since in general $|w^k_{B_1}(i)|^2 \neq |w^k_{B_2}(i)|^2 \neq \ldots \neq |w^k_{B_M}(i)|^2$.

[9]The above approach is commonly used in literature [12], [28] for decoupling the power allocation variables from the precoding coefficients and for the design of simple yet efficient resource allocation algorithms for multiple antenna systems.

[10]In [30], the authors used simulations to show that the duality gap is virtually zero for 8 subcarriers in an OFDMA system. In practical systems such as Long-Term-Evolution (LTE), the number of subcarriers can vary from 128 to 2048 [31]. In other words, the condition of a "sufficiently large number of subcarriers" is always satisfied in practice. Note that [30] does not provide any analytical evidence in support of their results. In contrast, the proof provided in our paper is rigorous as it is based on showing the existence of a saddle point of the Lagrangian function by using the concavity of the perturbation function.

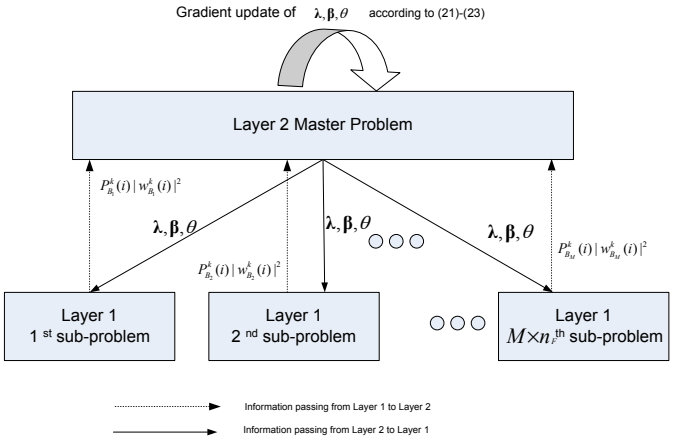


Fig. 2. Illustration of the dual decomposition of a large scale problem into a two-layer problem.

In the following, we solve the above dual problem iteratively by decomposing it into two layers via *dual decomposition*: Layer 1, the maximization over $\mathcal{P}$ in (18), consists of $M \times n_F$ subproblems with identical structure; Layer 2, the minimization over $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, and θ in (18), is the master dual problem to be solved by the gradient method, cf. Figure 2.

Layer 1 Solution (Power Allocation): By Theorem 2, the KKT conditions are the necessary and sufficient conditions for the optimal solution. Thus, the closed-form power allocation for the BSs serving user k in subcarrier i for a given parameter q is obtained as

$$P^k_{B_m}(i) = \left[\frac{\mathcal{B}/n_F(\alpha_k+\theta-\beta_m)}{\ln(2)\Omega^k(i)} - \frac{\sigma_z^2}{|\gamma^k(i)|^2}\right]^+ \text{ and}$$

$$P^k_{B_a}(i) = P^k_{B_m}(i)\,\forall a \neq m, \tag{19}$$

$$\text{where} \quad \Omega^k(i) = \Big(\sum_{c=1}^{M}(\lambda_c + q\varepsilon)|w^k_{B_c}(i)|^2\Big) \tag{20}$$

and $P^k_{B_a}(i) = P^k_{B_m}(i)\,\forall a \neq m$ is due to the imposed assumption after (13). The optimal power allocation solution in (19) is in the form of *multi-level water filling*. Note that if a user has a higher value of α_k (higher priority), a higher power will be allocated to the user since she has a higher water level $\frac{\mathcal{B}/n_F(\alpha_k+\theta-\beta_m)}{\ln(2)\Omega^k(i)}$ compared to other users. $\beta_m \geq 0$ controls the scheduled data rate via adjusting the water level of the power allocation in (19), such that the scheduled data rate will not exceed the backhaul capacity limit. $\Omega^k(i)$ represents the influence of the power consumption of other BSs on the joint transmission on subcarrier i for user k. On the other hand, although P_C and P_{BH} do not appear in (19), they influence the solution of the dual problem via the updating process of q. From Table I, both P_C and P_{BH} are used to update the value of q in each iteration. Then, the updated q is used to derive the solution of the dual problem in the next iteration.

Solution of Layer 2 (Master Problem): To solve the Layer 2 master minimization problem in (18), i.e, to find $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, and θ for a given $\mathcal{P}$, the gradient method can be used since the dual function is differentiable. The gradient update equations are given by:

$$\lambda_m(n+1) = \Big[\lambda_m(n) - \xi_1(n) \times \Big(P_{T_m} - \sum_{i=1}^{n_F}\sum_{k\in\mathcal{S}_\perp(i)}|w^k_{B_m}(i)|^2 P^k_{B_m}(i)\Big)\Big]^+, \forall m, \tag{21}$$

$$\theta(n+1) = \Big[\theta(n) - \xi_2(n) \times \Big(\sum_{m=1}^{M}\sum_{i=1}^{n_F}\sum_{k\in\mathcal{A}_m\cap\mathcal{S}_\perp(i)} C^k(i) - R_{\min}\Big)\Big]^+, \tag{22}$$

$$\beta_m(n+1) = \Big[\beta_m(n) - \xi_3(n) \times \Big(R_{\max_m} - \sum_{i=1}^{n_F}\sum_{k\in\mathcal{A}_m\cap\mathcal{S}_\perp(i)} C^k(i)\Big)\Big]^+, \forall m, \tag{23}$$

where index $n \geq 0$ is the iteration index and $\xi_u(m)$, $u \in \{1,2,3\}$, are positive step sizes. Then, the updated Lagrange multipliers in (21)-(23) are used for solving the Layer 1 subproblems in (18) via updating the resource allocation policies, cf. Figure 2. By Theorem 2, the duality gap is zero and it is guaranteed that the iteration between Layer 1 and Layer 2 converges to the optimal solution of (12) with respect to the power allocation variables in the main loop, if the chosen step sizes satisfy the infinite travel condition [32], [33].

