User-Centric Downlink Cooperative Transmission With Orthogonal Beamforming Based Limited Feedback

Di Su and Chenyang Yang

***Abstract*—With the increase of cell density and explosive growth of data traffic, user centric is becoming one of the design principles of next-generation cellular networks. One meaning of user centric lies in that no matter where a user is located, its demand in quality of service (QoS) will be guaranteed in high probability. One approach to achieve such an ambitious goal is allowing each user to select several preferred base stations to transmit cooperatively. In this paper, we propose a user-centric downlink cooperative transmission scheme with orthogonal beamforming based limited feedback, where the cooperative clusters of multiple users may overlap and per-cell codebooks are considered. To assist the central unit (CU) for scheduling users with guaranteed QoS and performing adaptive transmission, a method for each user to estimate its signal-to-noise-and-interference ratio is derived. Targeting to ensure the required QoS of multiple users, we propose a method to select the cooperative cluster at each user and provide a method to schedule users based on their service priorities and channel conditions at the CU, where the clusters are selected semidynamically. Simulation results show that the proposed scheme significantly increases the percentage of users with satisfactory QoS demands.**

***Index Terms*—User centric, downlink cooperation transmission, limited feedback, orthogonal beamforming.**

I. Introduction

THE design principle of future cellular systems is evolving towards user-centric [1]. While there exist various meanings for user-centric [1], [2], one implication is to allow user to participate in network coordination and to ensure the quality of service (QoS) required by each user no matter where the user is located. To achieve such an ambitious goal, many techniques can be employed, e.g., a specific beam can be formed for a user to satisfy its required QoS in massive multi-input multi-output (MIMO) systems [3]. Another example is to ensure the QoS of a user by exploiting context information, e.g., location or mobility pattern of the user [4]. When the base stations (BSs) in a dense network can share information via backhaul, say under the new network architecture of cloud radio access network (C-RAN) [2], another natural way is allowing each user to select several preferred BSs (or remote radio units) for transmission in a coordinated manner according to its QoS requirement and channel condition.

By sharing data and channel state information (CSI) among cooperative BSs, which is possible under the CRAN architecture, network multi-input-multi-output (MIMO) can provide high spectral efficiency for the system and support high data rate for each user if the CSI is of high quality [5].

Considering the training and/or feedback overhead to obtain CSI, the BSs in such a network should be divided into cooperative clusters for joint transmission. For user-centric cooperative transmission, the users may select different sets of BSs to satisfy their QoS requirements according to their own channel conditions, which inevitably leads to overlapped cooperative clusters [6]. To coordinate the possible conflicting requirements from multiple users and efficiently use the limited system resources, a central unit (CU) is necessary to schedule the users according to the service priority and signal to noise and interference ratio (SINR) of each user. Yet for the user-centric cooperative transmission network with overlapped clusters, it is challenging to estimate the SINR and select cooperative cluster at each user, and to compute the network MIMO precoding for multiple users with estimated or quantized channels. This is because each user is unaware of the channel conditions and preferred cooperative clusters of other co-scheduled users such that the SINR estimation is non-trivial, while each BS obtains different sets of channel vectors with different quality and the BSs jointly serve different sets of users such that existing network MIMO precoding is not applicable.

Limited feedback techniques are widely applied for reporting channel information to the BS and have been extensively studied [7]. In single cell limited feedback systems, each user can quantize its channel direction information (CDI) with a predetermined codebook and feed back the corresponding codeword to the BS for beamforming, and estimate its SINR and feed back to the BS for user scheduling and modulation and coding scheme (MCS) selection [8]. In multi-cell limited feedback network MIMO systems, the global CDI and SINR of each user can be quantized and estimated using a similar way if the cooperative clusters are non-overlapped [9], [10].

For user-centric network MIMO systems, orthogonal beamforming based limited feedback provides an easy-implemented joint precoding for multiple users with overlapped cooperative

Manuscript received June 10, 2014; revised October 18, 2014, February 6, 2015, and April 29, 2015; accepted June 1, 2015. Date of publication June 16, 2015; date of current version August 7, 2015. This work was supported by the National High Technology Research and Development Program of China under Grant 2014AA01A703. The associate editor coordinating the review of this paper and approving it for publication was G. Abreu.

D. Su was with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China. She is now with Datang Mobile, Beijing 100083, China (e-mail: disu@ee.buaa.edu.cn).

C. Yang is with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China (e-mail: cyyang@buaa.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2015.2445816

clusters, since each user is simply served with maximum ratio transmission (MRT). A popular orthogonal beamforming based limited feedback scheme in single cell systems is per user unitary and rate control (PU2RC) [11], where orthogonal codewords are employed as the beamforming vectors for the simultaneously served users. To report the global CDIs of multiple users with different and possibly overlapped cooperative clusters, per-cell codebooks [9] instead of globally generated codebooks is more desirable, then the codebooks designed for single cell systems can be reused to quantize each per-cell CDI. However, when using the per-cell codebooks, the users with orthogonal beamforming vectors are scheduled based on multiple codewords instead of one codeword as in single cell systems. As a result, the PU2RC can not be applied into network MIMO systems in a straightforward manner.

In this paper we design user-centric downlink BS cooperative transmission scheme with per-cell codebook based orthogonal beamforming, in order to satisfy the QoS demand of each user. Since the QoS requirement depends on the incoming traffic, we use *BT problem* to model a class of service, where a given number of bits needs to be transmitted within a given duration [12]. The major contributions are summarized as follows.

- We analyze the sufficient and necessary condition to ensure orthogonal beamforming with the per-cell codebook based feedback, and the maximal number of users with orthogonal beamforming vectors that can be co-scheduled in the same time-frequency resource.
- We derive a method to estimate the SINR at each user, and propose a method to select cooperative cluster at each user to maximize an average utilization efficiency of orthogonal codewords under the constraint of the QoS required by the user. A user scheduling method is provided for the CU to ensure the QoS requirements of all users.

Notations: $(\cdot)^T$ and $(\cdot)^H$ are respectively the transpose and Hermitian operations, $|\cdot|$ either represents the absolute value of a complex variable or represents the cardinality of a set. $\|\cdot\|$ denotes the norm.

II. System and Channel Model

Consider a downlink BS cooperative transmission network, where N_b BSs jointly serve multiple users in the same frequency. To focus on designing the user-centric cooperative transmission scheme with orthogonal beamforming based limited feedback, we assume that the BSs are connected with a CU via backhaul links with unlimited capacity and zero latency. Then, network MIMO can be employed.[1] Each BS is equipped with n_t transmit antennas. Each user located in the network is equipped with a single antenna, which is allowed to select at most N_b BSs as a cooperative cluster for joint transmission with network MIMO, where the BS with the strongest average channel gain is called local BS for the user.

Denote Ω_m as the set of the BSs selected by the mth user for cooperative transmission. Due to different QoS requirements and channel conditions, the cooperative clusters of different users may overlap. In other words, a BS may belong to the cooperative clusters of different users. The association of the mth user and the kth BS can be represented as the following indicator

$$\omega_{m,k} = \begin{cases} 1, & k \in \Omega_m, \\ 0, & k \notin \Omega_m. \end{cases} \tag{1}$$

When a BS belongs to the cooperative cluster of a user, it has both data and CSI of the user.

The downlink global CSI of each user, say the mth user, is,

$$\mathbf{g}_m = \left[\omega_{m,1}\alpha_{m,1}\mathbf{h}_{m,1}^H, \cdots, \omega_{m,N_b}\alpha_{m,N_b}\mathbf{h}_{m,N_b}^H\right]^H, \tag{2}$$

where $\alpha_{m,k}$ is the average channel gain from the kth BS to the mth user including the path loss and shadowing, $\mathbf{h}_{m,k} \in \mathbb{C}^{n_t \times 1}$ is the per-cell small scale channel vector whose entries are independent complex Gaussian random variables with zero mean and unit variance.

The global CDI vector of the mth user can be expressed as

$$\bar{\mathbf{g}}_m \triangleq \frac{\mathbf{g}_m}{\|\mathbf{g}_m\|} = \left[g_{m,1}\bar{\mathbf{h}}_{m,1}^H, \cdots, g_{m,N_b}\bar{\mathbf{h}}_{m,N_b}^H\right]^H, \tag{3}$$

where $g_{m,k} = \frac{\omega_{m,k}\alpha_{m,k}\|\mathbf{h}_{m,k}\|}{\sqrt{\sum_{i=1}^{N_b}\omega_{m,i}\alpha_{m,i}^2\|\mathbf{h}_{m,i}\|^2}}$ is the weighting factor reflecting the contribution of the per-cell CDI to the global CDI, $\bar{\mathbf{h}}_{m,k} \triangleq \frac{\mathbf{h}_{m,k}}{\|\mathbf{h}_{m,k}\|}$ is the kth per-cell CDI vector, and $\|\mathbf{h}_{m,k}\|$ is the per-cell small scale channel norm.

