User Grouping and Scheduling for Large Scale MIMO Systems with Two-Stage Precoding

Yi Xu[†‡], Guosen Yue[‡], Narayan Prasad[‡], Sampath Rangarajan[‡] and Shiwen Mao[†]

[†]Department of Electrical and Computer Engineering, Auburn University, Auburn, AL

[‡]NEC Laboratories America Inc., Princeton, NJ

Email: yzx0010@auburn.edu, yueguosen@gmail.com, {prasad, sampath}@nec-labs.com, smao@ieee.org

Abstract—In this paper, we consider the design of user grouping and scheduling for large-scale multiple-input multiple-output (MIMO) frequency-division-duplexing (FDD) systems. Based on a recently proposed two-stage precoding framework, we first propose an improved K-means user grouping scheme which allocates the users to different pre-beamforming groups using the second-order channel statistics, and then a user grouping scheme that considers both load balancing and precoding design. After user groups are so determined, we present a dynamic user scheduling scheme where second-stage precoding is designed based on instantaneous channel conditions. We demonstrate the efficacy of the proposed schemes through simulations.

I. INTRODUCTION

Last decades have witnessed the ever-increasing demand for higher data rates. To cater for this demand, many advanced physical layer techniques have been developed, e.g., multiple input multiple output with orthogonal frequency division multiplexing (MIMO-OFDM). However, with linear throughput improvement but exponential growth in the data traffic, the gap between the demand and supply has never become smaller. To solve the issue, the next technology we can resort to is large-scale MIMO, or the so called massive MIMO (we will use these terms interchangeably hereafter) which significantly increases the system throughput by employing large number of transmit antennas. As an emerging but promising technology, besides the throughput enhancement, large-scale MIMO also enjoys advantages of low-power, robust transmissions, simplified transceiver, and simple multiple-access layer [1], [2]. Recently, a lab demo system has demonstrated the benefits of the large-scale MIMO systems [3], [4].

In general, being equipped with more transmit antennas, the large-scale MIMO system can provide more degrees of freedom resulting in better reliability or higher throughput. However, due to the difficulties of acquiring channel state information at the transmitter side, it is still a challenge to simultaneously support a large number of users [2]. Therefore, most existing works on large scale MIMO systems consider time-division-duplexing (TDD). By exploiting the channel reciprocity in the TDD systems, the downlink channel can be estimated from uplink training. However, the frequency-division-duplexing (FDD) system does not have such privilege. Pilot based channel estimation and uplink channel feedback are required, which consume lots of spectrum resources. With many FDD deployments worldwide, it is important to investigate the large-scale MIMO design for FDD systems. Recently, a two-stage precoding scheme has been proposed in [5] to reduce pilot resources and the channel state information (CSI)

feedback in FDD systems. Firstly, the users in service are put into groups with each group of users having similar second-order channel statistics, i.e., transmit correlation. The same pre-beamforming, or the first-stage precoding is then used for each group of users semi-statically. Then, with reduced dimensions on the effective channel, simple channel feedback can be realized, thus the second-stage dyanmic precoding can be applied. Therefore, one important issue for such system design is user grouping. Some user grouping methods are recently presented in [6] including a K-means clustering based approach. In this paper, based on the same two-stage precoding framework, we consider the design of user grouping and scheduling for the large-scale MIMO systems. We propose an improved K-means user grouping scheme which results in an improved user sum-throughput. When the number of users is not too large, to achieve user fairness, we propose a joint design of user grouping with load balancing and precoding. After user groups are determined, we then present a dynamic user scheduling scheme with second-stage precoding designed based on the dynamic channel condition. We demonstrate the efficacy of the proposed schemes via numerical simulations

II. SYSTEM MODEL AND PRELIMINARIES

We consider a downlink system with M antennas at base station (BS) and each user terminal (UT) equipped with single antenna. The transmit antennas can have different geometries, e.g., placed along one axis to form uniform linear array (ULA), along a circle to form circular array, or in two or even more dimensions. The signals received by all users $\mathbf{y}$ are given as

$$\mathbf{y} = \mathbf{H}^H \mathbf{V} \mathbf{d} + \mathbf{z}, \qquad (1)$$

where $(\cdot)^H$ denotes the Hermitian of a matrix; $\mathbf{H}$, of dimension $M \times K$, is the actual channel between the BS and the users; $\mathbf{V}$ is the precoding matrix of dimension $M \times S$; $\mathbf{d}$ is the data vector of dimension S; and $\mathbf{z}$ is the zero mean circulant symmetric complex Gaussian noise vector. Note that throughout this paper, we use bold upper (low) case letter to denote a matrix (vector), and normal letter to denote a scalar.