V. Results and Discussions

In this section, we evaluate the system performance for the proposed resource allocation and scheduling algorithm using simulations. A multi-cell system with 3 cells is considered. The inter-site distance between each pair of BSs is 500 meters as suggested in the 3GPP specification [31]. The number of subcarriers is $n_F = 128$ with carrier center frequency 2.5 GHz, system bandwidth $\mathcal{B} = 1.25$ MHz, and $\alpha_k = 1, \forall k$. Note that by setting $\alpha_k = 1, \forall k$, we obtain the achievable maximum network capacity. Each subcarrier for RF transmission has a bandwidth of 9.7656 kHz and the noise variance is $\sigma_z^2 = -134$ dBm. The 3GPP urban path loss model is used [31]. The small scale fading coefficients of the BS-to-user links are generated as independent and identically distributed (i.i.d.) Rayleigh random variables with unit variances. We assume that all BSs have the same maximum transmit power, i.e., $P_{T_m} = P_T$, $\forall m$. Besides, a fully connected backhaul connection topology is considered for simulation purpose, i.e., there are $\delta = 6$ connections, cf. Figure 1. For the backhaul connections, we adopt the specifications of a commercial optical fiber modem [34] which supports three types of data rates for backhaul within a distance of 2.5 km: $R_1 = 11.184$ Mbit/s, $R_2 = 34.368$ Mbit/s, and $R_3 = 44.736$ Mbit/s[11]. The power consumption of each backhaul link is $P_{BH} = 15$ Watts as specified in [34]. The average system energy efficiency is obtained by counting the amount of data which are successfully decoded by the users and dividing it by the total power consumption averaged over both macroscopic and microscopic fading. We assume a static circuit power consumption of $P_C = 40$ dBm [23], a data rate requirement of $R_{\min} = 4$ bit/s/Hz/cell, and an

[11]The values of the backhaul capacities used in the paper are for illustration purpose. In practice, the choice of backhaul capacities should scale with the bandwidth and the number of subcarriers used in the RF transmission.

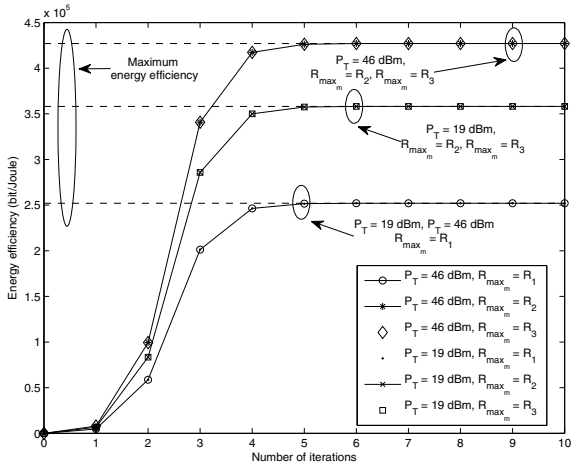


Fig. 3. Energy efficiency (bit-per-Joule) versus number of iterations with different maximum transmit power allowances per BS, P_T, and different backhaul capacities, $R_{\max_m}$, for $K = 45$ users. The dashed lines represent the maximum achievable energy efficiencies with respect to the power allocation variables for different cases.

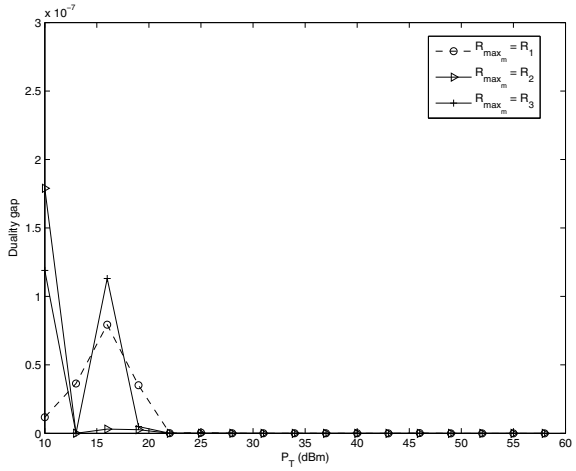


Fig. 4. Duality gap versus the maximum transmit power allowance at each BS, P_T, for different backhaul capacities $R_{\max_m}$.

orthogonality parameter of $\eta = 0.1$. Furthermore, we assume a power efficiency of 20% for the power amplifiers used in the RF, i.e., $\varepsilon = \frac{1}{0.2} = 5$. In the following results, the "number of iterations" refers to the number of outer loop iterations of Algorithm 1 in Table I. For each inner loop, we set the number of iterations to five.

A. Convergence of Iterative Algorithm 1 and Duality Gap

Figure 3 illustrates the evolution of the proposed iterative algorithm for different numbers of users and different maximum transmit powers at each BS. The results in Figure 3 were averaged over 100000 independent adaptation processes where each adaptation process involves a different realization of the path loss and the multipath fading[12]. Note that the maximum energy efficiency in the figure is with respect to power allocation optimization. It can be observed that the iterative algorithm converges to the optimal value within 5 iterations for all considered numbers of transmit antennas. On the other hand, the inner loop converges within 5 iterations on average. In other words, the overall algorithm takes in total around 25 iterations (inner loop + outer loop) to converge to the dual optimal for a given set of selected users and ZFBF transmission.

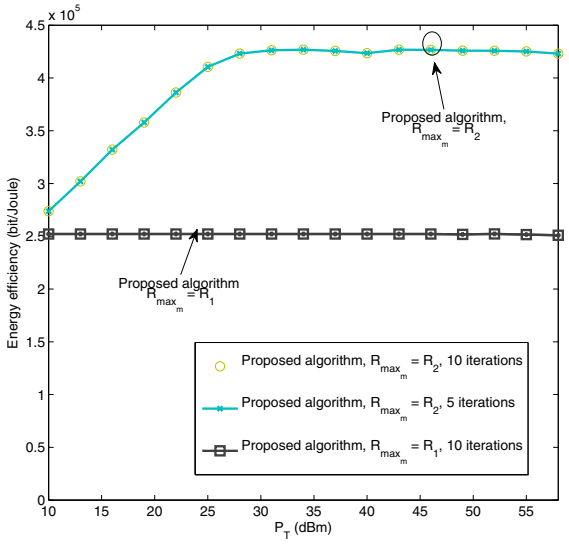
Figure 4 shows the duality gap, D − P, versus the maximum transmit power allowance at each BS, P_T, for different maximum backhaul capacities. The primal problem is solved by using the power allocation solution obtained with the dual decomposition. It can be seen that the duality gap is practically zero for the considered cases, despite the non-convexity of the primal problem. The few small non-zero spikes in the duality gap (in the order of 10^{-7}) are mainly due to a finite-precision arithmetic of computation and a finite number of iterations in solving the dual problem.

[12]In general, the number of iterations (outer loop) required for the algorithm to converge can be different for different channel realizations. As a results, we show the average energy efficiency over 100000 channel realizations.