After each user estimates its own global CSI via downlink training, the user selects its own cooperative BSs, and selects multiple per-cell codewords from a per-cell codebook to quantize the per-cell CDIs from these BSs to itself, and then sends the indices of the codewords to its local BS via uplink feedback. In this paper, we assume that each user perfectly knows its own downlink global CSI. Considering that the average channel gains are semi-static and do not need frequent feedback, we assume that they are perfectly available at the BSs. After the BSs receive the preferred cooperative cluster and per-cell CDIs reported by each user through feedback, they forward the information to the CU. The global CDI of the mth user can be reconstructed at the CU as [9]

$$\hat{\mathbf{g}}_m = \left[\hat{g}_{m,1}\hat{\mathbf{h}}_{m,1}^H, \cdots, \hat{g}_{m,N_b}\hat{\mathbf{h}}_{m,N_b}^H\right]^H, \tag{4}$$

where the estimated weighting factor is

$$\hat{g}_{m,k} = \frac{\omega_{m,k}\alpha_{m,k}}{\sqrt{\sum_{i=1}^{N_b}\omega_{m,i}\alpha_{m,i}^2}}, \tag{5}$$

and $\hat{\mathbf{h}}_{m,k}$ is the kth quantized per-cell CDI. When the mth user selects the kth BS for cooperation transmission, i.e., $\omega_{m,k} = 1$, $\hat{\mathbf{h}}_{m,k}$ is the codeword selected from a per-cell codebook for

[1] When considering the backhaul with limited capacity or large latency, other precoding needs to be devised, e.g., hybrid cooperative transmission incorporating both network MIMO and coordinated beamforming. Therefore, new problems need to be resolved, e.g., how to find such a hybrid precoding. These new problems deserve further investigation, which however are complicate even with non-overlapped clusters and perfect CSI, and hence are beyond the scope of this work.

quantizing the per-cell CDI vector from the kth BS to the mth user, and $\hat{g}_{m,k} \neq 0$. Otherwise, $\hat{\mathbf{h}}_{m,k} = \mathbf{0}$ and $\hat{g}_{m,k} = 0$.

The quantization accuracy of the global CDI is defined as follows,

$$\cos\theta_m \triangleq \left|\bar{\mathbf{g}}_m^H \hat{\mathbf{g}}_m\right|, \tag{6}$$

and the quantization error is $\sin\theta_m \triangleq \sqrt{1 - \left|\bar{\mathbf{g}}_m^H \hat{\mathbf{g}}_m\right|^2}$.

III. User-Centric Cooperative Transmission

In this section, we propose a user-centric BS cooperative transmission scheme with per-cell codebook based orthogonal beamforming.

While such a user-centric transmission scheme aims at satisfying the QoS of each user no matter where it is located, the spectral efficiency of the network should also be improved in order to support the QoS of more users. Considering the different roles of the CU and each user played in the network, the CU needs to work together with the users to achieve the goals respectively from the perspective of the whole network and of each user. Specifically, each user needs to choose the cooperative BSs according to its QoS requirement and channel condition, and feed back multiple per-cell codewords to its local BS as well as an estimated downlink SINR to assist the CU for spatial scheduling and MCS selection. On the other hand, the CU needs to select multiple users with orthogonal beamforming to achieve high sum rate of the network meanwhile guarantee the QoS of each user, and perform MCS selection for these users, after gathering the information reported by multiple users from each BS. Finally, with the decision of scheduling, beamforming and MCS selection sent by the CU, each BS can transmit data to the users.

We start by introducing the sufficient and necessary condition to ensure orthogonal beamforming with the per-cell codebook limited feedback. Then, we derive a method for estimating the SINR at each user. We proceed to develop a method for each user to select cooperative cluster based on its QoS requirement and channel condition. Since different traffic has different QoS provision, we take a kind of traffic modeled as the *BT problem* [12] as an example. Next, we provide a method of scheduling users at the CU to satisfy the QoS requirement of each user. Finally, we summarize the procedure of the cooperative transmission scheme.

A. Condition for Ensuring Orthogonality Among Beamforming Vectors

After gathering the per-cell codewords of each user in the network, the CU first computes the beamforming vectors for multiple users. Similar to the PU2RC proposed for single-cell systems [11], the CU only selects the users with orthogonal beamforming vectors. The beamforming vector of each user, say the mth user, is $\mathbf{w}_m = \hat{\mathbf{g}}_m$ [11]. In other words, the cooperative BSs jointly transmit to each user with MRT.

Note that the beamforming vector of the mth user equals to its quantized global CDI $\hat{\mathbf{g}}_m$. When $\mathbf{w}_m$ and $\mathbf{w}_k$ are orthogonal, the beamforming vector of user m may not be orthogonal to the true value of the global CDI of user k, $\bar{\mathbf{g}}_k$, due to the quantization error. The orthogonality between $\mathbf{w}_m$ and $\mathbf{w}_k$ can be ensured by the scheduling constraint to be introduced in the sequel.

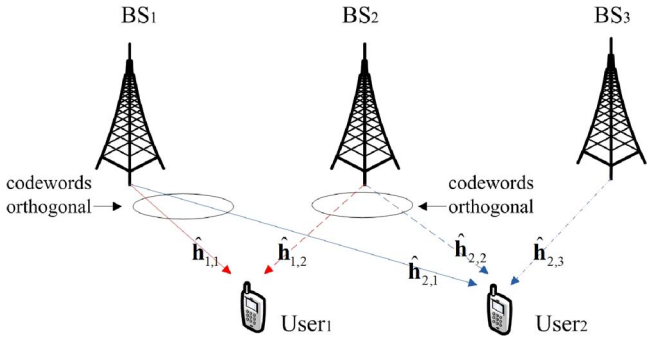


Fig. 1. Illustration of the constraint on user scheduling for per-cell codebook based orthogonal beamforming: $\hat{\mathbf{h}}_{1,1}^H \hat{\mathbf{h}}_{2,1} = 0$ and $\hat{\mathbf{h}}_{1,2}^H \hat{\mathbf{h}}_{2,2} = 0$.

In per-cell codebook based cooperative transmission systems, the users are scheduled based on their reconstructed global CDIs each consisting of multiple per-cell codewords instead of a single codeword. The following proposition shows the sufficient and necessary condition to ensure the beamforming vectors of the selected users being orthogonal.

Proposition 1: When the size of per-cell codebook is $B = \log_2(n_t)$, the beamforming vectors of any two users will be orthogonal *if and only if* the per-cell codewords selected for the links from any one BS to the two users are orthogonal.

The proof is trivial and hence is omitted. When $B > \log_2(n_t)$, the proposition is also true as will be shown by simulations in Section IV, which is however hard to prove rigorously.

The proposition indicates a constraint on user scheduling. As illustrated in Fig. 1, the co-scheduled two users should select orthogonal per-cell codewords for the links from the same BS in order for the CU to form the per-cell codebook-based orthogonal beamforming.

This suggests that the per-cell codebooks should consist of multiple sets of orthogonal codewords, where in each set the codewords are orthogonal. Fortunately, such kind of codebooks are available and widely applied in existing cellular networks, e.g., the codebooks generated by Householder transformation or DFT matrix in LTE systems [13].

For the per-cell channels from the same BS, say the kth BS, the total number of orthogonal codewords used by M co-scheduled users is $\sum_{m=1}^{M} \omega_{m,k}$. Since there exist no more than n_t mutually orthogonal codewords to quantize each per-cell CDI vector of size n_t, we have $\sum_{m=1}^{M} \omega_{m,k} \leq n_t, \forall\, k = 1, \cdots, N_b$. Then, the overall number of codewords employed for a cooperative cluster of any user in the network is no more than

$$\sum_{k=1}^{N_b} \sum_{m=1}^{M} \omega_{m,k} \leq N_b n_t, \tag{7}$$

which can also be expressed as

$$\sum_{m=1}^{M} |\Omega_m| \leq N_b n_t, \tag{8}$$

where $|\Omega_m| = \sum_{k=1}^{N_b} \omega_{m,k}$ is the number of per-cell codewords to quantize the global CDI of the mth user, which is equal to the number of the cooperative BSs selected by the user. The equality holds when the number of co-scheduled users M achieves its the maximal value.

Since $1 \le |\Omega_m| \le N_b$, the maximal number of co-scheduled users, denoted as $\bar{M}$, satisfies

$$n_t \le \bar{M} \le N_b n_t. \tag{9}$$

When every scheduled user selects all the N_b BSs for joint transmission, $\bar{M} = n_t$. When every scheduled user selects its local BS for transmission, $\bar{M} = N_b n_t$.

B. SINR Estimation at Each User

In single cell systems, the maximal number of users that can be supported simultaneously in downlink transmission, $\bar{M}$, is equal to the number of antennas at the BS, which is a critical fact for deriving an accurate SINR estimate [11]. From (9) we see that $\bar{M}$ is not always equal to $N_b n_t$ in the BS cooperative transmission system with the per-cell codebook-based orthogonal beamforming. As a result, the SINR estimation needs to be re-designed.

With orthogonal beamforming, the SINR experienced at the mth user of the BS cooperative transmission system is

$$\mathrm{SINR}_m = \frac{p_m \left|\mathbf{g}_m^H \hat{\mathbf{g}}_m\right|^2}{\sigma_m^2 + p_m \sum_{k\neq m}^{M} \left|\mathbf{g}_m^H \hat{\mathbf{g}}_k\right|^2}, \tag{10}$$

where σ_m^2 is the power of noise and inter-cluster interference, p_m is the power allocated to the user, and M is the number of co-scheduled users.

