Spectrum and Energy Efficient Cooperative Base Station Doze

Shengqian Han, *Member, IEEE,* Chenyang Yang, *Senior Member, IEEE,* and Andreas F. Molisch, *Fellow, IEEE*

***Abstract*—This paper aims to explore the potential of a high spectrum efficiency (SE) technology, coordinated multi-point (CoMP) transmission, for improving energy efficiency (EE) of downlink cellular networks. To this end, a traffic-aware mechanism, named *cooperative base station (BS) doze*, is introduced and optimized. The key idea is to allow BS idling by exploiting the delay tolerance of some users as well as the short-term spatio-temporal traffic fluctuations in the network, and to increase the opportunity of the idling by using CoMP transmission. The cooperative BS doze strategy involves BS time-slot doze pattern, and multicell user scheduling and cooperative precoding with different amount of data sharing, which are jointly optimized in a unified framework. To ensure various performance requirements of multiple users including delay tolerance and data rate, we maximize the network EE under different time-average rate constraints for different users, where the consumptions on transmit power, circuitry power and backhauling power are taken into account. We then propose a hierarchical iterative algorithm to solve the optimization problem. Simulations under practical power consumption parameters demonstrate that cooperative BS doze can provide substantial EE gain and support high data rate services with high achievable SE.**

***Index Terms*—Spectrum efficiency, energy efficiency, cooperative doze, coordinated multi-point (CoMP).**

I. INTRODUCTION

IN THE PAST years, energy efficiency (EE) has become an important design goal for cellular networks in addition to spectrum efficiency (SE) [1]. It has been widely recognized that SE can be effectively improved by coordinated multi-point (CoMP) transmission, where multiple base stations (BSs) cooperatively serve the users in multiple cells – a technique that is particularly useful in urban environments, where the dense deployment of BSs ensures that the users are always within communication distance with multiple BSs [2, 3]. However, since CoMP transmission requires considerably larger signal processing and backhauling energy consumptions than Non-CoMP [4], its high SE does not necessarily lead to a high EE.

Nonetheless, CoMP transmission has the potential for improving the EE. On one hand, the increased SE can shorten the transmission time to ensure the quality of service (QoS) requirements of users, which reduces the circuitry energy consumption. On the other hand, sharing data and channel state information (CSI) among coordinated BSs provides CoMP systems a large spatio-temporal resource pool, which can be allocated flexibly to save energy while accommodating various traffics in the network and various QoS requirements.

To improve the EE without sacrificing QoS, exploiting the spatial and temporal fluctuation of the traffic is an essential principle [5], based on which some traffic-aware transmission strategies have been proposed [5–10]. For example, switching the underutilized BSs to sleep mode during off-peak time of cellular networks is commonly recognized as a promising approach to reduce energy consumption. This is practically possible because the deployment of existing cellular networks, usually optimized for fully loaded traffics, leads to very inefficient usage of BSs during off-peak time. Moreover, the daily traffic variation due to user mobility and activities is predictable. The traffic pattern is highly stable over consecutive days and the off-peak time lasts hours each day exhibiting a pronounced diurnal behavior. This indicates that BS sleep strategies can operate on a long-term time scale, say, hours. Several BS sleep schemes have been studied, see e.g., [6–8] and references therein, to shut down the BSs with low traffics during the off-peak period.

Another example of the traffic-aware mechanism is based on a recent analysis on the traffic features in existing cellular networks, which shows that in fact there exist many BSs having no data to transmit if measured on a time scale of *milliseconds* [9]. This suggests that we can employ an all-day strategy operated on a short-term time scale, *BS doze*, to improve the efficiency of the network resource usage. BS doze, also known as BS micro-sleep or cell discontinuous transmission [10], is believed to be an efficient approach for saving circuitry energy, since it deactivates the energy-consuming hardware components when the BS operates in idle mode (i.e., there is nothing to transmit). The basic idea behind BS doze to improve the EE is to provide *on demand* service, where the transmission adapts to the short-term traffic fluctuations as well as the delay-tolerant traffics. In particular, within the delay tolerance, the BS can aggregate data and transmit them with high rate during a part of the time slots, while remaining in idle mode during other time slots. On the other hand, however, since higher SE needs higher transmit power, increasing the number of idle time slots will lead to an increase of transmit energy consumption [11, 12]. Therefore, BS doze essentially is a kind of resource allocation strategy to balance the transmit and circuitry energy consumptions.

The application of BS doze in practical networks is often limited by delay-sensitive services (e.g., voice and video

Manuscript received April 15, 2012; revised November 2, 2012; accepted February 3, 2013. This work was supported in part by the National Natural Science Foundation of China (No. 61120106002) and by the National Basic Research Program of China (No. 2012CB316003).

S. Han and C. Yang are with the School of Electronics and Information Engineering, Beihang University, Beijing, 100191, P. R. China (e-mail: sqhan@ee.buaa.edu.cn; cyyang@buaa.edu.cn).

A. F. Molisch is with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2565, USA (e-mail: molisch@usc.edu).

Digital Object Identifier 10.1109/JSAC.2014.141209.

teleconferencing), for which the BS needs to be active even when delay-tolerant applications are served simultaneously. We propose that this problem can be solved by using CoMP transmission since some coordinated BSs can help the dozed BSs to provide service. Owing to the large spatio-temporal resource pool contributed by the coordinated BSs, CoMP has the flexibility to assign the BSs and allocate the time slots for supporting the QoS requirements of various traffics. Moreover, CoMP can provide high SE, thereby the required number of active time slots is reduced and more circuitry energy can be saved. In other words, CoMP can be employed to increase the opportunity of BS doze. An example to illustrate the impact of CoMP transmission is shown in Fig. 3, which is obtained with the proposed BS doze strategy as explained in more detail in Section IV. It is shown that BS doze is possible for Non-CoMP only when all applications are delay-tolerant, while it works for CoMP regardless of the types of applications.

CoMP transmission and delay tolerant traffics have been separately considered to reduce energy consumption in the literature. The EE of CoMP transmission was evaluated in [4] with a minimum mean square error (MMSE) precoder, which showed that the benefits of CoMP depend on the extra power consumption from the complicated signal processing and the increased backhaul traffic. CoMP assisted BS sleep was studied in [7] and [13], where a multicell multi-user zero-forcing precoder and a single-user precoder were respectively considered to avoid the coverage hole caused by the BS sleep. In [14], CoMP assisted BS doze was investigated, where CoMP was used to improve the coverage again and the extra power consumption led by cooperative processing and backhauling was not considered. When accounting for delay tolerant traffics, from the viewpoint of minimizing transmit energy, it was shown in [15] that transmitting in the entire time duration with low data rate is the most energy efficient way. However, when circuitry energy consumption is taken into account, the tradeoff between delay tolerance and energy consumption becomes complicated. Accounting for both the transmit and circuitry energy consumptions, transmission and idle time allocation was optimized respectively for sensor networks in [11] and for relay networks in [12].

In this paper, we investigate the potential of exploiting BS cooperation, spatio-temporal traffic variation in the network and different QoS requirements of multiple users to improve the EE of downlink cellular networks. Note that although CoMP transmission can increase the opportunity of BS idling, it also induces extra signal processing and backhauling power consumptions, which largely depend on the amount of data shared among the BSs. Therefore, it is necessary to analyze the impact of data sharing strategies on the EE of the network. To this end, we consider CoMP with partial data sharing, i.e., the data of some users are shared only within a subset of coordinated BSs. The main contributions are summarized as follows:

- We propose a spectrum and energy efficient transmission strategy, named *cooperative BS doze*, which employs BS idling to improve the EE, and employs CoMP to improve the achievable SE and to increase the opportunity of BS idling. Specifically, the strategy exploits the increased SE and flexible user access supported by CoMP as well as exploits the diverse QoS demands of various traffics to increase the number of idle time slots and idle BSs for energy saving. A unified framework for optimizing the cooperative BS doze strategy is established. The BS time-slot doze pattern, and multi-user scheduling and precoding for CoMP systems with partial data sharing are jointly optimized, aimed at maximizing the network EE under the per-BS power constraints (PBPC) and per-user time-average rate constraints to meet the data rate and delay tolerance requirements of all users.
- To explore the potential of cooperative BS doze for improving the EE, we develop a hierarchical iterative algorithm to efficiently find a solution of the joint optimization problem. Simulation results using practical power consumption parameters show that cooperative BS doze can provide substantial EE gain over Non-CoMP BS doze, besides supporting high data rate services with the same transmit power constraint. By judiciously designing the transmission strategies as well as the data sharing strategies for CoMP systems accounting for the delay and data rate requirements, higher EE can be achieved. Moreover, the EE increases when the required SE increases.

II. System and Power Consumption Model

A. System Model

Consider a universal frequency reuse downlink CoMP network consisting of L cooperative cells each including an M-antenna BS. The radio resources are divided into a number of orthogonal time-frequency resource blocks (RBs), each including T time slots and with W Hz bandwidth, which are allocated to different users through time-division and frequency-division multiple access. Within each RB, multi-user multi-input multi-output (MU-MIMO) precoding based space-division and time-division multiple access techniques are employed to serve K single antenna users located in the L cells.

In this paper, we assume that BS doze is enabled in both frequency and time domain. Frequency BS doze is also known as bandwidth adaptation, which adjusts the usage of RBs according to the traffic variation [16]. This leads to a reduction of the maximal transmit power, so that the operating point of the power amplifier can be adapted for saving energy. Temporal BS doze saves energy by deactivating power-consuming hardware components when there are no data to transmit over the whole bandwidth. Although frequency and temporal BS doze are based on different principles, they perform similarly in energy saving in practice as analyzed in [16]. For simplifying the analysis, we assume that frequency BS doze in one RB saves $1/N$ energy of temporal BS doze in one time slot, where N denotes the number of RBs in the whole bandwidth. We refer to the frequency and temporal BS doze as BS doze in the following, and concentrate on the design of the cooperative BS doze strategy in one RB.

