A General Hybrid Precoding Method for mmWave Massive MIMO Systems

Yi XIE [1], *Bo LI* [1], *Zhongjiang YAN* [1], *Jiancun FAN* [2], *Mao YANG* [1]

[1] School of Electronics and Information, Northwestern Polytechnical University, 710072 Xi'an, Shaanxi, China
[2] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

xieyi@mail.nwpu.edu.cn, libo.npu@nwpu.edu.cn, zhjyan@nwpu.edu.cn, fanjc@xjtu.edu.cn, yangmao@nwpu.edu.cn

Submitted November 11, 2018 / Accepted March 3, 2019

Abstract. *Recently, hybrid precoding architectures have been proposed for the purpose of practical implementation of massive Multiple-Input Multiple-Output (MIMO) systems in the Fifth Generation (5G) networks. In this paper, a general precoding method is investigated for Millimeter Wave (mmWave) multi-user systems, which is composed of the designs in analog Radio Frequency (RF) and digital baseband matrices. In the general hybrid architecture, the analog part is constituted of independent analog sub-arrays with full connection inside. The analog precoding matrix is considered by maximizing Signal-to-Leakage-plus-Noise Ratio (SLNR) with only the long-term statistics of user groups. Due to the constant module constraint of RF chains, a supplemental matrix is introduced to reduce the performance loss. The digital precoding matrix performs Regularized Zero-Forcing (RZF) with the reduced amount of effective channels. Finally, simulation results demonstrate the performance improvement of the proposed precoding method. Meanwhile, trade-off between the performance and the complexity is handled well by the proposed method.*

Keywords

Massive MIMO, hybrid precoding architectures, beamforming, millimeter wave

1. Introduction

In order to meet dramatically increasing requirements of spectral efficiency in the fifth generation (5G) systems, massive multiple-input-multiple-output (MIMO) has been well studied as a promising technology. Massive MIMO systems is based on a large size antenna array at the base station (BS). In this way, signal energy can be focused into small zones in the direction of the desired users by beamforming techniques [1], [2]. In traditional MIMO systems, every antenna is linked to a radio frequency (RF) chain for the fully utilized degrees of freedom (DoF) and then simultaneously served multiple users. However, it is too difficult for such a great many RF chains to be installed at the physically constrained transceivers because of the limited space, the unaffordable cost and the tremendous energy consumption [3]. Particularly, in millimeter wave (mmWave) massive MIMO systems, it is impossible to have one RF chain for each antenna. Therefore, the hybrid analog and digital array architectures are considered as potential solutions.

To exploit the full benefits of multiple antennas in the hybrid architectures, channel state information (CSI) is usually necessary in the BS side. Unfortunately, huge overhead for full CSI acquisition is hardly possible for both frequency division duplex (FDD) and time division duplex (TDD) massive MIMO systems, even with channel estimation and channel reciprocity, respectively. Consequently, low complexity hybrid precoding is encouraged in the hybrid architecture for massive MIMO systems. The long-term statistics, for example the spatial covariance properties, are often applied for the reduction of the complexity in CSI acquisition. That kind of properties remain the same for a relatively long time and can be obtained more infrequently compared to the instantaneous CSI. In addition, users who share the common spatial scatterers are divided into a group. Then the common parameters of the group can be obtained with lower overhead. In the hybrid precoding, analog precoding is performed in the light of the long-term statistics of channels while the digital precoding is achieved with a much-reduced amount of instantaneous CSI.

Two start-of-art hybrid precoding architectures, which are named fully-connected architecture and sub-connected architecture, are shown in Fig. 1. The performance of fully-connected architecture in Fig. 1(a) with reduced CSI feedback has been investigated in [4–6] to serve multiple users. The analog part is for dimension reduction and is determined by channel covariance matrix while the digital part is designed by the traditional MIMO precoding methods. The sub-connected hybrid architecture in Fig. 1(b) on the other hand can reduce the hardware complexity in [7], [8], but is usually based on full CSI. Furthermore, a more general hybrid architecture is given in [9–11], aimed at striking balance between complexity and performance. However, full CSI is required and also limits its application in the actual network.

The related work is list on Tab. 1. In summary, there are still challenges need to be solved: 1) A flexible precoding method for the general hybrid architecture; 2) mmWave frequency band and multi-user interference elimination should be considered for the future networks; 3) Reduced amount of CSI is important for the system with large scale antennas. Therefore, for mmWave massive MIMO systems, a more general hybrid precoding with multi-users should be investigated.

In this paper, the goal is to provide a flexible precoding method with reduced CSI for the general hybrid architecture in multi-user massive MIMO systems. Based on long-term channel and interference statistics, the analog part is designed by maximizing signal-to-leakage-plus-noise ratio (SLNR). Afterwards, a constant module matrix and a supplemental matrix are extracted iteratively from the optimal result. The former one is set as the analog precoding matrix due to the phase shifter constraint while the latter one is added ahead of the digital part. In addition, the digital precoding matrix is formulated depends on the instantaneous CSI with low-dimension. The contributions are summarized as follows.