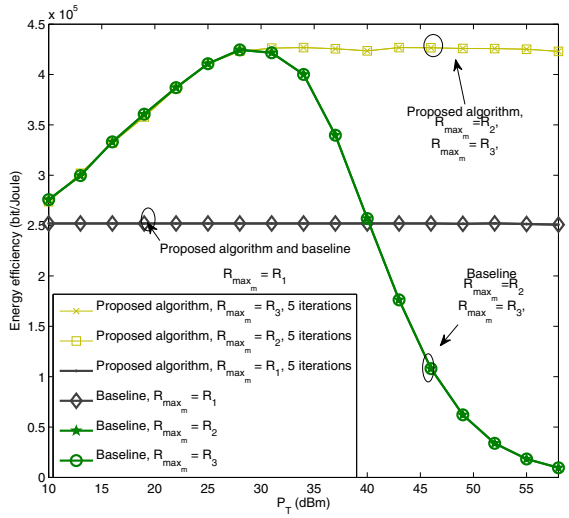
B. Energy Efficiency and Average Capacity versus Transmit Power

Figure 5(a) illustrates the energy efficiency versus the maximum transmit power allowance at each BS, P_T, for $K = 45$ users. The number of iterations for the proposed iterative resource allocation algorithm is 5 and 10 for different backhaul capacities. It can be seen that the performance difference between 5 iterations and 10 iterations is negligible which confirms the practicality of our proposed iterative resource allocation algorithm.

In Figure 5(b), we set the number of iterations in the proposed algorithm to 5 and study the trade-off between energy efficiency, maximum transmit power, and backhaul capacity. It can be observed that when both the maximum transmit power allowance at the power amplifier and the capacities of the backhaul links are large enough, e.g., $P_T \geq$ 30 dBm and $R_{\max_m} \geq R_2 \ \forall m$, the energy efficiency of the proposed algorithm approaches a constant value since the resource allocator is not willing to consume more power, when the maximum energy efficiency is achieved. Besides, further increasing the backhaul capacities from $R_{\max_m} = R_2 \ \forall m$ to $R_{\max_m} = R_3 \ \forall m$ is not beneficial for energy efficiency as the system performance is now confined by the capacity of the radio links. However, for the case of backhaul capacity $R_{\max_m} = R_1 \ \forall m$, the energy efficiency is quickly saturated even if the transmit powers at the BSs are low since the system capacity is always limited by the bottleneck of the backhaul connections. For comparison, Figure 5(b) also contains the energy efficiency of a baseline resource allocation scheme in which we maximize the weighted system capacity (bit/s/Hz) with constraints C1-C6 in (9) for a given selected users set and ZFBF transmission, instead of the energy efficiency. It can be observed that in the low transmit power regime with high backhaul capacity, i.e., $P_T < 30$ dBm and $R_{\max_m} \geq R_2 \, \forall m$, the baseline scheme has virtually the same performance as the proposed algorithm. In other words, this result suggests that in the low transmit power regime, transmitting with the maximum available power is the most energy-efficient



(a) Energy efficiency (bit-per-Joule) versus the maximum transmit power allowance at each BS, P_T, for the proposed resource allocation algorithm, different backhaul capacities, and different numbers of iterations with $K = 45$ users .



(b) Energy efficiency (bit-per-Joule) versus the maximum transmit power allowance at each BS, P_T, for the proposed resource allocation algorithm and different backhaul capacities with $K = 45$ users.

Fig. 5. Energy efficiency (bit-per-Joule) versus the maximum transmit power allowance at each BS, P_T, for different resource allocation algorithms and different backhaul capacities with $K = 45$ users.

option. However, the energy efficiency of the baseline scheme decreases dramatically in the high transmit power regime. This is because there is a diminishing return in the system capacity with respect to the increment of transmit power. Meanwhile, the total power consumption scales linearly with respect to the transmit power. Hence, the capacity gain is unable to compensate for the negative impact of the total power consumption in the RF amplifiers and results in a low energy efficiency. On the other hand, for $R_{\max_m} = R_1 \, \forall m$, the proposed algorithm and the baseline scheme achieve the same energy efficiencies as the degrees of freedom in the resource allocation are limited by the small backhaul capacities.

Figure 6 shows the average system capacity (bit/s/Hz/cell) versus the maximum transmit power P_T for $K = 45$ users and different backhaul capacities. We compare the system

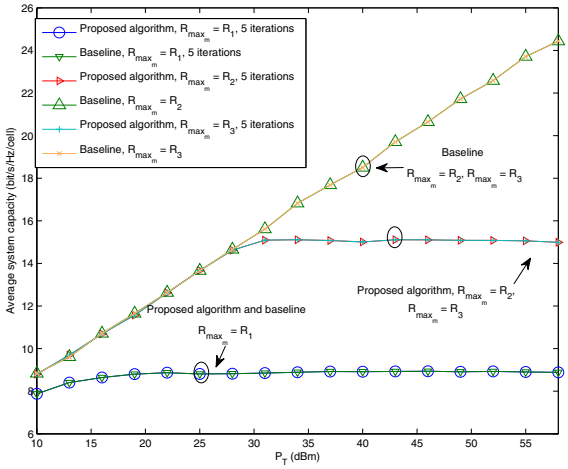


Fig. 6. Average system capacity (bit/s/Hz/cell) versus the maximum transmit power allowance at each BS, P_T, for different resource allocation algorithms and different backhaul capacities with $K = 45$ users. The number of iterations for the proposed algorithm is set to 5.

performance of the proposed algorithm again with the baseline scheme. The number of iterations in the proposed algorithm is set to 5. It can be observed that the average system capacity of the proposed algorithm approaches a constant in the high transmit power and high backhaul capacity regimes, i.e, $P_T \geq 30$ dBm and $R_{\max_m} \geq R_2 \ \forall m$. This is because the proposed algorithm clips the transmit power at the BSs in order to maximize the system energy efficiency. However, when the backhaul capacity is small, i.e., $R_{\max_m} = R_1 \, \forall m$, the maximum achievable average system capacities of both the proposed algorithm and the baseline scheme do not scale with the transmit power. We note that, as expected, the baseline scheme achieves a higher average system capacity than the proposed algorithm in the high transmit power regime for $R_{\max_m} \geq R_2 \ \forall m$ since the former scheme consumes all the available transmit power in all scenarios. However, the superior average system capacity of the baseline scheme comes at the expense of low energy efficiency as shown in Figure 5(b). On the other hand, increasing the backhaul capacity beyond $R_{\max_m} = R_2 \, \forall m$ is not beneficial for the average system capacity as the wireless communication links are the bottleneck links.