Based on the two-stage precoding approach in [5], the precoding is formed as multiplication of two precoding matrices, i.e., $\mathbf{V} = \mathbf{B}\mathbf{P}$. The first part $\mathbf{B}$ of dimension $M \times b$ is pre-beamforming matrix, which is designed based on the second order channel statistics, particularly, the transmit spatial correlation. The same pre-beamforming matrix is semi-statically applied to the users with the same or similar transmit correlation, which forms a user group. Therefore, pre-beamforming matrix is designed to suppress the interferences
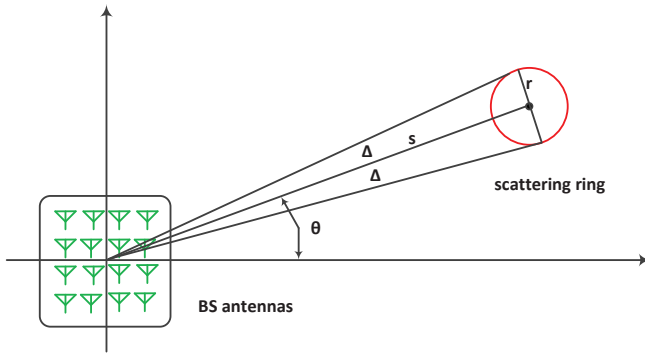
Fig. 1.   One ring channel model.

among the groups. We can see that the effective transmit size after the pre-beamforming is b, which is determined by dominant eigenmodes of the average transmit correlation of user groups. The second part $\mathbf{P}$ of dimension $b \times S$, is designed to suppress the interferences within each group with dynamical channel condition. Note that we have $S \leq b$ as the second-stage precoding is supposed to suppress the interference within the group.

Let $\tilde{\mathbf{H}} = \mathbf{B}^H\mathbf{H}$, ($\tilde{\mathbf{H}}$ is called the effective channel after pre-beamforming). The signal model can be rewritten as

$$\mathbf{y} = \mathbf{H}^H\mathbf{BPd} + \mathbf{z} = \tilde{\mathbf{H}}^H\mathbf{Pd} + \mathbf{z}. \quad (2)$$

We adopt the one-ring channel model in [5] as shown in Fig. 1, in which θ is the azimuth angle of the user location, s is the distance between the BS and the user, r is the radius of the scattering ring, and Δ is the angle spread, which can be approximated as

$$\Delta \approx \arctan(r/s). \quad (3)$$

Denote $\mathbf{R}$ as the channel covariance matrix of the transmitter of which the (m,p)-th entry is given by

$$[\mathbf{R}]_{m,p} = \frac{1}{2\Delta} \int_{-\Delta}^{\Delta} e^{j\mathbf{k}^T(\alpha+\theta)(\mathbf{u}_m - \mathbf{u}_p)} d\alpha, \quad (4)$$

where $\mathbf{k}(\alpha) = -\frac{2\pi}{\lambda}(\cos(\alpha), \sin(\alpha))^T$ is the vector for a planar wave impinging with Angle of Arrival (AoA) α, λ is the carrier wavelength, $\mathbf{u}_m$ and $\mathbf{u}_p$ are the position vectors of antenna m, p, and $(\cdot)^T$ denotes the transpose operation. It can be verified that $\mathbf{R}$ is a normal matrix. With eigen-decomposition, we have

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H, \quad (5)$$

where $\mathbf{U}$ is a unitary matrix comprising eigenvectors of $\mathbf{R}$ and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues of $\mathbf{R}$ as the diagonal entries. Furthermore, the actual channel is generated using the following model

$$\mathbf{h} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{w}, \quad (6)$$

where $\mathbf{w}$ is a vector of random numbers.

Let G denotes the number of groups, $\mathbf{H}_g = [\mathbf{h}_{g_1}, \ldots, \mathbf{h}_{g_{K_g}}]$, $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_G]$, $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_G]$, $\mathbf{P} = \mathrm{diag}\{\mathbf{P}_1, \ldots, \mathbf{P}_G\}$, and $\tilde{\mathbf{H}}_g = \mathbf{B}_g^H\mathbf{H}_g$. The signal vector received by the g-th group of users is represented as

$$\mathbf{y}_g = \tilde{\mathbf{H}}_g^H\mathbf{P}_g\mathbf{d}_g + \sum_{g' \neq g}\mathbf{H}_g^H\mathbf{B}_{g'}\mathbf{P}_{g'}\mathbf{d}_{g'} + \mathbf{z}_g, g = 1, \ldots, G. \quad (7)$$

From (7) we can see that the design of $\mathbf{B}_g$ is to achieve

$$\mathbf{H}_g^H\mathbf{B}_{g'} \approx 0, \forall g' \neq g. \quad (8)$$

Three different approaches are introduced in [5], which are briefly recapped as follows.

1) *Eigen-beamforming*: If the locations of all group members in a group are close, we assume they have similar transmit correlations. We can design $\mathbf{B}_g$ as

$$\mathbf{B}_g = \mathbf{V}_g, \quad (9)$$

where $\mathbf{V}_g$ is the unitary matrix after eigen-decomposing $\mathbf{R}_g$, which is the average of the transmit correlation matrices of users within the group g and can be viewed as the group center.