- To reduce the complexity of the analog part, the long-term channel and interference statistics are used to design the analog precoding, which is unchanged for a long time and can be obtained through previous channel property. The SLNR maximization is used to jointly consider the effects of signal and interference between different user groups.

- In the digital part, a supplemental matrix is added to each sub-array. It is to reduce the performance degradation caused by constant module constraint in RF chains. Then the digital part is performed with reduced CSI, which needs overhead in TDD and FDD systems.

- From the simulations, the proposed precoding method has performance enhancement compared to other work. The complexity and performance strike balances compared to the fixed hybrid architectures.

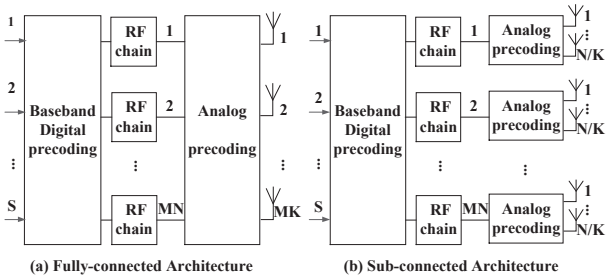

Fig. 1. Special hybrid precoding architectures.

The structure of the rest of this paper is organized as follows. The system model is formulated in Sec. 2. In Sec. 3, hybrid precoding design is described in detail, as well as the complexity comparison for the architecture. Simulation results are presented in Sec. 4 to demonstrate the performance of the proposed hybrid precoding method. Finally, conclusions are summarized in Sec. 5.

2. System Model

In this section, a multi-user massive MIMO system is considered with hybrid precoding at the BS, as shown in Fig. 2. M sub-arrays are equipped on the BS side, each with K antenna elements, N RF chains ($N \leq K$) and NK phase shifts (PSs). Then there are total NM RF chains, KM antenna elements, and NKM PSs. The fully-connected and the sub-connected architectures are special cases of the general model, corresponding to $M = 1$ and $M = N$, respectively. For simplicity, we assume that all the sub-arrays have the same structure. The extension to the irregularity case is straightforward.

The BS serves $S(\leq NM)$ users, each with a single antenna. Let $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{KM \times NM}$ be the analog RF precoding matrix, $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{NM \times S}$ be the digital baseband precoding matrix, and $\mathbf{P} \in \mathbb{C}^{S \times S}$ be the power allocation diagonal matrix. Denote $\mathbf{y} = [y_1 \ldots, y_s, \ldots y_S]^{\mathrm{T}}$ be the received signal vectors, where y_s corresponds to the received signal of user s.

Ref.	Architectures			Frequency band		User Number		Required CSI	
	Full	Sub.	General	MicroWave	mmWave	Single	Multiple	Full	Part
[4]	√			√			√		√
[5]	√			√			√		√
[6]	√	√			√		√		√
[7]		√			√		√	√	
[8]	√	√			√	√		√	
[9]			√		√	√		√	
[10]			√	√			√	√	
[11]			√		√		√		√
[12]	√				√		√		√

Tab. 1. The comparison of related work.

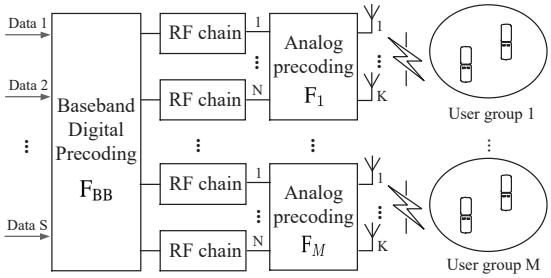

Fig. 2. The general hybrid precoding architecture.

The received signal vector can be written as

$$\mathbf{y} = \mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{P}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^{S\times 1}$ is the transmitted signal vector with $E[\mathbf{x}\mathbf{x}^{\mathrm{H}}] = \mathbf{I}$ and $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ is the complex additive white Gaussian noise (AWGN) vector. Let $\mathbf{H} \in \mathbb{C}^{S\times KM}$ be the downlink channel matrix and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_S]^{\mathrm{T}}$, where $\mathbf{h}_s$ is the channel vector for the s^{th} user.

Denote the antenna indexes to be $\{1, \ldots, MK\}$ and $T_m = \{(m-1)K + 1, \ldots, mK\}$, for $m \in \{1, \ldots M\}$, as the partitioned subset of antenna indexes connected to the m^{th} sub-array, such as

$$\begin{aligned} T_1 &= \{1, \ldots, K\}, \\ T_2 &= \{K+1, \ldots, 2K\}, \\ &\vdots \\ T_M &= \{(M-1)K + 1, \ldots, MK\}. \end{aligned}$$

For the hybrid architecture in Fig. 2, the analog RF precoding matrix, $\mathbf{F}_{\mathrm{RF}}$, is block-diagonal and can be expressed as

$$\mathbf{F}_{\mathrm{RF}} = \begin{bmatrix} \mathbf{F}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{F}_M \end{bmatrix} \tag{2}$$

where $\mathbf{F}_m \in \mathbb{C}^{K\times N}$, $m = \{1, \ldots M\}$ is a fully-connected sub-matrix. This is different from the fully-connected architecture and the sub-connected one.