Figure 7 depicts the average total power consumption, i.e., $\mathcal{E}\{U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S})\}$, versus the maximum transmit power P_T for the proposed algorithm and the baseline scheme for 5 iterations. In the considered transmit power regimes, the proposed algorithm consumes less power than the baseline scheme for the case of $R_{\max_m} \geq R_2 \, \forall m$. This is because the proposed algorithm clips the transmit power for energy efficiency maximization. However, when the backhaul capacity is the limiting factor, i.e., $R_{\max_m} = R_1 \, \forall m$, both the baseline scheme and the proposed algorithm consume almost the same amount of power since the transmit power usage is confined by the backhaul capacity instead of energy efficiency maximization.

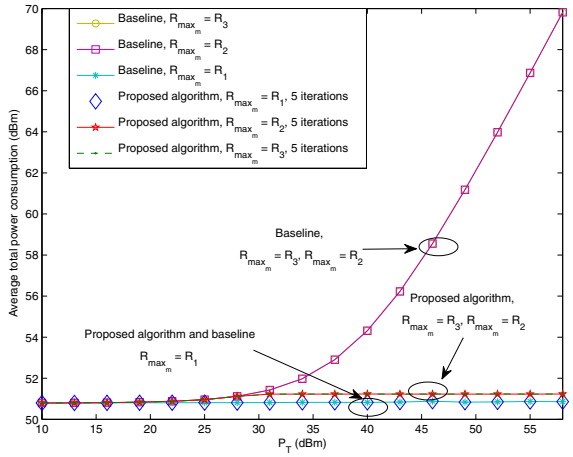


Fig. 7. Average total power consumption, $\mathcal{E}\{U_{TP}(\mathcal{P}, \mathcal{W}, \mathcal{S})\}$, versus the maximum transmit power allowance at each BS, P_T, for different backhaul capacities, 5 iterations, and $K = 45$ users.

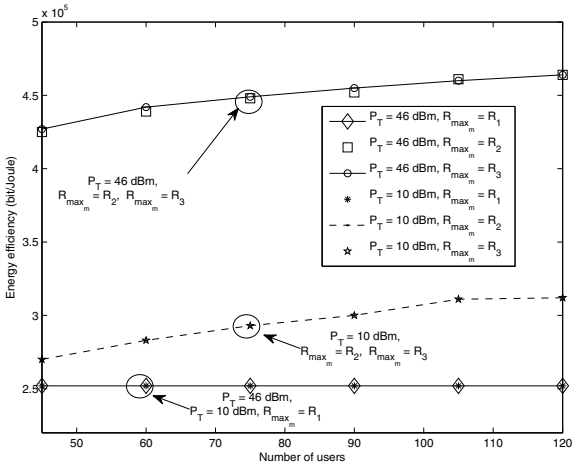


Fig. 8. Energy efficiency (bit-per-Joule) versus the number of users K for different maximum transmit power allowances at each BS, P_T, and different backhaul capacities.

C. Energy Efficiency and Average System Capacity versus Number of Users

Figures 8 and 9 depict the energy efficiency and the average system capacity versus the number of users, respectively. Different backhaul capacities, different maximum transmit power allowances P_T at the BSs, and 5 iterations of the proposed algorithm are considered. It can be observed that for $R_{\max_m} \geq R_2 \, \forall m$, both the energy efficiency and the average system capacity grow with the number of users since the proposed resource allocation and scheduling algorithm is able to exploit multiuser diversity (MUD) due to the semi-orthogonal user selection algorithm. In general, MUD introduces an extra power gain [35, Chapter 6.6] to the system which provides further energy savings. Yet, when the backhaul capacity is the performance limiting factor, i.e., $R_{\max_m} = R_1 \, \forall m$, the proposed algorithm is unable to take advantage of MUD since the performance gain due to joint BS transmission is limited by the small backhaul capacities. As a result, both the average system capacity and the energy efficiency remain constant when the backhaul is the bottleneck.

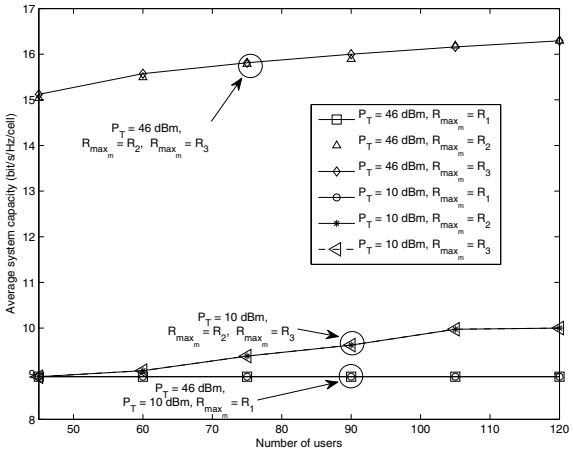


Fig. 9. Average system capacity (bit/s/Hz/cell) versus the number of users K for different maximum transmit power allowances at each BS, P_T, and different backhaul capacities.

