Cluster-by-Cluster Block Diagonal Digital Precoding for Multi-User MIMO Downlink Transmissions

Guanchong Niu, Qi Cao, Man-On Pun[‡]
The Chinese University of Hong Kong, Shenzhen
Guangdong, China, 518172

Abstract—Beam division multiple access (BDMA) has recently been proposed for massive multiple-input multiple-output (MIMO) systems by simultaneously transmitting multiple users' data streams via different beams. Meanwhile, the block hybrid precoding has been proposed to reduce the computational complexity. However, previous works mostly rely on a crucial condition that the number of RF chains must be not less than the number of data streams. In this paper, we propose a multipath BDMA based hybrid block precoding system to break the ceiling on minimal required RF chains where the data streams are processed cluster-by-cluster. To overcome the performance degradation arisen from block hybrid precoding scheme, a K-means based heuristic user clustering algorithm is proposed to minimize the inter-cluster interference. Then two digital precoding approaches are investigated to suppress the intra-cluster interference by separately considering the signal-to-interference-and-noise ratio (SINR) and signal-to-leakage-and-noise ratio (SLNR). Simulation results confirm the effectiveness of proposed block clustering precoding scheme compared to conventional hybrid beamforming scheme.

I. INTRODUCTION

To meet the ever-increasing demand of higher user data rates, it is envisioned that the next-generation cellular systems will be equipped with massive antenna arrays [1]. Capitalizing on the large number of antennas at the base-station (BS), beam division multiple access (BDMA) has recently been proposed to transmit multiple users' data-streams via different beams [2], [3]. In contrast to the more conventional multiple access schemes such as Code Division Multiple Access (CDMA) or Orthogonal Frequency Multiple Division Access (OFDMA) that multiplex users in code, time and frequency domains, BDMA separates users in the beam space by transmitting data to different users in orthogonal beam directions. In [2], BDMA was first proposed to decompose the multiuser multiple-input multiple-output (MU-MIMO) system into multiple single-user MIMO channels by multiplexing multiple users' data onto non-overlapping beams. BDMA is particularly attractive in practice as beamforming is commonly implemented in the analog domain using low-cost phase shifters. More recently, joint user scheduling and beam selection for BDMA was formulated under the Lyapunov-drift optimization framework before the optimal user-beam scheduling policy was derived in a closed form [3]. However, the assumption of non-overlapping orthogonal beams is hard to be satisfied, which handicaps the analog-only BDMA applications. In contrast,

fully digital precoding has been developed to eliminate the inter-user interference through digital signal processing techniques. However, fully digital precoding requires a dedicated radio frequency (RF) chain per each antenna [4].

Despite its many performance advantages, such a fully digital precoding design is prohibitively expensive for massive MIMO systems. To cope with the obstacle, hybrid digital and analog beamforming has been developed for massive MIMO transmissions by dividing the procoding process into two steps, namely analog and digital precoding [5], [6]. More specifically, the transmitted signals are first precoded digitally using a smaller number of RF chains followed by the analog precoding implemented with a much larger number of phase shifters. As a result, the hybrid analog-digital precoding architecture requires significantly less RF chains as compared to the fully digital precoding. To further reduce the computational complexity, the notion of *block diagonal* (BD) precoding was first introduced [7]. By converting the inverse of a large matrix into the inverse of multiple much smaller matrices, the BD precoding can be efficiently implemented with only marginal or no performance degradation as compared to the full digital precoding [7]. The BD design has been recently extended to the hybrid precoding [8], [9]. However, most existing hybrid BD precoding schemes were developed based on a crucial assumption, i.e. the number of RF chains must be no less than the total number of data streams to be transmitted. Some pioneering proposals have been explored to relax this constraint by exploiting the state-of-the-art fast-speed phase shifters and switches that are capable of changing their states symbol by symbol [10]. However, [10] requires users to recover their symbols via the compressive sensing technique, which makes the scheme in [10] not suitable for low-complexity receivers.

In this paper, we consider a multiuser massive MIMO downlink system in which the transmitter sends more data streams with less RF chains by exploiting the hybrid BD precoding architecture built upon the state-of-the-art fast-speed phase shifters and switches. Unlike [10], we consider low-complexity receivers that only perform analog beamforming. Our contributions are summarized as follows:

- To transmit more data streams with less RF chains, we propose a cluster-by-cluster hybrid precoding scheme that effectively decomposes a large digital precoding matrix into block diagonal matrices. More specifically, users are first divided into K clusters before their signals are precoded in the digital and analog domains cluster by cluster. As a result, the minimum number of RF chains is constrained by the number of data streams in each cluster, in lieu of the total number of data
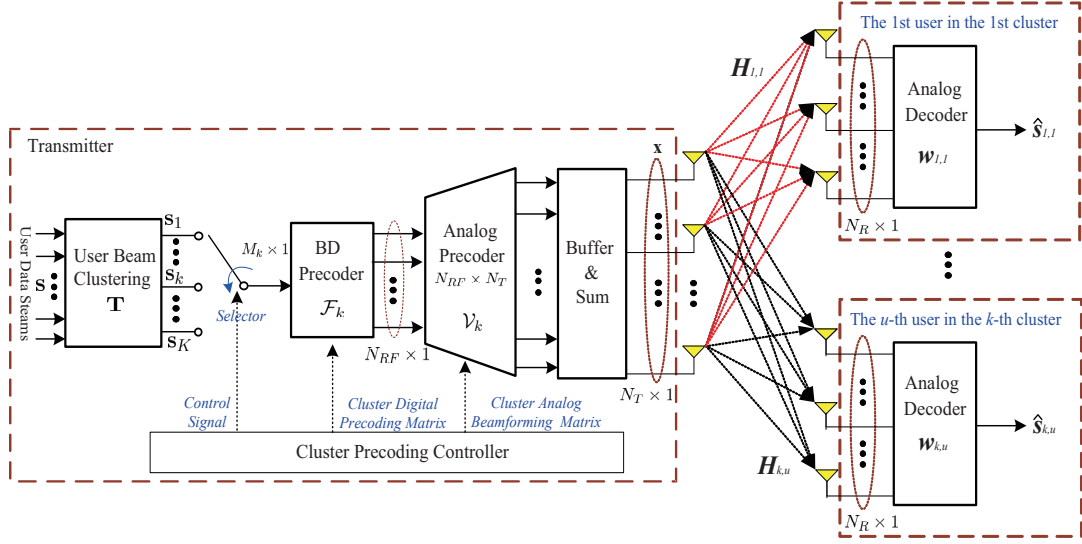
Fig. 1. Block diagram of the hybrid precoding system under consideration

streams in the system. After precoding, all precoded cluster signals are summed together before transmission.

- While analog beamforming in our proposed precoding scheme can suppress much inter-cluster interference, the residual inter-cluster interference as well as the intra-cluster interference have to be removed via digital precoding. In this work, we derive two distinguishing digital precoders, namely the zero-forcing and the signal-to-leakage-and-noise (SLNR)-based precoders. Both precoders show good performance in simulation.

- Most existing hybrid BD precoding schemes send multiple data streams to each user. As a result, it is a natural design to digitally precode each user's data streams while using analog beamforming to separate users. In sharp contrast, we consider the scenario where each scheduled user has only one data stream. Then, user clustering plays an important role in determining the system performance, therefore, we develop a greedy clustering algorithm to further improve the sum-rate of the whole hbrid BD precoding system.

Notation: Vectors and matrices are denoted by boldface letters. $\boldsymbol{I}_N$ denotes the identity matrix with size $N \times N$. $\boldsymbol{A}^T$ and $\boldsymbol{A}^H$ denote transpose and conjugate transpose of $\boldsymbol{A}$, respectively. $\boldsymbol{A}^\dagger$ being the pseudo inverse of $\boldsymbol{A}$ while $\|\boldsymbol{A}\|_0$, $\|\boldsymbol{A}\|$ and $|\boldsymbol{A}|$ stand for 0 norm, the Frobenius norm and determinant of $\boldsymbol{A}$, respectively. $\boldsymbol{A}(i,j)$ denotes the i row, j column element of $\boldsymbol{A}$; $|\mathcal{I}|$ is the cardinality of the enclosed set $\mathcal{I}$; Finally, $\mathbb{E}[\cdot]$ denotes the expectation of a random variable.

II. SYSTEM MODEL

We consider a MU-MIMO downlink system as shown in Fig.1 in which N_U out of N_{tot} users are scheduled for service. The base station (BS) equipped with N_{RF} RF chains and N_T antennas transmits N_U data streams to N_U receivers with N_R receive antennas. We assume only one data stream is designated to each scheduled receiver. Denoted by $\boldsymbol{s}(n)$ the n-th block of N_U data to be transmitted, $\boldsymbol{s}(n)$ has unit power with $\mathbb{E}\left[\boldsymbol{s}\boldsymbol{s}^H\right] = \frac{1}{N_U}\boldsymbol{I}_{N_U}$. In the sequel, we concentrate on a single block and omit the temporal index n for notational simplicity.

A. Transmitter

In our proposed cluster-by-cluster BD digital precoding scheme, the N_U users are first divided into K clusters with the cluster size being $0 < M_k \le N_U$ for $k = 1, 2, \cdots, K$. It is clear that $\sum_{k=1}^{K} M_k = N_U$. Accordingly, the data streams $\boldsymbol{s}$ can be rewritten in clusters as:

$$\boldsymbol{s} = \left[\mathbf{s}_1^T, \mathbf{s}_2^T, \cdots, \mathbf{s}_K^T\right]^T, \tag{1}$$

where $\mathbf{s}_k \in \mathcal{C}^{M_k \times 1}$ is the data vector transmitted to the users in the k-th cluster and modeled as:

$$\mathbf{s}_k = \left[s_{k,1}, s_{k,2}, \cdots, s_{k,M_k}\right]^T, \tag{2}$$

with $s_{k,u}$ being the data transmitted to the u-th user in k-th cluster for $u = 1, 2, \cdots, M_k$.