Based on the long-term statistics of massive MIMO channels, the spatial covariance properties, all users are clustered into M groups according to their similarity. The user groups are denoted by U_m, $m \in \{1, \ldots M\}$. The user grouping algorithms in [13], [12] help divide all users into M groups, each with $\tilde{S}$ users, where $\tilde{S}M = S$. Let $\mathbf{H}_m \in \mathbb{C}^{\tilde{S}\times KM}$ be the channel matrix of the m^{th} group and the corresponding whole channel matrix be $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_M]$. Denote $\mathbf{R}_s = \mathbb{E}\{\mathbf{h}_s^{\mathrm{H}}\mathbf{h}_s\}$ to be the long-term spatial covariance matrix of user s, where $\mathbf{h}_s$ is the channel vector. Then the spatial covariance property of the group m is calculated by the users in this group,

$$\overline{\mathbf{R}}_m = \frac{1}{\tilde{S}} \sum_{s=1}^{\tilde{S}} \mathbf{R}_s. \tag{3}$$

When the hybrid array architecture and user groups are given, the channel matrix can be written as

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \ldots & \mathbf{H}_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{M1} & \mathbf{H}_{M2} & \ldots & \mathbf{H}_{MM} \end{bmatrix} \tag{4}$$

where $\mathbf{H}_{gm}$, $g, m \in \{1, \ldots, M\}$ is the $\tilde{S} \times K$ channel matrix between the m^{th} sub-array and the g^{th} user group.

In addition, the accuracy of baseband precoding $\mathbf{F}_{\mathrm{BB}}$ in (1) depends on the available amount of CSI at the BS. In our model, to make sure that the data streams of one sub-array are routed to a selected user group, the baseband precoding matrix has a block diagonal form

$$\mathbf{F}_{\mathrm{BB}} = \begin{bmatrix} \mathbf{W}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{W}_M \end{bmatrix} \tag{5}$$

where $\mathbf{W}_m = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{\tilde{S}}]$, $m \in \{1, .., M\}$, with the size $\mathbb{C}^{N\times\tilde{S}}$. It is the baseband precoding sub-matrix for the m^{th} user group. Furthermore, $\mathbf{W}_m$ is a function of the corresponding sub-effective channel matrix.

The power allocation matrix $\mathbf{P} = \mathrm{diag}\{P_1, \ldots, P_S\}$ in (1) is chosen to satisfy the transmit power limitation, $\| \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{P} \|_F^2 \leq \mathrm{P_T}$, where $\mathrm{P_T}$ is the transmit power limitation.

3. Hybrid Precoding Design

From Sec. 2, the hybrid precoder can be divided into two parts, an analog precoder and a digital precoder. Based on channel statistics, the analog precoder is used to eliminate intergroup interference. Then the digital precoder, based on the instantaneous effective channel, is for multi-user interference cancellation and is obtained through real-time CSI. In addition, between these two parts, a supplemental matrix is added for loss of the RF constant module constraint. We will discuss their designs in this section.

3.1 Analog Precoding

The signal-to-interference-plus-noise ratio (SINR) of each user is optimized by the beamforming matrices. However, the optimal method generally needs to deal with M coupled variables and with high complexity. Instead, we design analog beamforming to maximize SLNR [14]. In our model, the M sub-arrays are assigned to M groups by an allocation index $\delta_{g,m}$, $g, m \in \{1, \ldots, M\}$. If sub-array m is assigned to serve user group g, then $\delta_{g,m} = 1$. Otherwise $\delta_{g,m} = 0$.

Assume that each user group is exclusively connected to a sub-array. Therefore, the data streams of a user group are only routed to the selected sub-array rather than all the sub-arrays. Then, the complexity of baseband to RF processing and the power consumption are reduced. The SLNR of group g is given by

$$\begin{aligned}
&\mathbb{E}\{\mathrm{SLNR}_{g,m}\} \\
&= \mathbb{E}\{\frac{\delta_{g,m}\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}\mathbf{H}_{g,m}^{\mathrm{H}}\mathbf{H}_{g,m}\mathbf{F}_m\}}{\sigma^2\mathbf{I}+\sum_{g'\neq g}(1-\delta_{g',m})\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}\mathbf{H}_{g',m}^{\mathrm{H}}\mathbf{H}_{g',m}\mathbf{F}_m\}}\} \\
&\overset{(a)}{\geq} \frac{\delta_{g,m}\mathbb{E}\{\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}\mathbf{H}_{g,m}^{\mathrm{H}}\mathbf{H}_{g,m}\mathbf{F}_m\}\}}{\sigma^2\mathbf{I}+\sum_{g'\neq g}(1-\delta_{g',m})\mathbb{E}\{\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}\mathbf{H}_{g',m}^{\mathrm{H}}\mathbf{H}_{g',m}\mathbf{F}_m\}\}} \\
&\overset{(b)}{=} \frac{\delta_{g,m}\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}\tilde{\mathbf{R}}_{g,m}\mathbf{F}_m\}}{\sigma^2\mathbf{I}+\sum_{g'\neq g}(1-\delta_{g',m})\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}\tilde{\mathbf{R}}_{g',m}\mathbf{F}_m\}} \\
&= \frac{\delta_{g,m}\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}\tilde{\mathbf{R}}_{g,m}\mathbf{F}_m\}}{\mathrm{Tr}\{\mathbf{F}_m^{\mathrm{H}}(\sigma^2\mathbf{I}+\sum_{g'\neq g}(1-\delta_{g',m})\tilde{\mathbf{R}}_{g',m})\mathbf{F}_m\}}. \qquad (6)
\end{aligned}$$