We consider that the users have different QoS requirements, which are characterized by packet delay and packet size. For the k-th user (denoted by MS_k), assume its packet delay as T_k time slots, within which it needs to successfully receive B_k bits of information. These two QoS requirements together can be represented by a *time-average data rate constraint* for MS_k during the T_k time slots. In contrast to an instantaneous data rate constraint, such a constraint allows BS doze. The values of T_k and B_k depend on the application type of MS_k. In general, small T_k can represent delay-sensitive applications such as voice and video teleconferencing, while large T_k and B_k can reflect delay-tolerant ones like web browsing and file transfers [17, 18].

We assume flat fading channel within each RB and assume block fading channels, i.e., the channels remain constant during each RB and are independent among different time blocks.[1] We assume perfect channel estimation at the BSs and perfect sharing of data and channel information among the coordinated BSs via noiseless and zero-latency backhaul links.[2] Denote $\mathbf{h}_{kb} \in \mathbb{C}^{M\times 1}$ as the composite channel from BS_b to MS_k, which is comprised of both large-scale and small-scale fading channels. Then the global channel of MS_k from all coordinated BSs can be expressed as $\mathbf{h}_k = [\mathbf{h}_{k1}^T, \ldots, \mathbf{h}_{kL}^T]^T$, where $(\cdot)^T$ denotes the transpose. We consider linear precoding, which provides good performance with low complexity. Denote $\mathbf{w}_{kt} = [\mathbf{w}_{kt,1}^T, \ldots, \mathbf{w}_{kt,L}^T]^T$ as the precoding vector for MS_k in the t-th time slot, where $\mathbf{w}_{kt,b} \in \mathbb{C}^{M\times 1}$ represents the precoder of MS_k at BS_b. Then the signal received at MS_k in the t-th time slot can be expressed as

$$y_{kt} = \underbrace{\mathbf{h}_k^H \mathbf{w}_{kt} x_{kt}}_{\text{desired signal}} + \mathbf{h}_k^H \underbrace{\textstyle\sum_{j=1, j\neq k}^{K} \mathbf{w}_{jt} x_{jt}}_{\text{inter-user interference}} + z_{kt}, \quad (1)$$

where x_{kt} is the data symbol for MS_k in the t-th time slot with $\mathbb{E}\{|x_{kt}|^2\} = 1$, and z_{kt} is the additive white Gaussian noise (AWGN) with zero mean and variance σ^2. Herein, $\mathbb{E}\{\cdot\}$ denotes the expectation and $(\cdot)^H$ denotes the conjugate transpose.

When we treat the inter-user interference as white noise, the instantaneous signal-to-interference plus noise ratio (SINR) and the achievable data rate of MS_k in the t-th time slot are respectively

$$\text{SINR}_{kt} = \frac{|\mathbf{h}_k^H \mathbf{w}_{kt}|^2}{\sum_{j\neq k} |\mathbf{h}_k^H \mathbf{w}_{jt}|^2 + \sigma^2}, \; R_{kt} = \log_2\left(1 + \text{SINR}_{kt}\right). \quad (2)$$

B. Power Consumption Model

A typical power consumption model for currently deployed BSs is presented in [1], which reflects the impact of power amplifier, radio frequency (RF) circuit, baseband processor, power supply and battery backup, and cooling. To capture the features of CoMP transmission, this model is extended similar to [4] as follows,

$$P_{\text{BS}}^{b,t} = aP_{\text{tx}}^{b,t} + P_{\text{sp}}^{b,t} + P_{\text{cc}}^{b,t} + P_{\text{bh}}^{b,t}, \quad (3)$$

where $P_{\text{BS}}^{b,t}$ is the total power consumption of BS_b in the t-th time slot, $P_{\text{tx}}^{b,t}$, $P_{\text{sp}}^{b,t}$, $P_{\text{cc}}^{b,t}$ and $P_{\text{bh}}^{b,t}$ denote the transmit power, the signal processing power, the circuitry power and the backhauling power in the t-th time slot, respectively, and the factor a reflects the impact of power amplifier, cooling, power supply and battery backup.

The transmit power of BS_b in the t-th time slot can be expressed as

$$P_{\text{tx}}^{b,t} = \sum_{k=1}^{K} \|\mathbf{w}_{kt,b}\|^2, \quad (4)$$

where $\|\cdot\|$ denotes Euclidian norm.

The signal processing power refers to the power consumption for channel estimation and spatial precoding. The value of $P_{\text{sp}}^{b,t}$ depends on the number of channel coefficients to be estimated and the dimension of precoders to be designed. For Non-CoMP transmission, each BS needs to estimate $\frac{KM}{L}$ channel coefficients and compute $\frac{KM}{L}$-dimension precoder for all $\frac{K}{L}$ users in each cell, where we assume the same number of users in each cell for simplicity. For CoMP transmission, each BS needs to estimate KM channel coefficients from it to all users in the coordinated cells [2]. The dimension of CoMP precoder depends on the data sharing strategy, because a BS only computes precoder for the users whose data are available at it. Let Q_b denote the dimension of the precoder to be computed at BS_b, and $Q_b \geq KM/L$. Then following [4] the signal processing power can be modeled as

$$P_{\text{sp}}^{b,t} = p_{\text{sp,c}} L + p_{\text{sp,p}} \left(\frac{LQ_b}{KM}\right)^2, \quad (5)$$

where $p_{\text{sp,c}}$ and $p_{\text{sp,p}}$ are the baseline processing powers for channel estimation and precoder computation consumed by a Non-CoMP BS, i.e., $L = 1$. It is easy to see that $P_{\text{sp}}^{b,t} = p_{\text{sp,c}} + p_{\text{sp,p}}$ when $L = 1$. CoMP transmission will consume more signal processing power. For instance, when all data are shared among L coordinated BSs, we have $Q_b = KM$ and $P_{\text{sp}}^{b,t} = p_{\text{sp,c}} L + p_{\text{sp,p}} L^2$.

The backhauling power consumption comes from sharing channel and data among coordinated BSs, which is modeled as

$$P_{\text{bh}}^{b,t} = \frac{p_{\text{bh}}(\rho_D^{b,t} + \rho_C^{b,t})}{C_{\text{bh}}}, \quad (6)$$

where p_{bh} denotes the power consumption of the backhaul equipment under the maximum rate C_{bh}, and $\rho_D^{b,t}$ and $\rho_C^{b,t}$ denote the backhaul traffic in the t-th time slot due to the data and channel sharing for BS_b. Considering practical system configurations and moderate Doppler speeds, it was shown in [19] that the backhaul capacity required for channel sharing is negligible compared with data sharing. Therefore, we only consider the backhauling power consumption for data sharing.

The circuitry power includes the power consumption of RF circuits, the baseband processing excluding signal processing, cooling, power supply and battery backup. The value of $P_{\text{cc}}^{b,t}$ depends on the BS's operation modes, including the active mode when there is a signal to transmit and the idle mode when there is nothing to transmit [1]. The operation mode of a BS can be identified by examining its transmit power. The circuitry power can be modeled by a piecewise function as $P_{\text{cc}}^{b,t} = P_{\text{cc},i}$ if $P_{\text{tx}}^{b,t} = 0$ and $P_{\text{cc}}^{b,t} = P_{\text{cc},a}$ if $P_{\text{tx}}^{b,t} > 0$, where $P_{\text{cc},i}$ and $P_{\text{cc},a}$ respectively denote the circuitry power in idle and active modes and $P_{\text{cc},a} > P_{\text{cc},i}$. By defining $\delta_P =$

[1] As most closed-loop transmission schemes, the cooperative BS doze strategy is mainly applicable for low-mobility scenarios.

[2] For highly interactive applications such as voice and video teleconferencing, the performance of CoMP will degrade if the backhaul links are noisy and with large latency.

$P_{\text{cc},a} - P_{\text{cc},i}$, the circuitry power can be rewritten as

$$P_{\text{cc}}^{b,t} = \delta_P \text{sign}(P_{\text{tx}}^{b,t}) + P_{\text{cc},i}, \tag{7}$$

where the function $\text{sign}(x) = 1$ if $x > 0$ and $\text{sign}(x) = 0$ if $x = 0$.

III. Optimization of Cooperative BS Doze

In this section, the cooperative BS doze strategy is first introduced and then optimized, aimed at maximizing the EE under the constraints of the QoS requirements of multiple users with hybrid traffics. Finally, an efficient algorithm is proposed to obtain the cooperative BS doze strategy.

A. Cooperative BS Doze Strategy and Problem Formulation

Cooperative BS doze is a strategy to save energy by supporting the service *on demand* with a flexible usage of the available resource pool contributed by multiple BSs. Its basic principle is to allow some BSs to switch into idle mode during some time slots, whenever the remaining coordinated BSs are able to support the QoS requirements of the users. It is essentially a spatio-temporal resource allocation strategy to maximize the EE by exploiting the short-term spatio-temporal traffic fluctuations and by exploiting the delay tolerance of some users. In particular, we need to determine which coordinated BSs should serve which group of users in which time slots with what form of precoding, when all other BSs are turned into idle mode in the remaining time slots. In other words, we need to find the time-slot doze pattern of each BS, select the users to be served in the same time slot, and design the joint precoding. To establish a unified framework to optimize the BS doze strategy, we consider linear precoding for all users in all time slots of a RB. Then we can say that MS_k is scheduled in the t-th time slot if $\|\mathbf{w}_{kt}\|^2 > 0$, and BS_b operates in idle mode in the t-th time slot if $\|\mathbf{w}_{kt,b}\|^2 = 0$ for all k.