2) *Approximate Block Diagonalization (BD)*: Firstly, find the group center for all groups $\{\mathbf{R}_g\}$. Then we form

$$\mathbf{\Xi}_g = [\mathbf{V}_1, \ldots, \mathbf{V}_{g-1}, \mathbf{V}_{g+1}, \ldots, \mathbf{V}_G]. \quad (10)$$

Perform Singular Value Decompostion (SVD) of $\mathbf{\Xi}_g$ to obtain $[\mathbf{E}_g^{(1)}, \mathbf{E}_g^{(0)}]$, such that $\mathrm{Span}(\mathbf{E}_g^{(0)}) = \mathrm{Span}^\perp(\{\mathbf{V}_{g'} : g' \neq g\})$. Find

$$\hat{\mathbf{R}}_g = (\mathbf{E}_g^{(0)})^H\mathbf{U}_g\mathbf{\Lambda}_g\mathbf{U}_g^H\mathbf{E}_g^{(0)}. \quad (11)$$

And perform SVD to $\hat{\mathbf{R}}_g$, such that

$$\hat{\mathbf{R}}_g = \mathbf{G}_g\mathbf{\Phi}_g\mathbf{G}_g^H. \quad (12)$$

Let $\mathbf{G}_g = [\mathbf{G}_g^{(1)}, \mathbf{G}_g^{(0)}]$, where $\mathbf{G}_g^{(1)}$ contains b_g dominant eigenmodes of $\hat{\mathbf{R}}_g$. Finally, $\mathbf{B}_g$ can be obtained as

$$\mathbf{B}_g = \mathbf{E}_g^{(0)}\mathbf{G}_g^{(1)}. \quad (13)$$

3) *DFT Matrix Approximation*: For large scale MIMO systems with ULA antennas, we have

$$\lim_{M \to \infty} \frac{1}{M}\left\|\mathbf{U}\mathbf{U}^H - \mathbf{F}_S\mathbf{F}_S^H\right\|^2 = 0, \quad (14)$$

where $\mathbf{F}_S$ is a submatrix of unitary DFT matrix whose (a,b)-th entry is given by $[\mathbf{F}]_{a,b} = \frac{e^{-j2\pi ab/M}}{\sqrt{M}}$. Thus, we can select certain columns of a DFT matrix to approximate the prebeamforming matrix.

For the second stage precoding, we apply conventional zero-forcing beamforming (ZFBF) or regularized zero-forcing beamforming (RZFBF). The precoding matrix is given by

$$\mathbf{P}_g = \tilde{\mathbf{H}}_g(\tilde{\mathbf{H}}_g^H\tilde{\mathbf{H}}_g + S_g\alpha\mathbf{I}_{S_g})^{-1}, \quad (15)$$

where α can be set as $\alpha = 0$ for ZFBF or $\alpha = \frac{\sum_g S_g}{p\sum_g b_g}$ for RZFBF, S_g is number of data streams in the g-th group and p is the total transmit power of the BS.

III.   User Grouping in Massive MIMO System

As aforementioned, the same prebeamforming matrix is formed for a user group, which is designed based on the group centers $\mathbf{R}_g, g = 1, \ldots, G$. Therefore it is important to design an effective user grouping method. User grouping also has effect on the user scheduling, since for each pre-beamforming group, only the users within its group can be scheduled.

In [6], a K-means clustering algorithm for user grouping is proposed. The metric of the K-means clustering algorithm to

group users is the chordal distance between the eigenvectors $\mathbf{U}_k$ of user's channel correlation $\mathbf{R}_k$ and that of the group center, given as

$$d_c(\mathbf{U}_k, \mathbf{V}_g) = \left\| \mathbf{U}_k \mathbf{U}_k^H - \mathbf{V}_g \mathbf{V}_g^H \right\|_F^2, \qquad (16)$$

where $\mathbf{U}_k$ is obtained by decomposing user k's covariance matrix $\mathbf{R}_k$ using $\mathbf{R}_k = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{U}_k^H$. User grouping is then formed via an iterative process. In each iteration, each user is assigned to the group with minimum distance. Then group center is updated with user associated to the group as

$$\mathbf{V}_g = \Upsilon \left\{ \frac{1}{|\mathcal{S}_g|} \sum_{k \in \mathcal{S}_g} \mathbf{U}_k \mathbf{U}_k^H \right\}. \qquad (17)$$

Note that $\Upsilon(\cdot)$ denotes the unitary matrix after eigendecompostion, $\mathcal{S}_g$ denotes the user set of group g, $|\mathcal{S}_g|$ denotes the size of group g.

We now propose an improved K-means clustering algorithm, which is described in Algorithm 1. Note that in Algorithm 1, $\mathbf{U}_{\pi(g)}$ is the unitary matrix of the user with index $\pi(g)$. The key idea is as follows. Instead of chordal distance, we propose to use the weighted likelihood as the grouping metric defined as

$$L(\mathbf{R}_k, \mathbf{V}_g) = \left\| (\mathbf{U}_k \mathbf{\Lambda}_k^{\frac{1}{2}})^H \mathbf{V}_g \right\|_F^2. \qquad (18)$$

The proposed likelihood metric uses the projection of the eigenspaces of the users to that of the group centers, so that users can be readily separated into different groups. It also takes into account the weight of different eigenmodes so that the user's group is mainly determined by the dominant eigenmodes. After obtaining $L(\mathbf{R}_k, \mathbf{V}_g)$ for each user and each group, assign each user to the group with maximum likelihood and update group center through

$$\mathbf{V}_g = \Upsilon \left\{ \frac{1}{|S_g|} \sum_{k \in S_g} \mathbf{R}_k \right\}. \qquad (19)$$

Notice that the updates of the group center and the total likelihood L_{tot} also consider the weights of eigenmodes in the proposed algorithm.

IV. User Scheduling in Massive MIMO System

Given the user grouping, we can find the prebeamforming matrix $\mathbf{B}_g, \forall g$. At a particular time slot, based on the instantaneous channel conditions of the users, we dynamically schedule a subset of users in each group for the transmissions in this time slot.