In (6), (a) comes from the Jensen's inequation and (b) comes from the definition $\tilde{\mathbf{R}}_{g,m} = \mathbb{E}(\mathbf{H}_{g,m}^{\mathrm{H}}\mathbf{H}_{g,m})$, where $\tilde{\mathbf{R}}_{g,m}$ is the correlation matrix between sub-array T_m and group U_g.

The optimal precoding matrix $\mathbf{F}_m^{(\mathrm{o})}$ is designed based on

$$\mathbf{F}_m^{(\mathrm{o})} = \arg\max_{\mathbf{F}_m\in\mathbb{C}^{K\times N}} \mathbb{E}\{\mathrm{SLNR}_{g,m}\}. \qquad (7)$$

From the Rayleigh-Ritz quotient result [14], the optimal result is proportional to a generalized eigenvector of the matrix pair $\{\delta_{g,m}\tilde{\mathbf{R}}_{g,m}, \sigma^2\mathbf{I}+\sum_{g'\neq g}(1-\delta_{g',m})\tilde{\mathbf{R}}_{g',m}\}$, and can be written as

$$\mathbf{\Gamma}_m^{(\mathrm{o})} \propto \mathrm{EV}.\{\delta_{g,m}\tilde{\mathbf{R}}_{g,m}, \sigma^2\mathbf{I}+\sum_{g'\neq g}(1-\delta_{g',m})\tilde{\mathbf{R}}_{g',m}\} \qquad (8)$$

where EV. denotes the generalized eigenvector operation. Then, the optimal precoding matrix can be obtained by the leading columns of $\mathbf{\Gamma}_m^{(\mathrm{o})}$ corresponding to the largest eigenvalues,

$$\mathbf{F}_m^{(\mathrm{o})}(:, 1:N) = \varsigma\mathbf{\Gamma}_m^{(\mathrm{o})}(:, 1:N) \qquad (9)$$

where ς is a normalization factor so that $\mathrm{Tr}(\mathbf{F}_m\mathbf{F}_m^{\mathrm{H}}) = 1/M$.

Furthermore, for the constant module constrain of RF chains, the following optimization problem should be solved,

$$\min_{|\mathbf{F}_m(i,j)|=\frac{1}{\sqrt{K}}} \| \mathbf{F}_m^{(\mathrm{o})} - \mathbf{F}_m \|_F^2 . \qquad (10)$$

On the basis of the demonstration in [7], the optimal precoding sub-matrix will be

$$\mathbf{F}_m = \frac{1}{\sqrt{K}}\exp\{\mathrm{j}\angle\mathbf{F}_m^{(\mathrm{o})}\} \qquad (11)$$

where $\angle\mathbf{F}_m^{(\mathrm{o})}$ represents the angle of $\mathbf{F}_m^{(\mathrm{o})}$.

3.2 Supplemental Matrix

Due to the constant module constraint in (11), the optimal solution in (8) and (9) for the problem in (7) is no longer valid. Therefore, a supplemental matrix $\mathbf{C}$ is added between the analog part and the digital part to compensate the performance loss, as shown in 3. The matrix is a block diagonal form, $\mathbf{C} = \mathrm{diag}\{\mathbf{C}_1, \ldots, \mathbf{C}_M\}$, where $\mathbf{C}_m \in \mathbb{C}^{N\times N}$ is the sub-supplemental matrix for sub-array m. Then, an orthonormal Procrustes problem is constructed and the supplemental matrix is obtained by solving the following optimization problem

$$\min_{\mathbf{C}_m\mathbf{C}_m'=\mathbf{I}} \| \mathbf{F}_m^{(\mathrm{o})} - \mathbf{F}_m\mathbf{C}_m \|_F^2 . \qquad (12)$$

After the mathematical operator, the singular value decomposition (SVD), $\mathbf{F}_m^{(\mathrm{o})H}\mathbf{F}_m = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}}$, the optimal result will be

$$\mathbf{C}_m = \mathbf{V}\mathbf{U}^{\mathrm{H}}. \qquad (13)$$

Based on the dictionary learning theory and alternating optimization theory in [17], the optimal analog precoding matrix and supplemental matrix are obtained by iterative refinement. The optimization algorithm is summarized as follows.

- Initialization:
  - Get the unconstraint analog precoding matrix $\mathbf{F}^{(\mathrm{o})}$ from (8) and (9).
  - Get the constraint analog precoding matrix $\mathbf{F}$ from (11).
  - Get the supplemental matrix $\mathbf{C}$ from (13).
- Iterations for groups:
  - Replace the unconstraint analog precoding matrix with $\mathbf{F}^{(\mathrm{o})} = \mathbf{F}^{(\mathrm{o})}\mathbf{C}^{\mathrm{H}}$.
  - Replace the constraint analog precoding matrix $\mathbf{F}$ by (11).
  - Replace the supplemental matrix $\mathbf{C}$ by (13).