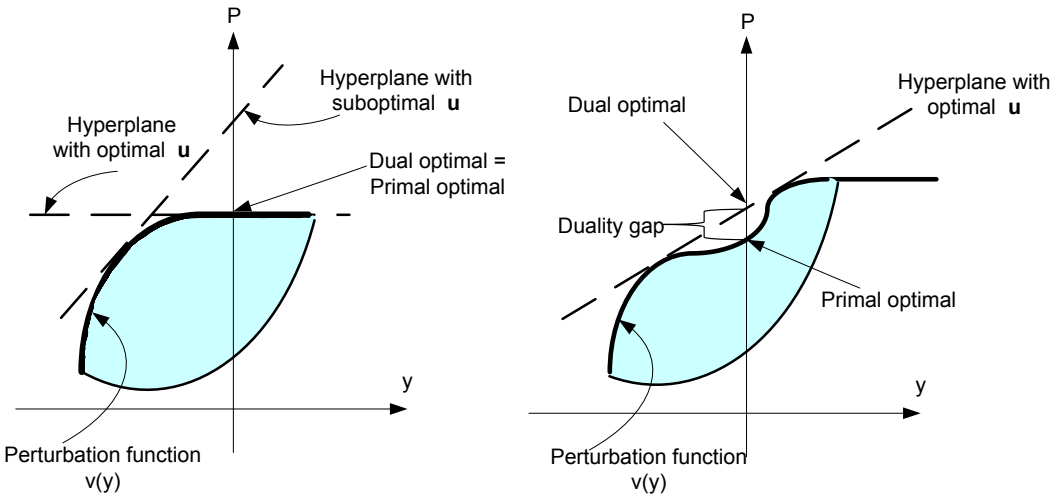
VI. Conclusions

In this paper, we formulated the resource allocation and scheduling design for multi-cell OFDMA networks with joint BS ZFBF transmission as a non-convex and combinatorial optimization problem, in which the circuit power dissipation, the limited backhaul capacity, and the system data rate requirement were taken into consideration. By exploiting the properties of fractional programming, the considered problem was transformed into an equivalent problem with a tractable iterative solution. In each iteration, a low complexity user selection and ZFBF are performed for maximization of the energy efficiency. Furthermore, we demonstrated that when the number of subcarriers is sufficiently large, the duality gap for the power allocation problem is practically zero despite the non-convexity of the primal problem. As a result, an efficient closed-form power allocation can be obtained in each iteration via dual decomposition. Simulation results showed that the proposed algorithm converges within a small number of iterations and unveiled a trade-off between energy efficiency, network capacity, and backhaul capacity. In particular, (1) in the low transmit power regime, an algorithm which achieves the maximum spectral efficiency may also achieve the maximum energy efficiency; (2) a high spectral efficiency does not necessarily result in a high energy efficiency; (3) spectral efficiency is always limited by the backhaul capacity; (4) energy efficiency increases with the backhaul capacity only until the maximum energy efficiency is achieved.

Interesting topics for future work include studying the impact of imperfect CSIT and multiple-antenna BSs.

Appendix- Proof of Theorem 2

As mentioned in the main text, the transformed optimization problem for given parameter q, selected user set, and ZFBF transmission is a non-convex optimization problem due to



(a) Zero duality gap when maximizing a concave problem. (b) Non-zero duality gap when maximizing a non-convex problem.

Fig. 10. Geometric interpretation of duality and perturbation function for concave and non-convex optimization problems with a 1-dimensional perturbation vector $y \in \mathbb{R}^1$ for illustration. The shaded areas represent the set of values of the primal problem for different perturbations $y \in \mathbb{R}^1$.

constraint C3. In general, a non-zero duality gap exists if we solve the transformed problem by solving its dual, cf. Figure 10(b). However, we demonstrate in the following that when a non-convex optimization problem satisfies certain conditions, the duality gap is always zero. Before introducing another important theorem for the proof of Theorem 2, we first introduce the concept of *perturbation function*. For the sake of notational simplicity and to avoid ambiguity, we use the following notations. We use $v\langle\cdot\rangle$ to denote a function v and $(\cdot)$ to denote brackets. Without loss of generality, the primal optimization problem in (12) can be written in general form as

$$\begin{aligned} \mathrm{P} = \max_{\mathbf{p}_i^k \geq \mathbf{0}} \quad & \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^k \rangle \\ \text{s.t.} \quad & \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \leq \mathbf{0}, \end{aligned} \tag{24}$$

where $f_i\langle\cdot\rangle : \mathbb{C}^K \rightarrow \mathbb{R}$ and $\mathbf{g}_i\langle\cdot\rangle : \mathbb{C}^K \rightarrow \mathbb{R}^L$ are arbitrary continuous functions. L and $\mathbf{0}$ are the total number of inequality constraints and a column vector with all zero elements, respectively. $\mathbf{p}_i^k \in \mathbb{R}^K$ represents a feasible solution vector of the primal problem in general form. Indeed, (24) is a general representation of an optimization problem and the physical meaning of vector $\mathbf{p}_i^k$ is not limited to transmit power. Note that we do not make any assumption on the concavity of functions $f_i\langle\cdot\rangle$ and $\mathbf{g}_i\langle\cdot\rangle$. Then, the perturbation function is defined as [36], [37]

$$\begin{aligned} v\langle \mathbf{y} \rangle = \max_{\mathbf{p}_i^k \geq \mathbf{0}} \quad & \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^k \rangle \\ \text{s.t.} \quad & \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \leq \mathbf{y}, \end{aligned} \tag{25}$$

where $\mathbf{y} \in \mathbb{R}^L$ is a *perturbation vector*. A geometrical interpretation of the perturbation function is given in Figure 10. The perturbation function $v\langle\mathbf{y}\rangle$ corresponds to the upper envelope of the shaded areas in Figure 10. Note that $v\langle\mathbf{y}\rangle$ is a *non-decreasing* function of $\mathbf{y}$ since a larger value of each element in $\mathbf{y}$ results in a larger feasible set. The Lagrangian function of (24) can be expressed as

$$\mathcal{L}\langle \mathbf{p}_i^k, \mathbf{u} \rangle = \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^k \rangle - \mathbf{u}^T \left(\mathbf{g}_i\langle \mathbf{p}_i^k \rangle \right) \tag{26}$$

where $\mathbf{u} \in \mathbb{R}^L, \mathbf{u} \geq \mathbf{0}$ is a vector of Lagrange multipliers. Thus, the corresponding dual problem is given by

$$\mathrm{D} = \min_{\mathbf{u} \geq \mathbf{0}} \max_{\mathbf{p}_i^k} \mathcal{L}\langle \mathbf{p}_i^k, \mathbf{u} \rangle. \tag{27}$$

Indeed, from a geometrical point of view, the dual problem is equivalent to finding the slope $\mathbf{u}$ of the supporting hyperplane of the perturbation function at $\mathbf{y} = \mathbf{0}$, i.e., $v\langle\mathbf{y}\rangle = v\langle\mathbf{0}\rangle$, such that its intercept on the P-axis is minimal, cf. Figure 10(a).

We are now ready to introduce the following theorem.

Theorem 3: If the perturbation function $v\langle\mathbf{y}\rangle$ is a concave function of $\mathbf{y}$, then the duality gap is zero despite the convexity of the primal problem[13], i.e., $\mathrm{D} = \mathrm{P}$.

Proof of Theorem 3:

The main idea of the proof is based on [37, Theorem 6.2.7] which states that a zero duality gap is equivalent to the existence of a saddle point of the Lagrangian function. Before proceeding to the proof of Theorem 3, let us first show how the concavity of the perturbation function $v\langle\mathbf{y}\rangle$ can be used to prove the existence of a saddle point of the Lagrangian function.