Since data sharing affects the signal processing and backhauling powers as well as the transmit power,[3] we consider partial data sharing among the BSs and express it as a constraint on the precoding vector for MS_k. In this way, we can incorporate the multicell precoding with different amount of data sharing into the unified framework. We denote $\mathcal{D}_k \subseteq \{1, \ldots, L\}$ as a subset of BSs who have the data of MS_k, and define $\tilde{\mathbf{D}}_k \in \mathbb{C}^{LM \times LM}$ as a block-diagonal matrix with block size M whose b-th diagonal block is $\mathbf{0}_M$ if $b \in \mathcal{D}_k$ and $\mathbf{I}_M$ if $b \notin \mathcal{D}_k$, i.e., $\tilde{\mathbf{D}}_k$ sorts out the BSs that do not transmit data to MS_k. Then the constraints on the precoding vector for MS_k can be expressed as

$$\tilde{\mathbf{D}}_k \mathbf{w}_{kt} = \bar{\mathbf{0}}_{LM}. \tag{8}$$

Herein, $\mathbf{I}_N$, $\mathbf{0}_N$ and $\bar{\mathbf{0}}_N$ denote $N \times N$ identity and zero matrices and an $N \times 1$ zero vector.

We define the EE as the ratio of total number of bits transmitted to all users to total energy consumed by all BSs in a RB. For notational simplicity, we assume that the time interval of a RB, T, is an integer multiple of T_k so that each RB can be divided into G_k sub-blocks for MS_k in time domain, where $G_k = \frac{T}{T_k}$, as illustrated in Fig. 1. Within each sub-block, B_k bits need to be delivered.[4] Therefore, the total number of bits to be transmitted in a RB can be obtained as

$$B_{\text{total}} = \sum_{k=1}^{K} G_k B_k. \tag{9}$$

This translates to the QoS constraint of MS_k in each sub-block

$$\sum_{t \in \mathcal{S}_{kg}} W \Delta_t R_{kt} = B_k, \quad g = 1, \ldots, G_k, \tag{10}$$

where $\mathcal{S}_{kg}$ denotes the index set of time slots in the g-th sub-block for MS_k, $|\mathcal{S}_{kg}| = T_k$, Δ_t is the duration of each time slot, R_{kt} is the data rate in the t-th time slot that can be obtained from (2), and $|\mathcal{S}|$ denotes the cardinality of a set $\mathcal{S}$. Note that R_{kt} can vary for different time slots, because the number of the cooperative BSs who are active in the time slots may differ.

The QoS constraint (10) captures the impact of the diverse delay tolerance of multiple users. We can set the delay tolerance of MS_k as T_k, i.e., B_k bits of message should be conveyed within T_k time slots. It reduces to an instantaneous data rate constraint when $T_k = 1$ for delay-sensitive applications and is a time-average rate constraint for the users with delay-tolerant traffics.

Next we compute the energy consumption. According to the power consumption model in Section II-B, in the t-th time slot, the total transmit power of all BSs can be expressed as $\sum_{b=1}^{L} P_{\text{tx}}^{b,t} = \sum_{k=1}^{K} \|\mathbf{w}_{kt}\|^2$, and the total circuitry power is $\sum_{b=1}^{L} P_{\text{cc}}^{b,t}$. To reflect the impact of partial data sharing, we use $\mathcal{U}_b \subseteq \{1, \ldots, K\}$ to denote the user set whose data are available at BS_b. Then BS_b needs to compute spatial precoders for $|\mathcal{U}_b|$ users, whose dimension Q_b is $|\mathcal{U}_b|M$, and the signal processing power consumption is $P_{\text{sp}}^{b,t} = p_{\text{sp,c}} L + p_{\text{sp,p}} \left(\frac{L|\mathcal{U}_b|}{K}\right)^2$. For the backhauling power consumption, because BS_b needs to receive the data of users in $\mathcal{U}_b$ from core networks, $\rho_D^{b,t} = \sum_{k \in \mathcal{U}_b} W R_{kt}$ and $P_{\text{bh}}^{b,t} = p_{\text{bh}} \sum_{k \in \mathcal{U}_b} W R_{kt} / C_{\text{bh}}$. Therefore, we can obtain the total energy consumption in one RB as

$$E_{\text{total}} = \Delta_t \Bigg(\underbrace{a \sum_{t=1}^{T} \sum_{k=1}^{K} \|\mathbf{w}_{kt}\|^2}_{\text{transmit power}} + \underbrace{T \sum_{b=1}^{L} \left(p_{\text{sp,c}} L + p_{\text{sp,p}} \left(\frac{L|\mathcal{U}_b|}{K}\right)^2 \right)}_{\text{signal processing power}} + \underbrace{\sum_{t=1}^{T} \sum_{b=1}^{L} P_{\text{cc}}^{b,t}}_{\text{circuitry power}} + \underbrace{\frac{W p_{\text{bh}}}{C_{\text{bh}}} \sum_{t=1}^{T} \sum_{b=1}^{L} \sum_{k \in \mathcal{U}_b} R_{kt}}_{\text{backhauling power}} \Bigg). \tag{11}$$

From (10), we have $\sum_{t=1}^{T} R_{kt} = \sum_{g=1}^{G_k} \sum_{t \in \mathcal{S}_{kg}} R_{kt} = \frac{G_k B_k}{W \Delta_t}$. Then the backhauling power in (11) can be rewritten as $\frac{p_{\text{bh}}}{\Delta_t C_{\text{bh}}} \sum_{b=1}^{L} \sum_{k \in \mathcal{U}_b} G_k B_k$. Further considering (9), the EE is obtained as (12) at the top of the following page, where the

[3]In general, to achieve the same data rate, the more the data is shared, the less transmit power is needed.

[4]The more general case can be easily included by using a time-proportioned rule as follows. We divide a RB into G_k sub-blocks for MS_k, where $G_k = \lceil \frac{T}{T_k} \rceil$ and $\lceil x \rceil$ denotes the smallest integer not smaller than x. Each of the first $G_k - 1$ sub-blocks includes T_k time slots, during which B_k bits are delivered. The final sub-block includes T_k' time slots, where $T_k' = T - (G_k - 1)T_k$, during which $\frac{T_k'}{T_k} B_k$ bits are delivered.

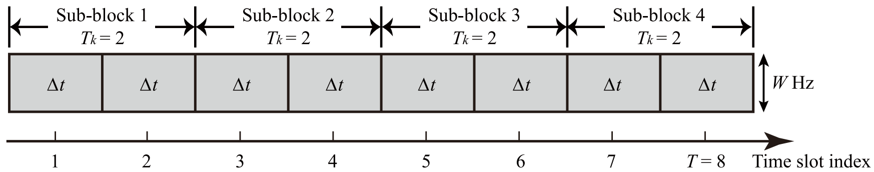


Fig. 1. Illustration of delay tolerance related parameters for MS_k with $T = 8$, $T_k = 2$, $G_k = 4$, and $\mathcal{S}_{k3} = \{5, 6\}$ as an example.

$$EE = \frac{B_{\text{total}}}{E_{\text{total}}} = \frac{W \cdot SE}{\frac{a}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\|\mathbf{w}_{kt}\|^2 + \sum_{b=1}^{L}\left(p_{\text{sp,c}}L + p_{\text{sp,p}}\left(\frac{L|\mathcal{U}_b|}{K}\right)^2\right) + \frac{1}{T}\sum_{t=1}^{T}\sum_{b=1}^{L}P_{\text{cc}}^{b,t} + \frac{Wp_{\text{bh}}}{C_{\text{bh}}}\sum_{b=1}^{L}\sum_{k\in\mathcal{U}_b}SE_k}. \tag{12}$$

per-RB SE of MS_k and the total per-RB SE of the system are respectively defined as

$$SE_k = \frac{G_k B_k}{WT\Delta_t} \text{ and } SE = \sum_{k=1}^{K} SE_k = \frac{\sum_{k=1}^{K} G_k B_k}{WT\Delta_t}. \tag{13}$$

For a given data sharing strategy, the optimization problem for cooperative BS doze aimed at maximizing EE while satisfying the QoS requirements can be formulated as follows,[5]

$$\max_{\mathbf{w}} \; EE \tag{14a}$$

$$s.\ t.\ \tilde{\mathbf{D}}_k \mathbf{w}_{kt} = \bar{\mathbf{0}}_{LM}, \quad k = 1, \ldots, K, \; t = 1, \ldots, T \tag{14b}$$

$$\sum_{t\in\mathcal{S}_{kg}} R_{kt} = \frac{B_k}{W\Delta_t}, \quad g = 1, \ldots, G_k, \quad k = 1, \ldots, K \tag{14c}$$

$$\sum_{k=1}^{K} \|\mathbf{w}_{kt,b}\|^2 \le P_0/N, \quad b = 1, \ldots, L, \quad t = 1, \ldots, T, \tag{14d}$$

where (14b) are the partial data sharing constraints, (14c) are the time-average rate constraints for multiple users that reflect the impact of their different delay tolerance and data rate demands, (14d) are the PBPC, and P_0 is the maximal transmit power per BS that is equally allocated to N RBs.