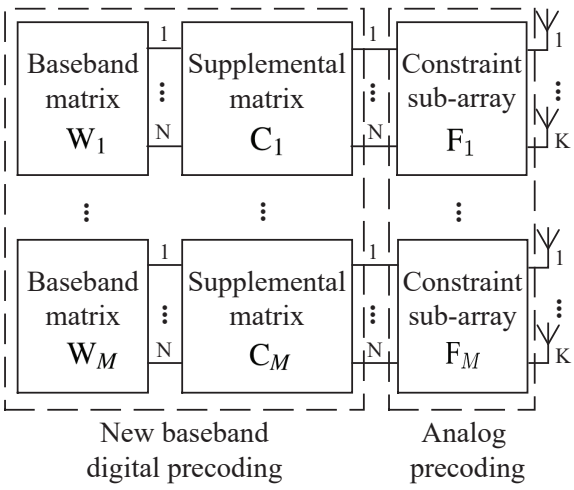

Fig. 3. The hybrid precoding architecture with supplemental matrix.

3.3 Digital Precoding

After computing the analog precoding part and the supplemental part, the effective channel matrix, $\mathbf{H}^{(\mathrm{e})}$, can be obtained. From (2) and (4), the effective channel matrix between group g and sub-array m can be written as

$$\mathbf{H}_{g,m}^{(\mathrm{e})} = \mathbf{H}_{gm}\mathbf{F}_m\mathbf{C}_m. \tag{14}$$

With the instantaneous CSI acquired at the BS side, the baseband precoding is simplified to a multi-user precoding problem and can be solved by existing algorithms. By applying regularized zero-forcing (RZF) method in digital baseband part, user groups are addressed separately at the baseband with their effective channel matrices $\mathbf{H}_{g,m}^{(\mathrm{e})} = \mathbf{H}_{gm}\mathbf{F}_m\mathbf{C}_m$, where $m \in \{1, \ldots, M\}$. Then

$$\mathbf{W}_m = (\mathbf{H}_{g,m}^{(\mathrm{e})H}\mathbf{H}_{g,m}^{(\mathrm{e})} + \beta\mathbf{I})^{-1}\mathbf{H}_{g,m}^{(\mathrm{e})H} \tag{15}$$

where β is the regularization factor.

3.4 Complexity Comparison

By fixing the numbers of total antenna elements and RF chains to MK and MN, the hardware complexities of the different architectures are compared in Tab. 2. By raising M, the number of antenna elements, PSs and RF chains in each sub-array decreases. Meanwhile, the numbers of deployed combiners and splitters connected to the antenna elements reduce: they are MK and MN when $M \neq S$ and become zeros when $M = S$. Hence, the hardware complexity and cost are reduced as the number of sub-arrays increases. Furthermore, effective RF information shrinks as M grows, which will degrade performance. Therefore, proper numbers of sub-arrays should be decided according to the tradeoff among performance requirements, hardware complexities, and deployment costs in different areas.

Architecture	Antennas /sub-array	RFs /sub-array	PSs /sub-array
	Total	Total	Total
Full connection	KM	NM	NKM^2
$M = 1$	KM	NM	NKM^2
General model	K	N	NK
$1 < M < S$	KM	NM	NKM
Sub-connection	K/N	1	K/N
$M = S$	KM	NM	KM

Tab. 2. Hardware complexity comparison with different M.

4. Numerical Results

The numerical results are presented in this section to verify the performance of the proposed method of hybrid precoding design. We use the software MATLAB V8.0 [18] as the simulation tool in this section. Users are distributed as Poisson Point Process (PPP) in a sector with the angle range of 120° and the radius of 50 m. The total number of antenna elements is set to be $KM = 64$. The Uniform Linear Array (ULA) with $\lambda/2$ between adjacent antennas is used inside each sub-array and $10^2\lambda$ between adjacent sub-arrays, where λ is the wavelength. The number of users is set to be as the number of RF chains, $S = MN$. We use $M = 4$ as an example.

The mmWave propagation channels have limited scattering clusters. The classic Saleh-Valenzuela geometric channel model is adopted here [9]. The channel vectors of user s can be obtained by

$$\mathbf{h}_s(t) = \sqrt{\frac{MK}{L}}\sum_{l=1}^{L}\alpha_l\mathbf{a}(\theta_l)^{\mathrm{H}} \tag{16}$$

where t, L, α_l and $\mathbf{a}(\theta_l)$ are the time index, the limited number of scattering clusters, the l^{th} channel gain and the l^{th} array response at the BS with the angle of departure θ_l, respectively. In this paper, we adopt $L = 8$, $\alpha_l \sim \mathcal{CN}(0, 1)$ and the carrier frequency at 28 GHz. The array response for ULA is

$$\mathbf{a}(\theta_l) = \frac{1}{\sqrt{MK}}[1, \mathrm{e}^{2\pi\mathrm{j}/\lambda d_t \sin(\theta_l)}, \ldots, \mathrm{e}^{2\pi\mathrm{j}/\lambda d_t (MK-1)\sin(\theta_l)}]^{\mathrm{T}} \tag{17}$$

where d_t is the space between two antenna elements. Through the uplink pilot, the BS estimates the statistical information of each user $\mathbf{R}_s$ [13].