Suppose $v\langle\mathbf{y}\rangle$ is a concave function with respect to $\mathbf{y}$, then there exists a hyperplane that supports the hypograph of $v\langle\mathbf{y}\rangle$ for any $\mathbf{y} \in \mathbb{R}^L$, cf. Figure 10. Thus, by the definition of

[13]It is obvious that if the primal problem is concave, then the perturbation function $v\langle\mathbf{y}\rangle$ will be a concave function of $\mathbf{y}$. However, the reverse is not necessarily true.

concavity [32, Chapter 3.1.3], there exists some vector $\tilde{\mathbf{u}}$ such that

$$v\langle \mathbf{y} \rangle \leq v\langle \mathbf{0} \rangle + \tilde{\mathbf{u}}^T(\mathbf{y}-\mathbf{0}), \tag{28}$$

where $\tilde{\mathbf{u}} \in \mathbb{R}^L$ is known as the sub-gradient of $v\langle\cdot\rangle$.

Next, let $\mathbf{p}_i^{k*}$ be the optimal solution of optimization problem (24). Then, $(\mathbf{p}_i^{k*}, \tilde{\mathbf{u}})$ is a saddle point of the Lagrangian function if $\tilde{\mathbf{u}} \geq \mathbf{0}$ and $(\mathbf{p}_i^{k*}, \tilde{\mathbf{u}})$ satisfies

$$\mathcal{L}\langle \mathbf{p}_i^{k*}, \mathbf{u} \rangle \geq \mathcal{L}\langle \mathbf{p}_i^{k*}, \tilde{\mathbf{u}} \rangle \geq \mathcal{L}\langle \mathbf{p}_i^{k}, \tilde{\mathbf{u}} \rangle, \forall\, \mathbf{p}_i^k \geq \mathbf{0}, \mathbf{u} \geq \mathbf{0}. \tag{29}$$

First, we prove $\tilde{\mathbf{u}} \geq \mathbf{0}$ by contradiction. In order to relate (26) and (28), we consider a vector $\mathbf{\Delta} \in \mathbb{R}^L, \mathbf{\Delta} \geq \mathbf{0}$ which corresponds to the condition $\sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \leq \mathbf{0}$ of the Lagrangian function in (26). By the non-decreasing property of the perturbation function, we have $v\langle \mathbf{y}+\mathbf{\Delta} \rangle \geq v\langle \mathbf{y} \rangle$. As a result, we obtain the following inequality

$$v\langle \mathbf{y} \rangle \leq v\langle \mathbf{y}+\mathbf{\Delta} \rangle \leq v\langle \mathbf{0} \rangle + \tilde{\mathbf{u}}^T(\mathbf{y}+\mathbf{\Delta}-\mathbf{0}) \tag{30}$$

which holds for arbitrary vectors $\mathbf{\Delta} \geq \mathbf{0}$ and $\mathbf{y}$. Now, we put $\mathbf{y}=\mathbf{0}$ into (30) which yields

$$v\langle \mathbf{0} \rangle \leq v\langle \mathbf{0} \rangle + \tilde{\mathbf{u}}^T\mathbf{\Delta}. \tag{31}$$

Suppose now, there exists one element in $\tilde{\mathbf{u}}$ which takes a negative value. Then, we can always choose a vector $\mathbf{\Delta}$ such that $\tilde{\mathbf{u}}^T\mathbf{\Delta} < 0$ which violates the inequality in (31). Thus, $\tilde{\mathbf{u}} \geq \mathbf{0}$ has to be true. Second, we prove $\tilde{\mathbf{u}}^T \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle = 0$. Again, we consider the hyperplane in (28) with input vector $\mathbf{y} = \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle$. Since $\mathbf{p}_i^{k*}$ is the optimal solution of the primal problem in (24), $\sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle \leq \mathbf{0}$ must hold. Therefore, $v\Big\langle \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle \Big\rangle = v\langle \mathbf{0} \rangle$ and $\tilde{\mathbf{u}}^T \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle = 0$ must be true for satisfying (28). Now, we are ready to prove the right hand side of (29). Let us first consider the following:

$$\begin{aligned} \mathcal{L}\langle \mathbf{p}_i^{k*}, \tilde{\mathbf{u}} \rangle &= \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^{k*} \rangle - \tilde{\mathbf{u}}^T\Big(\mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle\Big) = \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^{k*} \rangle \\ &= v\langle \mathbf{0} \rangle \geq v\langle \mathbf{y} \rangle - \tilde{\mathbf{u}}^T\mathbf{y},\ \forall \mathbf{y}. \end{aligned} \tag{32}$$

Suppose $\mathbf{p}_i^k$ is a feasible solution of the primal problem. Then, $\mathbf{p}_i^k$ is also a feasible solution of the perturbation function $\mathbf{v}\langle\cdot\rangle$ if we set the perturbation vector $\mathbf{y}$ such that $\mathbf{y} = \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle$. Then, we substitute $\mathbf{y} = \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle$ into (32) which yields

$$\begin{aligned} \mathcal{L}\langle \mathbf{p}_i^{k*}, \tilde{\mathbf{u}} \rangle = v\langle \mathbf{0} \rangle &\geq v\Big\langle \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \Big\rangle - \tilde{\mathbf{u}}^T \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \\ &\overset{(a)}{\geq} v\Big\langle \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \Big\rangle + \tilde{\mathbf{u}}^T \sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \\ &= \mathcal{L}\langle \mathbf{p}_i^k, \tilde{\mathbf{u}} \rangle, \end{aligned} \tag{33}$$

where (a) is due to $\sum_{i=1}^{n_F} \mathbf{g}_i\langle \mathbf{p}_i^k \rangle \leq \mathbf{0}$.

On the other hand, the left hand side inequality in (29) can be proved as follows:

$$\begin{aligned} &\mathcal{L}\langle \mathbf{p}_i^{k*}, \tilde{\mathbf{u}} \rangle \\ =\ & \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^{k*} \rangle - \tilde{\mathbf{u}}^T\Big(\mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle\Big) = \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^{k*} \rangle \\ \leq\ & \sum_{i=1}^{n_F} f_i\langle \mathbf{p}_i^{k*} \rangle - \mathbf{u}^T\Big(\mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle\Big) \\ =\ & \mathcal{L}\langle \mathbf{p}_i^{k*}, \mathbf{u} \rangle \quad \because \mathbf{g}_i\langle \mathbf{p}_i^{k*} \rangle \leq \mathbf{0}, \mathbf{u} \geq \mathbf{0}. \end{aligned} \tag{34}$$

Therefore, $\langle \mathbf{p}_i^{k*}, \tilde{\mathbf{u}} \rangle$ is a saddle point of the Lagrangian function and by [37, Theorem 6.2.5], the duality gap is zero.