B. Equivalent Optimization Problem

1) Equivalent Transformation I: When both the QoS requirements of MS_k, T_k and B_k, and the data sharing strategy are given,[6] the following total transmit and circuitry power consumption minimization problem subject to the minimal time-average data rate constraints and the PBPC

$$\min_{\mathbf{w}} \; a\sum_{t=1}^{T}\sum_{k=1}^{K}\|\mathbf{w}_{kt}\|^2 + \sum_{t=1}^{T}\sum_{b=1}^{L}P_{\text{cc}}^{b,t} \tag{15a}$$

$$s.\ t.\ \ (14b), (14d)$$

$$\sum_{t\in\mathcal{S}_{kg}} R_{kt} \ge \frac{B_k}{W\Delta_t}, \quad g = 1, \ldots, G_k, \quad k = 1, \ldots, K \tag{15b}$$

[5]In (14), the notation $\mathbf{w}$ is short for $\{\mathbf{w}_{kt}\}$, which denotes precoders $\mathbf{w}_{kt}$ for all k and t. The notations $q = \{q_{bt}\}$, $v = \{v_{kt}\}$, $\beta = \{\beta_{kt}\}$, $\theta = \{\theta_{kg}\}$, and $\theta^\star = \{\theta_{kg}^\star\}$ in the following are defined similarly.

[6]In practice, the values of T_k and B_k are predetermined according to the QoS requirement of MS_k, and the data sharing strategy is not able to change dynamically. Therefore, assuming that they are known on the time scale of BS dozing is reasonable.

is equivalent to problem (14), i.e., the two problems have identical globally optimal solutions.

The equivalence is shown as follows. First, it is easy to see the equivalence between the maximization of EE in (14a) and the minimization of total transmit and circuitry power consumption in (15a) when T_k, B_k and the data sharing strategy are given. Note that the objective function of problem (15) is the total consumption of transmit power and circuitry power, rather than the total transmit power only. Second, the global optimum of problem (15) is attained when the constraints in (15b) hold with equality as in (14c). Otherwise, suppose that $\sum_{t\in\mathcal{S}_{kg}} R_{kt} > \frac{B_k}{\Delta_t W}$ with the optimal precoder $\{\mathbf{w}_{kt}^\star\}$, then we can always find a new precoder $\{\mathbf{w}_{kt}'^\star\}$, defined as $\mathbf{w}_{kt}'^\star = c\mathbf{w}_{kt}^\star$ for $t \in \mathcal{S}_{kg}$ with $c < 1$ and $\mathbf{w}_{kt}'^\star = \mathbf{w}_{kt}^\star$ for $t \notin \mathcal{S}_{kg}$, which can meet all users' data rate requirements with properly selected c but consuming less transmit power and the same circuitry power. In addition, the same partial data sharing and power constraints make the two problems equivalent and yield the same globally optimal solutions.

2) Equivalent Transformation II: The circuitry power $P_{\text{cc}}^{b,t}$ in the objective function of problem (15) is not an explicit function of the precoder $\mathbf{w}_{kt}$. To consider its impact on the precoder design and facilitate the optimization, we introduce an auxiliary binary scalar variable $q_{bt} \in \{0, 1\}$ defined as $q_{bt} = \text{sign}(P_{\text{tx}}^{b,t})$, with which the circuitry power modeled in (7) can be expressed as $P_{\text{cc}}^{b,t} = \delta_P q_{bt} + P_{\text{cc},i}$. Since $P_{\text{tx}}^{b,t} = \sum_{k=1}^{K} \|\mathbf{w}_{kt,b}\|^2$, we can incorporate the auxiliary variable q_{bt} into a modified PBPC as follows

$$\sum_{k=1}^{K} \|\mathbf{w}_{kt,b}\|^2 \le q_{bt}P_0/N, \quad b = 1, \ldots, L, \quad t = 1, \ldots, T. \tag{16}$$

Then, problem (15) can be equivalently reformulated as follows

$$\min_{\mathbf{w},q} \; a\sum_{t=1}^{T}\sum_{k=1}^{K}\|\mathbf{w}_{kt}\|^2 + \sum_{t=1}^{T}\sum_{b=1}^{L}(\delta_P q_{bt} + P_{\text{cc},i}) \tag{17a}$$

$$s.\ t.\ \ (14b), (15b), (16)$$

$$q_{bt} \in \{0, 1\}, \quad b = 1, \ldots, L, \quad t = 1, \ldots, T. \tag{17b}$$

Note that q_{bt} actually reflects the time-slot doze pattern of BS_b, e.g., $q_{bt} = 0$ indicates that BS_b is in idle mode in the t-th time slot. On the other hand, it also implies that BS_b does not participate in CoMP transmission in this time slot. With the modified PBPC (16), the doze pattern indicator

q_{bt} is decoupled from the vector $\mathbf{w}_{kt}$ in problem (15). As a result, $\mathbf{w}_{kt}$ only includes user scheduling and joint precoding in problem (17).

Problem (17) is a combinatorial optimization problem involving binary variables q_{bt} and complex variables $\mathbf{w}_{kt}$. To solve this problem, we propose a hierarchical iterative algorithm to separately optimize q_{bt} and $\mathbf{w}_{kt}$. In the outer iteration, a low-complexity greedy dozing algorithm is used to find the doze pattern q_{bt}. When the doze pattern is given, in the inner iteration the joint precoder is found with an iterative weighted MMSE algorithm.

C. Inner Iteration: Iterative Weighted MMSE Algorithm

For a given value of q_{bt}, minimizing the objective function of problem (17) is equivalent to minimizing the total transmit power. When $T_k = 1$ for all k, the time-average rate constraints (15b) reduce to instantaneous rate constraints, which can be expressed as instantaneous SINR constraints and further converted into second-order cone constraints [20]. Therefore, problem (17) is convex since both the objective function and all constraints are convex.

When $T_k > 1$ for any k, however, the constraints (15b) become non-convex. One way to solve this problem is to convert the time-average rate constraints into the weighted sum mean square error (MSE) constraints. Along the same lines as the proof in [21], we can show that for a given q_{bt} problem (17) is equivalent to the following problem[7]

$$\min_{\mathbf{w},v,\beta} \sum_{t=1}^{T}\sum_{k=1}^{K}\|\mathbf{w}_{kt}\|^2 \tag{18a}$$

$$\begin{aligned} s.\ t.\ & (14b), (16) \\ & \sum_{t\in\mathcal{S}_{kg}} \beta_{kt}\epsilon_{kt} - \log_e \beta_{kt} \le \mu_k,\ g = 1,\ldots,G_k, \\ & \qquad k = 1,\ldots,K \end{aligned} \tag{18b}$$

in the sense that the two problems have identical globally optimal solutions, where v, ϵ_{kt} and β_{kt} are defined in Appendix A, and $\mu_k = T_k - \log_e 2 \cdot \frac{B_k}{\Delta_t W}$.

Considering that the expressions of the optimal v and β will be used later, we provide a brief proof for the equivalence in Appendix A for the reader's convenience.

Note that the constraints (18b) are still not jointly convex for $\mathbf{w}$, v and β, because the weighted MSE includes a fourth-order polynomial in $\mathbf{w}$ and v and the Hessian matrix of the fourth-order polynomial is not positive semi-definite. Nevertheless, the advantage of using the weighted sum MSE constraints lies in their convexity for each of the optimization variables, which allows us to find efficient suboptimal solutions. Furthermore, since both the objective function and other constraints are convex respectively for $\mathbf{w}$, v, and β, we can solve the problem by alternately optimizing a variable with the other two variables fixed, which is called iterative weighted MMSE algorithm and summarized as follows.

[7]The PBPC (16) is necessary in problem (18), because the objective function is the sum transmit power of all BSs. If we first solve the problem by omitting the PBPC and then check the feasibility of the solution as in single cell systems, some feasible problems will be regarded as infeasible. This will lead to lower achievable SE of the system.

1) Initialize by finding a feasible $\mathbf{w}$ satisfying all the constraints of problem (18).
2) Sequentially update $\mathbf{w}$, v, and β:
   - Given $\mathbf{w}$, the receive filter v can be updated in closed form by (22) in Appendix A.
   - Given $\mathbf{w}$ and v, the MSE ϵ_{kt} can be computed from (21) in Appendix A, with which the scalar weight β can be updated in closed form by (22) in Appendix A.
   - Given v and β, the vector of precoding and scheduling $\mathbf{w}$ can be updated by solving the following convex optimization problem:

$$\min_{\mathbf{w}} \sum_{t=1}^{T}\sum_{k=1}^{K}\|\mathbf{w}_{kt}\|^2 \tag{19a}$$

$$\begin{aligned} s.\ t.\ & \sum_{t\in\mathcal{S}_{kg}} \beta_{kt}\Bigg(1 - 2\Re\{v_{kt}^*\mathbf{h}_k^H\mathbf{w}_{kt}\} + \Bigg(\sum_{j=1}^{K}|\mathbf{h}_k^H\mathbf{w}_{jt}|^2 \\ & \qquad + \sigma^2\Bigg)|v_{kt}|^2\Bigg) - \log_e \beta_{kt} \le \mu_k \end{aligned} \tag{19b}$$

$$(14b), (16),$$

   which comes from (18) by substituting (21) into (18b).
3) Repeat step 2 until the required accuracy or the maximum number of iterations is reached. □

To show the convergence of the iterative weighted MMSE algorithm, let $\mathbf{w}^{(i)}$, $v^{(i)}$ and $\beta^{(i)}$ denote the optimized variables in the i-th iteration. Given $\mathbf{w}^{(i)}$, the update of v and β in the $(i+1)$-th iteration leads to a decrease of the value of the left-hand side of (19b), because it is respectively convex for v and β. This means that $\mathbf{w}^{(i)}$ is feasible in the $(i+1)$-th iteration when $v^{(i+1)}$ and $\beta^{(i+1)}$ are given. Therefore, the value of the objective function (19a) (i.e., the total transmit power) in the $(i+1)$-th iteration will not be larger than that in the i-th iteration. Further consider that the transmit power is lower bounded, which guarantees the convergence of the algorithm.