In [6], a MAX user scheduling algorithm is proposed. The key idea is to appoximate the user's signal to interference plus noise ratio (SINR) without knowing the exact second stage precoding matrix, and schedule the user with maximal SINR.

Different from this approach, we propose a dynamic user scheduling algorithm which schedules users in a greedy manner. In particular, at each step, the proposed algorithm only schedules the user with potentially maximum enhancement to the system throughput among all the unscheduled users. The proposed method is summarized in Algorithm 2.

Algorithm 1: Improved K-means Clustering Algorithm

1 Set $n = 0$, $L_{tot}^{(0)} = 1$, $L_{tot}^{(n)} = 0, \forall n > 0$ and $\mathcal{S}_g^{(0)} = \emptyset$ for $g = 1, \ldots, G$;

2 Randomly choose G different indices (denoted as $\pi(g), \forall g$) from the set $\{1, \ldots, K\}$ and set $\mathbf{V}_g^{(0)} = \mathbf{U}_{\pi(g)}, \forall g$;

3 **while** $\left| L_{tot}^{(n)} - L_{tot}^{(n-1)} \right| > \epsilon L_{tot}^{(n-1)}$ **do**

4 $\quad$ $n = n + 1$;

5 $\quad$ Let $\mathcal{S}_g^{(n)} = \emptyset$, $g = 1, \cdots, G$;

6 $\quad$ **for** $k = 1, \ldots, K$ **do**

7 $\quad\quad$ **for** $g = 1, \ldots, G$ **do**

8 $\quad\quad\quad$ Compute

$$L(\mathbf{R}_k, \mathbf{V}_g^{(n-1)}) = \left\| (\mathbf{U}_k \mathbf{\Lambda}_k^{\frac{1}{2}})^H \mathbf{V}_g^{(n-1)} \right\|_F^2$$

9 $\quad\quad$ **end**

10 $\quad\quad$ Find $g_k^* = \arg\max_{g'} L(\mathbf{R}_k, \mathbf{V}_{g'}^{(n-1)})$ and let $\mathcal{S}_{g_k^*}^{(n)} = \mathcal{S}_{g_k^*}^{(n)} \cup \{k\}$;

11 $\quad$ **end**

12 $\quad$ **for** $g=1, \ldots, G$ **do**

13 $\quad\quad$ $\mathbf{V}_g^{(n)} = \Upsilon \left\{ \frac{1}{|\mathcal{S}_g^{(n)}|} \sum_{k \in \mathcal{S}_g^{(n)}} \mathbf{R}_k \right\}$;

14 $\quad$ **end**

15 $\quad$ Compute $L_{tot}^{(n)} = \sum_{g=1}^{G} \sum_{k \in \mathcal{S}_g^{(n)}} L(\mathbf{R}_k, \mathbf{V}_g^{(n)})$

16 **end**

17 Assign $\mathbf{V}_g = \mathbf{V}_g^{(n)}$ and $\mathcal{S}_g = \mathcal{S}_g^{(n)}$.

Given the user grouping and scheduling, we can calculate the instantaneous SINR denoted as γ_{g_k} for user k in group g as

$$\gamma_{g_k} = \frac{\frac{p}{\sum_g S_g} \zeta_g^2 \left| \mathbf{h}_{g_k}^H \mathbf{B}_g \mathbf{P}_g(:,g_k) \right|^2}{1 + I_{in}(g,k) + I_{it}(g,k)},$$

$$I_{in}(g,k) = \frac{p}{\sum_g S_g} \zeta_g^2 \sum_{j \neq k} \left| \mathbf{h}_{g_k}^H \mathbf{B}_g \mathbf{P}_g(:,g_j) \right|^2,$$

$$I_{it}(g,k) = \frac{p}{\sum_g S_g} \sum_{g' \neq g} \zeta_{g'}^2 \sum_j \left| \mathbf{h}_{g_k}^H \mathbf{B}_{g'} \mathbf{P}_{g'}(:,g_j) \right|^2 \qquad (20)$$

where I_{in} denotes the intra group interferences, I_{it} denotes the inter group interferences, $\mathbf{P}_g(:,g_k)$ denotes the submatrix containing all the rows and the g_k-th column of $\mathbf{P}_g$ and ζ_g^2 is the scaling factor for satisfying certain power constraint, which can be obtained as

$$\zeta_g^2 = \frac{S_g}{\text{tr}(\mathbf{P}_g^H \mathbf{B}_g^H \mathbf{B}_g \mathbf{P}_g)}. \qquad (21)$$

Then the rate for the scheduled user g_k is given by

$$\eta_{g_k} = \log_2(1 + \gamma_{g_k}). \qquad (22)$$

System throughput r_{ws} is obtained as $r_{ws} = \sum_{g=1}^{G} \sum_{k \in \mathcal{K}_g} \eta_{g_k}$, where $\mathcal{K}_g$ is the scheduled user set in the gth group. Obviously, r_{ws} is a function of $\{\mathcal{K}_g\}$ and precoding for all co-scheduled users, denoted as $r_{ws}(\{\mathcal{K}_g\}, \{\mathbf{B}_g\}, \{\mathbf{P}_{g_k}\})$.