Since the transmit power allocated to the scheduled users $\{p_1, \cdots, p_M\}$, the value of M and the beamforming vectors of other scheduled users $\mathbf{w}_k$ (i.e., $\hat{\mathbf{g}}_k$), $k \neq m$ are unknown to each user, the SINR can not be computed at the user. To allow each user for estimating the SINR, we assume equal power allocation among users as in single cell systems [11]. Note that the maximal power constraint of each BS can be met with the equal power allocation when the number of co-scheduled users is large [14]. We further assume that the maximal number of users can be scheduled, which is valid when the number of candidate users is large such that the CU can always select $\bar{M}$ users with orthogonal beamforming vectors. Then, $p_m = \frac{P}{\bar{M}}$, and $M = \bar{M}$, and the SINR can be rewritten as

$$\mathrm{SINR}_m = \frac{\frac{P}{\bar{M}} \left|\mathbf{g}_m^H \hat{\mathbf{g}}_m\right|^2}{\sigma_m^2 + \frac{P}{\bar{M}} \sum_{k\neq m}^{\bar{M}} \left|\mathbf{g}_m^H \hat{\mathbf{g}}_k\right|^2} = \frac{P\|\mathbf{g}_m\|^2 \cos^2\theta_m}{\bar{M}\sigma_m^2 + P \cdot I_m}, \tag{11}$$

where $\cos\theta_m = |\bar{\mathbf{g}}_m^H \hat{\mathbf{g}}_m|$ is the quantization accuracy defined in (6), and

$$I_m \triangleq \sum_{k\neq m}^{\bar{M}} \left|\mathbf{g}_m^H \hat{\mathbf{g}}_k\right|^2 \tag{12}$$

is the power of the interference among the selected users. Except for I_m, all other terms in (11) can be obtained at the mth user.

In what follows, we derive a method to estimate the interference power. Substituting (2) and (4) into (12), we have

$$\begin{aligned} I_m &= \sum_{k\neq m}^{\bar{M}} \left| \sum_{i=1}^{N_b} \alpha_{m,i} \|\mathbf{h}_{m,i}\| \cdot \hat{g}_{k,i} \bar{\mathbf{h}}_{m,i}^H \hat{\mathbf{h}}_{k,i} \right|^2 \\ &= \sum_{k\neq m}^{\bar{M}} \left| \sum_{i=1}^{N_b} \hat{g}_{k,i} \beta_{k,i} \right|^2 \triangleq \sum_{k\neq m}^{\bar{M}} \mathbf{f}_k^H(\Omega_k) \mathbf{B}_k(\Omega_k) \mathbf{f}_k(\Omega_k), \end{aligned} \tag{13}$$

where $\beta_{m,i} \triangleq \alpha_{m,i} \|\mathbf{h}_{m,i}\| \bar{\mathbf{h}}_{m,i}^H \hat{\mathbf{h}}_{k,i}$, $\mathbf{f}_k(\Omega_k) \triangleq [\hat{g}_{k,1}, \cdots, \hat{g}_{k,N_b}]$ reflects the location of the kth user and $\|\mathbf{f}_k(\Omega_k)\| = 1$ from (5), and the element in the ith row and jth column of matrix $\mathbf{B}_k(\Omega_k)$ is $b_{i,j} = \beta_{k,i}^H \beta_{k,j} = \alpha_{m,i}\alpha_{m,j} \|\mathbf{h}_{m,i}\| \cdot \|\mathbf{h}_{m,j}\| \bar{\mathbf{h}}_{m,i}^H (\hat{\mathbf{h}}_{k,i} \hat{\mathbf{h}}_{k,j}^H) \bar{\mathbf{h}}_{m,j}$, which depends on the downlink channel of the mth user, the quantized per-cell CDIs and cooperative clusters of other co-scheduled users.

When the estimated SINR of a user is higher than the true value, the data rate selected by MCS is not achievable and an outage will occur. To reduce the outage probability in downlink transmission introduced by overestimating the SINR for user m, we find the maximal value of I_m by solving the following problem

$$\begin{aligned} \max_{\{\mathbf{f}_k(\Omega_k)\},\{\Omega_k\}} \quad & \sum_{k\neq m}^{\bar{M}} \mathbf{f}_k^H(\Omega_k) \mathbf{B}_k(\Omega_k) \mathbf{f}_k(\Omega_k) \\ \text{s.t.} \quad & \|\mathbf{f}_k(\Omega_k)\| = 1, |\Omega_k| \ge 1, \forall\, k \neq m. \end{aligned} \tag{14}$$

Since $\omega_{k,i}$ in Ω_k is binary as shown in (1), the feasible region of this optimization problem is not convex, making the problem hard to solve. To find a variable solution, we decouple the problem into two subproblems: estimating the SINR given cooperative clusters and selecting the clusters. In the sequel, to estimate the SINR, we first suppose that the cooperative clusters of all users are known at user m, and then derive the maximal interference power over all possible cooperative BS sets to get rid of such an unrealistic assumption. We address the cooperative BS set selection issue in next subsection. Then, the maximal interference power can be obtained from the following subproblem

$$\begin{aligned} \max_{\{\mathbf{f}_k\}} \quad & \sum_{k\neq m}^{\bar{M}} \mathbf{f}_k^H(\Omega_k) \mathbf{B}_k(\Omega_k) \mathbf{f}_k(\Omega_k) \\ \text{s.t.} \quad & \|\mathbf{f}_k(\Omega_k)\| = 1, k \neq m. \end{aligned} \tag{15}$$

From the Karush-Kuhn-Tucker (KKT) condition, the optimal solution of problem (15), $\mathbf{f}_k^{\mathrm{op}}(\Omega_k)$, should satisfy $\frac{\partial \mathcal{L}(\mathbf{f}_k, \lambda_k, \Omega_k)}{\partial \mathbf{f}_k(\Omega_k)}\big|_{\mathbf{f}_k(\Omega_k)=\mathbf{f}_k^{\mathrm{op}}(\Omega_k)=0}$, where $\mathcal{L}(\mathbf{f}_k, \lambda_k, \Omega_k) = \sum_{k\neq m}^{M} \mathbf{f}_k^H(\Omega_k) \mathbf{B}_k(\Omega_k) \mathbf{f}_k(\Omega_k) - \sum_{k\neq m}^{M} \lambda_k(\Omega_k)(\mathbf{f}_k^H(\Omega_k)\mathbf{f}_k(\Omega_k) - 1)$ is the Lagrangian function, and $\lambda_k(\Omega_k)$ is the Lagrange multiplier. From which we can further derive that $\mathbf{f}_k^{\mathrm{op}}(\Omega_k)$ should satisfy the following equation,

$$\mathbf{B}_k(\Omega_k) \cdot \mathbf{f}_k^{\mathrm{op}}(\Omega_k) = \lambda_k(\Omega_k) \mathbf{f}_k^{\mathrm{op}}(\Omega_k).$$

It follows that the solution of problem (15) is the eigenvector corresponding to the maximal eigenvalue $\lambda_k^{\max}(\Omega_k)$. Then, we obtain the corresponding interference power as

$$I_m^*(\Omega_k) = \sum_{k\neq m}^{\bar{M}} \left(\mathbf{f}_k^{\text{op}}\right)^H \mathbf{B}_k(\Omega_k)\mathbf{f}_k^{\text{op}} = \sum_{k\neq m}^{\bar{M}} \lambda_k^{\max}(\Omega_k), \tag{16}$$

which is a function of Ω_k, the cooperative cluster selected by the co-scheduled user k, $k \neq m$.

Because each user does not know the choice of other users, the value of $I_m^*(\Omega_k)$ can not be computed at the mth user. To circumvent this problem, we find the maximal value of $I_m^*(\Omega_k)$ over all possible cooperative BS sets by deriving an upper bound of it. From (16) we know that this can be achieved by finding the upper bounded of $\lambda_k^{\max}(\Omega_k)$, which is

$$\begin{aligned}\lambda_k^{\max}(\Omega_k) &\leq \text{trace}\left(\mathbf{B}_k(\Omega_k)\right) = \sum_{i=1}^{N_b} \beta_{k,i}^H \beta_{k,i} \\ &= \sum_{i=1}^{N_b} \alpha_{m,i}^2 \|\mathbf{h}_{m,i}\|^2 \left|\bar{\mathbf{h}}_{m,i}^H \hat{\mathbf{h}}_{k,i}\right|^2, \end{aligned} \tag{17}$$

where the equality holds when $|\Omega_k| = 1$, i.e., the kth user chooses a single BS for downlink transmission.

Substituting (17) into (16), we obtain

$$\begin{aligned} I_m^*(\Omega_k) &\leq \sum_{k\neq m}^{\bar{M}} \sum_{i=1}^{N_b} \alpha_{m,i}^2 \|\mathbf{h}_{m,i}\|^2 \left|\bar{\mathbf{h}}_{m,i}^H \hat{\mathbf{h}}_{k,i}\right|^2 \\ &= \sum_{i=1}^{N_b} \alpha_{m,i}^2 \|\mathbf{h}_{m,i}\|^2 \sum_{k\neq m}^{\bar{M}} \left|\bar{\mathbf{h}}_{m,i}^H \hat{\mathbf{h}}_{k,i}\right|^2 \\ &= \sum_{i=1}^{N_b} \alpha_{m,i}^2 \|\mathbf{h}_{m,i}\|^2 \sin^2\theta_{m,i} \triangleq I_m^{\max}, \end{aligned} \tag{18}$$

where $\sin\theta_{m,i}$ is the quantization error of the ith per-cell CDI at the mth user, which is

$$\sin\theta_{m,i} = \begin{cases} \sqrt{1 - \left|\bar{\mathbf{h}}_{m,i}^H \hat{\mathbf{h}}_{m,i}\right|^2}, & \omega_{m,i} = 1, \\ 1, & \omega_{m,i} = 0. \end{cases} \tag{19}$$

Substituting $I_m^{\max}$ into (11), the SINR estimate is

$$\widehat{\text{SINR}}_m = \frac{P\|\mathbf{g}_m\|^2 \cos^2\theta_m}{\bar{M}\sigma_m^2 + P \cdot \sum_{i=1}^{N_b} \alpha_{m,i}^2 \|\mathbf{h}_{m,i}\|^2 \sin^2\theta_{m,i}}. \tag{20}$$

Remark I: The estimated SINR in (20) is computed by finding the maximal interference under all possible combinations of the selected BSs of other co-scheduled users, which avoids the outage in transmission but may cause degradation in data rate. The SINR estimate is accurate when the number of scheduled users achieves the maximal value, i.e., $M = \bar{M}$, and all the scheduled users other than the mth user choose a single BS for transmission. Both conditions are met simultaneously when there exist sufficiently large number of users choosing single BS transmission, e.g., the users are located in the cell center or the users are with low QoS. Otherwise, the SINR will be underestimated. Such an underestimation is inevitable because each user has no information of the beamforming vectors of other co-scheduled users, which are essential for computing the SINR. In fact, this problem also exist in the PU2RC of single-cell systems [11], which has no effective solutions yet as far as the authors known.