Now we discuss how to find a feasible initial value of $\mathbf{w}$, which is crucial for the iterative algorithm. Such an initial vector should satisfy all the constraints. An iterative method was proposed in [21] to maximize the sum rate of multiple users, where finding a feasible initial vector is easy since only the maximal transmit power constraint was considered. For our problem, however, the initialization is non-trivial because the maximal weighted sum MSE constraints and the PBPC contradict each other. To tackle this difficulty, we introduce auxiliary scalar variables θ_{kg} to relax the maximal weighted sum MSE constraints, and find a feasible initial value from the following optimization problem:

$$\min_{\mathbf{w},v,\beta,\theta} \sum_{k=1}^{K}\sum_{g=1}^{G_k}\theta_{kg} \tag{20a}$$

$$\begin{aligned} s.\ t.\ & \sum_{t\in\mathcal{S}_{kg}} \beta_{kt}\epsilon_{kt} - \log_e \beta_{kt} \le \mu_k + \theta_{kg}, \\ & \qquad g = 1,\ldots,G_k,\quad k = 1,\ldots,K \end{aligned} \tag{20b}$$

$$\theta_{kg} \ge 0,\quad g = 1,\ldots,G_k,\quad k = 1,\ldots,K \tag{20c}$$

$$(14b), (16).$$

We can always find a feasible solution to problem (20), since for any $\mathbf{w}$ satisfying constraints (14b) and (16), which are easy to obtain, we can choose θ_{kg} with large enough value to ensure that the other constraints are satisfied. Let $\theta^\star$ be the optimal solution to problem (20). Then if the optimal value of the objective function $\Theta^\star = \sum_{k=1}^{K}\sum_{g=1}^{G_k}\theta_{kg}^\star$ is zero, which means $\theta_{kg}^\star = 0$ considering constraints (20c), the relaxed constraints (20b) become the original constraints (18b) and the corresponding $\mathbf{w}$ is a feasible initial value of problem (18). If $\Theta^\star > 0$, the maximal weighted sum MSE constraints cannot be satisfied, then we say problem (18) is infeasible for the given q, i.e., for this cooperative doze pattern the QoS of the users can not be supported.

Problem (20) is respectively convex for $\{\mathbf{w}, \theta\}$, v, and β, because both its objective function and constraints are linear for θ and respectively convex for $\mathbf{w}$, v, and β. Hence we can also apply the iterative weighted MMSE algorithm to solve it by replacing the objective function (19a) with (20a), changing the term μ_k in (19b) with $\mu_k + \theta_{kg}$, and considering the constraints (20c). Then in step 2 when v and β are given, we solve the new convex optimization problem to update $\mathbf{w}$ and θ.

D. Outer Iteration: Greedy Dozing Algorithm

The optimal solution for the doze pattern q_{bt} can be found via exhaustive search over 2^{LT} possible candidates, which however is of prohibitive complexity. In the following, we propose a low-complexity greedy dozing algorithm, which successively turns the BSs into idle mode until the total transmit and circuitry power consumption will increase if one more BS is chosen. The algorithm can be briefly summarized as follows.

1) Initialize by setting $q_{bt} = 1$ for all b, t and computing the value of the objective function of problem (17) with the iterative weighted MMSE algorithm, which is denoted as $P_{\text{out},0}$.
2) In the i-th iteration for $i = 1, \ldots, LT$, find the variables b' and t' so that setting $q_{b't'} = 0$ leads to the minimal power consumption:
   - For each b, t, assume $q_{bt} = 0$ and compute the value of the objective function of problem (17) with the iterative weighted MMSE algorithm, denoted by $P_{\text{out},i}^{bt}$.
   - Find b' and t' as $(b', t') = \arg\min_{(b,t)} P_{\text{out},i}^{bt}$ and set $P_{\text{out},i} = P_{\text{out},i}^{b't'}$.
   - The iteration stops when $P_{\text{out},i} > P_{\text{out},i-1}$; otherwise, set $q_{b't'} = 0$. □

IV. Simulation Results

In this section, we study the potential of cooperative BS doze for improving the EE through simulations. Unless otherwise specified, the power consumption model of [1, 4, 9] for LTE macro BSs is considered. The simulation setting and power consumption parameters are shown in Fig. 2. We assume that the considered three cells require the same per-RB SE, i.e., $SE/3$, and each cell has two uniformly distributed users with different SE requirements, which are respectively set as $SE_{2b-1} : SE_{2b} = 1 : 4$ for MS_{2b-1} and MS_{2b} in the b-th cell.

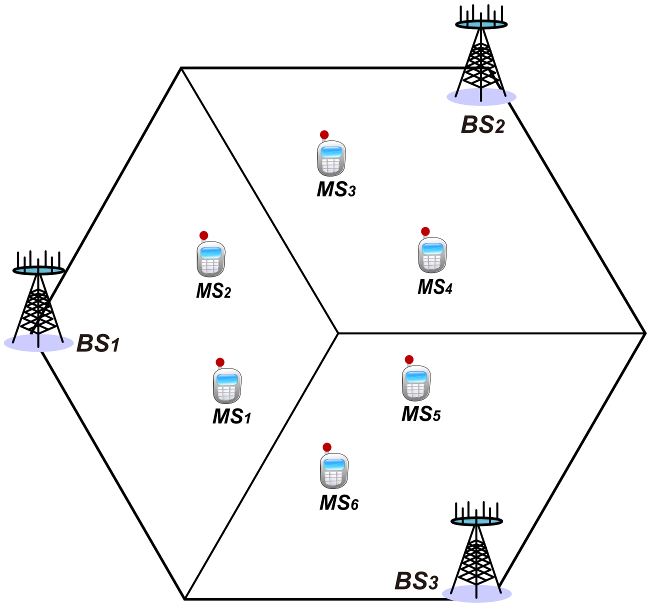


Fig. 2. A CoMP system consisting of 3 coordinated cells with 10 MHz bandwidth is simulated. The cell radius r is set to 250 m. The average receive signal-to-noise ratio (SNR) of the users located at cell boundary is set to 10 dB. The path loss model is set as $128.1+37.6\log_{10} d$, where d is the distance between the BS and the users in km. The shadowing follows lognormal distribution with a standard deviation of 8 dB, and the small-scale fading is flat Rayleigh fading. Each cell includes 2 uniformly distributed single-antenna users. The BS is equipped with 2 antennas, with a maximum transmit power of 40 W [1]. $p_{\text{sp,c}}$ = 5.8 W and $p_{\text{sp,p}}$ = 1.74 W [4], and the circuitry power consumption in active mode is 252.46 W [1]. The circuitry power consumption in idle mode is set to 10 W as an optimistic estimate in [9] and 150 W as a conservative estimate in [1]. All the power consumption parameters are measured for an LTE system with 10 MHz bandwidth [1, 4, 9]. For each RB with W Hz bandwidth, $1/N$ power consumption is considered in the simulations, where $NW = 10$ MHz. The backhauling power consumption is 50 W at the maximum rate of 100 Mbps [4], and the parameter a is set to 4.7 [1].

To evaluate the EE gain of cooperative BS doze, we also apply the proposed BS doze strategy to noncooperative networks. In this case, inter-cell interference (ICI) exists among the cells, which is modeled as AWGN to reflect the worst case of the interference. Given the transmit power of interfering BSs, in each cell the doze pattern, user scheduling and MU-MIMO precoding can be computed with the proposed hierarchical iterative algorithm by setting $L = 1$. Due to the existence of ICI, the transmit power variation of one BS will affect the transmit powers of other BSs because the specified QoS should be guaranteed. Therefore, in Non-CoMP systems the BS doze strategy needs to be optimized with one more hierarchical iteration. Before the inner and outer iterations, all the BSs need to update their transmit power iteratively.

A. Illustration of BS Time-Slot Doze Pattern

In Fig. 3, the BS time-slot doze patterns under Non-CoMP and CoMP transmission obtained from the proposed hierarchical iterative algorithm are shown for various kinds of traffics. We set $T_k = 1$ for all users with delay-sensitive applications, and set $T_k = 3$ for all users with delay-tolerant applications. When a mixture of delay-sensitive and

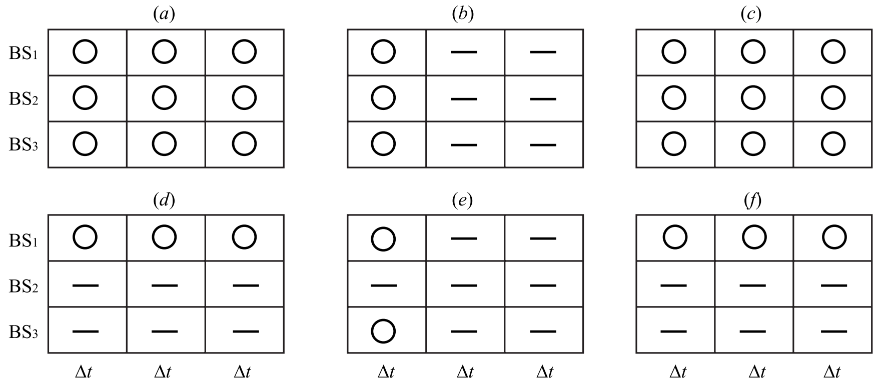


Fig. 3. An example of Non-CoMP BS doze and cooperative BS doze, obtained by using the proposed hierarchical iterative algorithm for one channel realization with $SE = 1$ bps/Hz, $L = 3$, and $T = 3$. (a) Non-CoMP doze for delay-sensitive applications, (b) Non-CoMP doze for delay-tolerant applications, (c) Non-CoMP doze for a mixture of delay-sensitive and delay-tolerant applications, (d) CoMP doze for delay-sensitive applications, (e) CoMP doze for delay-tolerant applications, and (f) CoMP doze for a mixture of delay-sensitive and delay-tolerant applications. "◯" and "–" denote active and idle modes, respectively.

delay-tolerant applications is considered for each cell, we set $T_{2b-1} = 1$ and $T_{2b} = 3$ for the two users located in the b-th cell for $b = 1, 2, 3$. Full data sharing is considered for CoMP transmission.