In other words, the concavity of the perturbation function $v\langle \mathbf{y} \rangle$ with respect to $\mathbf{y}$ is the key to proving that the duality gap is zero. The final step in proving Theorem 2 is to prove that $v\langle \mathbf{y} \rangle$ is a concave function of $\mathbf{y}$, i.e.,

$$v\langle \rho\mathbf{y} + (1-\rho)\mathbf{x} \rangle \geq \rho v\langle \mathbf{y} \rangle + (1-\rho)v\langle \mathbf{x} \rangle \tag{35}$$

for $0 \leq \rho \leq 1$, where $\mathbf{x} \in \mathbb{R}^L$ is another perturbation vector such that $\mathbf{x}-\mathbf{y} \neq \mathbf{0}$. Indeed, the concavity condition of the perturbation function is always satisfied in multi-carrier systems if frequency sharing is possible. For explaining the concept of frequency sharing, let $\mathbf{p}_{\mathbf{x}_i}^{k*}$ and $\mathbf{p}_{\mathbf{y}_i}^{k*}$ be the two optimal resource allocation policies with respect to the perturbation functions $v\langle \mathbf{x} \rangle$ and $v\langle \mathbf{y} \rangle$, respectively. Then, frequency sharing means that $\forall \rho \in [0,1]$, we can implement the optimal resource allocation policies $\mathbf{p}_{\mathbf{x}_i}^{k*}$ and $\mathbf{p}_{\mathbf{y}_i}^{k*}$ at the same time on a portion ρ of the subcarrier bandwidth and a portion $1-\rho$ of the subcarrier bandwidth, respectively. The basic idea for implementing frequency sharing in practice is to divide the total bandwidth $\mathcal{B}$ into a set of infinitesimally narrow subcarriers. As the number of subcarriers n_F in $\mathcal{B}$ increases, the bandwidth of each subcarrier becomes smaller and the channel gains in each subcarrier approaches a constant value. In the limiting case of $n_F \to \infty$, the channel gains of adjacent subcarriers are approximately the same[14] which facilitates frequency sharing. As a result, the original bandwidth of each subcarrier can be divided into two portions, i.e., $(1-\rho)$ and ρ, having the same channel gain. Then, for the perturbation function $v\langle \rho\mathbf{y}+(1-\rho)\mathbf{x} \rangle$, by construction, we implement the resource allocation policies $\mathbf{p}_{\mathbf{x}_i}^{k*}$ and $\mathbf{p}_{\mathbf{y}_i}^{k*}$ in portion one and portion two, respectively. Then, by exploiting the properties of the constraints[15] in (12), the constraints become a linear combination of the constraints in $v\langle \mathbf{x} \rangle$ and $v\langle \mathbf{y} \rangle$ due to the flatness of the channel over neighbouring subcarriers. Therefore, $v\langle \rho\mathbf{y} + (1-\rho)\mathbf{x} \rangle \geq \rho v\langle \mathbf{y} \rangle + (1-\rho)v\langle \mathbf{x} \rangle$ holds immediately[16] due to linearity. In other words, the perturbation function $v\langle \mathbf{y} \rangle$ is a concave function with respect to $\mathbf{y}$ under frequency sharing.

So, by combining Theorem 3 and the condition $n_F \to \infty$, Theorem 2 is proved.

REFERENCES

[1] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger, and F. Xu, "Overview of 3GPP LTE-advanced carrier aggregation for 4G wireless communications," *IEEE Commun. Mag.*, vol. 50, pp. 122–130, Feb. 2012.

[14] In practical systems such as LTE, the coherence bandwidth is in the order of 100 kHz [31] and the subcarrier spacing is in the order of 10 kHz. Thus, the spectrum of each subcarrier is virtually flat due to the highly correlated fading within each subcarrier.

[15] Note that the only condition required for this transformation is that each constraint has to be a linear combination of functions where the summation must be performed on a per subcarrier basis. The functions that are linearly combined can be arbitrary continuous functions (convex or non-convex) with respect to the optimization variables.

[16] In [38], frequency sharing was used to show the zero duality gap visually for multi-carrier systems. The proof provided in this paper is more rigorous as it is based on showing the existence of a saddle point of the Lagrangian function by using the concavity of the perturbation function.

[2] I.-K. Fu, Y.-S. Chen, P. Cheng, Y. Yuk, R. Yongho Kim, and J. S. Kwak, "Multicarrier technology for 4G WiMax system [WiMAX/LTE Update]," *IEEE Commun. Mag.*, vol. 48, pp. 50–58, Aug. 2010.

[3] C. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. Shellhammer, and W. Caldwell, "IEEE 802.22: the first cognitive radio wireless regional area network standard," *IEEE Commun. Mag.*, vol. 47, pp. 130–138, Jan. 2009.

[4] J. Zeng and H. Minn, "A novel OFDMA ranging method exploiting multiuser diversity," *IEEE Trans. Commun.*, vol. 58, pp. 945–955, Mar. 2010.

[5] P. Chan and R. Cheng, "Capacity maximization for zero-forcing MIMO-OFDMA downlink systems with multiuser diversity," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1880–1889, May 2007.

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

[7] R. Zhang, "Cooperative multi-cell block diagonalization with per-base-station power constraints," *IEEE J. Sel. Areas Commun.*, vol. 28, pp. 1435–1445, Dec. 2010.

[8] G. Dartmann, W. Afzal, X. Gong, and G. Ascheid, "Joint optimization of beamforming, user scheduling, and multiple base station assignment in a multicell network," in *Proc. 2011 IEEE Wireless Commun. Netw. Conf.*, pp. 209–214.

[9] D. W. K. Ng and R. Schober, "Resource allocation and scheduling in multi-cell OFDMA systems with decode-and-forward relaying," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 2246–2258, July 2011.

[10] 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.

[11] H. Zhang, L. Venturino, N. Prasad, P. Li, S. Rangarajan, and X. Wang, "Weighted sum-rate maximization in multi-cell networks via coordinated scheduling and discrete power control," *IEEE J. Sel. Areas Commun.*, vol. 29, pp. 1214–1224, June 2011.

[12] A. Chowdhery, W. Yu, and J. M. Cioffi, "Cooperative wireless multicell OFDMA network with backhaul capacity constraints," in *Proc. 2011 IEEE Intern. Commun. Conf.*, pp. 1–6.

[13] H. Galeana-Zapien and R. Ferrus, "Design and evaluation of a backhaul-aware base station assignment algorithm for OFDMA-based cellular networks," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 3226–3237, Oct. 2010.