Remark II: When the mth user only selects its local BS for transmission, say the 1st BS, the estimated SINR reduces to the estimated SINR employed in PU2RC [11], which is

$$\widehat{\text{SINR}}_m = \frac{P\|\alpha_{m,1}\mathbf{h}_{m,1}\|^2 \cos^2\theta_{m,1}}{\bar{M}\sigma_m^2 + P\|\alpha_{m,1}\mathbf{h}_{m,1}\|^2 \sin^2\theta_{m,1} + I_{\text{ICI}}},$$

where $I_{\text{ICI}} = P\sum_{i=2}^{N_b} \alpha_{m,i}^2 \|\mathbf{h}_{m,i}\|^2$ is the power of the inter-cell interference.

C. QoS-Oriented Cooperative Cluster Selection at Each User

1) Semi-Dynamic Cooperative Cluster Selection: To satisfy the QoS requirement of each user, the user demanding for high data rate prefers a large cooperative cluster, which occupies more orthogonal codewords. Since the total number of per-cell codewords in the per-cell codebook for each cooperative set is less than $N_b n_t$ as shown in (7) and (8), such a selfish selection may lead to the conflicting requirements from multiple users, which finally cause a rate loss of other users. This suggests that the selection of cooperative cluster should take into account the utilization efficiency of orthogonal codewords.

To this end, we define the rate per codeword as a metric of orthogonal codewords utilization efficiency (called *codeword efficiency* for short in the sequel), which is

$$\frac{R_m}{|\Omega_m|} = \frac{\log_2(1 + \widehat{\text{SINR}}_m)}{|\Omega_m|}, \tag{21}$$

where R_m is the downlink achievable rate of the mth user depending on the estimated SINR in (20), and $|\Omega_m|$ is the number of cooperative BSs selected by the user, which is also the number of codewords employed by the user.

Since the cooperative BS set is selected at each user, the user needs to feed the indices of selected BSs back to its local BS. To reduce the signalling overhead, it is highly desirable to select the cooperative clusters in a semi-dynamic manner. To this end, we use the average rate per codeword as a metric for cooperative cluster selection.

For the mth user with cooperative BS set Ω_m, the average rate per codeword can be obtained by taking the expectation over small scale fading channels, which is

$$\mathbb{E}\left[\frac{R_m}{|\Omega_m|}\right] = \frac{1}{|\Omega_m|}\mathbb{E}[R_m]. \tag{22}$$

In order to formulate an optimization problem for the cooperative cluster selection, we need to derive a closed form expression for the average rate per codeword.

The average rate per user can be approximated as $\mathbb{E}[R_m] = \mathbb{E}\left[\log_2(1 + \widehat{\text{SINR}}_m)\right] \approx \log_2\left(1 + \mathbb{E}[\widehat{\text{SINR}}_m]\right)$ if the variance of estimated SINR with (20) is small, which is true when the

number of antennas at each BS and the size of per-cell codebook are large. In practice, the approximation is accurate when $n_t \geq 2$ and $B \geq 2$ bits. Then the average rate per codeword can be approximated as

$$\mathbb{E}\left[\frac{R_m}{|\Omega_m|}\right] \approx \frac{\log_2\left(1+\mathbb{E}[\widehat{\mathrm{SINR}}_m]\right)}{|\Omega_m|}. \tag{23}$$

It is hard to derive the closed form expression of the average SINR, therefore we derive its lower bound instead. According to Gurland inequation [15], the average SINR estimate

$$\mathbb{E}[\widehat{\mathrm{SINR}}_m] \geq \mathbb{E}\left[P\|\mathbf{g}_m\|^2\cos^2\theta_m\right] \cdot \mathbb{E}\left[\frac{1}{\bar{M}\sigma_m^2 + P\sum_{i=1}^{N_b}\alpha_{m,i}^2\|\mathbf{h}_{m,i}\|^2\sin^2\theta_{m,i}}\right]. \tag{24}$$

Since $\mathbb{E}[x^{-p}] \geq \frac{1}{\mathbb{E}[x]^p}$ when both the random variable x and the constant p are positive [15] (see page 659), by setting $p=1$ we have

$$\mathbb{E}\left[\frac{1}{\bar{M}\sigma_m^2 + P\sum_{i=1}^{N_b}\alpha_{m,i}^2\|\mathbf{h}_{m,i}\|^2\sin^2\theta_{m,i}}\right] \geq \frac{1}{\bar{M}\sigma_m^2 + \mathbb{E}\left[P\sum_{i=1}^{N_b}\alpha_{m,i}^2\|\mathbf{h}_{m,i}\|^2\sin^2\theta_{m,i}\right]}. \tag{25}$$

Substituting (25) into (24), the average SINR is lower bounded by

$$\mathbb{E}[\widehat{\mathrm{SINR}}_m] \geq \frac{P\mathbb{E}\left[\|\mathbf{g}_m\|^2\cos^2\theta_m\right]}{\bar{M}\sigma_m^2 + P\sum_{i=1}^{N_b}\alpha_{m,i}^2\mathbb{E}\left[\|\mathbf{h}_{m,i}\|^2\sin^2\theta_{m,i}\right]}, \tag{26}$$

which is tight when the variance of normalized per-cell channel norm $\frac{\|\mathbf{h}_{m,i}\|^2}{n_t}$[2] and per-cell quantization error $\sin^2\theta_{m,i}$ are small, i.e., when the number of antennas at each BS and the number of bits for per-cell quantization are sufficiently large.

Since the per-cell channels are assumed independent and identically distributed (i.i.d.) with complex Gaussian entries, the per-cell channel norm, $\|\mathbf{h}_{m,i}\|$, is independent on the per-cell CDI, $\bar{\mathbf{h}}_{m,i}$. As a result, $\|\mathbf{h}_{m,i}\|$ is also independent on the quantization error of per-cell CDI. Then, we have

$$\mathbb{E}\left[\|\mathbf{h}_{m,i}\|^2\sin^2\theta_{m,i}\right] = \mathbb{E}\left[\|\mathbf{h}_{m,i}\|^2\right]\cdot\mathbb{E}[\sin^2\theta_{m,i}] = n_t\cdot\mathbb{E}[\sin^2\theta_{m,i}]. \tag{27}$$

Though both the global channel norm $\|\mathbf{g}_m\|$ and global quantization accuracy $\cos\theta_m$ are associated with per-cell channel norms, the value of $\cos\theta_m$ is mainly determined by the quantization of per-cell CDIs, especially when n_t is sufficiently large [16]. Consequently, we have

$$\mathbb{E}\left[\|\mathbf{g}_m\|^2\cos^2\theta_m\right] \approx \mathbb{E}\left[\|\mathbf{g}_m\|^2\right]\cdot\mathbb{E}[\cos^2\theta_m] = \left(n_t\cdot\sum_{i=1}^{N_b}\alpha_{m,i}^2\right)\cdot\mathbb{E}[\cos^2\theta_m]. \tag{28}$$

Substituting (27) and (28) into (26), we have

$$\mathbb{E}[\widehat{\mathrm{SINR}}_m] \gtrsim \frac{\left(P\sum_{i=1}^{N_b}\alpha_{m,i}^2\right)\cdot\mathbb{E}[\cos^2\theta_m]}{\bar{M}\sigma_m^2/n_t + P\sum_{i=1}^{N_b}\alpha_{m,i}^2\mathbb{E}[\sin^2\theta_{m,i}]} \triangleq \bar{\gamma}_m, \tag{29}$$

where $\bar{\gamma}_m$ is the lower bound of the average SINR.

To derive the closed form expression of $\bar{\gamma}_m$, we need the closed form expressions of $\mathbb{E}[\cos^2\theta_m]$ and $\mathbb{E}[\sin^2\theta_{m,i}]$, which depend on the codeword selection method. To simplify the analysis, we consider the simple but popular independent codeword selection, which is extended from single cell systems [17]. Then, the average quantization error of the per-cell CDI is [18],

$$\mathbb{E}[\sin^2\theta_{m,i}] = 2^{-\frac{B}{n_t-1}}, \tag{30}$$

and the average quantization accuracy of the global CDI can be approximated as [16]

$$\mathbb{E}[\cos^2\theta_m] \approx \left(1-2^{-\frac{B}{n_t-1}}\right)\cdot\left[\sum_{i\in\Omega_m}\hat{g}_{m,i}^4 + 2\sum_{i\in\Omega_m}\sum_{j>i,j\in\Omega_m}\hat{g}_{m,i}^2\hat{g}_{m,j}^2\cdot\frac{2^{2b}}{\pi^2}\sin^2\left(\frac{\pi}{2^b}\right)\right], \tag{31}$$

where B is the number of bits for quantizing each per-cell CDI, b is the number of bits for quantizing each phase ambiguity introduced by independent codeword selection for per-cell codebook based quantization [17].