When all users have delay-sensitive traffics, all BSs in the Non-CoMP network should be active in all time slots (see Fig. 3(a)). By contrast, only one BS remaining active in each time slot is enough to provide service to all users in the CoMP network (see Fig. 3(d)), since the traffic load is low, e.g., $SE = 1$ bps/Hz. When all users have delay-tolerant traffics, both Non-CoMP and CoMP prefer to complete the data transmission in one time slot and turn the BSs into idle mode in other time slots (see Fig. 3(b) and Fig. 3(e)). However, under Non-CoMP each BS needs to be active at least in one time slot (see Fig. 3(b)), while under CoMP the high SE and the BS cooperation allow one BS to remain idle in all time slots and ask other coordinated BSs to help serve its users (see Fig. 3(e)). When mixed applications are considered, all the BSs in the Non-CoMP network have to remain active for supporting the delay-sensitive applications, whereas BS doze still works for the CoMP network (see Fig. 3(c) and Fig. 3(f)).

B. Convergence and Optimality of The Proposed Algorithms

In Fig. 4(a) and Fig. 4(c), the convergence behavior of the inner iterative weighted MMSE algorithm is shown, where different initializations are considered. Fig. 4(a) considers $T = 1$ and $T_k = 1$ for all users. In this case, problem (17) is convex for the given q and hence the optimal value of the objective function of problem (19) can be obtained. It is shown that with the proposed inner iteration the objective function decreases monotonically and converges to the global optimum. Fig. 4(c) considers the case when problem (17) is non-convex for the given q with $T = 3$ and $T_k = 3$ for all users. Again the convergence of the inner iteration can be observed, and different initializations converge to the same value. With the precoder obtained from the inner iteration, we calculate the data rate achieved by MS_k in the t-th time slot, which is also shown in the table inside Fig. 4(c). From the table we can see the results of the implicit user scheduling, e.g., MS_1 is scheduled only in the second time slot, while MS_2 is served

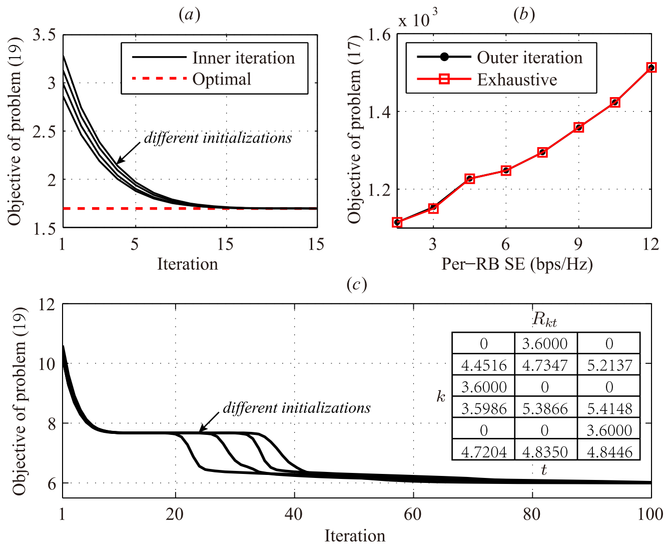


Fig. 4. Convergence behavior and optimality of the proposed hierarchical iteration algorithm with $L = 3$, $P_{\text{cc},i} = 150W$ and full data sharing for CoMP. (a) Convergence behavior of the inner iterative weighted MMSE algorithm, $SE = 18$ bps/Hz, $T = 1$, $T_k = 1$ for all k, and $q_{bt} = 1$ for all b, t. (b) Optimality of the outer greedy dozing algorithm, $T = 2$ and $T_k = 2$. (c) Convergence behavior of the inner iterative weighted MMSE algorithm, $SE = 18$ bps/Hz, $T = 3$, $T_k = 3$ for all k, and $q_{bt} = 1$ for all b, t.

in all time slots. The performance of the outer iteration, i.e., the greedy dozing algorithm, is evaluated in Fig. 4(b), where $L = 3$, $T = 2$ and $T_k = 2$ for all k. It shows that the greedy dozing algorithm performs very close to the exhaustive search through 2^6 possible candidates of q.

C. Impact of CoMP Transmission on Energy Saving

In Fig. 5 the power consumption breakdown for CoMP and Non-CoMP is depicted, where $T_k = 1$ is considered for all k to highlight the impact of CoMP transmission. In this case, all BSs need to be active in Non-CoMP systems. We consider one channel realization and obtain every part of power consumption in (11) with the obtained precoder, where full data sharing among the BSs is considered for CoMP.

The results help understand the reason of why cooperative BS doze can save energy. It is shown that the circuitry power

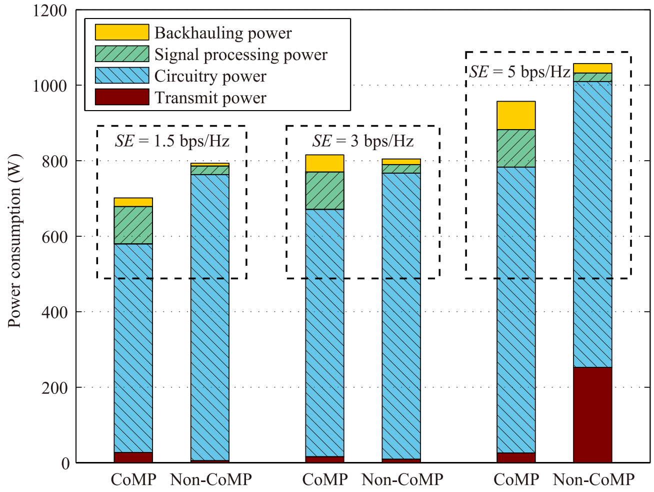


Fig. 5. Power consumption breakdown for CoMP and Non-CoMP transmission with $L = 3$, $T = 1$, $T_k = 1$ for all k, $SE = 1.5$, 3 and 5 bps/Hz, $P_{cc,i} = 150W$ and full data sharing for CoMP. Transmit power consumptions depend on the BS time-slot doze patterns and the required SE, which are less than those required for the full load Non-CoMP and CoMP systems (With the considered power consumption model, the transmit power respectively accounts for about 40% or 35% of the total power consumption of a full load Non-CoMP or CoMP BS.).

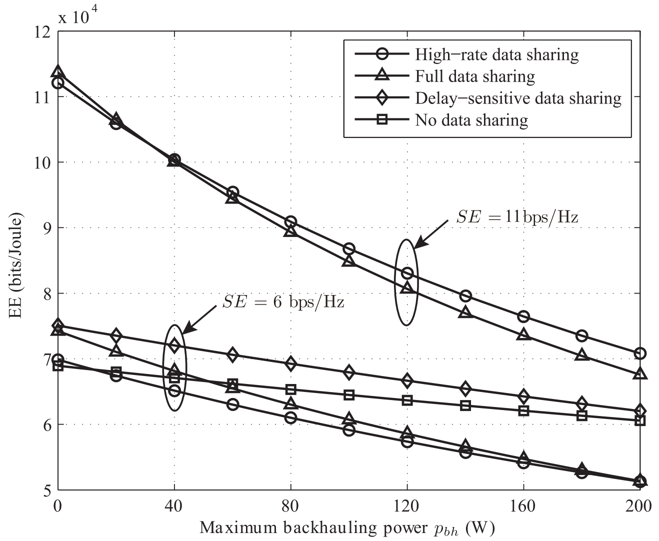


Fig. 6. Impact of data sharing on EE of cooperative BS doze with $L = 3$, $T = 3$, $SE = 6, 11$ bps/hz, and $T_k = \{1, 3\}$. The results of "Delay-sensitive data sharing" and "No data sharing" are not shown when $SE = 11$ bps/Hz, because they cannot support the SE under the PBPC.

consumption occupies a large portion of the overall power consumption, which calls for the BS doze. For instance, to achieve the per-RB SE of 1.5 bps/Hz, CoMP transmission enables one BS to operate in idle mode and all 6 users to be served by other two BSs. Although this leads to more transmit power consumption than Non-CoMP transmission, the increase of transmit power is much less than the decrease of circuitry power. When the SE is high so that all BSs need to be active, e.g., $SE = 5$ bps/Hz, the spectrum efficient CoMP can save more transmit power than Non-CoMP owing to the avoidance of ICI.

The gain of CoMP on saving the transmit power and circuitry power consumptions comes, however, at a cost of increased signal processing power and backhauling power consumptions. For example, CoMP consumes even more overall power than Non-CoMP when $SE = 3$ bps/Hz. The benefits and the cost of CoMP can be traded off by the partial data sharing among BSs as shown in the sequel.

D. Impact of Partial Data Sharing on Energy Saving

In Fig. 6 we analyze the EE of cooperative BS doze under various data sharing strategies, where $L = 3$ and $T = 3$. To show the impact of QoS requirements, we consider a mixture of delay-sensitive and delay-tolerant applications by setting the user with low data rate requirement in the b-th cell (i.e., MS_{2b-1}) as a delay-sensitive user with $T_{2b-1} = 1$ and the high-rate user MS_{2b} as a delay-tolerant user with $T_{2b} = 3$. Except for full data sharing and no data sharing, two other strategies that share only the data of delay-sensitive users and only the data of high-rate users are also considered. For each data sharing strategy, the EE can be computed based on (12). Since the backhauling power consumption has a large impact on EE, whose value depends on the employed backhaul technology [22], we consider different values of p_{bh}.