V. User Grouping with Joint Group Load Balancing and Precoding Design

When the number of users is not too large, with the previously discussed grouping approaches, some groups may

Algorithm 2: Greedy algorithm for dynamic user selection and beamforming with determined user grouping

---

1 User grouping $\{\mathcal{S}_g\}$ is given;
2 Initially set $\mathcal{U} = \{1, \ldots, K\}$, the weighted sum rate $r_{ws} = 0$ and $\mathcal{K}_g = \emptyset$ for $g = 1, \ldots, G$;
3 **while** *Termination conditions ($\sum_g |\mathcal{K}_g| = \sum_g b_g$, $\kappa(k^*, g_{k^*}) = 0$, or $\mathcal{U} = \emptyset$) are not satisfied* **do**
4   **for** $k \in \mathcal{U}$ **do**
5   Set $\mathcal{K}'_g = \mathcal{K}_g \cup \{k\}$ if $k \in \mathcal{S}_g$, and $\mathcal{K}'_{g'} = \mathcal{K}_{g'}, \forall g' \neq g$;
6   Perform ZFBF or RZFBF based on $\{\mathcal{K}'_g\}$ and $\{\mathbf{B}_g\}$;
7   Compute the gain $\kappa(k, g) = \max\{0, r_{ws}(\{\mathcal{K}'_g\}, \{\mathbf{B}_g\}) - r_{ws}(\{\mathcal{K}_g\}, \{\mathbf{B}_g\})\}$;
8   **end**
9   Obtain $(k^*, g_{k^*}) = \arg\max_{k \in \mathcal{U}} \kappa(k, g)$;
10   **if** $(k^*, g_{k^*}) > 0$ **then**
11   Let $\mathcal{U} \leftarrow \mathcal{U} \backslash k^*$;
12   Let $\mathcal{K}_{g_{k^*}} \leftarrow \mathcal{K}_{g_{k^*}} \cup \{k^*\}$;
13   **end**
14 **end**

---

end up with few users while some others may be so crowded with users which clearly affects the user fairness. We now form a user grouping method with joint group load balancing and precoding design considering proportional fairness. The user grouping with proportional fairness can be summarized as the following optimization problem.

$$\max_{x_{kg}} \quad \mathcal{J} = \sum_k \sum_g x_{kg} \log\left(\frac{\overline{\eta}_{g_k}}{\sum_i x_{ig}}\right)$$
$$s.t. \quad \sum_g x_{kg} = 1, \quad \forall\, k, \tag{23}$$

where $\mathcal{J}$ denotes the utility, $\overline{\eta}_{g_k}$ is the average user throughput, i.e., $\overline{\eta}_{g_k} = \log_2(1 + \overline{\gamma}_{g_k})$, $\overline{\gamma}_{g_k}$ is the average SINR when user k is assigned to group g, and x_{kg} is assignment indicator defined as

$$x_{kg} = \begin{cases} 1, & \text{if user } k \text{ is in group } g \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

We can see that one of the important issues is to obtain the average SINR $\overline{\gamma}_{g_k}$. However, without user grouping and scheduling information, we cannot calculate the exact SINR of each user. Moreover, over different time slots, different users will be scheduled based on the user grouping and the instantaneous channels. Thus, in order to solve (23), we need to find a way to approximate average SINR for each user in each group. Here we propose two methods. One is through the deterministic approximation for large-scale MIMO systems as shown in Algorithm 4 with following assumptions.

(i)   For the target user k assigned to group g, the transmit correlation is $\mathbf{R}_{g_k} = \mathbf{R}_k$;

(ii)   For intra-group co-scheduled users, their transmit correlations are same the group average, i.e., $\mathbf{R}_{g_{k'}} = \mathbf{R}_g$;

(iii)   For inter-group co-scheduled users, their transmit correlations are same as their group averages $\mathbf{R}_{g'_k} = \mathbf{R}_{g'}$.

Let $\bar{\mathbf{R}}_{g_k} = \mathbf{B}_g^H \mathbf{R}_{g_k} \mathbf{B}_g$. Based on above approximations, following [6], we obtain the large system approximation of

average SINR as

$$\overline{\gamma}_{g_k} = \frac{\frac{p}{\sum_g K_g} \zeta_g^2}{1 + \sum_{g'=1, g' \neq g}^{G} \frac{p}{\sum_g K_g} \varrho_{g', g_k} \zeta_{g'}^2}, \tag{25}$$

where

$$\varrho_{g', g_k} = \frac{1}{b_g} \sum_{l=1}^{K_{g'}} \frac{n_{g_k, g'_l}}{(m_{g'_l})^2}$$

$$\mathbf{n}'_{g_k} = (\mathbf{I}_{K_{g'}} - \mathbf{J}_{g'})^{-1} \mathbf{v}'_{g_k}$$

$$\{\mathbf{J}_g\}_{i,j} = \frac{\mathrm{tr}(\bar{\mathbf{R}}_{g_i} \mathbf{T}_g \bar{\mathbf{R}}_{g_j} \mathbf{T}_g)}{N b_g^2 (m_{g_k})^2}$$

$$\{\mathbf{v}_{g_k}\}_i = \frac{1}{N b_g} \mathrm{tr}(\bar{\mathbf{R}}_{g'_i} \mathbf{T}_{g'} \mathbf{B}_{g'}^H \mathbf{R}_{g_k} \mathbf{B}_{g'} \mathbf{T}_{g'})$$