The approximation in (31) is accurate when the following conditions are met simultaneously: (1) the number of antennas at each BS is sufficiently large, i.e., $n_t \to \infty$; (2) the per-cell codebooks are well designed and $B \to \infty$. In practice, $n_t \geq 2$ and $B = 4$ bits are enough for an accurate approximation [16].

Substituting (30) and (31) into (29), the lower bound of the average SINR can be approximated as

$$\bar{\gamma}_m \approx \frac{\left(P\sum_{i=1}^{N_b}\alpha_{m,i}^2\right)\cdot\eta_m\left(1-2^{-\frac{B}{n_t-1}}\right)}{\bar{M}\sigma_m^2/n_t + P\sum_{i=1}^{N_b}\alpha_{m,i}^2 2^{-\frac{B}{n_t-1}}}, \tag{32}$$

where $\eta_m \triangleq \sum_{i\in\Omega_m}\hat{g}_{m,i}^4 + 2\sum_{i\in\Omega_m}\sum_{j>i,j\in\Omega_m}\hat{g}_{m,i}^2\hat{g}_{m,j}^2\cdot\frac{2^{2b}}{\pi^2}\sin^2\left(\frac{\pi}{2^b}\right)$.

Considering (23) and (29), the lower bound of the average rate per codeword can be approximated as

$$\mathbb{E}\left[\frac{R_m}{|\Omega_m|}\right] \gtrsim \frac{\log_2(1+\bar{\gamma}_m)}{|\Omega_m|}, \tag{33}$$

which depends on the location of the mth user, as shown from the expression of $\bar{\gamma}_m$ in (32).

[2]Note that $\mathbf{g}_m = \sum_{i=1}^{N_b}\omega_{m,i}\alpha_{m,i}^2\|\mathbf{h}_{m,i}\|^2$. By dividing both the numerator and denominator of $\widehat{\mathrm{SINR}}_m$ with n_t, the numerator and denominator can be expressed as the function of normalized per-cell channel norm $\frac{\|\mathbf{h}_{m,i}\|^2}{n_t}$.

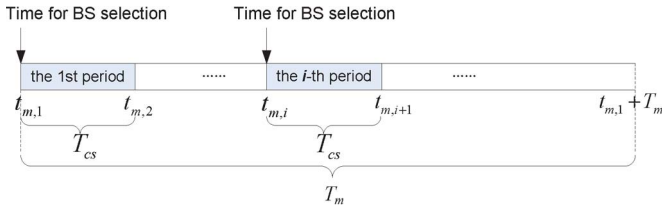


Fig. 2. Illustration of the cooperative BSs selection of the mth user.

The optimization problem for each user to select the preferred cooperative BSs is formulated as maximizing the lower bound of the average per codeword under the constraint of its QoS requirement. We take the *BT problem* [12] as an example, where the QoS requirement of a user, say the mth user, can be modeled as transmitting a given number of bits $\bar{D}_m$ within a given duration T_m. The values of T_m and $\bar{D}_m$ depend on specific application of the user. For instance, small value of T_m can reflect delay-sensitive applications such as voice and video teleconferencing, while large values of T_m and $\bar{D}_m$ can reflect delay-tolerant services like web browsing and file transfers [19]. To guarantee the QoS, the cooperative cluster is updated in a period of T_{cs}, which could be larger or smaller than T_m. Without loss of generality, we assume that $T_{cs} < T_m$. For notational simplicity, we assume that $\frac{T_m}{T_{cs}}$ is an integer. Then, the mth user will select and feed back its cooperative BS set $\frac{T_m}{T_{cs}}$ times within the required transmission duration. Denote the starting time of the mth user's request as $t_{m,1}$. The user updates the cooperative cluster every T_{cs} seconds at the time slot of $t_{m,i} = t_{m,1} + (i-1)T_{cs}$, $i = 1, \cdots, \frac{T_m}{T_{cs}}$, as shown in Fig. 2.

In the ith period, the user selects the cooperative cluster at the time slot $t_{m,i}$ by solving the following optimization problem,

$$\begin{aligned} \max_{\Omega_m} \quad & \frac{\log_2(1+\bar{\gamma}_m)}{|\Omega_m|} \\ \text{s.t.} \quad & T_{cs} \cdot \log_2(1+\widehat{\text{SINR}}_{m,t_{m,i}}) \geq D_{m,t_{m,i}} \\ & |\Omega_m| \geq 1, \end{aligned} \tag{34}$$

where $\widehat{\text{SINR}}_{m,t_{m,i}}$ is the SINR estimated with (20) based on the channel information at the time slot $t_{m,i}$, and $D_{m,t_{m,i}}$ is the number of bits necessary to be transmitted during the ith period, which is the remaining number of bits to be transmitted dividing the remaining time for transmission,

$$D_{m,t_{m,i}} = \frac{\bar{D}_m - \sum_{j=1}^{i-1} d_{m,t_{m,j}}}{\frac{T_m}{T_{cs}} - (i-1)}. \tag{35}$$

If the mth user requires the *best effort* service, we can simply set $D_{m,t_{m,i}} = 0$, indicating no QoS requirement.

Note that each user does not know future channel information at the moment of selecting the cooperative cluster. Therefore, the user has to judge whether its QoS requirement can be met based on the channel of the current time. In time-varying channels, the required number of bits in time slot $t_{m,i}$ may be failed to be conveyed to the mth user with the selected cooperative BS set Ω_m. To cope with this problem, we allow the user to recalculate the number of bits to be transmitted in the next period with (35), so as to ensure that all the bits can be transmitted within the given duration.

To find the solution of problem (34), we can exhaustively search among all the possible cooperative BS sets. The total number of combinations of the candidate BSs is $2^{N_b} - 1$, which leads to high complexity for cooperative cluster selection when N_b is large. Yet not arbitrary combination of candidate BSs can be a proper cooperative BS set, as will be discussed later. In the sequel, we develop a low complexity algorithm to select the cooperative BSs at each user.

2) Low-Complexity Cooperative Cluster Selection: To observe which kind of BSs are good candidate of a cooperative BS set, we analyze the lower bound of average rate per codeword in (33) for the user located at the cell center.

The global channel of a cell center user is dominated by the per-cell channel from its local BS, say the 1st BS. Then, $\alpha_{m,1} \gg \alpha_{m,i}$, $i = 2, \cdots, N_b$. Consequently, to maximize the lower bound of the average rate per codeword, the local BS should be in the cooperative set. For any possible cooperative set Ω_m, the estimated weighting factor in (5) meets $\hat{g}_{m,1} = 1$ and $\hat{g}_{m,i} = 0$, $i = 2, \cdots, N_b$, and the average global quantization accuracy in (31) becomes $\mathbb{E}[\cos^2\theta_m] \approx 1 - 2^{-\frac{B}{n_t-1}}$. Then, the lower bound of the average SINR in (32) with any possible Ω_m is

$$\bar{\gamma}_m \approx \frac{1 - 2^{-\frac{B}{n_t-1}}}{\bar{M}\sigma_m^2 / \left(n_t P \alpha_{m,1}^2\right) + 2^{-\frac{B}{n_t-1}}}. \tag{36}$$

Therefore, the lower bound of the average rate per codeword in (33) is an increasing function of $\alpha_{m,1}$.

This suggests that the per-cell channel with larger value of $\alpha_{m,i}$ has larger contribution to the lower bound of the average rate per codeword. Therefore, for a given number of cooperative BSs $|\Omega_m|$, we should always choose the BSs with the largest $\alpha_{m,i}$ to maximize the lower bound of average rate per codeword. This fact can be used to reduce the searching complexity in finding the solution of problem (34). For example, assuming that $N_b = 3$ and $\alpha_{m,1} > \alpha_{m,2} > \alpha_{m,3}$, we only need to consider three possible cooperative sets, $\Omega_m = \{1\}$, $\Omega_m = \{1, 2\}$ and $\Omega_m = \{1, 2, 3\}$. In this way, the number of candidate cooperative sets for searching the optimal solution of problem (34) can be reduced to N_b. The values of the objective function and the condition in (34) need to be computed for every candidate cooperative set, and the computational complexity[3] is on the order of $\mathcal{O}(k)$ if there are k BSs in the cooperative set. Then, the complexity to find the optimal solution with the low-complexity method is on the order of $\mathcal{O}\big(\sum_{k=1}^{N_b} k\big)$, which equals to $\mathcal{O}\big(\frac{N_b(N_b+1)}{2}\big)$.

Otherwise, we need to exhaustively search among all the $2^{N_b} - 1$ combinations of the candidate BSs to find the optimal solution of problem (34). The complexity is on the order of $\mathcal{O}\big(\sum_{k=1}^{N_b} k \cdot \binom{k}{N_b}\big)$, which equals to $\mathcal{O}(N_b 2^{N_b-1})$, where $\binom{k}{n} = \frac{n!}{(n-k)!k!}$ is combination number.

[3]Here we use the number of multiplications to reflect the computational complexity.

A low-complexity cooperative cluster selection algorithm is summarized in Algorithm 1.