Next, we start with modeling the digital precoding process. Denote by $\mathcal{F}_k$ of $N_{RF} \times M_k$ the digital precoder for the k-th cluster for $k = 1, 2, \cdots, K$, $\mathcal{F}_k$ can be written as:

$$\mathcal{F}_k = \left[\boldsymbol{f}_{k,1}, \boldsymbol{f}_{k,2}, \cdots, \boldsymbol{f}_{k,M_k}\right], \tag{3}$$

where $\boldsymbol{f}_{k,u}$ represents the digital precoding vector for the u-th user in k-th cluster. Thus, the overall digital precoding matrix can be modeled as a block diagonal matrix as follows:

$$\boldsymbol{F} = \begin{bmatrix} \mathcal{F}_1 & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \mathcal{F}_2 & \vdots & \vdots \\ \boldsymbol{0} & \cdots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \mathcal{F}_K \end{bmatrix}. \tag{4}$$

It is worth noting that inverting a BD matrix is less computationally expensive than a non-BD matrix of the same dimension. Therefore, the BD structure of $\boldsymbol{F}$ in Eq. (4) can potentially lead to reduced computational complexity.

Similarly, we model the corresponding analog precoder in clusters as

$$\boldsymbol{V} = \left[\boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2, \cdots, \boldsymbol{\mathcal{V}}_K\right], \tag{5}$$

where $\boldsymbol{\mathcal{V}}_k$ of $N_T \times N_{RF}$, the analog precoder for the k-th cluster for $k = 1, 2, \cdots, K$, is given by:

$$\boldsymbol{\mathcal{V}}_k = [\boldsymbol{v}_{k,1}, \boldsymbol{v}_{k,2}, \cdots, \boldsymbol{v}_{k,M_k}]. \tag{6}$$

with $\boldsymbol{v}_{k,u}$ being the analog beamforming vector for the u-th user in k-th cluster.

Finally, the resulting hybrid precoded signal $\boldsymbol{x} \in \mathbb{C}^{N_T \times 1}$ is transmitted to all N_U users.

$$\boldsymbol{x} = \boldsymbol{V} \cdot \boldsymbol{F} \cdot \boldsymbol{s} = \sum_{k=1}^{K} \boldsymbol{\mathcal{V}}_k \boldsymbol{\mathcal{F}}_k \mathbf{s}_k. \tag{7}$$

B. Channel Model

Denote by $\boldsymbol{H}_{k,u} \in \mathbb{C}^{N_R \times N_T}$, the *single-path* MIMO channel matrix between the transmitter and the u-th receiver in the k-th cluster is modeled using the Saleh-Valenzuela model [11].

$$\boldsymbol{H}_{k,u} = \sqrt{N_T N_R} \alpha_{k,u} \cdot \boldsymbol{a}_R(\phi_{k,u}^r, \theta_{k,u}^r) \cdot \boldsymbol{a}_T^H(\phi_{k,u}^t, \theta_{k,u}^t), \tag{8}$$

where $\alpha_{k,u}$, $\theta_{k,u}^r / \phi_{k,u}^r$ and $\theta_{k,u}^t / \phi_{k,u}^t$ are the complex path gain, azimuth/elevation angles of arrival(AoA) and azimuth/elevation angles of departure(AoD) of the u-th user in the k-th cluster, respectively. Furthermore, $\boldsymbol{a}(\phi, \theta)$ is the array response vector. For an uniform planar array (UPA) of size $P \times Q$ considered in this work, the array response vector is given by [12]

$$\boldsymbol{a}(\phi, \theta) = \frac{1}{\sqrt{N_T}} \Big[1, e^{j\kappa d(\sin\phi\sin\theta + \cos\theta)}, e^{j2\kappa d(\sin\phi\sin\theta + \cos\theta)},$$
$$\cdots, e^{j\kappa d((P-1)\sin\phi\sin\theta + (Q-1)\cos\theta)} \Big]^T, \tag{9}$$

where $\kappa = \frac{2\pi}{\lambda}$ is the wavenumber and d is the distance between two adjacent antennas. Even though in this paper single-path channel model is considered, it should be noted that the efficacy of our proposed transmission scheme in the sequel still holds for multi-path scenarios.

C. Receiver

The signal received by the u-th user in the k-th cluster is given by

$$\begin{aligned} \boldsymbol{y}_{k,u} = & \underbrace{\boldsymbol{H}_{k,u}\boldsymbol{\mathcal{V}}_k \boldsymbol{f}_{k,u} s_{k,u}}_{\text{Desired Signal}} + \underbrace{\boldsymbol{H}_{k,u}\boldsymbol{\mathcal{V}}_k \sum_{\substack{i=1 \\ i \neq u}}^{M_K} \boldsymbol{f}_{k,i} s_{k,i}}_{\text{Intra-cluster Interference}} \\ & + \underbrace{\boldsymbol{H}_{k,u} \sum_{\substack{j=1 \\ j \neq k}}^{K} \boldsymbol{\mathcal{V}}_j \boldsymbol{\mathcal{F}}_j \mathbf{s}_j}_{\text{Inter-cluster Interference}} + \underbrace{\boldsymbol{n}_{k,u}}_{\text{Noise}} \end{aligned} \tag{10}$$

where $\boldsymbol{n}_u$ is complex additive white Gaussian noise with zero mean and variance equal to σ^2.

Assuming the receivers are all low-cost terminals that perform analog beamforming only in decoding, the decoded signal by the u-th user in k-th cluster denoted by $\hat{s}_{k,u}$ is given by :

$$\hat{s}_{k,u} = \boldsymbol{w}_{k,u}^H \boldsymbol{H}_{k,u}\boldsymbol{\mathcal{V}}_k \boldsymbol{f}_{k,u} s_{k,u} + \boldsymbol{w}_{k,u}^H \tilde{\boldsymbol{n}}_{k,u}, \tag{11}$$

where $\boldsymbol{w}_{k,u}$ of length N_R is the analog beamforming vector employed by the u-th receiver with the power constraint of $|\boldsymbol{w}_{k,u}|^2 = 1$ and

$$\tilde{\boldsymbol{n}}_{k,u} = \boldsymbol{H}_{k,u}\boldsymbol{\mathcal{V}}_k \sum_{\substack{i=1 \\ i \neq u}}^{M_K} \boldsymbol{f}_{ki} s_{ki} + \boldsymbol{H}_{k,u} \sum_{\substack{j=1 \\ j \neq k}}^{K} \boldsymbol{\mathcal{V}}_j \boldsymbol{\mathcal{F}}_j \mathbf{s}_j + \underbrace{\boldsymbol{w}_{k,u}^H \boldsymbol{n}_{k,u}}_{\text{Noise}}. \tag{12}$$

Note that the first term in Eq. (11) stands for the desired signal while the second term is the sum of its own receiver noise and interference from intra-cluster users and other clusters' users.

D. Cluster-by-Cluster(CbC) hybrid precoding

For notational simplicity, we denote by $\boldsymbol{g}_{k,u}^{(j)H}$ the effective analog beamforming gain vector observed by the u-th user in the k-th cluster from the j-th cluster for $j, k = 1, 2, \cdots, K$.

$$\boldsymbol{g}_{k,u}^{(j)H} = \boldsymbol{w}_{k,u}^H \boldsymbol{H}_{k,u}\boldsymbol{\mathcal{V}}_j. \tag{13}$$

Assuming that the power allocated to each user is uniform, the signal-to-noise ratio can be represented as $\gamma = \frac{P}{\sigma^2 N_U}$ where P is the total power in transmitter and σ^2 is the noise power. Then channel capacity of the u-th user can be computed as

$$R_{k,u} = \log \left(1 + \frac{\gamma |\boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}|^2}{\gamma \sum_{\substack{i=1 \\ i \neq u}}^{M_k} \left(|\boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{ki}|^2 + \sum_{\substack{j=1 \\ j \neq k}}^{K} \|\boldsymbol{g}_{k,u}^{(j)H} \boldsymbol{\mathcal{F}}_j\|^2 \right) + 1} \right). \tag{14}$$

Subsequently, the system sum-rate capacity can be computed as a function of $\boldsymbol{W}$, $\boldsymbol{V}$ and $\boldsymbol{F}$:

$$R_{tot}(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{F}) = \sum_{k=1}^{K} \sum_{u=1}^{M_k} R_{k,u}. \tag{15}$$

For conventional hybrid beamforming with sufficient RF chains, the digital beamforming vectors can be designed to eliminate inter-user interference, i.e.

$$\sum_{\substack{i=1 \\ i \neq u}}^{M_k} \left(|\boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{ki}|^2 + \sum_{\substack{j=1 \\ j \neq k}}^{K} \|\boldsymbol{g}_{k,u}^{(j)H} \boldsymbol{\mathcal{F}}_j\|^2 \right) = 0. \tag{16}$$

In contrast, since the proposed BD precoding scheme requires less RF chains, i.e. $N_{RF} \leq N_U$, it can only achieve interference-free asymptotically as N_T grows large. Thus, the capacity of the proposed BD precoding is limited by the residual interference in the system. Given K clusters, we can derive the optimal analog and block digital precoding matrices by

$$P_1 : \max_{\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{F},} R_{tot}(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{F}) \tag{17}$$
$$\begin{aligned} s.t. \quad & C_1 : \|\boldsymbol{v}_{k,u}\|_2^2 = 1; \\ & C_2 : \|\boldsymbol{w}_{k,u}\|_2^2 = 1; \\ & C_3 : \|\boldsymbol{\mathcal{V}}_k \boldsymbol{f}_{k,u}\|^2 = 1; \\ & C_4 : \boldsymbol{V} = [\boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2, \cdots, \boldsymbol{\mathcal{V}}_K]; \\ & C_5 : \boldsymbol{F} = \mathrm{diag}(\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2, \cdots, \boldsymbol{\mathcal{F}}_K); \\ & C_6 : \max\{M_k\}_{k=1}^{K} \leq N_{RF}; \end{aligned}$$

3

where $k = 1, 2, \cdots, K$ and $u = 1, 2, \cdots, M_k$ in C_1, C_2 and C_3. C_1 and C_2 confine all analog beamforming vectors to phase-only style and C_3 ensures the power applied on each signal by transmit precoders is of unit power. C_4 defines the structure of digital prcoder, and C_5 specifics the structure of BD digital precoding matrix. Finally, C_6 depicts that the maximal data streams in each cluster is upper bounded by the number RF chains.

The problem P_1 is challenging due to its non-convex and combinatorial nature. Thus, it is analytically intractable to derive the optimal solution. Instead, we consider a two-stage suboptimal solution: In the first stage, we focus on the analog and digital precoders design to minimize the inter-user interference ; After fixing the analog and digital precoders, we leverage user clustering to further improve system sum-rate in the second stage.

III. PROPOSED BLOCK HYBRID BEAMFORMING FOR RF CHAINS REDUCTION

In this section, we will first optimize the analog precoder before deriving two digital precoders, namely the block zero-forcing (BZF) and block signal-to-leakage-and-noise ratio (SLNR) maximization (BSM) precoders. When designing the digital and analog precoders, we assume the user clustering is given. That is to say we firstly solve the precoder optimization problem for a particular user clustering, and then handle the user clustering algorithm, knowing the precoders design. Even though the optimization problems in these two stages are all highly non-convex, we attempt at closed-form solutions for precoder design and low complexity solution for user clustering algorithm.

A. Analog Beamforming Design

In this subsection, we first focus on the analog beamforming design on both transmitter and receiver sides. According to the theory for infinite antenna theory, *i.e.* as the number of transmit antennas goes to infinity, distinct array response vectors are asymptotically orthogonal, we have

$$\lim_{N \to +\infty} \boldsymbol{a}_T^H(\phi_{k,u}^t, \theta_{k,u}^t) \cdot \boldsymbol{a}_T(\phi_{\ell,v}^t, \theta_{\ell,v}^t) = \delta(k-\ell)\delta(u-v). \quad (18)$$

Hence, to orthogonalize transmitted signals, the transmit analog beamforming vector for user u in cluster k is designed as

$$\boldsymbol{v}_{k,u} = \boldsymbol{a}_T^H(\phi_{k,u}^t, \theta_{k,u}^t), \quad (19)$$

so that we expect the inter-user interference can be completely eliminated. Now, the equivalent channel from the transmitter to user u in cluster k is $\boldsymbol{H}_{k,u}\boldsymbol{v}_{k,u} = \sqrt{N_T N_R}\alpha_{k,u} \cdot \boldsymbol{a}_R(\phi_{k,u}^r, \theta_{k,u}^r)$ and the equivalent channels to other users are all equal to zero vectors. On the receiver side, we use maximum ratio combing (MRC) to design the analog beamforming vector *i.e.*

$$\boldsymbol{w}_{k,u} = \boldsymbol{a}_R^H(\phi_{k,u}^r, \theta_{k,u}^r), \quad (20)$$

In practice, the number of transmitter antenna is finite, which means after transmit analog beamforming, the residual inter-user interference inherently exists. Thereby we need digital precoders to further suppress these interferences.

B. Digital Precoder Design

Ideally the digital precoder can be solved by the optimization problem $P1$ given analog design by Eq.(19) and Eq.(20): However, due to the non-complexity of Eq. (14), we failed to find the optimal closed-form solution of $\boldsymbol{F}$. Nevertheless, we provide two intuitive approaches that lead to high sum-rate.

1) Blcok Zero Forcing (BZF): For a given number of users $K > N_{RF}$, the digital precoder is shaped by a block diagonal matrix as aforementioned. We firstly design the digital precoder for each block as the inverse of the effective channel of the block. Specifically, we have

$$\boldsymbol{\mathcal{F}}_k^{BZF} = \boldsymbol{\mathcal{G}}_k^H(\boldsymbol{\mathcal{G}}_k\boldsymbol{\mathcal{G}}_k^H)^{-1} \quad (21)$$

with $N_{RF} \geq M_k$, where $\boldsymbol{\mathcal{G}}_k = [\boldsymbol{g}_{k,1}^{(k)}, \boldsymbol{g}_{k,2}^{(k)}, \cdots, \boldsymbol{g}_{k,M_k}^{(k)}]^H$.

To satisfy the power constraint C_3 in P_1, power normalization is performed on each $\boldsymbol{f}_{k,u}$ derived from $\boldsymbol{\mathcal{F}}_k^{BZF} = \left[\boldsymbol{f}_{k,1}^{BZF}, \boldsymbol{f}_{k,2}^{BZF}, \cdots, \boldsymbol{f}_{k,M_k}^{BZF}\right]$ as

$$\bar{\boldsymbol{f}}_{k,u}^{BZF} = \frac{\boldsymbol{f}_{k,u}^{BZF}}{\|\boldsymbol{V} \cdot \boldsymbol{f}_{k,u}^{BZF}\|}. \quad (22)$$

For a special case $K = 1$, this problem can be simply solved by conventional zero-forcing.

2) Block SLNR Maximization(BSM): Although the BZF has good performance as shown in the simulation results, it requires to acknowledge all the channel information for each user. The conventional SLNR maximization(SM) is considered as a criterion to reduce co-channel interference(CCI) and noise. For the u-th user in cluster k, the received power of desired data streams can be expressed as

$$S_{k,u} = \gamma \left| \boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}^{BSM} \right|^2. \quad (23)$$

In the meanwhile, the power leaked to other users is

$$L_{k,u} = \gamma \left(\sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{i=1}^{M_j} \left| \boldsymbol{g}_{j,i}^{(k)H} \boldsymbol{f}_{k,u}^{BSM} \right|^2 + \sum_{\substack{t=1 \\ t \neq u}}^{M_k} \left| \boldsymbol{g}_{k,t}^{(k)H} \boldsymbol{f}_{k,u}^{BSM} \right|^2 \right). \quad (24)$$

Therefore, in contrast with the SINR expression in Eq.(14), the SLNR for the u-th user in k-th cluster can be expressed as

$$\phi_{k,u} = \frac{\left| \boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}^{BSM} \right|^2}{\sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{i=1}^{M_j} \left| \boldsymbol{g}_{j,i}^{(k)H} \boldsymbol{f}_{k,u}^{BSM} \right|^2 + \sum_{\substack{t=1 \\ t \neq u}}^{M_k} \left| \boldsymbol{g}_{k,t}^{(k)H} \boldsymbol{f}_{k,u}^{BSM} \right|^2 + \frac{1}{\gamma}}. \quad (25)$$

Preceding studies (e.g. [13]) show that the digital precoder maximizing SLNR $\phi_{k,u}$ is tractable i.e.,

$$\boldsymbol{f}_{k,u}^{BSM} = \mathbf{u}_{max} \left(\left(\frac{1}{\gamma} \boldsymbol{I} + \boldsymbol{Q}_{k,u} \right)^{-1} \boldsymbol{g}_{k,u}^{(k)} \boldsymbol{g}_{k,u}^{(k)H} \right), \quad (26)$$

where $\boldsymbol{Q}_{k,u}$ is the covariance of the leakage given arisen from u-th user in k-th cluster:

$$Q_{k,u} = \sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{i=1}^{M_j} g_{j,i}^{(k)} g_{j,i}^{(k)H} + \sum_{\substack{t=1 \\ t \neq u}}^{M_k} g_{k,t}^{(k)} g_{k,t}^{(k)H}. \qquad (27)$$

Still, we need to normalize the digital precoder as with Eq.(22). Note that after obtaining the digital precoders, we must align all of them in clusters to get the entire BD digital precoding matrix in Eq.(4).

Here, which users should be gathered into a cluster becomes an important question that surely would steer the performance of the whole multi-user system. In what follows, a heuristic user clustering algorithm is provided.

C. User Clustering

In this subsection, we propose a user clustering algorithm designated to BZF precoder, which also shows considerably good performance for BSM. Recalling that BZF digital precoder can eliminate the intra-cluster interference, we will cluster users to minimize the inter-cluster interference.

Ignoring the intra-cluster interference, we approximate the capacity of u-th user in k-th cluster by

$$\tilde{R}_{k,u} \approx \log \left(1 + \frac{\|w_{k,u}^H H_{k,u} \mathcal{V}_k\|^2}{\sum_{j=1, j \neq k}^{K} M_j \|w_{k,u}^H H_{k,u} \mathcal{V}_j\|^2 + \frac{1}{\gamma}} \right). \qquad (28)$$

The approximation is based on that each channel is independently and identically Rayleigh distributed, so that we can have $\mathcal{F}_k^{BZF} = \mathcal{G}_k^H (\mathcal{G}_k \mathcal{G}_k^H)^{-1} \approx \mathbf{I}_{M_k}$.

We propose a Greedy Clustering Algorithm 1 by exchanging the index of data stream. In this greedy clustering algorithm, the users are clustered to different centers by maximizing the sum-rate capacity. The rows indexed from 20 to 22 in Algorithm 1 take into account the constraint on the number of RF chains required for each cluster.

IV. SIMULATION RESULTS

In this section, we use computer simulation to compare the performance of sum-rate capacity for RF chains-reduction algorithm. Unless specified otherwise, we consider a transmitter equipped with an 12×12 UPA (*i.e.* $N_T = 144$) and $N_U = 16$ users each equipped with a 8×8 UPA (*i.e.* $N_R = 64$). The channels are multi-path with the azimuth AoAs/AoDs being uniformly distributed over $[0, 2\pi]$ and the elevation AoAs/AoDs being uniformly distributed in $[-\pi/2, \pi/2]$, respectively. For each computer experiment, we compute the average over 500 realizations.

We firstly compare the two proposed digital precoding approaches with conventional algorithm. As shown in Fig. 2, the dash line "ZF" is the conventional zero-forcing precoding system where 2×8 RF chains are required to serve 2×8 users. The BDMA is the analog-only precoding system. For the sake of fairness, the sum-rate capacities of dash lines are divided by 2. The solid lines represent the proposed RF chains reduction hybrid beamforming systems where only 8 RF chains are used to serve 16 users using CbC. The performance of block zero-forcing

---

Algorithm 1 Greedy clustering algorithm for block hybrid beamforming system

Initialization:
1: Denote all user index as $\mathcal{X}$
2: Initialize the user index set in k-th cluster as $\mathcal{I}_k = \emptyset$, $k = 1, 2, \cdots, K$
3: Initialize the user index with cluster index set : $\mathcal{I} = \emptyset$
4: Cluster index set: $\mathcal{K}$

Procedures:
5: Assign a user index with largest channel gain x^* corresponding to $\mathcal{I}_1$, *i.e.* $\mathcal{I}_1 \leftarrow (1, x^*)$, $\mathcal{I} \leftarrow \mathcal{I}_1$ and $\mathcal{X} \setminus x^*$,
6: **while** $\mathcal{X}$ is not empty **do**
7:     **for** k in $\mathcal{K}$ **do**
8:         **for** x in $\mathcal{X}$ **do**
9:             Update the auxiliary set $\bar{\mathcal{I}}_k$ by adding (k, x) to $\mathcal{I}_k$
10:             Update the auxiliary set $\bar{\mathcal{I}}$ by adding $\bar{\mathcal{I}}_k$ to $\mathcal{I}$
11:             Calculate the sum-rate capacity by Eq. (28)

$$R(k, x) = \sum_{(j,u) \in \bar{\mathcal{I}}} \tilde{R}_{j,u}$$

12:         **end for**
13:         Find the user index (k^*, x^*) with maximum $R(k, x)$
14:         Update $\mathcal{I}_k \leftarrow (k^*, x^*)$, $\mathcal{I} \leftarrow \mathcal{I}_k$ and $\mathcal{X} \setminus x^*$
15:     **end for**
16:     **if** *cardinality*$(\mathcal{I}_k) == N_{RF}$ **then**
17:         $\mathcal{K} \setminus k$
18:     **end if**
19: **end while**

---

without clustering can be significantly increased by the proposed block zero-forcing (BZF) and block SLNR maximization (BSM) with clustering. The performances of BZF and BSM are very similar.
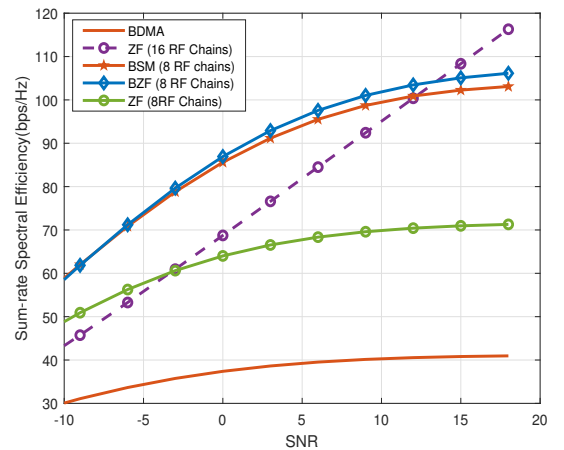

Fig. 2. Sum-rate capacity comparison with different algorithm.