In Fig. 4, the achievable rates are compared between of the proposed method and the existing methods, where the number of RF chains is set by $NM = 16$. In the figure, the existing precoding methods for the two architectures in [4] and [15] are denoted by Ref. 1 and Ref. 2, respectively. The modified MMSE method proposed in [16] is used for the general architecture and expressed as Ref. 3. From this figure we can see, the performance of the achievable rate with the proposed method is consistent with the existing one in a fully-connected architecture. That is because the proposed method treats all users as one group and exploits the second order statistics of channels. In addition, the proposed method has a higher achievable rate with the general architecture and the sub-connected architecture. The improvements are obtained by jointly considering the signal, leakage, and noise power together instead of just leakage power or signal power in other methods. Furthermore, together with Tab. 3 for complexity comparison, we can see that the flexible hybrid architecture is a compromise among the performance, complexity, and cost.

Figure 5 illustrates the performance of the achievable rate with the proposed method in different numbers of RF chains. The transmit SNR is $P_t/\delta^2 = 40$ dB. Due to high spatial correlation, the dominant rank of the spatial correlation matrix is far less than the number of antennas. According to the calculation in [10], the average dominant rank of the given system is about 24. In the figure, the ideal case stands for the performance without RF constant module constraint, the w. sup. cases stand for the performance with supplemental matrices, and the w/o. sup. cases stand for the performance without supplemental matrices. from the figure we can see, by raising the number of RF chains to the DoF, the number of served users grows, so does the achievable rates. In addition, the performance is improved by means of adding the supplemental matrix. Furthermore, the proposed method outperforms the existing method.

In Fig. 6, the performance with or without constant module constraint is compared, where $M = 2$ and $N = 8$. The coordinated analog precoding in [11] is used for the performance comparison and expressed as Ref. 4. The classical block diagonalization (BD) in [12] for a full connection is donated as BD. The BD method is used as a reference to see the performance of our method. Compared to the Ref. 4, the proposed method has better performance both with constant module constraint (w. constraint) and without constant module constraint (w./o. constraint). Without the constraint, the RF precoding keeps orthogonality and does not lose any gains. However, the RF orthogonality is broken by the constraint. Therefore, the performance deteriorates.

	K	N	PSs	Combiner (size)	Splitter (size)
$M = 1$	64	16	1024	64 (16×1)	16 (1×64)
$M = 4$	16	4	256	64 (4×1)	16 (1×16)
$M = 16$	4	1	64	0	16 (1×4)

Tab. 3. Complexity comparison based on the simulation parameters in Fig. 4.

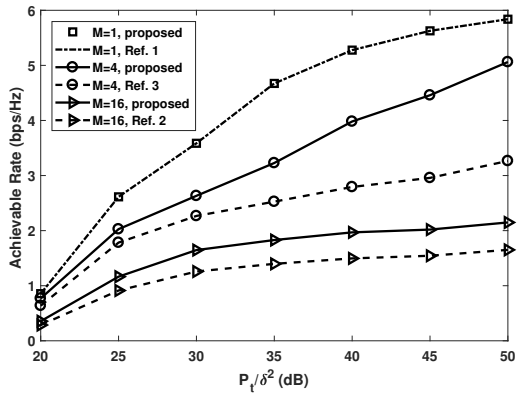

Fig. 4. Performance comparison with different number of sub-arrays.

In Fig. 7, the performance of the achievable rate with the proposed method in different numbers of users is shown. The transmit SNR is $P_t/\delta^2 = 40$ dB. Due to the spatial correlation statistics, the maximal number is set to 24. The two numbers of sub-arrays are considered here, $M = 2$ and $M = 4$. The BD method is used as a reference. By raising the number of users, the achievable rates increase. The proposed method performs better than the previous work both in $M = 2$ and $M = 4$ cases. That is because the eigen beamforming in Ref. 4 causes interference between user groups while the proposed method considers the signal and interference together. As the BD curve shows, the rate rises at first and then becomes flat because of the rank of the channel matrix. If more users are added in the model, the performance will go down eventually.

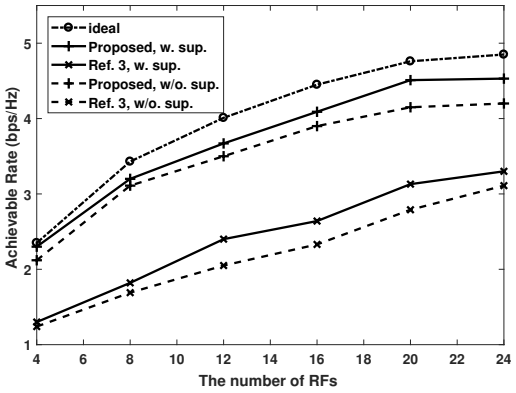

Fig. 5. Performance comparison with the supplemental matrix.