Algorithm 1 Low-Complexity Cooperative Cluster Selection Algorithm at Each User

1: **Initialize**: The mth user computes the number of bits to be transmitted in the ith period for selecting cooperative set, $D_{m,t_{m,i}}$, with (35) at time slot $t_{m,i}$.
2: The user sorts the average channel gains in a descending order as $\alpha_{m,n_1} \geq \cdots \geq \alpha_{m,n_{N_b}}$, where n_k is the index of BS.
3: The user computes the lower bound of the average rate per codeword $\frac{\log_2(1+\bar{\gamma}_m^{(k)})}{|\Omega_m^{(k)}|}$ with (33) for every possible cooperative set $\Omega_m^{(k)} = \{n_1, \cdots, n_k\}$, $k = 1, \cdots, N_b$.
4: The user sorts the lower bounds in a descending order as

$$\frac{\log_2\left(1+\bar{\gamma}_m^{(j_1)}\right)}{\left|\Omega_m^{(j_1)}\right|} \geq \cdots \geq \frac{\log_2\left(1+\bar{\gamma}_m^{(j_{N_b})}\right)}{\left|\Omega_m^{(j_{N_b})}\right|},$$

where $\Omega_m^{(j_k)} = \{n_1, \cdots, n_{j_k}\}$, and j_k is the index of the possible cooperative BS set.
5: The user computes the downlink SINRs for $\Omega_m^{j_k}$, $k = 1, \cdots, N_b$ with (20), denoted as $\widehat{\mathrm{SINR}}_m^{(j_k)}$, $k = 1, \cdots, N_b$.
6: The selected cooperative set of the user is $\Omega_m^{j_k}$ if

$$T_{cs} \cdot \log_2\left(1+\widehat{\mathrm{SINR}}_{m,t_{m,i}}^{(j_k)}\right) \geq D_{m,t_{m,i}},$$

$$T_{cs} \cdot \log_2\left(1+\widehat{\mathrm{SINR}}_{m,t_{m,i}}^{(j_l)}\right) < D_{m,t_{m,i}}, \forall\, l = 1, \cdots, k-1.$$

7: **return** The selected cooperative set of the user is $\Omega_m = \Omega_m^{(j_k)} = \{n_1, \cdots, n_{j_k}\}$.

D. QoS-Guaranteed Scheduling at the CU

After gathering the cooperative BS cluster and multiple per-cell codewords selected by each user, the CU schedules the users for each BS to serve in the downlink simultaneously. To satisfy the QoS of each user meanwhile serve as many users as possible, the CU needs to assess the service priority of each user.

The considered service is delay-guaranteed, which requires that a given number of bits are successfully transmitted before a given deadline. Therefore, we can employ the early deadline first (EDF) policy to schedule the users with earliest deadline [20]. At the current scheduling time t, the EDF scheduling metric can be expressed as

$$\beta_{m,t}^{\mathrm{EDF}} = \frac{1}{t_{m,1} + T_m - t}, \tag{37}$$

which reflects the degree of urgency for the mth user to be served at the time slot t. Intuitively, the closer the deadline approaches, the larger this EDF metric becomes, which indicates higher service priority of the user.

Besides the expiration time, the remaining number of bits to be transmitted for the mth user can also reflect the degree of urgency for the service, which is expressed as

$$\beta_{m,t}^{\mathrm{BIT}} = \bar{D}_m - D'_{m,t}, \tag{38}$$

where $D'_{m,t}$ is the number of bits successfully transmitted to the mth user until the time slot t. A large value of $\beta_{m,t}^{\mathrm{BIT}}$ indicates a high priority of the service for the mth user.

In practice, the traffic in a network may be mixed. For example, some users require *real-time traffic*, while other users may require *best effort traffic*. In fact, the mixed traffic will reduce into real-time traffic or best effort traffic if there is only one type of users, which will not affect the following algorithm. To guarantee the QoS of the users with *real-time traffic* and support as high sum rate as possible, the CU schedules the users following three rules: (1) first select the user with the earliest deadline, i.e., the one with the largest value of $\beta_{m,t}^{\mathrm{EDF}}$; (2) for the users having the same expiration time, first schedule the one with most non-transmitted bits, i.e., the one with the largest value of $\beta_{m,t}^{\mathrm{BIT}}$; (3) for the users having the same expiration time and the same numbers of non-transmitted bits, or for the users without QoS requirement, first schedule the one with highest SINR. If the mth user requires *best effort* service, we can simply set $\beta_{m,t}^{\mathrm{EDF}} = 0$ and $\beta_{m,t}^{\mathrm{BIT}} = 0$. Recall the necessary and sufficiency condition to support orthogonal beamforming, the scheduler should satisfy the constraint in *Proposition 1*.

The QoS-guaranteed scheduling at the CU is summarized in Algorithm 2.

Algorithm 2 QoS-Guaranteed Scheduling at the CU

1: **Initialize**: For the candidate users with indices $1, \cdots, U$, the set of scheduled users at current time slot t, $\mathcal{S}_t$, is set as an empty set, and the counter is set as $u = 1$;
2: The CU sorts the candidate users according to the descending order of $\beta_{m,t}^{\mathrm{EDF}}$, $\beta_{m,t}^{\mathrm{BIT}}$ and $\widehat{\mathrm{SINR}}_{m,t}$, $1 \leq m \leq U$. The indices of users meet

$$\begin{aligned}
&i > j, \quad \text{if} \quad \beta_{m_i,t}^{\mathrm{EDF}} > \beta_{m_j,t}^{\mathrm{EDF}};\\
&i > j, \quad \text{if} \quad \beta_{m_i,t}^{\mathrm{EDF}} = \beta_{m_j,t}^{\mathrm{EDF}}, \quad \beta_{m_j,t}^{\mathrm{BIT}} > \beta_{m_j,t}^{\mathrm{BIT}};\\
&i > j, \quad \text{if} \quad \beta_{m_i,t}^{\mathrm{EDF}} = \beta_{m_j,t}^{\mathrm{EDF}}, \quad \beta_{m_i,t}^{\mathrm{EDF}} = \beta_{m_j,t}^{\mathrm{EDF}},\\
&\qquad \text{and} \quad \widehat{\mathrm{SINR}}_{m_i,t} > \widehat{\mathrm{SINR}}_{m_j,t};\\
&i < j, \quad \text{if} \quad \beta_{m_i,t}^{\mathrm{EDF}} < \beta_{m_j,t}^{\mathrm{EDF}};\\
&i < j, \quad \text{if} \quad \beta_{m_i,t}^{\mathrm{EDF}} = \beta_{m_j,t}^{\mathrm{EDF}}, \quad \beta_{m_i,t}^{\mathrm{BIT}} < \beta_{m_j,t}^{\mathrm{BIT}};\\
&i < j, \quad \text{if} \quad \beta_{m_i,t}^{\mathrm{EDF}} = \beta_{m_j,t}^{\mathrm{EDF}}, \quad \beta_{m_i,t}^{\mathrm{EDF}} = \beta_{m_j,t}^{\mathrm{EDF}},\\
&\qquad \text{and} \quad \widehat{\mathrm{SINR}}_{m_i,t} < \widehat{\mathrm{SINR}}_{m_j,t}.
\end{aligned}$$

3: **while** $u \leq U$ **do**
4: if $u = 1$, then $\mathcal{S}_t = \mathcal{S}_t \bigcup \{m_1\}$;
5: otherwise $\mathcal{S}_t = \mathcal{S}_t \bigcup \{m_u\}$, if the user m_u meets $\mathbf{w}_{m_u} \perp \mathbf{w}_i$, $\forall\, i \in \mathcal{S}_t$;
6: $u = u + 1$;
7: **end while**
8: **return** The set of scheduled users at time slot t, $\mathcal{S}_t$.

E. User-Centric Cooperative Transmission Scheme

The proposed user-centric cooperative transmission scheme with orthogonal beamforming based limited feedback is summarized as follows.

- **Processing at each user**: (a) Each user estimates its own global CSI through downlink training. Then based on the global CSI, each user selects its cooperative BS set with Algorithm 1. (b) Each user selects the per-cell codewords for the links from the selected cooperative BSs, estimates the SINR using (20) with the selected cooperative BS set, and then feeds back the codewords indices, the estimated SINR, and the indices of selected cooperative BSs to its local BS.
- **Processing at the CU**: (a) After gathering the selected codewords, cooperative clusters and the estimated SINRs from all candidate users, the CU first reconstructs the global CDIs (i.e., the beamforming vectors) for all the candidate users with (4). (b) The CU schedules the users based on the service priority and estimated SINR of each user according to Algorithm 2. (c) The CU selects the MCS for the scheduled users based on the reported SINRs.

IV. Simulation Results

In this section we evaluate the performance of the proposed user-centric cooperative transmission scheme by simulations.

For comparison, three existing schemes are simulated: 1) all users only select its local BS for transmission (with legend "Non Coop."); 2) all users select all the candidate BSs in N_b cells for cooperative transmission (with legend "Full Coop."); and 3) the cooperative clusters of different users are non-overlapped (with the legend "Fix Cluster"). The proposed user-centric cooperative transmission scheme is with legend "Prop. Scheme." With the "Non Coop.," "Full Coop." and "Fix Cluster" schemes, the users also estimate the downlink SINRs with (20), and the CU also employs the QoS-guaranteed scheduling, Algorithm 2.