In Fig. 3, we can see that the sum-rate capacity will increase as more RF chains added. The upper bound is the zero-forcing precoding system with 16 RF chains for 16 users. The conventional ZF is the hybrid precoding system with 8 RF chains for 8
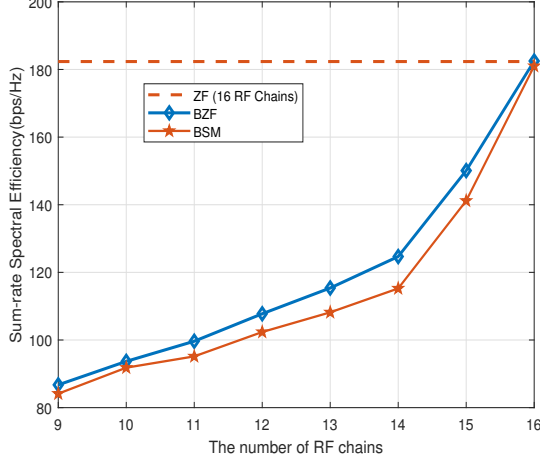
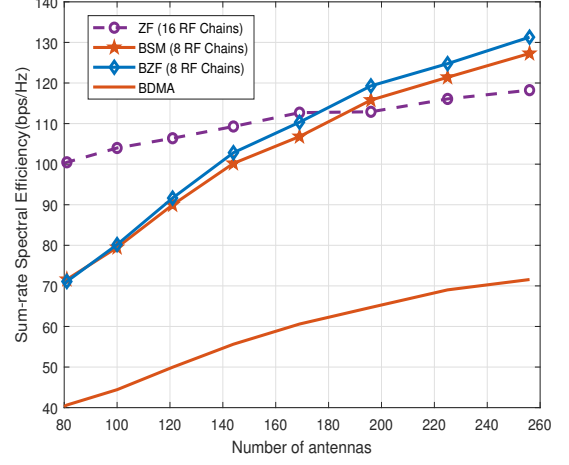Fig. 3. Different number of RF chains.

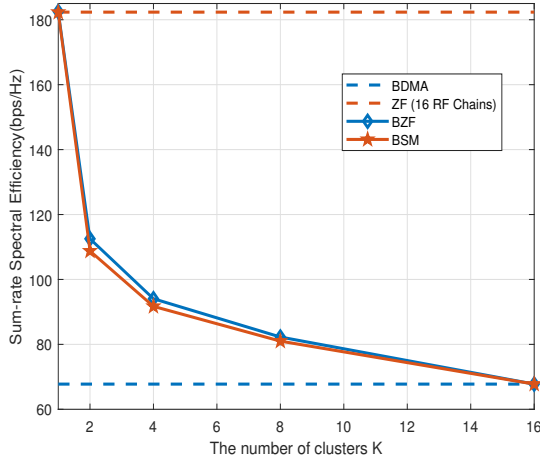Fig. 5. The sum-rate capacity for different number of antennas in transmitter.

users.

Fig. 4. The comparison for different number of clusters.

The Fig. 4 shows that the BZF and BSM are lower bounded by the BDMA and upper bounded by conventional zero-forcing. The system is converted as a conventional hybrid beamforming system when the number of RF chains is not less than data streams. For 16 clusters system, the system is simplified as an analog-only beamforming system.

Finally, the performance for different systems with increasing number of antennas is shown in Fig. 5. By varying the number of antennas, we investigate the sum-rate capacity improvement. The capacity of BZF and BSM will significantly increase as the increment of antennas. The reason is that the residual interference from Eq. (18) will be reduced for more antennas.

V. CONCLUSION

In this work, we have developed block clustering precoding scheme for mmWave massive MIMO systems by jointly performing hybrid analog-digital precoding and user-beam clustering.

First, we have modeled the block hybrid precoder design to reduce the required RF chains by processing the data streams cluster-by-cluster and the analog precoder can be solved by greedily maximizing the sum-SIR. Furthermore, two approaches are proposed, namely BZF and BSM, to obtain the digital precoder. Simulation results have confirmed that the proposed block clustering precoding scheme can achieve better sum-rate capacity with less RF chains compared to the conventional hybrid precoding scheme. The two approaches BZF and BSM can be considered to have the equivalent performance.

REFERENCES

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[2] C. Sun, X. Gao, S. Jin, M. Matthaiou, Z. Ding, and C. Xiao, "Beam Division Multiple Access Transmission for massive MIMO communications," *IEEE Transactions on Communications*, vol. 63, no. 6, pp. 2170–2184, June 2015.

[3] Z. Jiang, S. Chen, S. Zhou, and Z. Niu, "Joint user scheduling and beam selection optimization for beam-based massive MIMO downlinks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2190–2204, April 2018.

[4] T. E. Bogale and L. B. Le, "Beamforming for multiuser massive mimo systems: Digital versus hybrid analog-digital," *arXiv preprint arXiv:1407.0446*, 2014.

[5] S. Han, I. Chih-Lin, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, 2015.

[6] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, 2014.

[7] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser mimo channels," *IEEE transactions on signal processing*, vol. 52, no. 2, pp. 461–471, 2004.

[8] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser mimo systems," *IEEE transactions on communications*, vol. 64, no. 1, pp. 201–211, 2016.

[9] A. Liu and V. Lau, "Phase only rf precoding for massive mimo systems with limited rf chains," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4505–4515, 2014.

[10] N. Garcia, H. Wymeersch, and E. G. Larsson, "Mimo with more users than rf chains," *arXiv preprint arXiv:1709.05200*, 2017.

[11] T. S. Rappaport, R. W. Heath Jr, R. C. Daniels, and J. N. Murdock, *Millimeter wave wireless communications*. Pearson Education, 2014.

[12] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.

[13] J. Wang, S. Jin, X. Gao, K.-K. Wong, and E. Au, "Statistical eigenmode-based SDMA for two-user downlink," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5371–5383, 2012.