$$\zeta_g^2 = \frac{S_g}{\Gamma_g}$$

$$\Gamma_g = \frac{1}{b_g} \sum_{k=1}^{K_g} \frac{q_{g,k}}{(m_{g,k})^2}$$

$$\mathbf{q}_g = (\mathbf{I}_{K_g} - \mathbf{J}_g)^{-1} \mathbf{v}_g$$

$$\{\mathbf{v}_g\}_i = \frac{1}{N b_g} \mathrm{tr}(\bar{\mathbf{R}}_{g_i} \mathbf{T}_g \mathbf{B}_g^H \mathbf{B}_g \mathbf{T}_g) \tag{26}$$

and the deterministic approximations of the ZFBF precoding m_{g_k} for target user, $\overline{m}_g$ for intra-group users, and $m_{g'}$ for inter group user are given, respectively, as

$$m_{g_k} = \frac{1}{N b_g} tr(\overline{\mathbf{R}}_{g_k} \mathbf{T}_g), \quad \overline{m}_g = \frac{1}{N b_g} tr(\overline{\mathbf{Y}}_g \mathbf{T}_g)$$

$$\mathbf{T}_g = \left(\mathbf{I}_{b_g N} + \frac{1}{b_g} \frac{\overline{\mathbf{R}}_{g_k}}{m_{g_k}} + \frac{K_g - 1}{b_g} \frac{\overline{\mathbf{Y}}_g}{\overline{m}_g}\right)^{-1} \tag{27}$$

$$m_{g'} = \frac{1}{N b_{g'}} tr(\overline{\mathbf{Y}}_{g'} \mathbf{T}_{g'})$$

$$\mathbf{T}_{g'} = \left(\mathbf{I}_{b_{g'} N} + \frac{K_{g'}}{b_{g'}} \frac{\overline{\mathbf{Y}}_{g'}}{m_{g'}}\right)^{-1} \tag{28}$$

We also present another approach, which is a simplified version of the above method. We can obtain the average SINR approximation based on following assumptions.

(i)   We assume conjugate precoding for the target user.

(ii)   There are no intra-group co-scheduled users.

(iii)   Identity precoding for inter-group co-scheduled users is assumed.

We can obtain the asymptotic SINR approximation as

$$\overline{\gamma}_{g_k} = \frac{\frac{p}{G} \left| tr(\mathbf{B}_g^H \mathbf{R}_{g_k} \mathbf{B}_g) \right|}{1 + \frac{(G-1)p}{G \sum_{g' \neq g} b_{g'}} \sum_{g' \neq g} \left| tr(\mathbf{B}_{g'}^H \mathbf{R}_{g_k} \mathbf{B}_{g'}) \right|}. \tag{29}$$

After obtaining average SINR, following [7], the procedures to solve the user grouping problem in (23) are presented in Algorithm 5. We note here that recently developed alternate approaches [8] can also be used to solve (23).

Algorithm 3: User grouping with joint group load balancing and precoding design algorithm

---

1 Perform K-means Clustering Algorithm or Algorithm 1 to obtain user group ID x_{ij}.
2 **while** $\mathcal{J}^{*(n-1)} - \mathcal{J}^{*(n-2)} > \epsilon \mathcal{J}^{*(n-2)}$ **do**
3   **for** $g \in G$ **do**
4   Find $\mathbf{V}_g^{*(n)}$ using (17) or the proposed weighted likelihood (19)
5   **end**
6   **for** $g \in G$ **do**
7   Find $\mathbf{B}_g$ using approximate BD approach
8   **end**
9   **for** $k = 1, \ldots, K$ **do**
10   **for** $g = 1, \ldots, G$ **do**
11   Find γ_{g_k} using Algorithm 4 or (29)
12   **end**
13   **end**
14   Optimize (23) using Algorithm 5 ;
15   Update x_{ij} and $\mathcal{J}^{*(n)}$;
16 **end**

Algorithm 4: Large Scale SINR Approximation

---

1 **for** $g = 1, \ldots, G$ **do**
2 $\mathbf{Y}_g = \frac{1}{|\mathcal{S}_g|} \sum_{k \in \mathcal{S}_g} \mathbf{U}_k \mathbf{U}_k^H$,
3 $\overline{\mathbf{Y}}_g = \mathbf{B}_g^H \mathbf{Y}_g \mathbf{B}_g$;
4   **while** $m_g^{(n-1)} - m_g^{(n-2)} > \epsilon m_g^{(n-2)}$ **do**
5   Solve (28) iteratively to find m_g and $\mathbf{T}_g$.
6   **end**
7 **end**
8 **for** $k = 1, \ldots, K$ **do**
9   **for** $g = 1, \ldots, G$ **do**
10   Solve (27) iteratively to find m_{g_k} and updated $\mathbf{T}_g$ for target user k.
11   Solve (26) and (25) with assumptions: (i) For target user, the transmit correlation is R_{g_k}; (ii)For intra group co-scheduled users, the transmit correlation is same as the average of the group;(iii)For inter group co-scheduled users, the transmit correlation is same as the average of the group.
12   **end**
13 **end**