We consider a dense small cell network scenario, where multiple clusters of "small BSs" are randomly distributed in a "cell" with radius of 250 m with a CU. Four small BSs and 10 single-antenna users are randomly distributed in each cluster with radius of 50 m, and the minimal distance between the small BSs is 20 m. The carrier frequency for the small BSs is 3.5 GHz. The distance between the adjacent clusters of small BSs is set large enough such that we can only consider one cluster of small BSs to evaluate the performance of the proposed scheme. Then, $N_b = 4$. The hybrid automatic repeat request with chasing combining [21] are used, where the number of maximal retransmission times is four. Other simulation parameters are listed in Table I, which are set according to [13]. All the results are obtained from 100 Monte-Carlo trials, where the locations of the users and the small BSs are updated in each trial, and the transmission time is set as 1 s in each trial where the small scale channels vary with time following Jake's model. Unless otherwise specified, all the simulation results are obtained with the above set-up.

TABLE I
Simulation Parameters

Maximal transmit power of small BS	30 dBm
Number of transmit antenna at each BS	2
System bandwidth	10 MHz
Path loss model in dB	$36.8 + 36.7 \log 10(D)$, D is the BS-user distance.
Receiver noise	-95 dB
Moving speed of users	3 km/h
Per-cell codebook	2-bit householder codebook
Per-cell codeword selection	Independent codeword selection [10]
Phase ambiguity codebook	1-bit uniformed scalar codebook
Feedback period for SINR and CDI	5 ms
User scheduling period	5 ms
Cooperative cluster selection period	50 ms
Feedback error and delay	None

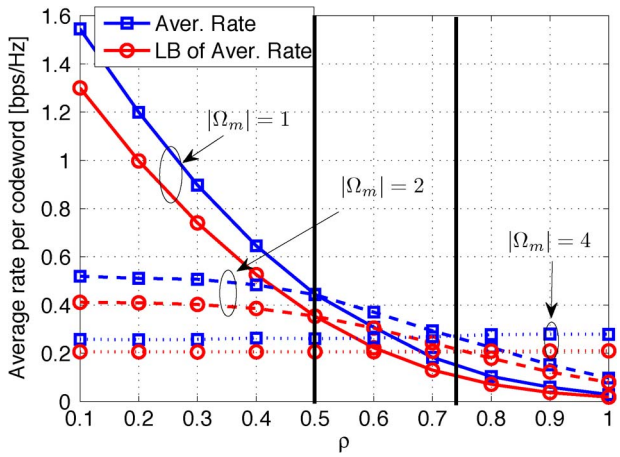


Fig. 3. Tightness of the lower bound of the average rate per codeword.

A. Tightness of the Bounds

We first evaluate the tightness of the lower bound of the average rate per codeword in (33), which depends on the location of a user. To observe the impact of the location, we model the average channel gains as $\alpha_{m,i}^2 = \frac{\rho^i(1-\rho)}{1-\rho^{N_b}}$, $i = 1, \cdots, N_b$, where $\rho \in (0, 1]$. When $\rho = 1$, the distances from the N_b BSs to the mth user are equal, i.e., it is the exact cell edge user; when $\rho \to 0$, the user is close to one BS and far away from the others, i.e., it is a cell center user.

As analyzed, the lower bound of the average rate per codeword depends on the number of selected cooperative BSs $|\Omega_m|$. In the following, we show the results when $|\Omega_m| = 1, 2, 4$, where the average rate per codeword is obtained by simulations (with legend "Aver. Rate"), and its lower bound is computed with (33) (with legend "LB of Aver. Rate"). As shown in Fig. 3, the lower bound of average rate per codeword in (33) is tight, especially for cell edge users or for large cooperative clusters. In fact, further simulation results show that the performance achieved by the BS selection algorithm using this lower bound is very close to that using the simulated average rate per codeword, which are omitted for conciseness.

As shown in (9), the maximal number of users able to be co-scheduled is not necessarily equal to the total number of antennas at the BSs in the cooperative cluster. This fact comes

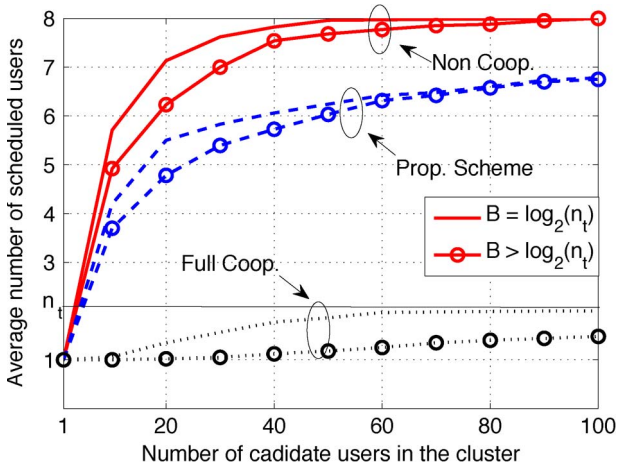


Fig. 4. Number of scheduled users versus number of candidate users.

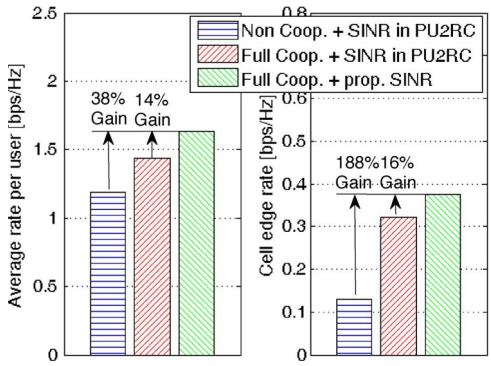


Fig. 5. Average rate per-user and cell edge rate with different SINR estimation methods.

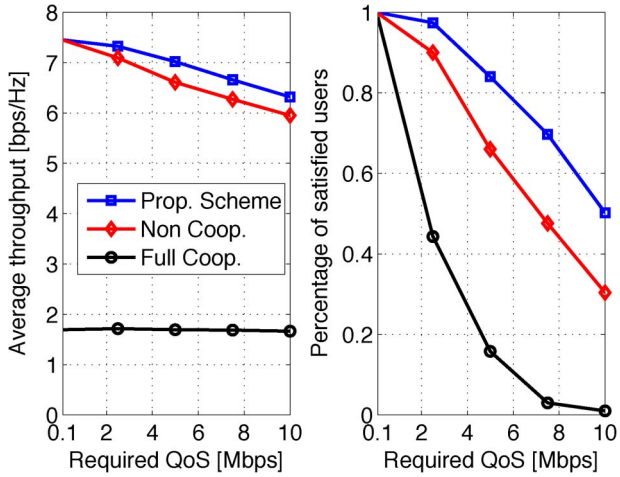


Fig. 6. Average throughput and percentage of satisfied users with different transmission schemes, where the users have the same QoS requirement.

from *Proposition 1*, which holds when the per-cell codebook size B equals to $\log_2(n_t)$, i.e., 1 bit in the simulation because $n_t = 2$. To demonstrate that the result is also true when $B > \log_2(n_t)$, we consider $B = 2$ bits, and the results of $B = \log_2(n_t)$ are also provided for comparison. To allow the maximal number of scheduled users to be scheduled, we consider the cases with a large user pool, where 100 candidate users are distributed in each cluster of N_b small cells. We set $\bar{D}_m = 5$ Mbits and $T_m = 1$ s, i.e., the equivalent time-average spectrum efficiency required by the mth user is 0.5 bps/Hz in the bandwidth of 10 MHz.

In Fig. 4, we show the number of co-scheduled users with different transmission schemes. For "Full Coop." transmission, the maximal number of scheduled users is equal to n_t, i.e., 2, which is much smaller than the total number of antennas at the BSs in the cluster $N_b n_t$, i.e., 8. When the users are allowed to select the cooperative BS sets based on the required QoS, the maximal number of scheduled users $\bar{M}$ is ranging from the minimal value of n_t to the maximal value of $N_b n_t$. This validates the result in (9), which indicates that the theoretical results obtained with $B = \log_2(n_t)$ is also true when $B > \log_2(n_t)$.

B. Evaluation of the SINR Estimation Method

In Fig. 5 we evaluate the performance of the proposed SINR estimation method. To decouple the impact of the SINR estimation from the cooperative cluster selection, we show the rate per user of "Full Coop." transmission without the QoS requirement of each user. Specifically, we consider three transmission schemes: full cooperation using the estimated SINR in [11] (with legend "Full Coop. + SINR in PU2RC") and using the proposed SINR estimate in (20) (with legend "Full Coop. + prop. SINR"), and non-cooperative transmission with the estimated SINR in [11] (with legend "Non Coop. + SINR in PU2RC"). We provide the results both for the average rate per user and the cell edge rate, which is the 5% point of the cumulative distribution function of the data rate per user. It shows that with the proposed SINR estimation method, the rate of each user is improved evidently, and the performance gain of cooperative transmission over non-cooperative transmission is more significant.

C. Evaluation of the User-Centric Cooperative Transmission Scheme

Finally, we evaluate the system performance of the proposed user-centric cooperative transmission. Simulation results show that the performance of the low complexity cooperative cluster selection is very close to that by exhaustive searching from problem (34), which are omitted for conciseness. In the sequel, we only show the results of using Algorithm 1.

In Fig. 6(a), we show the average throughput with different transmission schemes, i.e., the sum rate of all the co-scheduled users, where all the candidate users have the same QoS requirement ranging from 0.1 Mbps to 10 Mbps (i.e., with $T_m = 1$ s and D_m from 0.1 Mbits to 10 Mbits. When all the required number of bits are successfully transmitted before the required deadline, we call a user is satisfied, and show the "percentage of satisfied users" in Fig. 6(b). Note that we may need to control the percentage of satisfied users to a pre-determined value for practical systems, which can be accomplished by user access or offloading but are not taken into account in the simulation.