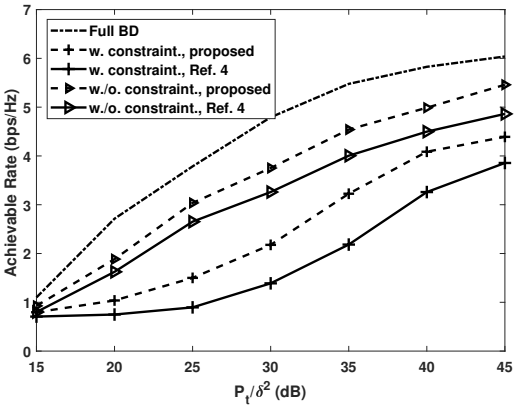

Fig. 6. Performance comparison with and without constant module constraint.

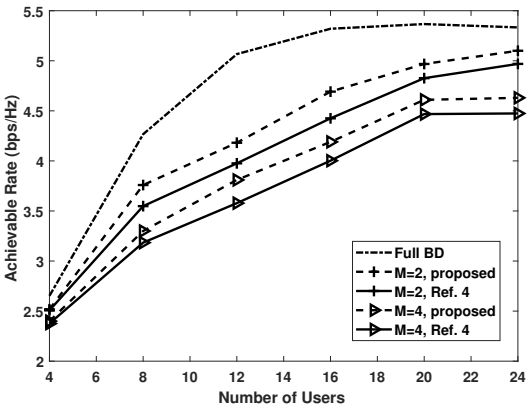

Fig. 7. Performance comparison with the number of users.

5. Conclusions

In this paper, a general hybrid precoding method is designed for mmWave massive MIMO systems. All users are grouped and served by independent analog sub-arrays. We design the high-dimensional analog precoding matrix with the knowledge of long-term statistics of channels, the spatial covariance matrix. By maximizing the SLNR function, the optimal precoding matrix is split into an analog part and a supplemental part. Then a baseband digital precoding with a low dimension is obtained through the knowledge of the instantaneous effective channel matrix. Finally, simulation results demonstrate the feasibility and superiority of the proposed method compared to the existing ones. Meanwhile, the proposed method also reduces the complexity.

Acknowledgments

This work was supported in part by the National Natural Science Foundations of CHINA (Grant No. 61771392, 61771390, 61871322, 61501373 and 61271279), the National Science and Technology Major Project (Grant No. 2015ZX03002006-004, and 2016ZX03001018-004), and the Fundamental Research Foundation of NPU (Grant No. 3102017ZY018).

References

[1] LARSSON, E. G., EDFORS, O., TUFVESSON, F., et al. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 2014, vol. 52, no. 2, p. 186–195. DOI: 10.1109/MCOM.2014.6736761

[2] MEKKAWY, T., YAO, R., TSIFTSIS, T. A., et al. Joint beamforming alignment with suboptimal power allocation for a two way untrusted relay network. *IEEE Transactions on Information Forensics & Security*, 2018, vol. 13, no. 10, p. 2464–2474, DOI: 10.1109/TIFS.2018.2819132.

[3] ZHANG, J. A., HUANG, X., DYADYUK, V., et al. Massive hybrid antenna array for millimeter-wave cellular communications. *IEEE Wireless Communications*, 2015, vol. 22, no. 1, p. 79–87. DOI: 10.1109/MWC.2015.7054722

[4] PARK, S., PARK, J., YAZDAN, A., et al. Exploiting spatial channel covariance for hybrid precoding in massive MIMO systems. *IEEE Transactions on Signal Process*, 2017, vol. 65, no. 14, p. 3818–3832. DOI: 10.1109/TSP.2017.2701321

[5] KIM, D., LEE, G., SUNG, Y. Two-stage beamformer design for massive MIMO downlink by trace quotient formulation. *IEEE Transactions on Communications*, 2015, vol. 63, no. 6, p. 2200–2211. DOI: 10.1109/TCOMM.2015.2429646

[6] LIN, C., LI, G. Y. Terahertz communications: An array-of-subarrays solution. *IEEE Communications Magazine*, 2016, vol. 54, no. 12, p. 124–131. DOI: 10.1109/MCOM.2016.1600306CM

[7] GAO, X., DAI, L., HAN, S., et al. Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays. *IEEE Journal on Selected Areas in Communications*, 2016, vol. 34, no. 4, p. 998–1009. DOI: 10.1109/JSAC.2016.2549418

[8] YU, X., SHEN, J. C., ZHANG, J., et al. Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems. *IEEE Journal of Selected Topics in Signal Processing*, 2016, vol. 10, no. 3, p. 485–500. DOI: 10.1109/JSTSP.2016.2523903

[9] ZHANG, D., WANG, Y., LI, X., et al. Hybridly connected structure for hybrid beamforming in mmwave massive MIMO systems. *IEEE Transactions on Communications*, 2018, vol. 66, no. 2, p. 662–674. DOI: 10.1109/TCOMM.2017.2756882