Since there is a delay-sensitive user in every cell, when no data is shared among the BSs, all BSs need to be active all the time to provide service to the users. When circuitry power consumption dominates the total power consumption, i.e., for lower SE (e.g., $SE = 6$ bps/Hz), sharing data of delay-sensitive users provides the possibility of cooperative BS doze and avoids the large backhauling power consumption, and hence achieves the highest EE. On the other hand, when transmit power consumption dominates the total power consumption, i.e., for high SE, sharing data of high-rate users is helpful to reduce transmit power consumption. For instance, the SE of 11 bps/Hz cannot be supported by only sharing the data of delay-sensitive low-rate users because the maximal transmit power per BS is constrained, but it can be achieved by sharing the data of high-rate users. Moreover, only sharing the data of high-rate users can provide higher EE than full data sharing due to the reduction of signal processing and backhauling power consumptions.

E. SE-EE Relationship of Cooperative BS Doze

In Fig. 7, we show the relationship between SE and EE of cooperative BS doze. Since both CoMP and delay-tolerant traffics can provide opportunities for BS doze, we consider the following four strategies to distinguish their impacts on energy saving. One baseline strategy is Non-CoMP transmission that simply regards all users as delay-sensitive applications such that a constant data rate is provided in every time slot, which is denoted by "Non-CoMP, ignore delay tolerance" in the legend. Another baseline strategy is Non-CoMP transmission that exploits the delay tolerance, i.e., Non-CoMP BS doze, denoted with legend "Non-CoMP, consider delay tolerance". Similar strategies are compared for CoMP transmission. We consider a mixture of delay-sensitive and delay-tolerant applications as described in the previous subsection. Both full data sharing and only delay-sensitive low-rate data sharing are considered for CoMP.

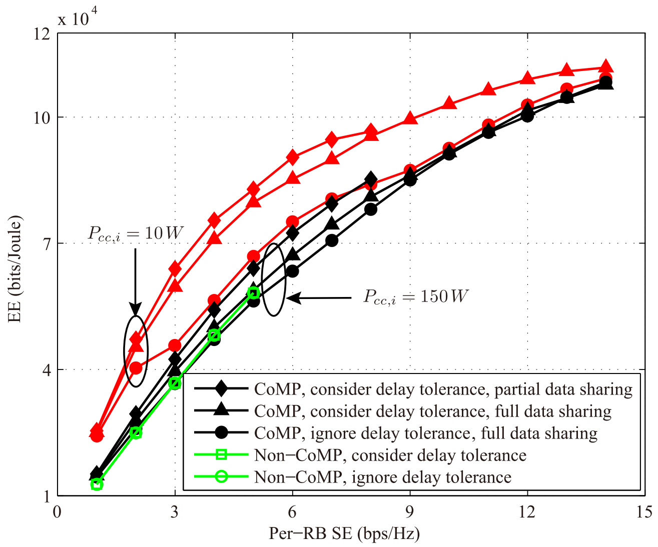


Fig. 7. SE-EE relationship for cooperative BS doze, $L = 3$, $T = 3$ and $T_k = \{1, 3\}$.

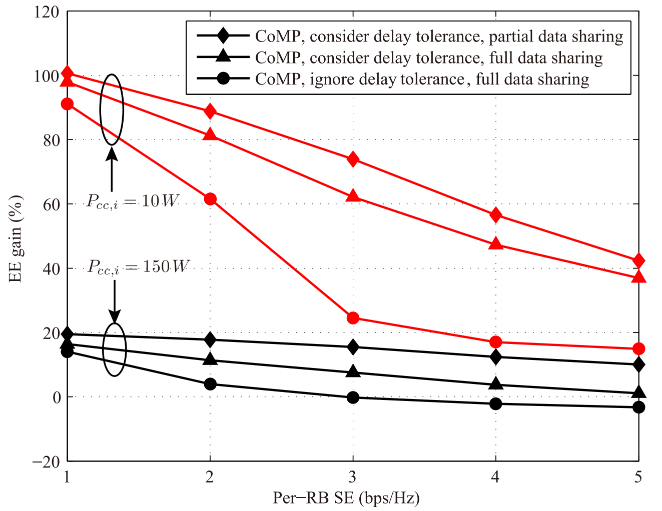


Fig. 8. EE gain of cooperative BS doze over Non-CoMP BS doze, $L = 3$, $T = 3$ and $T_k = \{1, 3\}$. The per-RB SE higher than 5 bps/Hz cannot be supported by Non-CoMP transmission.

It can be observed that compared with Non-CoMP, CoMP transmission supports much higher SE under the same transmit power constraint. For Non-CoMP, considering delay tolerance can reduce the transmit power consumption by adjusting the data rates in different time slots rather than using a constant data rate. However, the EE improvement is marginal because there is a delay-sensitive user in each cell, which forces all BSs to operate in active mode in all time slots. For CoMP, to ensure the QoS of the delay-sensitive user in each cell, at least one BS should be active in each time slot. When the SE is very low, one active BS is enough to support the services, so that exploiting the delay tolerance of other users cannot turn more BSs into idle mode. When the SE is high, most BSs should participate the cooperative transmission, so that exploiting the delay tolerance for BS doze contributes a minor EE gain. When the SE is medium, if the delay tolerance is taken into account for optimization, the high-rate delay-tolerant users will be served by only few BSs in *few time slots* with the spectrum efficient CoMP technique, and the low-rate delay-sensitive users will be served by few BSs in *each time slot*. By contrast, if ignoring the delay tolerance, more BSs need to be active in each time slot to serve both the delay-sensitive and delay-tolerant users. This leads to a large increase of circuitry power consumption and hence a noticeable performance degradation compared with the cases considering delay tolerance.

Another observation is that the EE increases with the SE. There does not exist an SE-EE trade-off region, as shown in many existing results in the literature, e.g., [23]. This arises from the simulation setting: both the system configuration and the power consumption model are for LTE macro BS. Under such a simulation setup, all the considered systems operate in the *circuitry power consumption dominant region*. In such a scenario, if the SE requirement of the systems is high, i.e., more users need high data rate transmission, the EE of the CoMP system will be also high. The same is true for Non-CoMP systems, which however can only support a maximal SE of 5 bps/Hz.

When the delay tolerance is ignored, CoMP does not necessarily outperform Non-CoMP in term of EE, depending on the value of idle circuitry power consumption. To observe their relationship more clearly, in Fig. 8 we show the EE gain of CoMP over Non-CoMP, defined as the ratio of the EE difference between CoMP and the case “Non-CoMP, consider delay tolerance” to the EE of the Non-CoMP case. It is shown that if ignoring the delay tolerance, CoMP will be inferior to Non-CoMP for high SE and with large value of $P_{\text{cc},i}$, e.g., $P_{\text{cc},i} = 150$ W. Nevertheless, if considering the delay tolerance, CoMP will always provide higher EE than Non-CoMP for any required SE. The EE can be further improved when the judiciously designed data sharing strategies are employed. If the idle circuitry power consumption can be significantly reduced (e.g., $P_{\text{cc},i} = 10$ W), which is challenging but not impossible [9], up to 40–100% EE gain can be achieved depending on the required SE.

V. Conclusions

In this paper, we studied the potential of CoMP transmission for improving the EE of downlink cellular networks, by optimizing a cooperative BS doze strategy. A CoMP system not only achieves high SE, but also provides high EE when we effectively exploit the short-term traffic fluctuations in the network as well as judiciously allocate the abundant network spatio-temporal resources to accommodate the traffics with various delay and data rate requirements. A unified framework was developed to jointly optimize the BS time-slot doze pattern, user scheduling and cooperative precoding with different amount of data sharing that maximizes the network EE with guaranteed QoS requirements for different traffics. A hierarchical iterative algorithm was proposed to reveal the potential of cooperative BS doze.

Simulation results showed that despite of suffering from the increased signal processing and backhauling power consumptions, cooperative BS doze can improve the EE significantly thanks to the increased idling opportunities, i.e., the increased number of idle time slots due to the high achievable SE and the increased number of idle BSs due to the flexible user access provided by CoMP transmission. The EE gain can be further enhanced by advanced data sharing strategies. Sharing the data of delay-sensitive services can create more idling opportunities and save more circuitry power consumption, while sharing the data of high-rate services can reduce transmit power consumption through cooperative transmission. If the delay tolerance is ignored, CoMP does not necessarily outperform Non-CoMP in term of EE, depending on the value of idle circuitry power consumption. If we exploit the diverse QoS demands and the dynamic traffic features, cooperative BS doze will provide both higher EE and higher SE than the Non-CoMP BS doze, and the EE increases when the required SE grows under the prevalent system setting. For the practical hybrid traffic scenarios with delay-sensitive and delay-tolerant applications, the cooperative BS doze strategy achieves much higher SE and provides an EE gain up to 40–100%. Although the gain may reduce if the backhaul links among the coordinated BSs are noisy and with high latency, the great potential of cooperative BS doze for improving the EE motivates further study in feasible solutions in future work.

