Hybrid Architecture and Beamforming Optimization for Millimeter Wave Systems

TANG Yuanqi[1], ZHANG Huimin[1], ZHENG Zheng[2], LI Ping[2], ZHU Yu[1]
(1. Department of Communication Science and Engineering, Fudan University, Shanghai 200433, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202303013
https://kns.cnki.net/kcms/detail/34.1294.TN.20230726.1546.002.html, published online July 27, 2023
Manuscript received: 2023-02-21

Abstract: Hybrid beamforming (HBF) has become an attractive and important technology in massive multiple-input multiple-output (MIMO) millimeter-wave (mmWave) systems. There are different hybrid architectures in HBF depending on different connection strategies of the phase shifter network between antennas and radio frequency chains. This paper investigates HBF optimization with different hybrid architectures in broadband point-to-point mmWave MIMO systems. The joint hybrid architecture and beamforming optimization problem is divided into two sub-problems. First, we transform the spectral efficiency maximization problem into an equivalent weighted mean squared error minimization problem, and propose an algorithm based on the manifold optimization method for the hybrid beamformer with a fixed hybrid architecture. The overlapped subarray architecture which balances well between hardware costs and system performance is investigated. We further propose an algorithm to dynamically partition antenna subarrays and combine it with the HBF optimization algorithm. Simulation results are presented to demonstrate the performance improvement of our proposed algorithms.

Keywords: hybrid beamforming; hybrid architecture; weighted mean square error; manifold optimization; dynamic subarrays

Citation (Format 1): TANG Y Q, ZHANG H M, ZHENG Z, et al. Hybrid architecture and beamforming optimization for millimeter wave systems [J]. ZTE Communications, 2023, 21(3): 93 - 104. DOI: 10.12142/ZTECOM.202303013
Citation (Format 2): Y. Q. Tang, H. M. Zhang, Z. Zheng, et al., "Hybrid architecture and beamforming optimization for millimeter wave systems," *ZTE Communications*, vol. 21, no. 3, pp. 93 - 104, Sept. 2023. doi: 10.12142/ZTECOM.202303013.

1 Introduction

Millimeter-wave (mmWave) communication has emerged as a key technology for fifth-generation (5G) wireless networks to cope with the dilemma between scarce sub-6 GHz spectrum resources and people's rapidly growing demand for higher data transmission[1-4]. MmWave's short wavelength makes it convenient to arrange large-scale antenna arrays at the transceiver end to compensate for the high propagation loss, and thus massive multiple-input and multiple-output (MIMO) becomes attractive in mmWave systems. However, conventional fully digital beamforming (FDBF) requires a separate radio frequency (RF) chain for each antenna and will result in huge hardware costs and power consumption in massive MIMO systems[5]. Therefore, hybrid beamforming (HBF) that requires very few RF chains has become a research hotspot recently[6-8].

This work was supported by ZTE Industry-University-Institute Cooperation Funds, the Natural Science Foundation of Shanghai under Grant No. 23ZR1407300, and the National Natural Science Foundation of China under Grant No. 61771147.

HBF has different hybrid architectures depending on different connection strategies between antennas and RF chains. The fully-connected (FC) architecture and the partially-connected (PC) architecture are two conventional hybrid architectures. Most previous works considered the FC architecture. In Ref. [9], the authors applied the orthogonal matching pursuit algorithm to design the column vectors of the analog precoding matrix based on codebooks. The authors in Ref. [10] proposed an HBF algorithm based on the coordinate update iteration method in the narrowband point-to-point MIMO system. The authors in Ref. [11] proposed an alternating minimization algorithm based on manifold optimization (MO) by minimizing the Frobenius norm between the HBF matrix and the FDBF matrix. The authors in Ref. [12] took the mean square error minimization as the optimization goal and the designed HBF algorithms based on MO and generalized eigenvalue decomposition (EVD).

On the other hand, the PC architecture, where each antenna is only connected to only one RF chain instead of all RF chains, can reduce the power consumption and hardware costs compared with the FC architecture at the cost of certain system performance loss. A low-complexity HBF optimization al-

gorithm based on the positive semi-definite relaxation for the PC architecture has been proposed in Ref. [11]. The HBF optimization algorithms based on element iteration and MO for the PC architecture in the broadband system have also been proposed in Ref. [13].

To achieve a good compromise between hardware costs and system performance, other fixed hybrid architectures have also attracted research attention recently. The authors in Ref. [14] proposed a partially-fully connected architecture that combines the FC and PC architectures and designed algorithms based on continuous interference cancellation and matrix factorization. The authors in Ref. [15] designed an alternative minimization algorithm for this architecture. An overlapped (OL) subarray architecture and a heuristic unified low-rank sparse recovery algorithm were proposed in Ref. [16]. The authors in Ref. [17] proposed a generalized subarray-connected architecture, and developed a successive interference cancellation-based HBF algorithm along with an exhaustive search algorithm to maximize the system energy efficiency.