[10] WU, K., WU, L., ZHANG, J. Multiuser hybrid analogue/digital beamforming for massive multiple-input-multiple-output. *IET Commununications*, 2016, vol. 10, no. 12, p. 1464–1472. DOI: 10.1049/iet-com.2015.0520

[11] SONG, N., WEN, P., SUN, H., et al. Multi-panel based hybrid beamforming for multi-user massive MIMO. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*. Singapore, 2017, p. 1–6. DOI: 10.1109/GLOCOM.2017.8254880

[12] ADHIKARY, A., SAFADI, E. A. Joint Spatial Division and Multiplexing for mm-Wave Channels. *IEEE Journal on Selected Areas in Communications*, 2014, vol. 32, no. 6, p. 1239–1255. DOI: 10.1109/JSAC.2014.2328173

[13] XU, Y., YUE, G., MAO, S. User grouping for massive MIMO in FDD systems: New design methods and analysis. *IEEE Access*, 2014, vol. 2, p. 947–959. DOI: 10.1109/ACCESS.2014.2353297

[14] SADEK, M., TARIGHAT, A., SAYED, A. H. A leakage-based precoding scheme for downlink multi-user MIMO channels. *IEEE Transactions on Wireless Communications*, 2007, vol. 6, no. 5, p. 1711–1721. DOI: 10.1109/TWC.2007.360373

[15] LIANG, L., DAI, Y., XU, W., et al. How to approach zero-forcing under RF chain limitations in large mmWave multiuser systems? In *Proceedings of the IEEE/CIC International Conference on Communications in China (ICCC)*. Shanghai (China), 2014, p. 518–522. DOI: 10.1109/ICCChina.2014.7008332

[16] JOROUGHI, V., VAZQUEZ, M. A., PEREZ-NEIRA, A. I. Generalized multicast multibeam precoding for satellite communications. *IEEE Transactions on Wireless Communications*, 2017, vol. 16, no. 2, p. 952–966. DOI: 10.1109/TWC.2016.2635139

[17] RUSU, C., MÉNDEZ-RIAL, R., GONZÁLEZ, N., et al. Low complexity hybrid precoding strategies for millimeter wave communication systems. *IEEE Transactions on Wireless Communications*, 2016, vol. 15, no. 12, p. 8380–8393. DOI: 10.1109/TWC.2016.2614495

[18] Matlab V8.0. https://www.mathworks.com/products/matlab.html

About the Authors . . .

Yi XIE received the B.E. degree in Telecommunication Engineering from Xi'an Polytechnic University, China, in 2011 and M.S. degree in Telecommunication Engineering from Xidian University, China, in 2014. She is currently working toward the Ph.D. degree in School of Electronics and Information at Northwestern Polytechnical University, China. From 2016 to 2017, she was a visiting scholar in the Department of Electrical and Computer Engineering, Louisiana State University, USA. Her research interests include massive MIMO systems and millimeter wave communication in next generation cellular systems.

Bo LI received the B.E., M.S., and Ph.D degree in Telecommunication Engineering from Xidian University in 1994, 1997, and 2002, respectively. From 1997 to 1998, he was a visiting scholar in the Department of Engineering, Shizuoka University. From 2002 to 2004, he was a Postdoctor in the University of Trento. From July 2007 to Dec. 2007, he was a visiting Professor in the Institut National des Sciences Appliquees (INSA), LYON. He is currently a Professor in the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. His research interests are in the area of wireless networking and communications, including the next generation cellular network (5G) and WLAN (e.g. IEEE 802.11ax and 11ay), the MAC and higher layer technologies, non-orthogonal multiple access for 5G.

Zhongjiang YAN (corresponding author) received the B.E. and Ph.D degree in Telecommunication Engineering from Xidian University in July 2006 and 2011, respectively. From Sept. 2010 to Dec. 2011, he was a visiting Ph.D. student in the Department of Electrical and Computer Engineering, University of Alberta. In Dec. 2011, he joined School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, where he is currently an Associate Professor. His research interests are in the area of wireless networking and communication, including protocols design and their FPGA implement of the media access control layer, radio resource management, and traffic scheduling strategy of the wireless networks, e.g., 5G, WLAN and etc.

Jiancun FAN received the B.S. and Ph.D. degrees in Electrical Engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2012, respectively. From 2009 to 2011, he was a Visiting Scholar with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. He is currently an Associate Professor with the Department of Information and Communications Engineering, Xi'an Jiaotong University. His general research interests include statistical signal processing and wireless communications, with emphasis on cross-layer optimization for spectral- and energy-efficient networks, multiple antenna MIMO communication systems, and practical issues in LTE, and 5G systems.

Mao YANG received the B.E. and M.S. degree in Information and Telecommunication Engineering from Xidian University, China, in 2006 and 2009, and the Ph.D degree in Electronic Engineering from Tsinghua University, China, in 2014. He is current an Associate Professor of School of Electronics and Information at Northwestern Polytechnical University, China. His research interests are in the area of wireless networking and communications, including the next generation cellular network (5G) and WLAN (e.g. IEEE 802.11ax and 11ay), the MAC and higher layer technologies, non-orthogonal multiple access for 5G, software-defined wireless networking, and wireless network virtualization.