VI. SIMULATION

We now perform numerical simulations to evaluate the proposed schemes. System configurations are provided in Table I. In particular, we consider a $120°$ sector. For each user drop, the azimuth angle θ_k, angle spread Δ_k and distance s_k for user k are randomly generated within the interval $[\theta_{\min}, \theta_{\max}]$, $[\Delta_{\min}, \Delta_{\max}]$ and $[s_{\min}, s_{\max}]$, respectively. We average over 100 user drops for the whole simulation, and in each user drop, we evaluate the performance with 200 channel realizations. We fix the number of groups $G = 8$. For the antenna configuration, we consider the ULA case and place 8, 16 or 40 antennas along the y-axis with spacing 0.5λ. So from (4), the (m, p)-th entry of the covariance matrix is given by

$$[\mathbf{R}]_{m,p} = \frac{1}{2\Delta} \int_{-\Delta}^{\Delta} e^{-j2\pi D(m-p)\sin(\alpha+\theta)} d\alpha. \tag{30}$$

In the simulation, to find the first and second stage precoding matrices, we adopt the approximate BD approach and the

---

Algorithm 5: Optimization Algorithm for (23)

---

1 $n = 0$, $\mu^{(1)} = 0$;
2 **while** *the optimization has not converged* **do**
3 $n \leftarrow n + 1$;
4   **for** $k = 1, \ldots, K$ **do**
5   **for** $g = 1, \ldots, G$ **do**
6   Compute $\overline{\gamma}_{g_k}$ and $\overline{\eta}_{g_k}$;
7   **end**
8   Assign user k to group g^* where $g^* = \arg\max_g (\log(\overline{\eta}_{g_k}) - \mu_g^{(n)})$, and let $x_{kg^*}^{(n)} = 1$, $x_{kg}^{(n)} = 0$ for $g \neq g^*$;
9   **end**
10   **for** $g = 1, \ldots, G$ **do**
11   Each group chooses a step size $\delta^{(n)}$ and computes $K_g^{(n+1)} = \min\{K, e^{(\mu_g^{(n)}-1)}\}$, $\mu_g^{(n+1)} = \mu_g^{(n)} - \delta^{(n)}(K_g^{(n)} - \sum_k x_{kg}^{(n)})$.
**12**   **end**
13 **end**

TABLE I.   SYSTEM CONFIGURATION IN THE SIMULATIONS

Parameter	Value	Parameter	Value
θ_{min}	$-60°$	M	8, 16, 40
θ_{max}	$60°$	D	0.5
Δ_{min}	$5°$	G	8
Δ_{max}	$15°$	r_g^*	M/G
s_{min}	20 (m)	ϵ	10^{-3}
s_{max}	120 (m)	p	10, 20 dB

regularized ZF precoding approach, respectively.

The sum rates for K-means and proposed improved grouping methods are shown Fig. 2 with MAX scheduling methods. We set $p = 10$ dB. As we can see that the proposed improved K-means clustering method provides higher throughput than the K-means clustering method for both $M = 16$ and $M = 40$ cases. It can also be seen that, with more users in the system, the throughput gain for the proposed method becomes larger. So it is beneficial to consider the weights of different eigenmodes. The sum-rates with greedy dynamic scheduling are shown in Fig. 3. We can see that with dynamic scheduling, the proposed improved K-means method still outperforms the existing method. If we cross-check Fig. 2 and Fig. 3 we can see that the proposed dynamic user scheduling algorithm further enhances the system throughput.