[14] S. Ali and V. Leung, "Dynamic frequency allocation in fractional frequency reused OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 4286–4295, Aug. 2009.

[15] Y. Chen, S. Zhang, S. Xu, and G. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, pp. 30–37, June 2011.

[16] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. Grant, H. Haas, J. Thompson, I. Ku, C.-X. Wang, T. A. Le, M. Nakhai, J. Zhang, and L. Hanzo, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, pp. 46–54, June 2011.

[17] H. Bogucka and A. Conti, "Degrees of freedom for energy savings in practical adaptive wireless systems," *IEEE Commun. Mag.*, vol. 49, pp. 38–45, June 2011.

[18] G. Miao, N. Himayat, and G. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Trans. Commun.*, vol. 58, pp. 545–554, Feb. 2010.

[19] Z. Hasan, G. Bansal, E. Hossain, and V. Bhargava, "Energy-efficient power allocation in OFDM-based cognitive radio systems: a risk-return model," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 6078–6088, Dec. 2009.

[20] X. Xiao, X. Tao, Y. Jia, and J. Lu, "An energy-efficient hybrid structure with resource allocation in OFDMA networks," in *Proc. 2011 IEEE Wireless Commun. Netw. Conf.*, pp. 1466–1470.

[21] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks—part II: algorithm development," vol. 4, pp. 625–634, Mar. 2005.

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

[23] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Proc. 2010 Future Netw. Mobile Summit*, pp. 1–8.

[24] A. Leon-Garcia and I. Widjaja, *Communication Networks: Fundamentals Concepts and Key Architectures*, 2nd edition. McGraw-Hill Science/Engineering/Math, 2003.

[25] W. Dinkelbach, "On nonlinear fractional programming," *Manag. Science*, vol. 13, pp. 492–498, Mar. 1967. Available: http://www.jstor.org/stable/2627691

[26] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. PP, pp. 1–13.

[27] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, pp. 528–541, Mar. 2006.

[28] J. Mao, J. Gao, Y. Liu, and G. Xie, "Simplified semi-orthogonal user selection for MU-MIMO systems with ZFBF," *IEEE Trans. Wireless Commun. Lett.*, vol. 1, pp. 42–45, Feb. 2012.

[29] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 1478–1491, Sep. 2007.

[30] K. Seong, M. Mohseni, and J. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proc. 2006 IEEE Intern. Symp. Inf. Theory*, pp. 1394–1398.

[31] "3rd Generation Partnership Project;Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects (Release 9) ," 3GPP TR 36.814 V9.0.0 (2010-03), Tech. Rep.

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

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

[34] "E3, T3 and HSSI Manageable Fiber Optic Modems," RADirect, Tech. Rep. Available: http://www.rad-direct.com/datasheet/fomie3t3.pdf.

[35] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, 1st edition. Cambridge University Press, 2005.

[36] C. Goh and X. Q. Yang, *Duality in Optimization and Variational Inequalities*, 1st edition. Taylor & Francis, 2002.

[37] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd edition. Wiley-Interscience, 2006.

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

Derrick Wing Kwan Ng (S'06) received the bachelor degree with first class honors 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 received his Ph.D. degree from the University of British Columbia (UBC) in 2012. In the summer of 2011 and spring of 2012, he was a visiting scholar at the Centre Tecnològic de Telecomunicacions de Catalunya - Hong Kong (CTTC-HK). He is now working as a postdoctoral fellow in the Institute for Digital Communications, University of Erlangen-Nürnberg, Germany. His research interests include cross-layer optimization for wireless communication systems, resource allocation in OFDMA wireless systems, and communication theory.

Dr. Ng received the Best Paper Awards at the IEEE Wireless Communications and Networking Conference (WCNC) 2012, the IEEE Global Telecommunication Conference (Globecom) 2011, and 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, the IEEE International Conference on Communications (ICC) 2011, and the IEEE Globecom 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 HKUST in 2006. He has served as an editorial assistant to the Editor-in-Chief of the TRANSACTIONS ON COMMUNICATIONS since Jan. 2012. He is currently an Editor of the IEEE COMMUNICATIONS LETTERS. He has been a TPC member of various conferences, including the WCNC 2013, ICC 2012 workshop on Green Communications and Networking, the IEEE Globecom'12 Workshop on Heterogeneous, Multi-hop, Wireless and Mobile Networks, the PIMRC2012-MAC, the ISIEA 2012, and the PEMOS 2012, etc.

Ernest S. Lo (S'02-M'08) is the Founding Director and Chief Representative of the Centre Tecnolgic de Tecnològic de Catalunya - Hong Kong (CTTC-HK). He was a Croucher Postdoc Fellow at Stanford University, and received his Ph.D., M.Phil., and B.Eng. (1st Hons.) from the Hong Kong University of Science and Technology. He has a broad spectrum of research interests, including channel coding, resource allocation, and wireless system and architectural design, all with a goal of finding new resources and inventing new technologies for realizing a flexible, spectrally-efficient, and energy-efficient wireless multiuser network. He contributed to the standardization of the IEEE 802.22 cognitive radio WRAN system and holds a few pending and granted US and China patents. Some of them were transferred to other companies.

Dr. Lo has received three Best Paper Awards, one at the IEEE International Conference on Communications (ICC) 2007, Glasgow, and another two at the IEEE Global Communications Conference (GLOBECOM), 2011, Houston, and the IEEE Wireless Communications and Networking Conference (WCNC) 2012, Paris, respectively. He served as an Editorial Assistant of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS when it was founded and has been a TPC member of various conferences, including the IEEE ICC'10,11,12, IEEE GLOBECOM'10,11,12, and IEEE ICCC'12, etc. He was honored as an Exemplary Reviewer of the IEEE Communications Letters.

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. Since January 2012 he is an Alexander von Humboldt Professor and the Chair for Digital Communication at the Friedrich Alexander University (FAU), Erlangen, Germany. His research interests fall into the broad areas of Communication Theory, Wireless Communications, and Statistical Signal Processing.

Dr. Schober received the 2002 Heinz MaierLeibnitz 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, the 2008 Charles McDowell Award for Excellence in Research from UBC, a 2011 Alexander von Humboldt Professorship, and a 2012 NSERC E.W.R. Steacie Fellowship. 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 WCNC 2012, IEEE Globecom 2011, IEEE ICUWB 2006, the International Zurich Seminar on Broadband Communications, and European Wireless 2000. Dr. Schober is a Fellow of the Canadian Academy of Engineering and a Fellow of the Engineering Institute of Canada. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS.