sensors

MDPI

Article

Block Diagonal Hybrid Precoding and Power Allocation for QoS-Aware BDMA Downlink Transmissions

Guanchong Niu [1,2,3], Qi Cao [1,3] and Manon Pun [1,2,3,*]

1 School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China; 216019007@link.cuhk.edu.cn (G.N.); Caoqi@cuhk.edu.cn (Q.C.)
2 Shenzhen Research Institute of Big Data, Shenzhen 518172, China
3 Shenzhen Key Laboratory of IoT Intelligent Systems and Wireless Network Technology, Shenzhen 518172, China
* Correspondence: simonpun@cuhk.edu.cn

Received: 7 July 2020; Accepted: 7 August 2020; Published: 11 August 2020

Abstract: Beam Division Multiple Access (BDMA) with hybrid precoding has recently been proposed for multi-user multiple-input multiple-output (MU-MIMO) systems by simultaneously transmitting multiple digitally precoded users' data-streams via different beams. In contrast to most existing works that assume the number of radio frequency (RF) chains must be greater than or equal to that of data-streams, this work proposes a novel BDMA downlink system by first grouping transmitting data-streams before digitally precoding data group by group. To fully harvest the benefits of this new architecture, a greedy user grouping algorithm is devised to minimize the inter-group interference while two digital precoding approaches are developed to suppress the intra-group interference by maximizing the signal-to-interference-and-noise ratio (SINR) and the signal-to-leakage-and-noise ratio (SLNR), respectively. As a result, the proposed BDMA system requires less RF chains than the total number of transmit data-streams. Furthermore, we optimize the power allocation to satisfy each user's quality of service (QoS) requirement using the D.C. (difference of convex functions) programming technique. Simulation results confirm the effectiveness of the proposed scheme.

Keywords: BDMA; hybrid beamforming; block diagonal precoder; power allocation

1. Introduction

To meet the explosive demand for higher user data rates, it is envisioned that the next-generation cellular systems will be equipped with massive antenna arrays. Capitalizing on a large number of antennas at the base station (BS), Beam Division Multiple Access (BDMA) has recently been proposed as a promising method for 5G communications [1–4]. Different beams are allowed to transmit multiple users' data-streams from BS. In contrast to the more conventional multiple access schemes such as Code Division Multiple Access (CDMA), Time Division Multiple Access (TDMA) 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. BDMA was first proposed in [1] to decompose the multi-user multiple-input multiple-output (MU-MIMO) system into multiple single-user MIMO channels by multiplexing multiple users' data onto non-overlapping beams. Since beamforming is commonly implemented in the analog domain using low-cost phase shifters, BDMA becomes particularly attractive in practice in recent years. Moreover, joint beam selection and user scheduling were formulated under the Lyapunov-drift optimization framework before the optimal user-beam scheduling policy for BDMA was derived in a closed-form [2]. However, the assumption of non-overlapping orthogonal beams is

generally difficult to be satisfied in practice. As a result, analog-only BDMA applications are heavily handicapped by the non-orthogonal inter-user interference among beams.

In the meantime, digital precoding has been widely investigated as an effective signal processing method to suppress the inter-user interference for MU-MIMO. It is well known that the classical fully digital precoding requires a dedicated radio frequency (RF) chain for each antenna. However, power consumption and the high hardware cost render the fully digital precoding impractical for massive MIMO systems [5–7]. To address this challenge, hybrid digital and analog beamforming has been proposed for massive MIMO transmissions by separating the precoding process into two steps, namely analog and digital precoding [8,9]. More specifically, the transmitted signals are first precoded digitally using a smaller number of RF chains followed by the analog precoder exploited by a much larger number of low-cost phase shifters [10,11]. As a result, the hybrid analog-digital precoding architecture requires significantly fewer RF chains as compared to the fully digital precoding [12]. It has been reported in the literature that the hybrid beamforming structure is capable of achieving performance compared to the fully digital beamforming scheme if the number of RF chains at each end is greater than or equal to twice the number of the data-streams [13]. Therefore, the hybrid precoded massive MU-MIMO system can benefit from the interference suppression supplied by the digital precoding while harvesting large antenna beamforming gains by implementing the massive antennas available in the systems [14]. This hybrid structure is particularly attractive for millimeter wave (mmWave) MIMO systems to support the transmissions of Gbps-order data throughput by exploiting the vast vacant spectrum available at RF of 6 GHz or above [15]. Furthermore, the notion of block diagonal (BD) precoding was first introduced to the conventional fully digital schemes to reduce the precoding complexity in [16]. By dividing the inverse of a large matrix into the inverse of multiple much smaller matrices, the BD precoding can be efficiently exploited with only marginal or no performance degradation as compared to the fully digital precoding [16]. In recent years, the BD design has been extended to the hybrid precoding for MU-MIMO [17]. However, most existing hybrid BD precoding schemes were constructed 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 works have investigated to relax this limitation by implementing the state-of-the-art fast-speed phase shifters and switches that can change their states symbol by symbol [18]. However, [19] requires users to resume their symbols via the compressive sensing technique, which makes the scheme impractical for low-complexity receivers.

Meanwhile, power allocation is also an important problem in co-channel interference management for multi-user wireless networks. In many MIMO applications, it is desirable to design a system satisfying the quality of service (QoS) constraint for each user by adjusting the power allocated to different users [20]. Since the objective function is highly non-convex, the problem is usually very difficult and complicated, especially for the coupled analog and digital precoding constraints [21]. Therefore, most existing works maximize the sum-rate capacity by implementing the water-filling algorithm without considering the QoS requirement for each user. For instance, [22] alternatively optimized the power allocation for sum-rate maximization by using the water-filling algorithm, assuming that the analog precoders are strictly orthogonal among distinct users. However, the water-filling algorithm cannot satisfy the per-user QoS constraint as users with poor channel conditions are not allocated any transmit power by the water-filling algorithm. To cope with this problem, the signal-to-interference-and-noise ratio (SINR)-balanced power allocation has been proposed to achieve identical SINR for all users [23–26]. However, the system performance using the SINR-based power allocation is limited by the user with the worst channel conditions.

In this work, we propose a downlink BDMA scheme empowered by BD digital precoding and global power allocation over multipath channels. Compared to the existing BDMA works [1,2], our proposed BDMA schemes can substantially suppress multi-user interference without requiring perfectly orthogonal beams as residual interference can be greatly removed by digital precoding. Furthermore, in sharp contrast to the conventional hybrid precoding schemes, the proposed scheme can

use fewer RF chains than the number of transmit data-streams by exploiting the hybrid BD precoding architecture built upon the state-of-the-art fast-speed phase shifters [27] and switches [18]. Furthermore, an iterative algorithm for power allocation is proposed to satisfy per-user QoS requirement based on the difference of convex functions (D.C.) programming technique.

The main contributions of this paper are summarized as follows:

- A block diagonal hybrid precoding scheme is proposed by exploiting the state-of-the-art fast-speed phase shifters and switches. The resulting scheme can use fewer RF chains than the number of transmit data-streams by jointly performing hybrid analog-digital precoding and user-beam grouping.
- Furthermore, we develop a greedy grouping algorithm to minimize the inter-group interference while maximizing the intra-group interference. Then, the intra-group interference is eliminated by two proposed digital precoders, namely the SINR- and SLNR-based precoders.
- In contrast to most works in the literature that used the single-path channel model, we analyze the sum-rate capacity using the multipath channel model.
- Finally, for given analog and digital precoders, an optimized power allocation scheme is derived to satisfy per-user QoS requirement by using the D.C. programming technique.

The rest of this paper is organized as follows: In section 2, we present the block diagonal system model with reduced RF chains before formulating the optimization problem. By allocating the power uniformly to each user, the analog and digital precoders are derived in Section 3. After that, the performance of the proposed system is analyzed in Section 5. Finally, to satisfy the QoS constraint, Section 6 proposes a QoS-aware power allocation algorithm based on the D.C. programming technique followed by extensive numerical results presented in Section 7.

Notation: In this paper, we use uppercase boldface and lowercase boldface letters to denote matrices and vectors, respectively. $\boldsymbol{I}_N$ denotes the identity matrix with size $N \times N$. $\boldsymbol{A}^T$ and $\boldsymbol{A}^H$ denote the transpose and conjugate transpose of $\boldsymbol{A}$, respectively. $\boldsymbol{A}^\dagger$ is the pseudo inverse of $\boldsymbol{A}$ while $\|\boldsymbol{A}\|$ stands for the L2 norm of $\boldsymbol{A}$ and $|A|$ denotes the absolute value of A. $[\boldsymbol{a}]_i$ denotes the i-th element of $\boldsymbol{a}$. $|\mathcal{I}|$ is the cardinality of the enclosed set $\mathcal{I}$. $\mathcal{X}^2(k)$ represents the chi-square distribution with k degrees of freedom. $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$ is the inner product of $\boldsymbol{A}$ and $\boldsymbol{B}$. $\mathcal{X} \leftarrow x$ stands for the addition of element x to set $\mathcal{X}$ while $\mathcal{X} \setminus x$ removal of element x from $\mathcal{X}$. Finally, $\mathbb{E}[\cdot]$ denotes the expectation of the enclosed random variable.

2. System Model and Problem Formulation

In this paper, we consider a MU-MIMO downlink system as shown in Figure 1 in which N_U users are scheduled for service. N_{RF} RF chains and N_T antennas are equipped on BS which transmits N_U data-streams to N_U receivers with N_R receive antennas at each time slot. In a practical MIMO system, the number of RF chains is typically much smaller than the number of antennas, i.e., $N_{RF} \ll N_T$. We assume that only one data stream is designated to each scheduled receiver for transmission. 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 can concentrate on a single block and omit the temporal index n for notation simplicity.

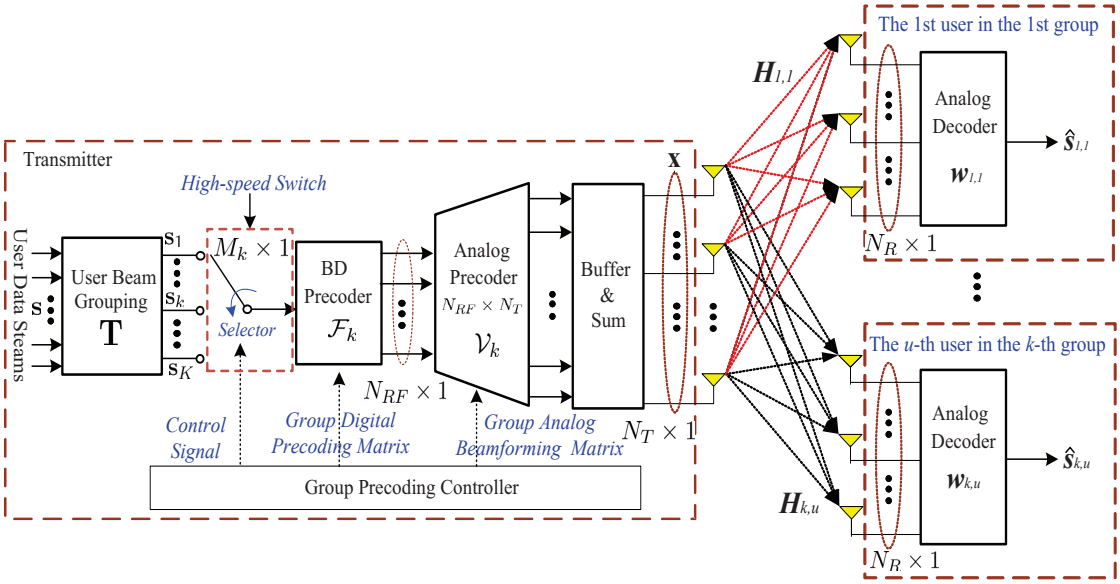

Figure 1. Block diagram of the hybrid precoding system under consideration.

2.1. Transmitter

In our proposed group-by-group BD digital precoding system shown in Figure 1, the N_U users are first divided into K groups with the group size being M_K, where $0 < M_k \leq N_U$ 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 groups 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 group 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 the k-th group for $u = 1, 2, \cdots, M_k$.

Next, we focus on modeling the digital precoding process. Denoted by $\boldsymbol{\mathcal{F}}_k$ of $N_{RF} \times M_k$ the digital precoder for the k-th group 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 the k-th group. Thus, the overall digital precoding matrix can be expressed as a block diagonal matrix as follows:

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

Clearly, 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 Equation (4) can potentially lead to reduced computational complexity.

Similarly, we model the corresponding analog precoder in groups 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 group for $k = 1, 2, \cdots, K$, is given by:

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

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

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

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

2.2. Channel Models

We denote $\boldsymbol{H}_{k,u} \in \mathbb{C}^{N_R \times N_T}$ the MIMO multipath channel matrix between the transmitter and the u-th receiver in the k-th group using the Saleh-Valenzuela model [8]:

$$\boldsymbol{H}_{k,u} = D_{k,u} \sum_{l=1}^{L_{k,u}} \alpha_{k,u,l} \boldsymbol{a}_R(\phi^r_{k,u,l}, \theta^r_{k,u,l}) \boldsymbol{a}_T^H(\phi^t_{k,u,l}, \theta^t_{k,u,l}), \tag{8}$$

where $L_{k,u}$ is the total number of multipath components between the transmitter and the u-th user in the k-th group. Furthermore, $\alpha_{k,u,l}$, $\theta^r_{k,u,l}/\phi^r_{k,u,l}$ and $\theta^t_{k,u,l}/\phi^t_{k,u,l}$ are the complex path gain, angles of arrival (AoA) and azimuth/elevation angles of departure (AoD) of the l-th path of the u-th user in the k-th group, respectively. Furthermore, $\boldsymbol{a}(\phi, \theta)$ is the array response vector. Finally, $D_{k,u} = \sqrt{\frac{N_T N_R}{L_{k,u}}}$ is a constant parameter. For a uniform planar array (UPA) of size $P \times Q$ considered in this work, the array response vector is given by [8]:

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

where $\kappa = \frac{2\pi}{\lambda}$ is the wavenumber and d is the distance between two adjacent antennas.

2.3. Receiver

Consequently, we formulate the receiver structure of the u-th user in the k-th group. The received signal is represented by

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

where $\boldsymbol{n}_{k,u}$ is the complex additive white Gaussian noise with zero mean and variance equal to $\sigma^2_{k,u}$.

Assuming that 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

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

Please note that the first term in Equation (11) stands for the desired signal while the second term the sum of inter- and intra-group interference as well as receiver thermal noise.

2.4. Group-By-Group 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 group from the j-th group for $j, k = 1, 2, \cdots, K$.

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

Let $\boldsymbol{p}$ be the transmitted power vector, where the power allocated to the u-th user in the k-th group is denoted by $p_{k,u}$ with

$$\sum_{k=1}^{K} \sum_{u=1}^{M_k} p_{k,u} \leq N_U. \tag{14}$$

Then, the resulting channel capacity can be computed as

$$R_{k,u} = \log_2 \left(1 + \frac{p_{k,u} |\boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}|^2}{\zeta_{k,u} + \sigma_{k,u}^2} \right), \tag{15}$$

with

$$\zeta_{k,u} = \sum_{\substack{i=1 \\ i \neq u}}^{M_k} p_{k,i} |\boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,i}|^2 + \sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{t=1}^{M_j} p_{j,t} |\boldsymbol{g}_{k,u}^{(j)H} \boldsymbol{f}_{j,t}|^2, \tag{16}$$

and $\sigma_{k,u}^2$ is the noise power.

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

$$R_{tot}(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{F}, \boldsymbol{p}) = \sum_{k=1}^{K} \sum_{u=1}^{M_k} R_{k,u}. \tag{17}$$

It is worth noting that the digital beamforming vectors can be designed to eliminate user interference for the conventional hybrid beamforming with sufficient RF chains, i.e.,

$$\sum_{\substack{i=1 \\ i \neq u}}^{M_k} p_{k,i} |\boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,i}|^2 + \sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{t=1}^{M_j} p_{j,t} |\boldsymbol{g}_{k,u}^{(j)H} \boldsymbol{f}_{j,t}|^2 = 0. \tag{18}$$

In contrast, it can only achieve interference-free asymptotically as N_T grows a very large number since the proposed BD precoding scheme requires fewer RF chains, i.e., $N_{RF} < N_U$. Thus, the capacity of the proposed BD precoding scheme is constrained by the residual inter- and intra-group interference

in the system. Given K groups, we can derive the optimal analog and block digital precoding matrices by

$$\begin{aligned}
\mathcal{P}_1: \quad & \underset{\boldsymbol{W},\boldsymbol{V},\boldsymbol{F},\boldsymbol{p}}{\text{maximize}} \quad R_{tot}(\boldsymbol{W},\boldsymbol{V},\boldsymbol{F},\boldsymbol{p}) \qquad (19)\\
\text{subject to} \quad & C_1: |[\boldsymbol{v}_{k,u}]_i|^2 = 1/N_T, i = 1,2,\cdots,N_T;\\
& C_2: |[\boldsymbol{w}_{k,u}]_j|^2 = 1/N_R, j = 1,2,\cdots,N_R;\\
& C_3: \|\boldsymbol{\mathcal{V}}_k \boldsymbol{f}_{k,u}\|^2 = 1;\\
& C_4: \boldsymbol{V} = [\boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2, \cdots, \boldsymbol{\mathcal{V}}_K];\\
& C_5: \boldsymbol{F} = \text{diag}(\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2, \cdots, \boldsymbol{\mathcal{F}}_K);\\
& C_6: \max\{M_k\}_{k=1}^{K} \leq N_{RF};\\
& C_7: \sum_{k=1}^{K}\sum_{u=1}^{M_k} p_{k,u} \leq N_U;\\
& C_8: R_{k,u} \geq \lambda_{k,u},
\end{aligned}$$

where $k = 1,2,\cdots,K$ and $u = 1,2,\cdots,M_k$ in C_1, C_2 and C_3.

In problem $\mathcal{P}_1$, C_1 and C_2 confine the analog beamforming vectors to the phase-only structure in transmitter and receiver while C_3 ensures that each precoded signal is of unit power. Furthermore, C_4 and C_5 define the analog and digital precoder, respectively. C_6 constrains the maximum number of data-streams in each group to be within the number of RF chains. Finally, C_7 defines the downlink transmitted power constraint while C_8 guarantees the minimal data rate $\lambda_{k,u}$ for each user.

The problem $\mathcal{P}_1$ is challenging due to its non-convex and combinatorial nature. Thus, it is analytically intractable to derive a closed-form optimal solution. Instead, we consider a two-stage suboptimal solution: In the first stage, we focus on the analog and digital precoder design while assuming uniform power allocation; After fixing the analog and digital precoders, we derive the QoS-aware optimal power allocation in the second stage.

3. Proposed Block Hybrid Beamforming for RF Chains Reduction

In this section, we will first ignore the constraints C_7 and C_8 in $\mathcal{P}_1$ by uniformly allocating the power to each user while assuming that the user grouping is given.

3.1. Analog Beamforming Design

We begin with the analog beamforming design for both transmitter and receiver. It is well known that distinct array response vectors are asymptotically orthogonal as the number of antennas in an antenna array goes to infinity [1], i.e.,

$$\lim_{N\to+\infty} \boldsymbol{a}_T^H(\phi_{k,u}^t, \theta_{k,u}^t) \cdot \boldsymbol{a}_T(\phi_{\ell,v}^t, \theta_{\ell,v}^t) = \delta(k-\ell)\delta(u-v). \qquad (20)$$

However, since the antenna number is finite in practice, the residual interference must be considered in the analog precoding design. Recalling the channel model presented in Equation (8), we can asymptotically orthogonalize the transmitted signals by optimizing the design of $\boldsymbol{w}_{k,u}$ and $\boldsymbol{v}_{k,u}$:

$$\begin{aligned}\{\boldsymbol{w}_{k,u}^{*}, \boldsymbol{v}_{k,u}^{*}\} = & \arg\max_{\tilde{\boldsymbol{w}}_{k,u}, \tilde{\boldsymbol{v}}_{k,u}} \sum_{k=1}^{K} \sum_{u=1}^{M_k} \log_2\left(1 + \mathrm{SINR}\left(\tilde{\boldsymbol{w}}_{k,u}, \tilde{\boldsymbol{v}}_{k,u}\right)\right) \\ \text{subject to} \quad & \tilde{\boldsymbol{v}}_{k,u} \in \mathcal{A}_{k,u}^{T}; \\ & \tilde{\boldsymbol{w}}_{k,u} \in \mathcal{A}_{k,u}^{R}; \\ & \max\{M_k\}_{k=1}^{K} < N_{RF},\end{aligned} \tag{21}$$

where

$$\mathcal{A}_{k,u}^{T} = \left[\boldsymbol{a}_T(\phi_{k,u,1}^{t}, \theta_{k,u,1}^{t}), \cdots, \boldsymbol{a}_T(\phi_{k,u,L_{k,u}}^{t}, \theta_{k,u,L_{k,u}}^{t})\right], \tag{22a}$$

$$\mathcal{A}_{k,u}^{R} = \left[\boldsymbol{a}_R(\phi_{k,u,1}^{r}, \theta_{k,u,1}^{r}), \cdots, \boldsymbol{a}_R(\phi_{k,u,L_{k,u}}^{r}, \theta_{k,u,L_{k,u}}^{r})\right]. \tag{22b}$$

Furthermore, $\mathrm{SINR}_{k,u}$ is given by

$$\mathrm{SINR}\left(\tilde{\boldsymbol{w}}_{k,u}, \tilde{\boldsymbol{v}}_{k,u}\right) = \frac{|\tilde{\boldsymbol{w}}_{k,u}^{H} \boldsymbol{H}_{k,u} \tilde{\boldsymbol{v}}_{k,u}|^2}{\sum_{j=1, j\neq k}^{N_U} \|\tilde{\boldsymbol{w}}_{k,u}^{H} \boldsymbol{H}_{k,u} \tilde{\boldsymbol{\mathcal{V}}}_j\|^2 + \sum_{t\neq u}^{M_k} |\tilde{\boldsymbol{w}}_{k,u}^{H} \boldsymbol{H}_{k,u} \tilde{\boldsymbol{v}}_{k,t}|^2 + \frac{1}{\gamma}}, \tag{23}$$

with $\gamma = \frac{1}{\sigma_{k,u}^2}$. The optimal analog beamforming precoder can be straightforwardly found by exhaustively searching in the feasible sets of $\mathcal{A}_{k,u}^{T}$ and $\mathcal{A}_{k,u}^{R}$.

3.2. Digital Precoder Design

In this section, two digital precoding schemes are proposed to maximize the system sum-rate by minimizing the intra-group interference.

3.2.1. Block Zero-Forcing (Bzf) Scheme

In contrast to the conventional ZF hybrid beamforming scheme [28] that requires $N_U \leq N_{RF}$, zero-forcing digital precoding scheme is first proposed to transmit data-streams group by group. More specifically, the digital precoder for each block is designed as the inverse of the effective channel of the block:

$$\boldsymbol{\mathcal{F}}_k^{\mathrm{BZF}} = \boldsymbol{\mathcal{G}}_k^{H} (\boldsymbol{\mathcal{G}}_k \boldsymbol{\mathcal{G}}_k^{H})^{-1}, \tag{24}$$

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 constraint C_3 in $\mathcal{P}_1$, power normalization is performed on each $\boldsymbol{f}_{k,u}$ derived from $\boldsymbol{\mathcal{F}}_k^{\mathrm{BZF}} = \left[\boldsymbol{f}_{k,1}^{\mathrm{BZF}}, \boldsymbol{f}_{k,2}^{\mathrm{BZF}}, \cdots, \boldsymbol{f}_{k,M_k}^{\mathrm{BZF}}\right]$ as

$$\bar{\boldsymbol{f}}_{k,u}^{\mathrm{BZF}} = \frac{\boldsymbol{f}_{k,u}^{\mathrm{BZF}}}{\|\boldsymbol{\mathcal{V}}_k \cdot \boldsymbol{f}_{k,u}^{\mathrm{BZF}}\|}. \tag{25}$$

Subsequently, this scheme is referred to as the block zero-forcing (BZF) scheme. It is worth noting that BZF degenerates to [28] if $K = 1$, i.e., all users are grouped into one independent group. On the other hand, BZF becomes the analog-only BDMA if $K = N_U$, i.e., each user forms one group and only analog beamforming is performed.

3.2.2. Block SLNR Maximization (BSM) Scheme

Instead of the received interference elimination, we can alternatively devise the digital precoder to suppress the interference leakage by maximizing SLNR [29]. More specifically, we denote by $P_{k,u}^{\text{Desired}}$ the desired signal power transmitted to the u-th user in the k-th group

$$P_{k,u}^{\text{Desired}} = \gamma \left| \boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2. \tag{26}$$

Furthermore, if we define leakage signal as the transmitted signal that is intended to a specific user but leaked to other users, the leakage signal power from the u-th user in the k-th group can be given as

$$P_{k,u}^{\text{Leakage}} = \gamma \left(\sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{i=1}^{M_j} \left| \boldsymbol{g}_{j,i}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2 + \sum_{\substack{t=1 \\ t \neq u}}^{M_k} \left| \boldsymbol{g}_{k,t}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2 \right). \tag{27}$$

Finally, the SLNR for the u-th user in the k-th group can be formulated as

$$\Gamma_{k,u} = \frac{\left| \boldsymbol{g}_{k,u}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2}{\sum\limits_{\substack{j=1 \\ j \neq k}}^{K} \sum\limits_{i=1}^{M_j} \left| \boldsymbol{g}_{j,i}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2 + \sum\limits_{\substack{t=1 \\ t \neq u}}^{M_k} \left| \boldsymbol{g}_{k,t}^{(k)H} \boldsymbol{f}_{k,u}^{\text{BSM}} \right|^2 + \frac{1}{\gamma}}. \tag{28}$$

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 [29]:

$$\boldsymbol{R}_{k,u}^{\text{Leakage}} = \left(\frac{1}{\gamma} \boldsymbol{I}_{N_{RF}} + \boldsymbol{Q}_{k,u} \right)^{-1} \boldsymbol{g}_{k,u}^{(k)} \boldsymbol{g}_{k,u}^{(k)H}, \tag{29}$$

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

$$\boldsymbol{Q}_{k,u} = \sum_{\substack{j=1 \\ j \neq k}}^{K} \sum_{i=1}^{M_j} \boldsymbol{g}_{j,i}^{(k)} \boldsymbol{g}_{j,i}^{(k)H} + \sum_{\substack{t=1 \\ t \neq u}}^{M_k} \boldsymbol{g}_{k,t}^{(k)} \boldsymbol{g}_{k,t}^{(k)H}. \tag{30}$$

In contrast to the conventional hybrid precoding algorithms with complexity $\mathcal{O}(N_U^3)$, the proposed group precoding schemes can reduce the complexity to $\mathcal{O}(N_{RF}^3)$.

From the above derivation, it is apparent that the user grouping algorithm plays an important role on the amount of inter-group interference, and subsequently the system performance. Thus, a heuristic algorithm for user grouping is investigated in the next section.

4. User Grouping Algorithm

Since the intra-group interference is eliminated by the digital precoding, we will focus on using the user grouping to maximize the intra-group interference while minimizing the inter-group interference. More specifically, we propose to group N_U users into K groups with minimal inter-group interference. Since the total number of possible group combinations is large, a greedy grouping algorithm is proposed in Algorithm 1.

Algorithm 1 Greedy User Grouping Algorithm

Input:

$\mathcal{X}$: the universal group and user index set;
$\boldsymbol{a}_T(\phi_x^t, \theta_x^t)$: Array response vector of index x;
$\mathcal{I}_k = \emptyset$: the user index set for the k-th group;
$k = 1$: group 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:
    Solve the optimal analog precoder by Equation (21)
    Let $\boldsymbol{A}$ be the analog precoders of grouped users
    for x in $\mathcal{X}$ **do**
      Compute $S(x) = \|\boldsymbol{a}_T^H(\phi_x^t, \theta_x^t) \cdot \boldsymbol{A}\|^2$
    end for
    Find the user index x^* with *maximum* $S(x)$
    Update $\boldsymbol{A} \leftarrow x^*$, $\mathcal{I}_k \leftarrow x^*$ 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 $S(x) = \|\boldsymbol{a}_T^H(\phi_x^t, \theta_x^t) \cdot \boldsymbol{A}\|^2$
      end for
      Find the user index x^* with *minimum* $S(x)$
      Update $\boldsymbol{A} \leftarrow x^*$, $\mathcal{I}_k \leftarrow x^*$ and $\mathcal{X} \setminus x^*$
    end if
end while

In this algorithm, for $k = 1$, we first group users who cause most interference to each other into Group 1 detailed in Stage 1. This selection is motivated by the observation that most interference can be eliminated by the digital precoder applied among each group. When the size of Group 1 reaches the number of RF chains, the user whose array response vector is most orthogonal to Group 1 is selected as the first member of Group 2 as shown in Stage 2, which is designed to minimize the inter-group interference. This process repeats until all users are assigned to different groups.

It is worth noting that the grouping problem is NP-hard. The greedy algorithm is proposed to find a suboptimal partition with complexity $\mathcal{O}(N_U^2)$.

5. Performance Analysis

We first investigate the capacity for the conventional analog-only BDMA scheme.

$$\mathbb{E}[R_{k,u}] = \mathbb{E}\left[\log_2\left(1 + \mathrm{SINR}\right)\right], \tag{31}$$

$$= \mathbb{E}\left[\log_2\left(1 + \frac{|\boldsymbol{w}_{k,u}^{*H}\boldsymbol{H}_{k,u}\boldsymbol{v}_{k,u}^*|^2}{1/\gamma + I_{k,u}}\right)\right], \tag{32}$$

where $I_{k,u}$ is the received interference represented as

$$I_{k,u} = \sum_{j \neq k}^{K} \| \boldsymbol{w}_{k,u}^{*H} \boldsymbol{H}_{k,u} \boldsymbol{\mathcal{V}}_j \|^2 + \sum_{t \neq u}^{M_k} | \boldsymbol{w}_{k,u}^{*H} \boldsymbol{H}_{k,u} \boldsymbol{v}_{k,t} |^2. \tag{33}$$

Proposition 1. *If the optimal analog beamformers are designed as* $\boldsymbol{w}_{k,u}^{*} = \boldsymbol{a}_R(\phi_{k,u,l^*}^{r}, \theta_{k,u,l^*}^{r})$ *and* $\boldsymbol{v}_{k,u}^{*H} = \boldsymbol{a}_T^H(\phi_{k,u,l^*}^{t}, \theta_{k,u,l^*}^{t})$*, respectively, the following approximation holds:*

$$| \boldsymbol{w}_{k,u}^{*H} \boldsymbol{H}_{k,u} \boldsymbol{v}_{j,i} |^2 \approx | D_{k,u} \alpha_{k,u,l^*} \boldsymbol{v}_{k,u}^{*H} \boldsymbol{v}_{j,i} |^2. \tag{34}$$

The proof is given in Appendices A and B.

From Proposition 1, Equation (31) can be rewritten as

$$\mathbb{E}[R_{k,u}] = \mathbb{E}\left[\log_2 \left(1 + \frac{Z}{1/\gamma + Y} \right) \right], \tag{35}$$

where

$$Z = | \boldsymbol{w}_{k,u}^{*H} \boldsymbol{H}_{k,u} \boldsymbol{v}_{k,u}^{*} |^2, \tag{36}$$

$$\approx D_{k,u}^2 \alpha_{k,u,l^*}^2, \tag{37}$$

and

$$\begin{aligned} Y &\approx D_{k,u}^2 \alpha_{k,u,l^*}^2 \left(\sum_{j \neq k}^{K} \left\| \boldsymbol{v}_{k,u}^{*H} \boldsymbol{\mathcal{V}}_j \right\|^2 + \sum_{t \neq u}^{M_k} \left| \boldsymbol{v}_{k,u}^{*H} \boldsymbol{v}_{k,t} \right|^2 \right), \\ &\approx Z \left(\sum_{j \neq k}^{K} \left\| \boldsymbol{v}_{k,u}^{*H} \boldsymbol{\mathcal{V}}_j \right\|^2 + \sum_{t \neq u}^{M_k} \left| \boldsymbol{v}_{k,u}^{*H} \boldsymbol{v}_{k,t} \right|^2 \right). \end{aligned} \tag{38}$$

Capitalizing on the Extreme Value Theory [30,31], we can derive the cumulative distribution function (CDF) of Z as

$$F_Z(z) = (1 - e^{-\frac{z}{C}})^{L_{k,u}}, \tag{39}$$

where $C = 2N_T N_R / L_{k,u}$. The detailed derivation is shown in Appendix C.

The residual interference of distinct beams is negligible as compared to the desired signal. Thus, Y can be upper bounded by

$$Y \leq Z \cdot \mathbb{E}\left[\sum_{j \neq k}^{K} \left| \boldsymbol{v}_{k,u}^{*H} \boldsymbol{\mathcal{V}}_j \right|^2 + \sum_{t \neq u}^{M_k} \left| \boldsymbol{v}_{k,u}^{*H} \boldsymbol{v}_{k,t} \right|^2 \right], \tag{40}$$

$$\approx Z(N_U - 1) \cdot \mathbb{E}\left[| \boldsymbol{v}_{k,u}^{*H} \boldsymbol{v}_{j,i} |^2 \right], \tag{41}$$

$$= Z(N_U - 1)T, \tag{42}$$

where $0 \leq T \leq 1$ with T being the expected residual interference power between distinct beams. In our proposed system, the beams will be selected and grouped to reduce the residual interference. Clearly, $T = 0$ if the number of antennas goes to infinity or the steering vectors of different users are strictly orthogonal. In contrast, $T = 1$ if different users have same AoDs. The value of T can be numerically derived.

Finally, the CDF of the SINR lower bound can be given by

$$F_X(x) = (1 - e^{-\frac{x}{C\gamma(1 - T(N_U - 1)x)}})^{L_{k,u}}. \tag{43}$$

The detailed derivation can be found in Appendix D.

Using the CDF above, the lower and upper bounds of the sum-rate capacity can be derived as

$$\int_0^\infty \log_2(1+x)dF_X(x) \leq \mathbb{E}[R_{k,u}] \leq \int_0^\infty \log_2(1+z)dF_Z(z). \tag{44}$$

It is analytically intractable to obtain a closed-form solution to Equation (44). We will show the numerical results in simulation section.

Please note that the upper bound is achieved if the interference from other users can be eliminated. Furthermore, since the number of transmitter antennas is finite in practice, the analog beamforming vectors shown in Equations (21) and (23) inevitably incur residual inter-user interference. Therefore, digital precoders are required to further suppress the residual interference.

6. Proposed QoS-Aware Power Allocation Algorithm Based on D.C. Programming

For given analog and digital precoders, we investigate the QoS-aware power allocation $\boldsymbol{p}$ in $\mathcal{P}_1$ by using the D.C. programming technique in this section.

We begin with reformulating $\mathcal{P}_1$ as

$$\begin{aligned} &\underset{\boldsymbol{p}}{\text{maximize}} \quad \sum_{k=1}^{K}\sum_{u=1}^{M_k} R_{k,u}(\boldsymbol{p}) \\ \text{subject to} \quad &C_1: \sum_{k=1}^{K}\sum_{u=1}^{M_k} p_{k,u} \leq P; \\ &C_2: R_{k,u} \geq \lambda_{k,u}, \end{aligned} \tag{45}$$

Following the procedures in [32], the problem above can be cast as a D.C. programming problem:

$$\underset{\boldsymbol{p}}{\text{maximize}} \quad f(\boldsymbol{p}) - g(\boldsymbol{p}) \tag{46}$$

where

$$\begin{aligned} f(\boldsymbol{p}) =& \sum_{k=1}^{K}\sum_{u=1}^{M_k} \log_2\left(\sum_{k=1}^{K}\sum_{u=1}^{M_k} p_{k,u}\left|\boldsymbol{g}_{k,u}^{(k)H}\boldsymbol{f}_{k,u}\right|^2 + \sigma_{k,u}^2\right), \\ g(\boldsymbol{p}) =& \sum_{k=1}^{K}\sum_{u=1}^{M_k} \log_2\left(\sum_{\substack{j=1\\ j\neq k}}^{K}\sum_{t=1}^{M_j} p_{j,t}\left|\boldsymbol{g}_{k,u}^{(j)H}\boldsymbol{f}_{j,t}\right|^2 + \sigma_{k,u}^2\right). \end{aligned}$$

For given analog and digital precoders, both $f(\boldsymbol{p})$ and $g(\boldsymbol{p})$ are concave in $\boldsymbol{p}$, i.e., Equation (46) is a D.C. function. Starting from a feasible $\boldsymbol{p}^{(0)}$, the optimal $\boldsymbol{p}^{(n+1)}$ at the n-th iteration is generated as the optimal solution of a convex problem:

$$\max_{\boldsymbol{p}} f(\boldsymbol{p}) - g(\boldsymbol{p}^n) - \langle \bigtriangledown g(\boldsymbol{p}^{(n)}), \boldsymbol{p} - \boldsymbol{p}^{(n)}\rangle, \tag{47}$$

which can be efficiently solved by any existing convex programming software, such as CVX [33]. The computational complexity of Equation (47) is $\mathcal{O}(N_{RF}^3)$ in each iteration [32].

As $g(\boldsymbol{p}^n)$ is concave, its gradient $\bigtriangledown g(\boldsymbol{p}^{(n)})$ is also super-gradient:

$$\begin{aligned} &f(\boldsymbol{p}^{(n+1)}) - g(\boldsymbol{p}^{(n+1)}) \geq \\ &f(\boldsymbol{p}^{(n+1)}) - \left[g(\boldsymbol{p}^{(n)}) + \langle \bigtriangledown g(\boldsymbol{p}^{(n)}), \boldsymbol{p}^{(n+1)} - \boldsymbol{p}^{(n)}\rangle\right]. \end{aligned} \tag{48}$$

The proof is given in Appendix E.

Finally, since $\boldsymbol{p}^{(n+1)}$ is the solution to Equation (47), it follows that

$$f(\boldsymbol{p}^{(n+1)}) - g(\boldsymbol{p}^{(n)}) - \langle \nabla g(\boldsymbol{p}^{(n)}), \boldsymbol{p}^{(n+1)} - \boldsymbol{p}^{(n)} \rangle, \tag{49a}$$

$$\geq \quad f(\boldsymbol{p}^{(n)}) - g(\boldsymbol{p}^{(n)}) - \langle \nabla g(\boldsymbol{p}^{(n)}), \boldsymbol{p}^{(n)} - \boldsymbol{p}^{(n)} \rangle, \tag{49b}$$

$$= \quad f(\boldsymbol{p}^{(n)}) - g(\boldsymbol{p}^{(n)}). \tag{49c}$$

Therefore, the $(n+1)$-th solution is always better than the previous one. The iterative process terminates after $|f(p^{(n+1)}) - g(p^{(n+1)}) - (f(p^{(n)}) - g(p^{(n)}))| \leq \epsilon$ is achieved with a pre-defined threshold $\epsilon > 0$.

7. Simulation Results

In this section, we will use computer simulation to evaluate 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 an 8×8 UPA (i.e., $N_R = 64$). The number of paths is set to $L_{k,u} = 4$ and the additive Gaussian noise power $\sigma^2_{k,u} = -30$ dBm for each user. We consider the azimuth AoA/AoD's uniformly distributed over $[0, 2\pi]$ while the elevation AoA/AoD's uniformly distributed over $[-\pi/2, \pi/2]$, respectively. For each computer experiment, we compute the average over 100 realizations.

In Figure 2, we first set $K = 1$, i.e., no grouping. As a result, 16 RF chains are required to support 16 data-streams. As shown in Figure 2, BZF slightly outperforms BSM as it can eliminate more multi-user interference even in multipath environment. It is observed that even in the high SNR regime, BDMA suffers from inter-beam interference and has poor performance.

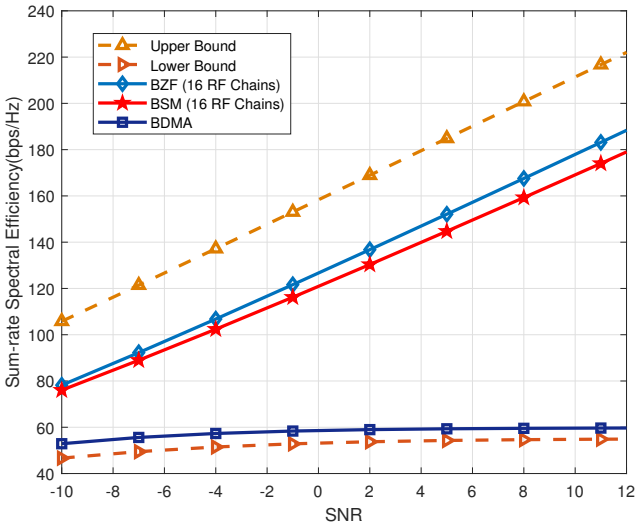

Figure 2. Performance with $K = 1$ (no grouping) over multipath channels.

Next, we evaluate the two proposed BD precoding schemes with $K = 2$ groups and 8 RF chains. The 16 users are grouped into $K = 2$ groups. The curves labeled as "BZF" and "BSM" stands for the proposed BD precoding schemes where only 8 RF chains are used to transmit 16 data-streams. It is observed that BZF and BSM have comparable performance. Furthermore, the curve labeled as "Conventional Hybrid BF (8 RF Chains)" is the sum-rate for the conventional hybrid beamforming system with 8 RF chains serving 8 users. Finally, BDMA is the analog-only precoding system that has the worst performance. Inspection of Figure 3 reveals that the proposed BZF and BSM have much better sum-rate performance than the conventional hybrid precoding algorithm over the SNR

range $[-10, 10]$ dB. When the SNR is larger than 12 dB, the system becomes interference-limited. Thus, the performance of BZF and BSM tends to saturate beyond this point.

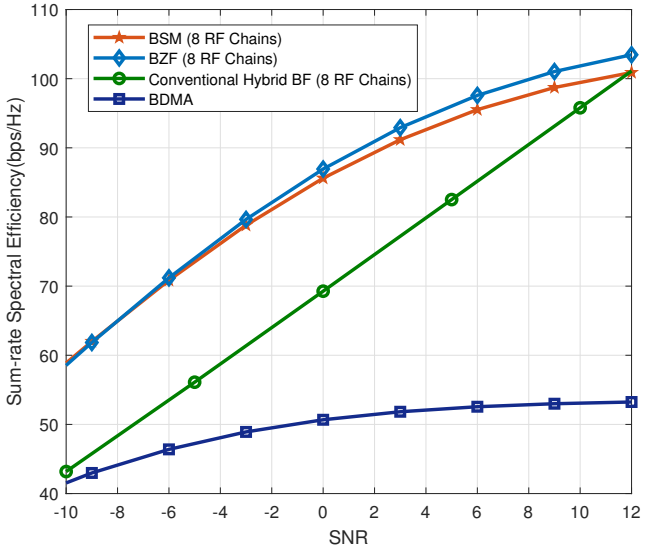

Figure 3. Performance with $K = 2$ groups over multipath channels with $N_u = 16$ users and 8 RF chains.

In Figure 4, we investigate the sum-rate capacity improvement as a function of the number of RF chains while fixing the SNR at 5 dB. The upper bound is the conventional ZF precoding system with 16 RF chains for 16 users. Interestingly, the performance improvement generated by an additional RF chain increases only marginally as the number of RF chains grows from eight to 14.

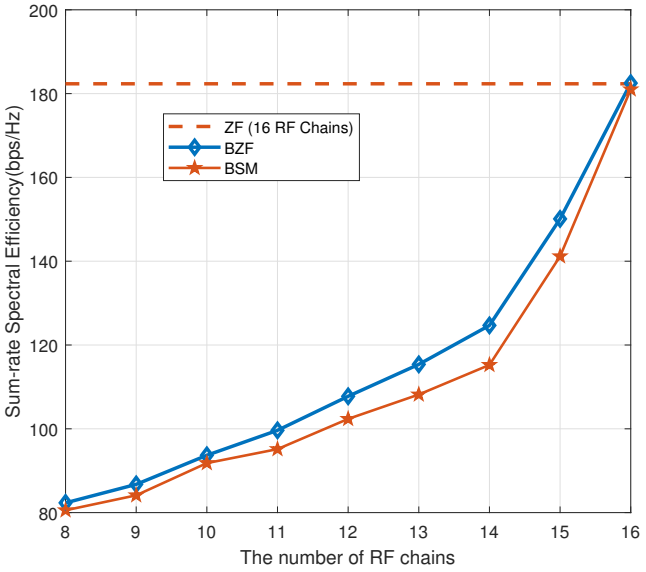

Figure 4. Sum-rate capacity improvement as a function of the number of RF chains.

Next, we vary the number of groups while fixing the total number of users at $N_u = 16$ and SNR $= 5$ dB. Figure 5 shows that BZF and BSM are lower bounded by 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 the conventional BDMA.

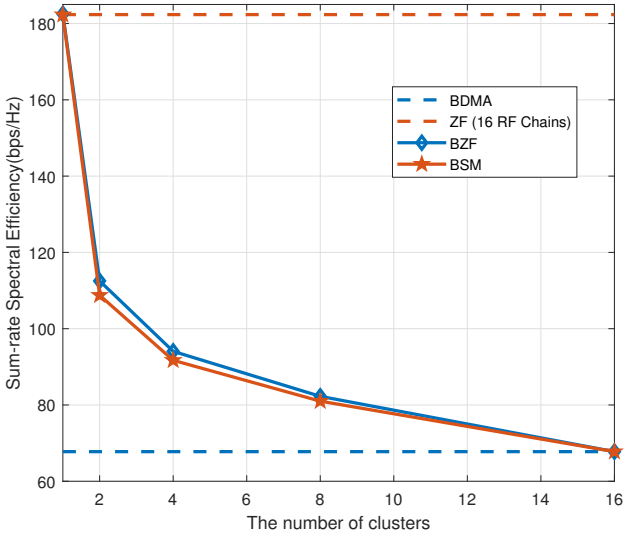

Figure 5. Sum-rate capacity as a function of the number of groups.

We then investigate the sum-rate performance as the number of transmit antennas increases. Figure 6 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-group interference is asymptotically removed as indicated in Equation (20).

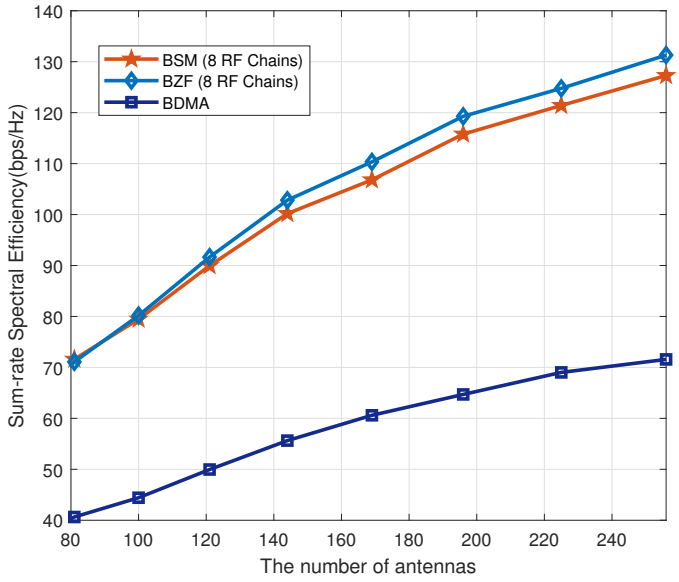

Figure 6. Sum-rate capacity as a function of the number of transmit antennas.

Finally, we evaluate the performance of the power allocation generated with the D.C. programming technique. We assume that the minimum QoS threshold for each user is set to 3 bps/Hz. Figures 7 and 8 show the performance achieved by our proposed QoS-aware power allocation algorithm. The curve labeled as "Water-filling Power Allocation" is obtained by allocating user power via the water-filling algorithm without taking into account the QoS requirement. The curve labeled as "QoS-Aware Power Allocation" shows the performance of the proposed power allocation algorithm. Compared to the curve labeled as "Uniform Power Allocation", the proposed algorithm has demonstrated significant advantages in terms of the sum-rate capacity. Furthermore, Figure 8 depicts

the CDF of the user data rate. It is evident that all users served by the QoS-aware power allocation satisfy the minimum QoS requirement (i.e., 3 bps/Hz). In contrast, the water-filling-based power algorithm suffers from an outage rate of about 20% where outage is defined as the user data rate being below the minimum required data rate.

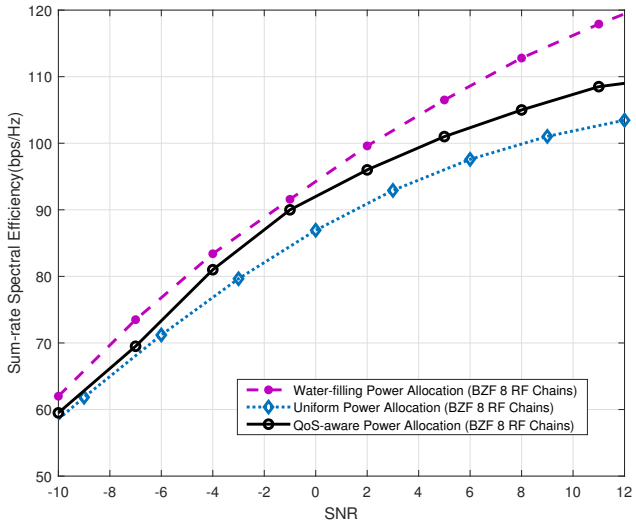

Figure 7. Performance achieved by the QoS-aware power allocation algorithm with $K = 2$.

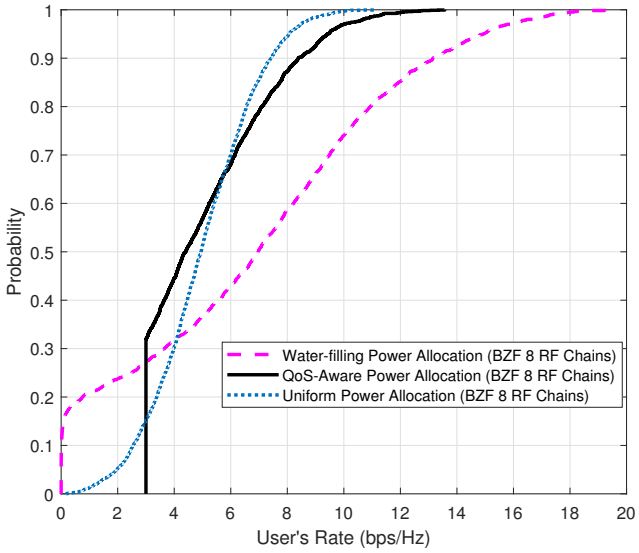

Figure 8. The CDF of user's rate comparison for QoS-ware power allocation schemes.

8. Conclusions

In this paper, we have developed block diagonal hybrid precoding schemes with optimized power allocation for mmWave massive MIMO systems by jointly performing hybrid analog-digital precoding and user-beam grouping. The proposed system requires fewer RF chains as compared to the conventional hybrid precoding systems by digitally precoding data-streams group by group. Although the intra-group interference is eliminated by the digital precoding, a greedy grouping algorithm has been derived to minimize the inter-group interference by carefully grouping users with orthogonal beams to different groups. Furthermore, two digital precoding schemes have been proposed to suppress the intra-group interference based on SINR and SLNR, respectively. In addition, the upper

and lower bounds of the system sum-rate capacity have been derived based on the multipath channel model. Finally, QoS-aware power allocation has been proposed by using the D.C. programming technique. Simulation results have demonstrated the good performance of the proposed grouped BDMA block diagonal hybrid precoding system.

Author Contributions: Conceptualization, G.N. and Q.C.; methodology, G.N. and Q.C.; software, G.N.; validation, G.N., Q.C. and M.P.; formal analysis, G.N.; investigation, G.N.; resources, G.N.; data curation, G.N.; writing—original draft preparation, G.N.; writing—review and editing, G.N.; visualization, G.N.; supervision, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported, in part, by the Shenzhen Science and Technology Innovation Committee under Grant No. ZDSYS20170725140921348 and JCYJ20190813170803617 and by the National Natural Science Foundation of China under Grant No. 61731018. The APC was funded by Shenzhen Research Institute of Big Data.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Proposition 1

Proof. First, if $\boldsymbol{v}_{j,i} = \boldsymbol{v}_{k,u}^* = \boldsymbol{a}_T^H(\phi_{k,u,l^*}^t, \theta_{k,u,l^*}^t)$, we show that the signal power received from other paths is much smaller than that from the optimal path,

$$\begin{aligned} &|\boldsymbol{w}_{k,u}^{*H}\boldsymbol{H}_{k,u}\boldsymbol{v}_{j,i}|^2 \\ =&\left|D_{k,u}\boldsymbol{w}_{k,u}^{*H}\sum_{l=1}^{L_{k,u}}\alpha_{k,u,l}\boldsymbol{a}_R(\phi_{k,u,l}^r,\theta_{k,u,l}^r)\boldsymbol{a}_T^H(\phi_{k,u,l}^t,\theta_{k,u,l}^t)\boldsymbol{v}_{k,u}^*\right|^2, \\ =&D_{k,u}^2\alpha_{k,u,l^*}^2|\boldsymbol{v}_{k,u}^{*H}\boldsymbol{v}_{k,u}^* + I_1|^2, \end{aligned} \tag{A1}$$

where

$$\begin{aligned} &\boldsymbol{v}_{k,u}^{*H}\boldsymbol{v}_{k,u}^* = 1.0, \\ &I_1 = \sum_{l\neq l^*}^{L_{k,u}}\frac{\alpha_{k,u,l}}{\alpha_{k,u,l^*}}\boldsymbol{w}_{k,u}^{*H}\boldsymbol{a}_R(\phi_{k,u,l}^r,\theta_{k,u,l}^r)\boldsymbol{a}_T^H(\phi_{k,u,l}^t,\theta_{k,u,l}^t)\boldsymbol{v}_{k,u}^*. \end{aligned} \tag{A2}$$

It can be easily shown that $I_1 \ll 1.0$, which means the desired signal in Equation (A1) can be well approximated by $|D_{k,u}\alpha_{k,u,l^*}\boldsymbol{v}_{k,u}^{*H}\boldsymbol{v}_{k,u}^*|^2$. In other words, the vast majority of desired signal is received from the selected optimal path.

Similarly, for $\boldsymbol{v}_{j,i} \neq \boldsymbol{a}_T^H(\phi_{k,u,l^*}^t, \theta_{k,u,l^*}^t)$, the interference from another user is given by

$$\begin{aligned} &|\boldsymbol{w}_{k,u}^{*H}\boldsymbol{H}_{k,u}\boldsymbol{v}_{j,i}|^2 \\ =&\left|D_{k,u}\boldsymbol{w}_{k,u}^{H}\sum_{l=1}^{L_{k,u}}\alpha_{k,u,l}\boldsymbol{a}_R(\phi_{k,u,l}^r,\theta_{k,u,l}^r)\boldsymbol{a}_T^H(\phi_{k,u,l}^t,\theta_{k,u,l}^t)\boldsymbol{v}_{j,i}\right|^2, \\ =&D_{k,u}^2|I_2 + I_3|^2, \end{aligned} \tag{A3}$$

where

$$\begin{aligned} I_2 =&\alpha_{k,u,l}\boldsymbol{a}_T^H(\phi_{k,u,l}^t,\theta_{k,u,l}^t)\boldsymbol{v}_{j,i}, \\ I_3 =&\sum_{l\neq l^*}^{L_{k,u}}\alpha_{k,u,l}\boldsymbol{w}_{k,u}^{*H}\boldsymbol{a}_R(\phi_{k,u,l}^r,\theta_{k,u,l}^r)\boldsymbol{a}_T^H(\phi_{k,u,l}^t,\theta_{k,u,l}^t)\boldsymbol{v}_{j,i}. \end{aligned}$$

Since analog beamformers are distinct for different receivers, we have $I_3 \ll I_2$ by Equation (20). Finally, it is straightforward to derive

$$\begin{aligned}
&|\boldsymbol{w}_{k,u}^{*H}\boldsymbol{H}_{k,u}\boldsymbol{v}_{j,i}|^2 \\
=&|\boldsymbol{w}_{k,u}^{*H}D_{k,u}\sum_{l=1}^{L_{k,u}}\alpha_{k,u,l}\boldsymbol{a}_R(\phi_{k,u,l}^r,\theta_{k,u,l}^r)\boldsymbol{a}_T^H(\phi_{k,u,l}^t,\theta_{k,u,l}^t)\boldsymbol{v}_{j,i}|^2, \\
\approx&|\boldsymbol{w}_{k,u}^{*H}D_{k,u}\alpha_{k,u,l^*}\boldsymbol{a}_R(\phi_{k,u,l^*}^r,\theta_{k,u,l^*}^r)\boldsymbol{a}_T^H(\phi_{k,u,l^*}^t,\theta_{k,u,l^*}^t)\boldsymbol{v}_{j,i}|^2, \\
=&|D_{k,u}\alpha_{k,u,l^*}\boldsymbol{a}_T^H(\phi_{k,u,l^*}^t,\theta_{k,u,l^*}^t)\boldsymbol{v}_{j,i}|^2, \\
=&|D_{k,u}\alpha_{k,u,l^*}\boldsymbol{v}_{k,u}^{*H}\boldsymbol{v}_{j,i}|^2. \qquad \text{(A4)}
\end{aligned}$$

A numerical example is given in Appendix B. □

Appendix B. A Numerical Example for Proposition 1

A numerical example is given below. We set $L_{k,u} = 2$ with $\phi_{k,u,1} = \phi_{k,u,1} = \pi/2$, $\theta_{k,u,1} = \pi/3$ and $\theta_{k,u,2} = \pi/5$. Recalling the given UPA array response vector, the number of antennas in transmitter is $N_T = 144$ and $N_R = 64$ in receiver. The optimal analog precoder in transmitter and receiver is given by $\boldsymbol{v}_{k,u}^* = \boldsymbol{a}_T(\phi_{k,u,1},\theta_{k,u,1})$ and $\boldsymbol{w}_{k,u}^* = \boldsymbol{a}_R(\phi_{k,u,1},\theta_{k,u,1})$, respectively. Furthermore, we examine another receiver whose optimal analog precoder is given by $\boldsymbol{v}_{j,i}^* = \boldsymbol{a}_T(\phi_{k,u,1},\theta_{j,i,1})$ with $\phi_{j,i,1} = \pi/2$ and $\theta_{j,i,1} = \pi/4$. The complex path gains are $\alpha_{k,u,1} = \alpha_{k,u,2} = 1$.

Figure A1 shows that I_1 is much less than 1.0, which confirms that the desired signal in Equation (A1) can be well approximated by $|\boldsymbol{w}_{k,u}^{*H}\boldsymbol{a}_R(\phi_{k,u,1},\theta_{k,u,1})\boldsymbol{a}_T^H(\phi_{k,u,1},\theta_{k,u,1})\boldsymbol{v}_{k,u}^*|^2$. Similarly, from Figure A2, it can be seen that the second term I_3 is much smaller than I_2. Thus, the interference term can be approximated by $|\boldsymbol{w}_{j,i}^{*H}\boldsymbol{a}_R(\phi_{k,u,1},\theta_{k,u,1})\boldsymbol{a}_T^H(\phi_{k,u,1},\theta_{k,u,1})\boldsymbol{v}_{j,i}^*|^2$.

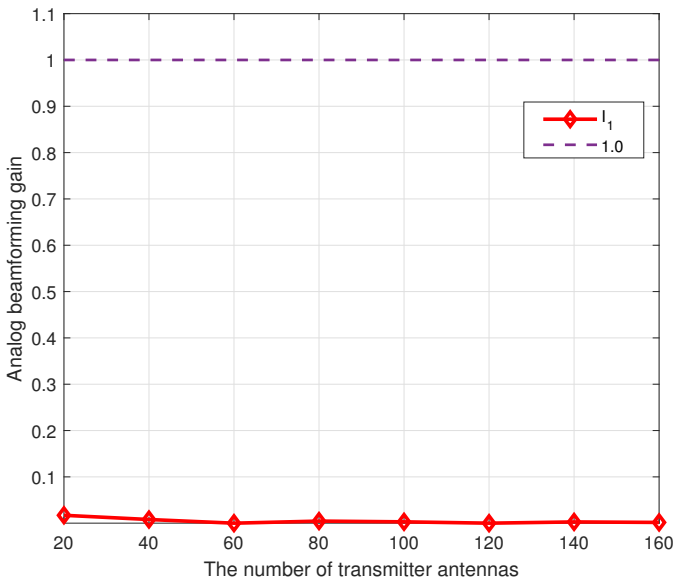

Figure A1. Desired signal from different paths.

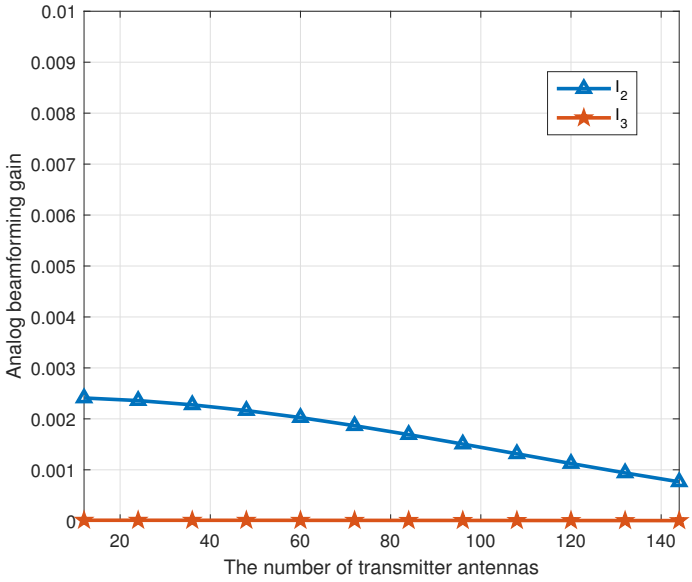

Figure A2. Other users' interference from different paths.

Appendix C. The Derivation for Equation (39)

From Proposition 1, Z can be approximated as

$$\begin{aligned} Z &= \max |\boldsymbol{w}_{k,u}^{*H} \boldsymbol{H}_{k,u} \boldsymbol{v}_{k,u}^{*}|^2, \\ &\approx \max |D_{k,u}\alpha_{k,u,l} \boldsymbol{w}_{k,u}^{*H} \boldsymbol{a}_R(\theta_{k,u,l}) \boldsymbol{a}_T^H(\theta_{k,u,l}) \boldsymbol{v}_{k,u}^{*}|^2, \\ &= \max_{0<l<L_{k,u}} |D_{k,u}\alpha_{k,u,l}|^2. \end{aligned} \tag{A5}$$

Since $\alpha_{k,u,l}$ has the distribution of $\mathcal{X}^2(2)$, the CDF of Z can be derived in a straightforward manner [31]:

$$F_Z(z) = (1 - e^{-\frac{z}{\mathbb{C}}})^{L_{k,u}}. \tag{A6}$$

Appendix D. The Derivation for Equation (43)

Recalling Equation (35), the SINR can be represented as

$$\mathrm{X} = \frac{Z}{1/\gamma + Y}. \tag{A7}$$

To derive the CDF of SINR, we must first calculate the CDF of Y. From Proposition 1, Y can be represented as

$$\begin{aligned} Y \approx & Z\mathbb{E}\left[\sum_{j\neq k}^{K} \left|\boldsymbol{v}_{k,u}^{*H} \boldsymbol{\mathcal{V}}_j\right|^2 + \sum_{t\neq u}^{M_k} \left|\boldsymbol{v}_{k,u}^{*H} \boldsymbol{v}_{k,t}\right|^2\right], \\ \approx & Z(N_U - 1)\mathbb{E}\left[\left|\boldsymbol{v}_{k,u}^{*H} \boldsymbol{v}_{j,i}\right|^2\right]. \end{aligned} \tag{A8}$$

As the AoDs and AoAs are i.i.d for UPA antennas, the expectation can be calculated as

$$\begin{aligned}&\mathbb{E}\left[\left|\boldsymbol{v}_{k,u}^{*H}\boldsymbol{v}_{j,i}\right|^2\right] \triangleq T\\ =&\frac{1}{4\pi^2 N_T^2}\int_0^{2\pi}\cdots\int_0^{2\pi}\left|B(\theta_{k,u}\theta_{j,i}\phi_{k,u}\phi_{j,i})\right|^2 d\theta_{k,u}\cdots d\phi_{j,i},\end{aligned} \tag{A9}$$

where $B(\cdot)$ is given by

$$\begin{aligned}&B(\theta_{k,u}\theta_{j,i}\phi_{k,u}\phi_{j,i})\\ =&\sum_{q=0}^{Q-1}\sum_{p=0}^{P-1}e^{j\kappa d\left[p\left(\sin\phi_{k,u}\sin\theta_{k,u}-\sin\phi_{j,i}\sin\theta_{j,i}\right)+q\left(\cos\theta_{k,u}-\cos\theta_{j,i}\right)\right]},\\ =&\frac{\left(1-e^{j\kappa dn\left(\sin\phi_{k,u}\sin\theta_{k,u}-\sin\phi_{j,i}\sin\theta_{j,i}\right)}\right)^P}{1-e^{j\kappa dn\left(\sin\phi_{k,u}\sin\theta_{k,u}-\sin\phi_{j,i}\sin\theta_{j,i}\right)}}\times\frac{\left(1-e^{j\kappa dn\left(\cos\theta_{k,u}-\cos\theta_{j,i}\right)}\right)^Q}{1-e^{j\kappa dn\left(\cos\theta_{k,u}-\cos\theta_{j,i}\right)}}.\end{aligned} \tag{A10}$$

As shown in Figure A3, the value of T can be numerically estimated. As the number of antennas increases, the value of T decreases gradually, which confirms Equation (20).

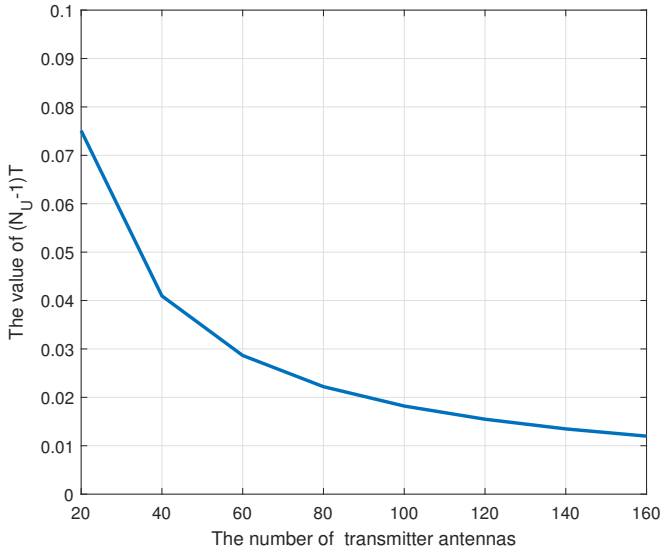

Figure A3. The value of $(N_U-1)T$ for different number of transmitter antennas.

For a given CDF of Z in Equation (39), the CDF of SINR can be computed as

$$\begin{aligned}F_X(x) &= P(X\le x)\\ &= P\left(\frac{Z}{1/\gamma+Z(N_U-1)T}\le x\right),\\ &= P\left(Z\le\frac{x}{\gamma(1-x(N_U-1)T)}\right),\\ &= \left(1-e^{-\frac{x}{C\gamma(1-T(N_U-1)x)}}\right)^{L_{k,u}},\end{aligned} \tag{A11}$$

where we have $\gamma>0$ in the derivation above.

Appendix E. Proof for Equation (48)

Suppose $f(\boldsymbol{p})$ is a concave function on a convex neighborhood $\mathcal{C}$ and differentiable at $\boldsymbol{p}$. Then, for every $\boldsymbol{y} \in \mathcal{C}$, we have the following inequality based on the definition of concavity:

$$\begin{aligned} & f((1-\lambda)\boldsymbol{p} + \lambda \boldsymbol{y}), \\ =& f\left(\boldsymbol{p} + \lambda(\boldsymbol{y} - \boldsymbol{p})\right), \\ \geq& f(\boldsymbol{p}) + \lambda\left(f(\boldsymbol{y}) - f(\boldsymbol{p})\right), \end{aligned} \tag{A12}$$

where $0 < \lambda \leq 1$.

Rearranging the terms and dividing both sides by λ, we have

$$\frac{f\left(\boldsymbol{p} + \lambda(\boldsymbol{y} - \boldsymbol{p})\right) - f(\boldsymbol{p})}{\lambda} \geq f(\boldsymbol{y}) - f(\boldsymbol{p}). \tag{A13}$$

Letting $\lambda \to 0$, it can be shown that the left hand side of Inequality (A13) converges to $f'(\boldsymbol{p}) \cdot (\boldsymbol{y} - \boldsymbol{p})$. Finally, we have Equation (48) as:

$$f(\boldsymbol{p}) + f'(\boldsymbol{p}) \cdot (\boldsymbol{y} - \boldsymbol{p}) \geq f(\boldsymbol{y}). \tag{A14}$$

References

1. Sun, C.; Gao, X.; Jin, S.; Matthaiou, M.; Ding, Z.; Xiao, C. Beam Division Multiple Access Transmission for massive MIMO communications. *IEEE Trans. Commun.* **2015**, *63*, 2170–2184. [CrossRef]
2. Jiang, Z.; Chen, S.; Zhou, S.; Niu, Z. Joint user scheduling and beam selection optimization for beam-based massive MIMO downlinks. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 2190–2204. [CrossRef]
3. Sun, C.; Gao, X.; Ding, Z. BDMA in multicell massive MIMO communications: Power allocation algorithms. *IEEE Trans. Signal Process.* **2017**, *65*, 2962–2974. [CrossRef]
4. Dalela, P.K.; Bhave, P.; Yadav, P.; Yadav, A.; Tyagi, V. Beam division multiple access (BDMA) and modulation formats for 5G: Heir of OFDM? In Proceedings of the 2018 International Conference on Information Networking (ICOIN), Chiang Mai, Thailand, 10–12 January 2018; pp. 450–455.
5. Liang, L.; Xu, W.; Dong, X. Low-complexity hybrid precoding in massive multiuser MIMO systems. *IEEE Wirel. Commun. Lett.* **2014**, *3*, 653–656. [CrossRef]
6. Molisch, A.F.; Win, M.Z.; Choi, Y.-S.; Winters, J.H. Capacity of MIMO systems with antenna selection. *IEEE Trans. Wirel. Commun.* **2005**, *4*, 1759–1772. [CrossRef]
7. Rangan, S.; Rappaport, T.S.; Erkip, E. Millimeter-wave cellular wireless networks: Potentials and challenges. *Proc. IEEE* **2014**, *102*, 366–385. [CrossRef]
8. Alkhateeb, A.; El Ayach, O.; Leus, G.; Heath, R.W. Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE J. Sel. Top. Signal Process.* **2014**, *8*, 831–846. [CrossRef]
9. Molisch, A.F.; Ratnam, V.V.; Han, S.; Li, Z.; Nguyen, S.L.H.; Li, L.; Haneda, K. Hybrid beamforming for massive MIMO: A survey. *IEEE Commun. Mag.* **2017**, *55*, 134–141. [CrossRef]
10. Li, Z.; Han, S.; Sangodoyin, S.; Wang, R.; Molisch, A.F. Joint optimization of hybrid beamforming for multi-user massive MIMO downlink. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 3600–3614. [CrossRef]
11. Zhang, X.; Molisch, A.F.; Kung, S.-Y. Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection. *IEEE Trans. Signal Process.* **2005**, *53*, 4091–4103. [CrossRef]
12. Ratnam, V.V.; Molisch, A.F.; Bursalioglu, O.Y.; Papadopoulos, H.C. Hybrid beamforming with selection for multiuser massive MIMO systems. *IEEE Trans. Signal Process.* **2018**, *66*, 4105–4120. [CrossRef]
13. Han, S.; Chih-Lin, I.; Xu, Z.; Wang, S. Reference signals design for hybrid analog and digital beamforming. *IEEE Commun. Lett.* **2014**, *18*, 1191–1193. [CrossRef]
14. Sun, S.; Rappaport, T.S.; Heath, R.W.; Nix, A.; Rangan, S. MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both? *IEEE Commun. Mag.* **2014**, *52*, 110–121. [CrossRef]

15. Rappaport, T.S.; Sun, S.; Mayzus, R.; Zhao, H.; Azar, Y.; Wang, K.; Wong, G.N.; Schulz, K.; Samimi, M.; Gutierrez, F. Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access* **2013**, *1*, 335–349. [CrossRef]
16. Spencer, Q.H.; Swindlehurst, A.L.; Haardt, M. Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels. *IEEE Trans. Signal Process.* **2004**, *52*, 461–471. [CrossRef]
17. Liu, A.; Lau, V. Phase only RF precoding for massive MIMO systems with limited RF chains. *IEEE Trans. Signal Process.* **2014**, *62*, 4505–4515. [CrossRef]
18. Nosrati, H.; Aboutanios, E.; Smith, D. Switch-based hybrid precoding in mmWave massive MIMO systems. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019; pp. 1–5.
19. Garcia, N.; Wymeersch, H.; Larsson, E.G. MIMO with more users than RF chains. *arXiv* **2017**, arXiv:1709.05200.
20. Palomar, D.P.; Lagunas, M.A.; Cioffi, J.M. Optimum linear joint transmit-receive processing for MIMO channels with QoS constraints. *IEEE Trans. Signal Process.* **2004**, *52*, 1179–1197. [CrossRef]
21. Yih, C.-H.; Geranotis, E. Centralized power allocation algorithms for OFDM cellular networks. In Proceedings of the IEEE Military Communications Conference (MILCOM 2003), Boston, MA, USA, 13–16 October 2003; pp. 1250–1255.
22. Sohrabi, F.; Yu, W. Hybrid digital and analog beamforming design for large-scale antenna arrays. *IEEE J. Sel. Top. Signal Process.* **2016**, *10*, 501–513. [CrossRef]
23. Koskie, S.; Gajic, Z. A Nash game algorithm for SIR-based power control in 3G wireless CDMA networks. *IEEE/ACM Trans. Netw.* **2005**, *13*, 1017–1026. [CrossRef]
24. Negro, F.; Cardone, M.; Ghauri, I.; Slock, D.T. SINR balancing and beamforming for the MISO interference channel. In Proceedings of the 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications, Toronto, ON, Canada, 11–14 September 2011; pp. 1552–1556.
25. Xie, M.; Lok, T.-M. SINR balancing via base station association, beamforming, and power control in downlink multicell MISO systems. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 1811–1821. [CrossRef]
26. Sifaou, H.; Kammoun, A.; Sanguinetti, L.; Debbah, M.; Alouini, M.-S. Max–min SINR in large-scale single-cell MU-MIMO: Asymptotic analysis and low-complexity transceivers. *IEEE Trans. Signal Process.* **2017**, *65*, 1841–1854. [CrossRef]
27. Lee, W.; Valdes-Garcia, A. Continuous true-time delay phase shifter using distributed inductive and capacitive miller effect. *IEEE Trans. Microw. Theory Tech.* **2019**, *67*, 3053–3063. [CrossRef]
28. El Ayach, O.; Rajagopal, S.; Abu-Surra, S.; Pi, Z.; Heath, R.W. Spatially sparse precoding in millimeter wave MIMO systems. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 1499–1513. [CrossRef]
29. Wang, J.; Jin, S.; Gao, X.; Wong, K.-K.; Au, E. Statistical eigenmode-based SDMA for two-user downlink. *IEEE Trans. Signal Process.* **2012**, *60*, 5371–5383. [CrossRef]
30. de Haan, L.; Ferreira, A. *Extreme Value Theory: An Introduction*, 1st ed.; Springer: New York, NY, USA, 2010.
31. Sharif, M.; Hassibi, B. On the capacity of MIMO broadcast channels with partial side information. *IEEE Trans. Inf. Theory* **2005**, *51*, 506–522. [CrossRef]
32. Kha, H.H.; Tuan, H.D.; Nguyen, H.H. Fast global optimal power allocation in wireless networks by local DC programming. *IEEE Trans. Wirel. Commun.* **2012**, *11*, 510–515. [CrossRef]
33. Grant, M.; Boyd, S. CVX: Matlab Software for Disciplined Convex Programming, Version 2.1. Available online: http://cvxr.com/cvx (accessed on 15 March 2014) .