For user grouping with joint load balancing and precoding design scheme, due to space limitation, we only present the results using (29) for approximating the SINR. Here we set $p = 20$ dB. Figure 4 shows the resulting utility metric of the optimization in (23) which is solved by Algorithm 5. We can see that the proposed scheme greatly enhances the utility with user fairness compared the scheme without considering load balancing. Note that negative values of utility are resulted from the $\log$ function of achievable rate over the number of group members. Figure 5 shows that the average number of group members among groups. We can see that difference between the most loaded group and the least loaded group is only about 0.6 for the proposed scheme, but 3 for the scheme without considering load balancing. Therefore, the proposed scheme strikes a much better balance as the users are somehow evenly distributed among different groups.
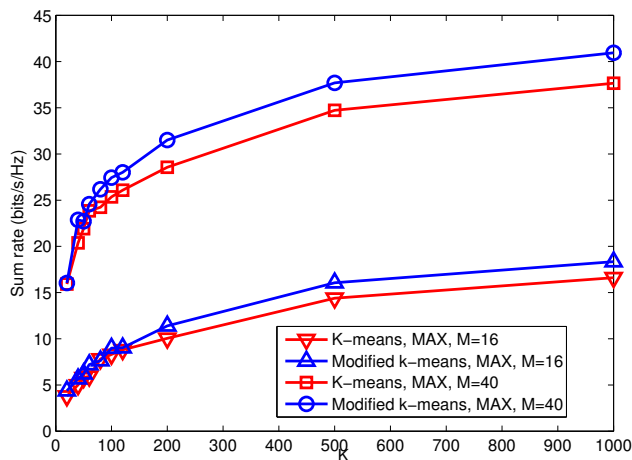
Fig. 2. System sum rate Vs. number of users when $M = 16$ or $M = 40$.
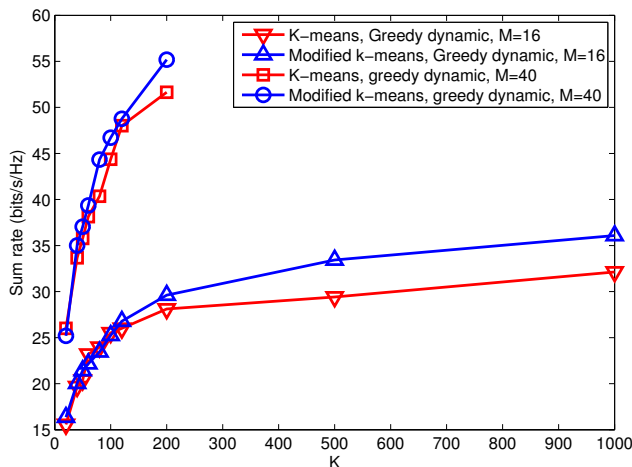
Fig. 3. System sum rate Vs. number of users when $M = 16$ or $M = 40$.
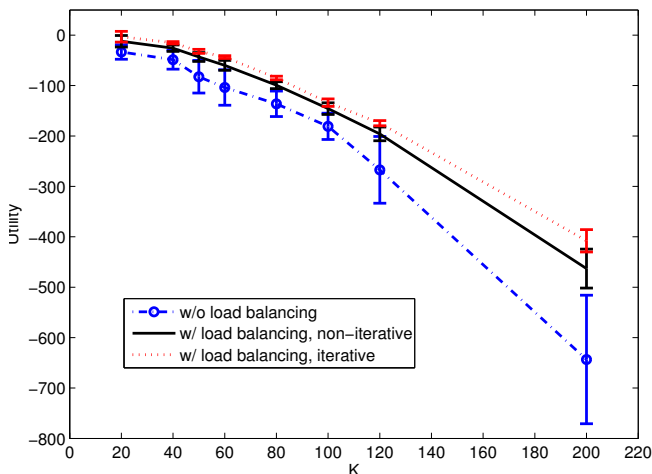
Fig. 4. Utility Vs. number of users, for User grouping with joint group load balancing and precoding design when $M = 40$.

VII. CONCLUSIONS

In this paper, based on a two-stage precoding framework for large-scale MIMO systems with FDD, we have proposed
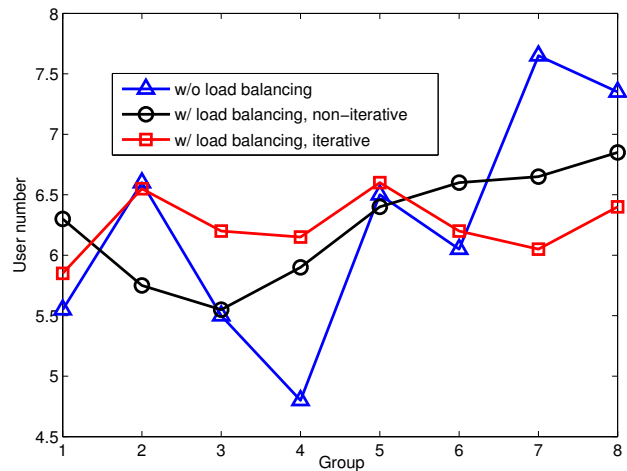
Fig. 5. Number of group members Vs. Group, for User grouping with joint group load balancing and precoding design when $M = 40$.

an improved K-means clustering user grouping scheme which provides a large sum-throughput gain. To achieve user fairness, we have proposed a user grouping method with joint loading balancing and precoding design. Moreover, we have also presented dynamic user scheduling scheme after the user groups being determined. With numerical simulations, we have demonstrated that the proposed schemes provide better performances than the existing schemes in the literature.

REFERENCES

[1] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for Next Generation Wireless Systems", available online: http://arxiv.org/abs/1304.6690

[2] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling Up MIMO Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, iss. 1, pp. 40–60, Jan. 2013.

[3] C. Shepard, H. Yu, N. Anand, L. E. Li, T. L. Marzetta, R. Yang, and L. Zhong, "Argos: Practical Many-Antenna Base Stations", in Proc. *ACM MobiCom'12*, Istanbul, Turkey, pp. 53-64, Aug. 2012.

[4] C. Shepard, H. Yu, and L. Zhong, "ArgosV2: A Flexible Many-Antenna Research Platform," to appear in *ACM Mobicom'13*, Miami, FL, Sep. 2013.

[5] Ansuman Adhikary, Junyoung Nam, Jae-Young Ahn, and Giuseppe Caire, "Joint Spatial Division and Multiplexing-The Large-Scale Array Regime," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[6] Ansuman Adhikary, Giuseppe Caire, "Joint Spatial Division and Multiplexing: Opportunistic Beamforming and User Grouping", available online: http://arxiv.org/abs/1305.7252

[7] Qiaoyang Ye, Beiyu Rong, Yudong Chen, Mazin Al-Shalash, Constantine Caramanis, and Jeffrey G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, iss. 6, pp. 2706–2716, Jun. 2013.

[8] N. Prasad, M. Arslan, and S. Rangarajan, "Exploiting Cell dormancy and Load Balancing in LTE HetNets: Optimizing the Proportional Fairness Utility," *IEEE ICC*, Jun. 2014.