APPENDIX A
PROOF OF THE EQUIVALENCE BETWEEN (17) AND (18) WITH GIVEN q

Proof: Let v_{kt} and β_{kt} denote the receive filter at MS_k and the scalar weight for the MSE of MS_k's data in the t-th time slot, respectively. The MSE of MS_k's data stream can be obtained from (1) as

$$\begin{aligned}\epsilon_{kt} &\triangleq \mathbb{E}\{|v_{kt}^* y_{kt} - x_{kt}|^2\}\\ &= 1 - 2\Re\{v_{kt}^*\mathbf{h}_k^H\mathbf{w}_{kt}\} + \left(\sum_{j=1}^{K}|\mathbf{h}_k^H\mathbf{w}_{jt}|^2 + \sigma^2\right)|v_{kt}|^2, \quad (21)\end{aligned}$$

where the expectation is taken over the data and noise, $(\cdot)^*$ denotes the complex conjugate, and $\Re\{\cdot\}$ denotes the real part of a complex number.

Since problem (18) is respectively convex over v_{kt} and β_{kt}, given other variables, the optimal v_{kt} and β_{kt} can be obtained by the stationary principle of the Lagrangian function, i.e., setting the first order derivatives of the Lagrangian function to zero, which are

$$v_{kt}^\star(\mathbf{w}) = \frac{\mathbf{h}_k^H\mathbf{w}_{kt}}{\sum_{j=1}^{K}|\mathbf{h}_k^H\mathbf{w}_{jt}|^2 + \sigma^2} \text{ and } \beta_{kt}^\star(\mathbf{w}, v) = \epsilon_{kt}^{-1} \;\; \forall\, k, t. \quad (22)$$

Substituting $v_{kt}^\star(\mathbf{w})$ in (22) into (21), the minimum MSE can be obtained as

$$\epsilon_{kt}^\star(\mathbf{w}) = 1 - \frac{|\mathbf{h}_k^H\mathbf{w}_{kt}|^2}{\sum_{j=1}^{K}|\mathbf{h}_k^H\mathbf{w}_{jt}|^2 + \sigma^2}. \quad (23)$$

With (22) and (23), the constraints (18b) can be written with respect to $\mathbf{w}_{kt}$ as

$$\begin{aligned}&\sum_{t\in\mathcal{S}_{kg}} 1 - \log_e\left(1 - \frac{|\mathbf{h}_k^H\mathbf{w}_{kt}|^2}{\sum_{j=1}^{K}|\mathbf{h}_k^H\mathbf{w}_{jt}|^2 + \sigma^2}\right)^{-1}\\ &= T_k - \sum_{t\in\mathcal{S}_{kg}} \log_e\left(1 + \frac{|\mathbf{h}_k^H\mathbf{w}_{kt}|^2}{\sum_{j\neq k}|\mathbf{h}_k^H\mathbf{w}_{jt}|^2 + \sigma^2}\right)\\ &= T_k - \log_e 2 \cdot \sum_{t\in\mathcal{S}_{kg}} R_{kt} \;\leq\; T_k - \log_e 2 \cdot \frac{B_k}{\Delta_t W} \quad (24)\end{aligned}$$

for $g = 1, \ldots, G_k$ and $k = 1, \ldots, K$, which are equivalent to the constraints (15b). By discarding the constants in (17a) for a given q, problem (17) and problem (18) are equivalent and yield the same globally optimal solutions. ■

REFERENCES

[1] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[2] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[3] P. Marsch and G. Fettweis, *Coordinated Multi-Point in Mobile Communications*. Cambridge University Press, 2011.

[4] A. J. Fehske, P. Marsch, and G. P. Fettweis, "Bit per Joule efficiency of cooperating base stations in cellular networks," in *Proc. IEEE GLOBECOM Workshops*, 2010.

[5] Z. Niu, "TANGO: traffic-aware network planning and green operation," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 25–29, Oct. 2011.

[6] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Proc. IEEE GLOBECOM*, 2010.

[7] S. Han, C. Yang, G. Wang, and M. Lei, "On the energy efficiency of base station sleeping with multicell cooperative transmission," in *Proc. IEEE PIMRC*, 2011.

[8] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1525–1536, Sept. 2011.

[9] T. Bohn, *et al.*, "D4.1: Most promising tracks of green radio technologies," INFSO-ICT-247733 EARTH, Tech. Rep., Dec. 2010. [Online]. Available: https://www.ict-earth.eu/publications/deliverables/deliverables.html

[10] L. M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Godor, G. Auer, and L. Van Der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 66–72, Nov. 2010.

[11] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1089–1098, Aug. 2004.

[12] C. Sun and C. Yang, "Energy efficiency analysis of one-way and two-way relay systems," *EURASIP Journal on Wireless Communications and Networking*, 2012.

[13] D. Cao, S. Zhou, C. Zhang, and Z. Niu, "Energy saving performance comparison of coordinated multi-point transmission and wireless relaying," in *Proc. IEEE GLOBECOM*, 2010.

[14] J. Xu, L. Qiu, and C. Yu, "Improving network energy efficiency through cooperative idling in the multi-cell systems," *EURASIP J. Wireless Communications and Networking*, 2011.

[15] B. Prabhakar, E. Uysal Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. IEEE INFOCOM*, 2001.

[16] A. Ambrosy, O. Blume, H. Klessig, and W. Wajda, "Energy saving potential of integrated hardware and resource management solutions for wireless base stations," in *Proc. IEEE PIMRC*, 2011.

[17] M. Alasti, B. Neekzad, J. Hui, and R. Vannithamby, "Quality of service in WiMAX and LTE networks," *IEEE Commun. Mag.*, vol. 48, no. 5, pp. 104–111, May 2010.

[18] M. Salah, N. A. Ali, A.-E. Taha, and H. Hassanein, “Evaluating uplink schedulers in LTE in mixed traffic environments,” in *Proc. IEEE ICC*, 2011.

[19] D. Samardzija and H. Huang, “Determining backhaul bandwidth requirements for network MIMO,” in *Proc. EUSIPCO*, 2009.

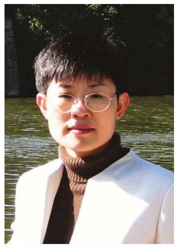
[20] W. Yu and T. Lan, “Transmitter optimization for the multi-antenna downlink with per-antenna power constraints,” *IEEE Trans. Signal Processing*, vol. 55, no. 6, pp. 2646–2660, June 2007.

[21] Q. Shi, M. Razaviyayn, Z. Luo, and C. He, “An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[22] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, “Impact of backhauling power consumption on the deployment of heterogeneous mobile networks,” in *Proc. IEEE GLOBECOM*, 2011.

[23] Z. Xu, C. Yang, G. Li, S. Zhang, Y. Chen, and S. Xu, “Energy–efficient configuration of spatial and frequency resources in MIMO-OFDMA systems,” *IEEE Trans. Commun.*, 2012, early access.

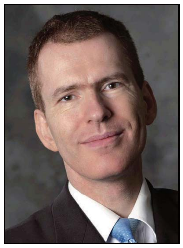
Shengqian Han (S’05-M’12) received his B.S. degree in communication engineering and Ph.D. degree in signal and information processing from Beihang University (formerly Beijing University of Aeronautics and Astronautics), Beijing, China, in 2004 and 2010 respectively. From 2010 to 2012, he held a postdoctoral research position with the School of Electronics and Information Engineering, Beihang University. He is currently a Lecturer of the same school. His research interests are multiple antenna techniques, cooperative communication and energy efficient transmission in the areas of wireless communications and signal processing.



Chenyang Yang (SM’08) received the M.S.E and Ph.D. degrees in electrical engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics), Beijing, China, in 1989 and 1997, respectively.

She is currently a Full Professor with the School of Electronics and Information Engineering, Beihang University. She has published various papers and filed many patents in the fields of signal processing and wireless communications. Her recent research interests include signal processing in network MIMO, cooperative communication, energy efficient transmission and interference management.

Prof. Yang was the Chair of the IEEE Communications Society Beijing chapter from 2008 to 2012. She has served as Technical Program Committee Member for many IEEE conferences, such as the IEEE International Conference on Communications and the IEEE Global Telecommunications Conference. She currently serves as an Associate editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor-in-Chief of the *Chinese Journal of Communications*, and an Associate editor-in-chief of the *Chinese Journal of Signal Processing*. She was nominated as an Outstanding Young Professor of Beijing in 1995 and was supported by the First Teaching and Research Award Program for Outstanding Young Teachers of Higher Education Institutions by Ministry of Education (P.R.C. “TRAPOYT”) during 1999-2004.



Andreas F. Molisch (S’89-M’95-SM’00-F’05) received the Dipl. Ing., Ph.D., and habilitation degrees from the Technical University of Vienna, Vienna, Austria, in 1990, 1994, and 1999, respectively. He subsequently was with AT&T (Bell) Laboratories Research (USA); Lund University, Lund, Sweden, and Mitsubishi Electric Research Labs (USA). He is now a Professor of electrical engineering with the University of Southern California, Los Angeles. His current research interests are the measurement and modeling of mobile radio channels, ultra-wideband communications and localization, cooperative communications, multiple-input multiple-output systems, and wireless systems for health care. He has authored, coauthored, or edited four books (among them the textbook Wireless Communications, Wiley-IEEE Press), 14 book chapters, some 140 journal papers, and numerous conference contributions, as well as more than 70 patents and 60 standards contributions. Dr. Molisch has been an Editor of a number of journals and special issues, General Chair, Technical Program Committee Chair, or Symposium Chair of multiple international conferences, as well as Chairman of various international standardization groups. He is a Fellow of the IET, an IEEE Distinguished Lecturer, and a member of the Austrian Academy of Sciences. He has received numerous awards, most recently the 2011 James Evans Avant-Garde award of the IEEE Vehicular Technology Society, the Donald Fink Prize of the IEEE, and the Eric Sumner Award of the IEEE.