Clustered Block Diagonal Digital Precoding for BDMA 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 BDMA based hybrid block precoding system to break the ceiling on minimal required RF chains where the data streams are processed cluster-by-cluster. 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). To overcome the performance degradation arisen from block hybrid precoding scheme, a greedy user clustering algorithm is proposed to minimize the inter-cluster interference. 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. 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 [1], [2]. 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 [1], 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 [2]. 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.

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 [3]. 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 [4]. 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 fully digital precoding [4]. The BD design has been recently extended to the hybrid precoding [5]. 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 [6]. However, [6] requires users to recover their symbols via the compressive sensing technique, which makes the scheme in [6] 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 [6], 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

This work was supported, in part, by the CUHKSZ President's Fund under Grant No. PF.01.000211, the Shenzhen Science and Technology Innovation Committee under Grant No. ZDSYS20170725140921348 and the National Natural Science Foundation of China under Grant No. 61731018.

‡ Corresponding author, email: SimonPun@cuhk.edu.cn.

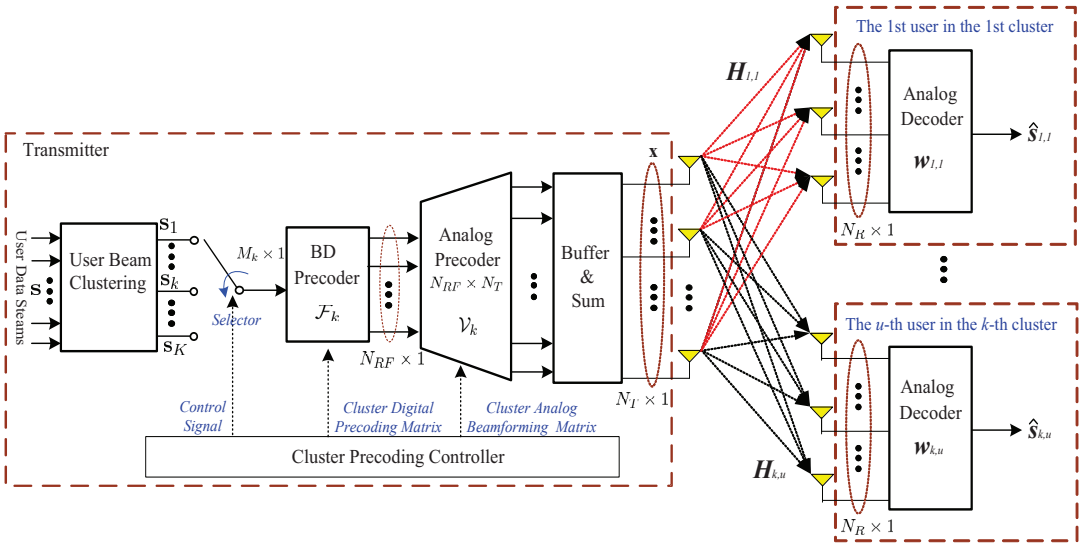

Fig. 1. Block diagram of the hybrid precoding system under consideration

total number of data 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, since we consider the scenario where each scheduled user has only one data stream, user clustering plays an important role in determining the system performance. To cope with this challenge, we develop a greedy clustering algorithm to further improve the sum-rate capacity 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. $\boldsymbol{A}^\dagger$ is the pseudo inverse of $\boldsymbol{A}$ while $\|\boldsymbol{A}\|$ stands for the norm-2 of $\boldsymbol{A}$ and $|A|$ denotes the absolute value of A. $\boldsymbol{A}(i,j)$ denotes the i-th row, j-th 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 system shown in Fig. 1, the N_U users are first divided into K clusters with the cluster size being $0 < M_k \leq N_U$ 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 $\boldsymbol{\mathcal{F}}_k$ of $N_{RF} \times M_k$ the digital precoder for the k-th cluster for $k = 1, 2, \cdots, K$, $\boldsymbol{\mathcal{F}}_k$ can be written as:

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

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 expressed as a block diagonal matrix as follows:

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

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

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

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 [3].

$$\boldsymbol{H}_{k,u} = \sqrt{N_T N_R}\alpha_{k,u} \cdot \boldsymbol{a}_R(\phi^r_{k,u}, \theta^r_{k,u}) \cdot \boldsymbol{a}^H_T(\phi^t_{k,u}, \theta^t_{k,u}), \tag{8}$$

where $\alpha_{k,u}$, $\theta^r_{k,u}/\phi^r_{k,u}$ and $\theta^t_{k,u}/\phi^t_{k,u}$ 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 [3]

$$\boldsymbol{a}(\phi,\theta) = \frac{1}{\sqrt{N_T}} \Big[1, e^{j\kappa d(\sin\phi\sin\theta+\cos\theta)}, e^{j\kappa d(2\sin\phi\sin\theta+2\cos\theta)}, \cdots, e^{j\kappa d((P-1)\sin\phi\sin\theta+(Q-1)\cos\theta)}\Big]^T, \tag{9}$$

where $\kappa = \frac{2\pi}{\lambda}$ is the wavenumber and d is the distance between two adjacent antennas. In the sequel, our proposed schemes focus on the single-path channel model. However, extension of our proposed schemes to the multi-path scenarios is straightforward.

C. Receiver

In this sequel, we investigate the receiver structure of the u-th user in the k-th cluster, unless specified otherwise. The signal received is given by

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

where $\boldsymbol{n}_{k,u}$ 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 denoted by $\hat{s}_{k,u}$ is given by :

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

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

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

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

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

Assuming that the power allocated to each user is uniform, the signal-to-noise ratio (SNR) can be represented as

$$\gamma = \frac{P}{\sigma^2 N_U}, \tag{14}$$

where P is the total transmission power and σ^2 is the noise power.

Then, the resulting channel capacity can be computed as

$$R_{k,u} = \log\left(1 + \frac{\gamma|\boldsymbol{g}^{(k)H}_{k,u}\boldsymbol{f}_{k,u}|^2}{\gamma\sum_{\substack{i=1\\i\neq u}}^{M_k}\left(|\boldsymbol{g}^{(k)H}_{k,u}\boldsymbol{f}_{k,i}|^2 + \sum_{\substack{j=1\\j\neq k}}^{K}\|\boldsymbol{g}^{(j)H}_{k,u}\boldsymbol{\mathcal{F}}_j\|^2\right) + 1}\right). \tag{15}$$

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

$$R_{tot}(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{F}) = \sum_{k=1}^{K}\sum_{u=1}^{M_k} R_{k,u}. \tag{16}$$

For conventional hybrid beamforming with sufficient RF chains, the digital beamforming vectors can be designed to completely eliminate inter- and intra-cluster interference, *i.e.*

$$\sum_{\substack{i=1\\i\neq u}}^{M_k}\left(|\boldsymbol{g}^{(k)H}_{k,u}\boldsymbol{f}_{k,i}|^2 + \sum_{\substack{j=1\\j\neq k}}^{K}\|\boldsymbol{g}^{(j)H}_{k,u}\boldsymbol{\mathcal{F}}_j\|^2\right) = 0. \tag{17}$$

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 inter- and intra-cluster interference in the system.

Given K clusters, we can derive the optimal analog and block digital precoding matrices by

$$\begin{aligned} P_1: \quad & \max_{\boldsymbol{W},\boldsymbol{V},\boldsymbol{F},} \quad R_{tot}(\boldsymbol{W},\boldsymbol{V},\boldsymbol{F}) \\ s.t. \quad & C_1: \|\boldsymbol{v}_{k,u}\|_2^2 = 1; \\ & C_2: \|\boldsymbol{w}_{k,u}\|_2^2 = 1; \\ & C_3: \|\boldsymbol{\mathcal{V}}_k \boldsymbol{f}_{k,u}\|^2 = 1; \\ & C_4: \boldsymbol{V} = [\boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2, \cdots, \boldsymbol{\mathcal{V}}_K]; \\ & C_5: \boldsymbol{F} = \text{diag}(\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2, \cdots, \boldsymbol{\mathcal{F}}_K); \\ & C_6: \max\{M_k\}_{k=1}^K \leq N_{RF}; \end{aligned} \tag{18}$$

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 the phase-only structure while C_3 ensures that each precoded signal is of unit power. Furthermore, C_4 and C_5 define the structure of clustered analog precoder and BD digital precoding matrix, respectively. Finally, C_6 constrains the maximal data streams in each cluster to be within 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 its optimal solution in closed form. 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 that the user clustering is given. This approach enables us to first solve the precoder optimization problem for a particular user clustering before handling the user clustering algorithm. 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, as the number of transmit antennas goes to infinity, distinct array response vectors are asymptotically orthogonal, *i.e.*

$$\lim_{N\to+\infty} \boldsymbol{a}_T^H(\phi_{k,u}^t, \theta_{k,u}^t) \cdot \boldsymbol{a}_T(\phi_{\ell,v}^t, \theta_{\ell,v}^t) = \delta(k-\ell)\delta(u-v). \tag{19}$$

Recalling the channel model presented in Eq. (8), we can asymptotically orthogonalize transmitted signals by setting the transmit analog beamforming vector for the u-th user in the k-th as

$$\boldsymbol{v}_{k,u} = \boldsymbol{a}_T(\phi_{k,u}^t, \theta_{k,u}^t). \tag{20}$$

As a result, the inter-user interference can be asymptotically eliminated if a large number of beamforming antennas is employed. Using the analog beamforming vector in Eq. (20), the equivalent channel from the transmitter to user u in cluster k becomes $\boldsymbol{H}_{k,u}\boldsymbol{v}_{k,u} = \sqrt{N_T N_R}\alpha_{k,u} \cdot \boldsymbol{a}_R(\phi_{k,u}^r, \theta_{k,u}^r)$ and the equivalent channels to other users are all equal to zero vectors. Subsequently, the maximum ratio combing (MRC) is employed at the receiver and the resulting receive analog beamforming vector is given by:

$$\boldsymbol{w}_{k,u}^H = \boldsymbol{a}_R^H(\phi_{k,u}^r, \theta_{k,u}^r). \tag{21}$$

Unfortunately, the number of transmitter antenna is finite in practice. As a result, the analog beamforming vectors shown in Eq. (20) and Eq. (21) inevitably incur residual inter-user interference. Thereby digital precoders are required to further suppress the residual interference.

B. Digital Precoder Design

Ideally the digital precoder can be derived from $P1$ for any given analog design specified in Eq. (20) and Eq. (21). However, due to the high complexity of Eq. (15), it is analytically intractable to derive the optimal solution of $\boldsymbol{F}$ in closed form. In this section, two digital precoding schemes are proposed to maximize the system sum-rate.

1) Block Zero-Forcing (BZF) Scheme: In contrast to the conventional zero-forcing hybrid beamforming scheme [7] that requires $N_U \leq N_{RF}$, the first proposed scheme applies zero-forcing digital precoding cluster by cluster. More specifically, the digital precoder for each block is designed as the inverse of the effective channel of the block:

$$\boldsymbol{\mathcal{F}}_k^{\text{BZF}} = \boldsymbol{\mathcal{G}}_k^H(\boldsymbol{\mathcal{G}}_k\boldsymbol{\mathcal{G}}_k^H)^{-1} \tag{22}$$

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 $\boldsymbol{\mathcal{F}}_k^{\text{BZF}} = \left[\boldsymbol{f}_{k,1}^{\text{BZF}}, \boldsymbol{f}_{k,2}^{\text{BZF}}, \cdots, \boldsymbol{f}_{k,M_k}^{\text{BZF}}\right]$ as

$$\bar{\boldsymbol{f}}_{k,u}^{\text{BZF}} = \frac{\boldsymbol{f}_{k,u}^{\text{BZF}}}{\|\boldsymbol{\mathcal{V}}_k \cdot \boldsymbol{f}_{k,u}^{\text{BZF}}\|}. \tag{23}$$

In the sequel, this scheme is referred to as the block zero-forcing (BZF) scheme. It is worth noting that BZF degenerates to [7] if $K = 1$, *i.e.* all users are grouped into one single cluster. On the other hand, BZF becomes the analog-only BDMA if $K = N_U$, *i.e.* each user forms one cluster and only analog beamforming is performed.

2) Block SLNR Maximization (BSM) Scheme: Instead of eliminating the received interference, we can alternatively design the digital precoder to reduce co-channel interference (CCI) by maximizing SLNR. More specifically, we denote by $P_{k,u}^{\text{Desired}}$ the desired signal power received by the u-th user in the k-th cluster

$$P_{k,u}^{\text{Desired}} = \gamma \left| \boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2, \tag{24}$$

where γ is the SNR defined in Eq. (14).

Further, if we define leakage signal as the transmitted signal that is intended to a specific user but leaked to other users,

then the leakage signal power due to the u-th user in the k-th cluster can be expressed as

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

Finally, the SLNR for the u-th user in k-th cluster can be written as

$$\Gamma_{k,u} = \frac{\left| \boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2}{\sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{i=1}^{M_j} \left| \boldsymbol{g}_{j,i}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2 + \sum_{\substack{t=1 \\ t \neq u}}^{M_k} \left| \boldsymbol{g}_{k,t}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2 + \frac{1}{\gamma}}. \tag{26}$$

Denoted by $\boldsymbol{f}_{k,u}^{\text{BSM}}$ the optimal digital precoder maximizing SLNR, it has been shown that $\boldsymbol{f}_{k,u}^{\text{BSM}}$ turns out to be the eigenvector associated with the largest eigenvalue of the following matrix [8]:

$$\boldsymbol{R}_{k,u}^{\text{Leakage}} = \left(\frac{1}{\gamma} \boldsymbol{I}_{N_{RF}} + \boldsymbol{Q}_{k,u} \right)^{-1} \boldsymbol{g}_{k,u}^{(k)} \boldsymbol{g}_{k,u}^{(k)H}, \tag{27}$$

where $\boldsymbol{Q}_{k,u}$ is the leakage covariance matrix related to the u-th user in the k-th cluster and given as:

$$\boldsymbol{Q}_{k,u} = \sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{i=1}^{M_j} \boldsymbol{g}_{j,i}^{(k)} \boldsymbol{g}_{j,i}^{(k)H} + \sum_{\substack{t=1 \\ t \neq u}}^{M_k} \boldsymbol{g}_{k,t}^{(k)} \boldsymbol{g}_{k,t}^{(k)H}. \tag{28}$$

It is apparent from the above derivation that the user clustering algorithm plays an important role in determining the amount of inter-cluster interference, and subsequently the system performance. In the next section, the user clustering algorithm is investigated.

C. User Clustering

Recalling that the BZF scheme can completely remove intra-cluster interference via the zero-forcing digital precoding described in Eq. (22), we will focus on designing the user clustering algorithm for the BZF scheme by minimizing the inter-cluster interference. Since the infinite number of antennas can't be achieved in practical, the residual interference will be generated by the non-orthogonality of users' array response vectors. Motivated by this observation, we propose to cluster N_U users into K clusters whose array response vectors are as orthogonal as possible. Since the total number of possible cluster combinations can be rather large, a greedy clustering algorithm is proposed in Algorithm 1 by grouping users with most orthogonal transmitted beams to distinct clusters.

In this algorithm, for $k = 1$, we first group users with the most non-orthogonal array response vectors to cluster 1 detailed in *Stage 1*. When the cluster size is equal to the number of RF chains, the user whose array response vector is most orthogonal to cluster 1 is selected as the first member of cluster 2 as shown in *Stage 2*. By same manner, the users can be clustered to minimize the inter-cluster interference. Despite that Algorithm 1 is designed for the BZF scheme, our simulation shows that it also works well for the BSM scheme.

Algorithm 1 Greedy User Clustering Algorithm

Initialization:

$\mathcal{X}$: the universal user index set;

$\boldsymbol{a}_T(\phi_x^t, \theta_x^t)$: Array response vector of user index x;

$\mathcal{I}_k = \emptyset$: the user index set for the k-th cluster;

$k = 1$: cluster index;

Initialize $\mathcal{I}_1 \leftarrow x^*$ with x^* being the user index of the largest channel gain and $\mathcal{X} \setminus x^*$;

Procedures:

while $\mathcal{X}$ is not empty **do**

Stage 1:

for x in $\mathcal{X}$ **do**

$\boldsymbol{A} = [\boldsymbol{a}_T(\phi_{k,1}^t, \theta_{k,1}^t), \boldsymbol{a}_T(\phi_{k,2}^t, \theta_{k,2}^t), \cdots, \boldsymbol{a}_T(\phi_{k,|\mathcal{I}_k|}^t, \theta_{k,|\mathcal{I}_k|}^t)]$

Compute

$$p(x) = \|\boldsymbol{a}_T^H(\phi_x^t, \theta_x^t) \cdot \boldsymbol{A}\|^2$$

end for

Find the user index x^* with *maximum* $p(x)$

Update $\mathcal{I}_k \leftarrow x^*$ and $\mathcal{X} \setminus x^*$

Stage 2:

if $|\mathcal{I}_k| = N_{RF}$ **then**

Update $k \leftarrow k + 1$

for x in $\mathcal{X}$ **do**

Compute

$$p(x) = \|\boldsymbol{a}_T^H(\phi_x^t, \theta_x^t) \cdot \boldsymbol{A}\|^2$$

end for

Find the user index x^* with *minimum* $p(x)$

Update $\mathcal{I}_k \leftarrow x^*$ and $\mathcal{X} \setminus x^*$

end if

end while

IV. SIMULATION RESULTS

In this section, we will use computer simulation to compare the sum-rate performance of the proposed block diagonal digital precoding schemes. Unless specified otherwise, we consider a transmitter equipped with a 12×12 UPA (*i.e.* $N_T = 144$) and $N_U = 16$ users each equipped with a 8×8 UPA (*i.e.* $N_R = 64$). We consider the azimuth AoAs/AoDs uniformly distributed over $[0, 2\pi]$ while the elevation AoAs/AoDs uniformly distributed over $[-\pi/2, \pi/2]$, respectively. For each computer experiment, we compute the average over 500 realizations.

We first compare the two proposed digital precoding approaches against the conventional ZF algorithm reported in [3]. As shown in Fig. 2, the solid lines represent the proposed block diagonal digital precoding systems where only 8 RF chains are used to serve 16 users. The performances of BZF and BSM are very similar. Furthermore, the dashed line labeled as "ZF (16 RF Chains)" is the the sum-rate for the conventional ZF precoding system with 16 RF chains serving 16 users. For the sake of fairness, its sum-rate has been divided by a factor of 2. In contrast, the dashed line labeled as "ZF (8 RF Chains)" is the sum-rate for the conventional ZF precoding system with 8 RF chains serving 8 users. Finally, BDMA is the analog-only precoding system that has the worst performance. Inspection of Fig. 2 has revealed that the proposed BZF and BSM have much better sum-rate performance than the conventional ZF algorithms.

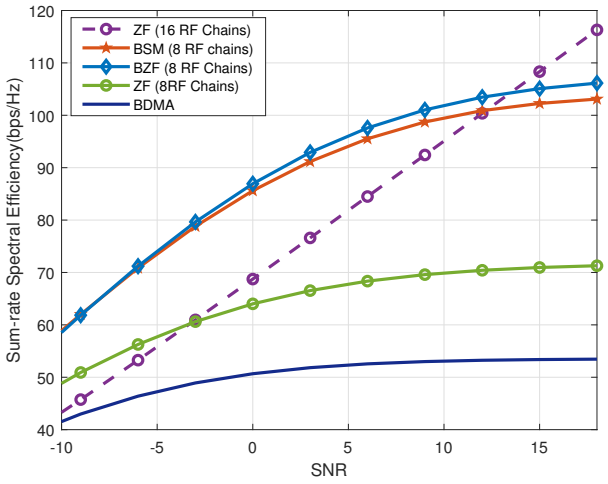

Fig. 2. Sum-rate capacity comparison for different algorithms

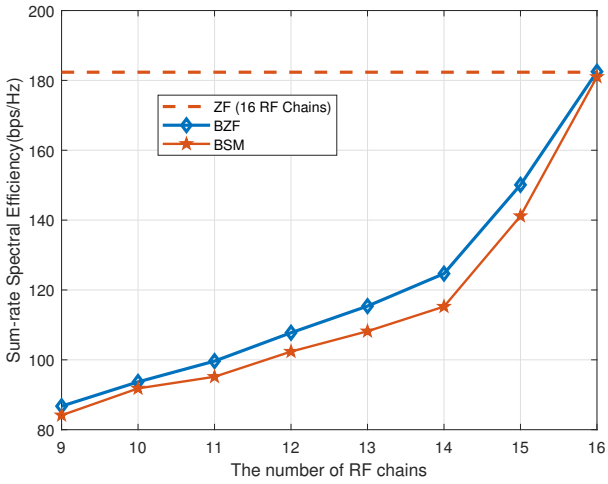

Fig. 3. Sum-rate capacity improvement as a function of the number of RF chains.

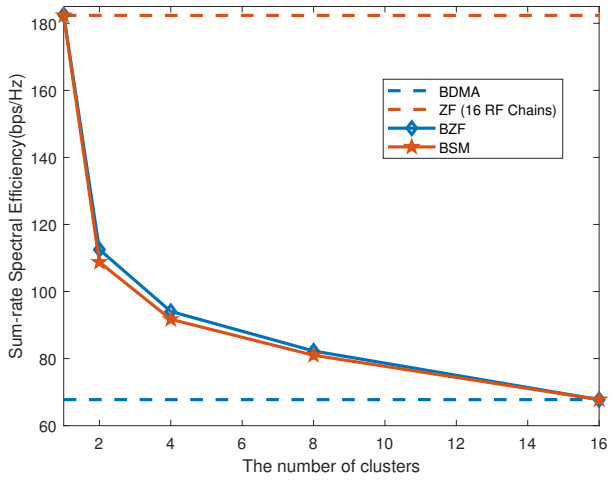

Fig. 4. Sum-rate capacity comparison for different number of clusters.

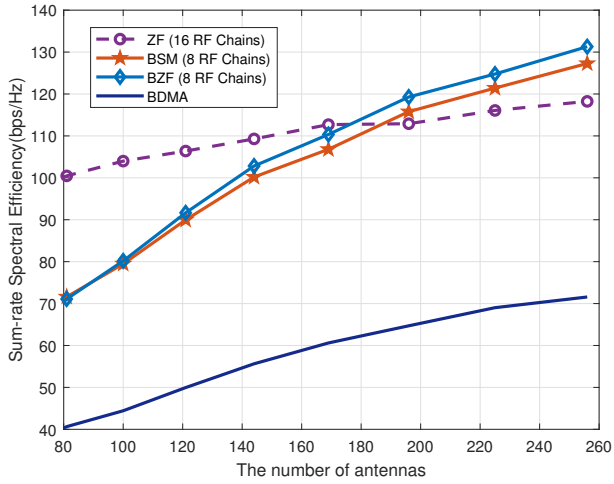

Fig. 5. Sum-rate capacity comparison for different number of antennas in transmitter.

In Fig. 3, we investigates the sum-rate capacity improvement as a function of the number of RF chains. The upper bound is the ZF precoding system with 16 RF chains for 16 users.

Next, we vary the number of cluster while fixing the total number of users to 16. Fig. 4 shows that the BZF and BSM are lower bounded by the BDMA and upper bounded by the conventional ZF system with 16 RF chains. When $K = 1$, the system degenerates back to the conventional ZF system with $N_{RF} = M_1 = 16$. On the other hand, if $K = 16$, the system becomes BDMA.

Finally, we investigate the sum-rate performance as the number of transmit antennas increases. Fig. 5 shows that the capacity of BZF and BSM has been significantly increased as the number of transmit antennas increases. This is because that the inter-cluster interference is asymptotically removed as indicated in Eq. (19).

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.