It shows that the proposed user-centric transmission scheme outperforms existing schemes in terms of both throughput and percentage of satisfied users, especially for high QoS

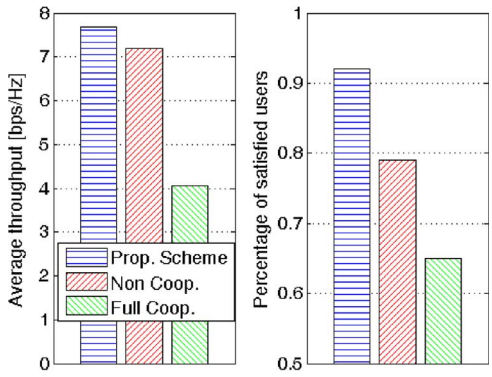


Fig. 7. Throughput and percentage of satisfied users with different transmission schemes, where the users have random QoS requirements, $N_b = 4$.

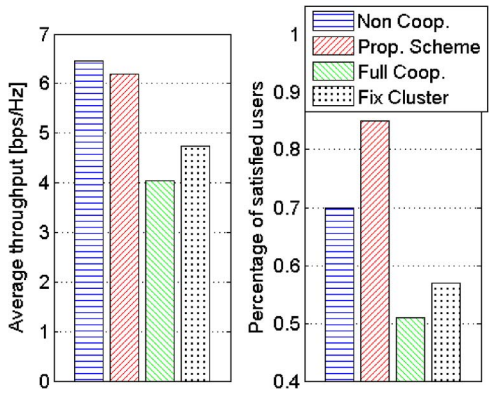


Fig. 8. Throughput and percentage of satisfied users with different transmission schemes, where the users have random QoS requirements, $N_b = 10$.

requirement. With "Full Coop." transmission, because the CU can at most schedule two users in the N_b cells to serve simultaneously, the performance is the worst. Given that all the schemes under comparison employs the same user scheduling method (i.e., Algorithm 2), the performance gain of the proposed scheme comes from the user-centric BS selection, as explained as follows. As implied by the analysis in Section III-A, there is a trade off between the number of scheduled users with orthogonal beamforming vectors and the number of selected cooperative BSs by each user. This suggests that although choosing more cooperative BSs can increase the data rate of each user, it becomes harder to select more users with orthogonal beamforming vectors, which leads to a degradation in the sum rate of the system. An extreme case is the full cooperative scheme, where all available BSs participate in the cooperative transmission. By contrast, choosing less cooperative BSs allows more users to be scheduled simultaneously, which however may not guarantee the QoS of each user. An extreme case is the non-cooperative scheme. By using the user-centric BS selection, the cooperative cluster can be selected with a proper size required by each user, which achieves a balance between the required QoS in data rate of each user and the throughput of the system.

In practical systems different users need different services. In the following we consider a more realistic scenario. Suppose that the users need the real-time service or best effort service with a probability of 50%. For the users with real-time service, the value of T_m is uniformly distributed in a range from 0.5 s to 1 s, and $\bar{D}_m$ is uniformly distributed in a range from 1 Mbits to 10 Mbits. For the best effort users, we still consider the codeword-efficient cooperative BS selection, the cooperative cluster can also be found by solving the optimization problem in (34) where $D_{m,t_{m,i}}$ is set to 0, and the metrics for the service priority in scheduling are set to 0, i.e., $\beta_{m,t}^{\mathrm{EDF}} = 0$ and $\beta_{m,t}^{\mathrm{BIT}} = 0$. The CU schedules both best effort and QoS-guaranteed users at the same time using Algorithm 2. As shown in Fig. 7, the percentage of satisfied users of the proposed scheme is higher than those of the other two schemes, and again the "Full Coop." transmission performs the worst.

To show the impact of the constraint implied by *Proposition 1* on scheduling users, in Fig. 8 we show the results for a denser small cell network, where the simulation setup is the same as Fig. 7 except for the value of N_b. Except for "Non Coop." and "Full Coop.," we also compare with a more advanced BS cooperative scheme with fixed and non-overlapped clusters (with the legend "Fix Cluster"). In this scheme, the BSs in the network are divided into several non-overlapped cooperative clusters. Since the BSs are randomly placed in each simulation trail, we let at most four adjacent BSs to form a cooperative cluster, which does not depends on the QoS requirement and channel condition of each user (the value for the maximal size of fixed clusters will not affect the result of performance comparison). Compared with existing schemes, we can see that the proposed scheme can satisfy the demands of more users without causing significant loss in throughput of the network. Because the number of co-scheduled users using the proposed scheme becomes lesser than that of using "Non Coop." for larger value of N_b, there exists a minor loss in the average sum rate of the proposed scheme. Since the signaling overhead increases with the number of cooperative BSs, the signaling overhead of the proposed scheme is higher than "Non Coop.," and lower than "Full Coop.". With the increasing of QoS requirement, i.e., larger $\bar{D}_m$ and smaller T_m, the signaling overhead of the proposed scheme will also increase. When the sizes of the clusters of the proposed scheme is the same as those of "Fix Cluster," the overhead of our scheme will be slightly higher than the "Fix Cluster" since the clusters are formed semi-dynamically with average channel gains.

V. Conclusion

In this paper we proposed a user-centric downlink BS cooperative transmission scheme with limited feedback, where the per-cell codebook based orthogonal beamforming is employed. We analyzed the sufficient and necessary condition for scheduling users with orthogonal beamforming, and derived a method to estimate the SINR at the user side. We proposed a method for each user to select preferred cooperative cluster based on its average channel gains and its required QoS, and provided a user scheduling method according to the service priority of the users such that the QoS of the users can be satisfied. Simulation results demonstrated that the proposed scheme can increase the percentage of satisfied users significantly.

REFERENCES

[1] F. Boccardi, R. W. Heath Jr., A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–81, Feb. 2014.

[2] C.-L. I *et al.*, "Toward green and soft: A 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[3] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[4] H. Abou-zeid and H. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 92–99, Oct. 2013.

[5] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Commun. Mag.*, vol. 13, no. 4, pp. 56–61, Aug. 2006.

[6] D. Liu *et al.*, "Semi-dynamic cooperative cluster selection for downlink coordinated beamforming systems," in *Proc. IEEE WCNC*, 2014, pp. 1194–1199.

[7] D. J. Love *et al.*, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[8] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 1, pp. 528–541, Mar. 2006.

[9] Y. Cheng, V. K. N. Lau, and Y. Long, "A scalable limited feedback design for network MIMO using per-cell product codebook," *IEEE Trans. Wireless Commun.*, vol. 9, no. 10, pp. 3093–3099, Oct. 2010.

[10] X. Hou and C. Yang, "Codebook design and selection for multi-cell cooperative transmission limited feedback systems," in *Proc. IEEE VTC Spring*, 2011, pp. 1–5.

[11] K. Huang, J. G. Andrews, and R. W. Heath, "Performance of orthogonal beamforming for SDMA with limited feedback," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 152–164, Jan. 2009.

[12] J. Lee and N. Jindal, "Energy-efficient scheduling of delay constrained traffic over fading channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1866–1875, Apr. 2009.

[13] *3rd Generation Partnership Project; Technical specification group radio access network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (Release 11)*, 3GPP TS 36.213 V11.3.0, Jun. 2013. [Online] Available: www.3gpp.org

[14] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[15] J. Kuang, *Applied Inequalities*, 3rd ed. Shandong, China: Shandong Science and Technology Press, 2004.

[16] D. Su and C. Yang, "What should be fed back for per-cell codebook-based limited feedback coordinated multi-point systems?" *EURASIP J. Wireless Commun. Netw.*, vol. 2013, no. 1, p. 232, Sep. 2013:232.

[17] D. Su, X. Hou, and C. Yang, "Quantization based on per-cell codebook in cooperative multi-cell systems," in *Proc. IEEE WCNC*, 2011, pp. 1753–1758.

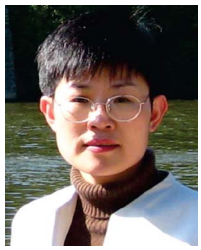
[18] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

[19] 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.

[20] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, 2nd Quart. 2013.

[21] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband*. San Diego, CA, USA: Academic, 2008.

Di Su received the B.S. degree in electronic engineering and the Ph.D. degree in signal and information processing from Beihang University (formerly Beijing University of Aeronautics and Astronautics), Beijing, China, in 2008 and 2014, respectively. She is currently working in the industry with Datang Mobile, Beijing. Her research interests include limited feedback techniques, cooperative communication networks, and signal processing.



Chenyang Yang received the Ph.D. degree in electrical engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics), Beijing, China, in 1997. Since 1999, she has been a Full Professor with the School of Electronics and Information Engineering, Beihang University. 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 the Ministry of Education during 1999–2004. She has authored or coauthored over 200 international journal and conference papers and filed over 60 patents in the fields of green communication, coordinated multipoint transmission, interference management, cognitive radio, and relay. Her recent research interests include green radio and interference control for 5G wireless systems. Dr. Yang was the Chair of the IEEE Communications Society Beijing Chapter during 2008–2012 and the MDC Chair of APB of the IEEE Communications Society during 2011–2013. She has served as a TPC Member for numerous IEEE conferences. She was an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. She is currently a Guest Editor of the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS and an Associate Editor-in-Chief of the *Chinese Journal of Communications* and the *Chinese Journal of Signal Processing*.