Since the hardware costs and power consumption of switches in mmWave massive MIMO systems are relatively small[18-19], the dynamic hybrid architecture becomes a promising approach to achieving a better balance between hardware costs and system performance. The authors in Ref. [20] proposed a greedy algorithm with low complexity to partition the antennas over RF chains. A low complexity algorithm to design the optimal partition using statistical channel state information was proposed in Ref. [21]. The authors in Ref. [22] considered the scenario of ultra-wideband mmWave and terahertz frequency band and decomposed the precoding problem into multiple subproblems under the FC architecture.

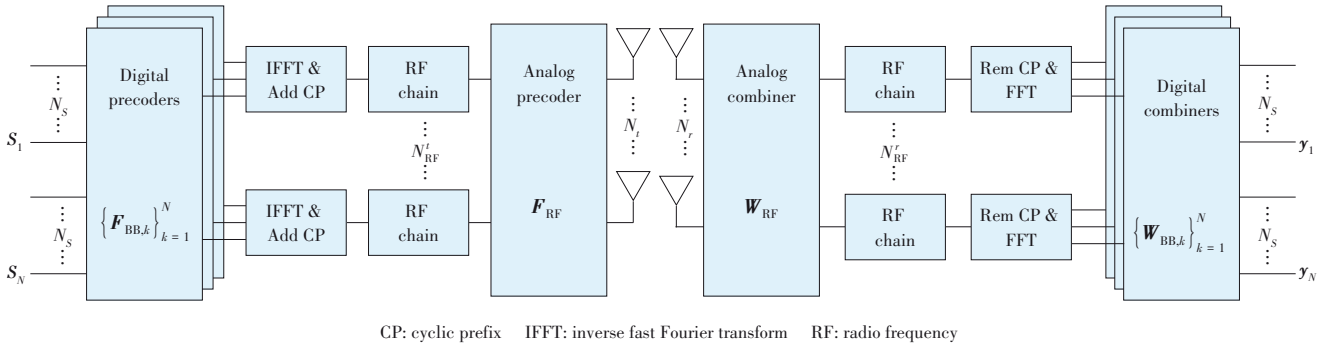
In this paper, we investigate the HBF algorithms with different hybrid architectures for broadband mmWave massive MIMO systems, aiming at maximizing the spectral efficiency. Based on the equivalence between the spectral efficiency maximization (SEM) problem and the weighted minimum mean square error minimization (WMMSE) problem, we design the beamforming optimization algorithm to directly tackle the original SEM optimization problem instead of the conventional indirect design approach of approximating the FDBF matrix with the HBF matrix. We adopt the alternating minimization method to decompose the joint transmitting and receiving HBF optimization problem into two sub-problems. It shows that both the digital precoding and combining optimization sub-problems have closed-form optimal solutions. To further optimize the analog precoder and combiner, we apply the MO method to deal with the constant modulus constraint. In contrast to Ref. [11], where the MO method was applied to solve the matrix approximation problem with the objective of minimizing the Frobenius norm between the FDBF matrix and the HBF matrix of the FC architecture, in our work, the MO method is applied to solve the HBF problem with the WMMSE objective and for arbitrary hybrid architectures by introducing the Hadamard product of the analog precoder and a connection matrix. Apart from the conventional FC and PC architectures, we consider the OL architecture and the PC architecture with dynamic subarrays (PC-dynamic architecture). In particular, we simulate three specific types of fixed OL architectures with a uniform planar array (UPA) and find that our proposed HBF optimization algorithm could achieve a compromise between hardware costs and system performance compared with conventional fixed architectures. Besides, for the PC-dynamic architecture, we derive a lower bound of the original WMMSE objective, based on which, and with some approximations we formulate an eigenvalue maximization problem. Then, we propose a greedy partition algorithm to optimize the dynamic partition of subarrays. Simulation results show that the PC-dynamic architecture with the proposed dynamic partition algorithm can achieve significant performance improvement over the fixed PC architecture.

We denote matrices and vectors by boldface capitals and lower-case letters respectively. $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and the complex conjugate transpose of a matrix or vector, respectively. $\mathrm{tr}(\cdot)$ and $\|\cdot\|_F$ represent the trace and the Frobenius norm of a matrix, respectively. $\mathbb{E}[\cdot]$ is the statistical expectation, $\odot$ is the Hadamard product of two matrices, $\boldsymbol{I}_N$ denotes the $N \times N$ identity matrix, and $CN(0, \boldsymbol{K})$ represents the circularly symmetric complex Gaussian distribution with zero mean and covariance matrix $\boldsymbol{K}$.

2 System Model and Problem Formulation

2.1 System Model

In this paper, we consider the downlink of a broadband mmWave MIMO-orthogonal frequency division multiplexing (OFDM) system with HBF, as shown in Fig. 1. The transmitter first precodes N_s data streams, denoted by the vector $\boldsymbol{s}_k \in \mathbb{C}^{N_s \times 1}$, and at the k-th subcarrier uses a digital precoder $\boldsymbol{F}_{\mathrm{BB},k} \in \mathbb{C}^{N_{\mathrm{RF}}^t \times N_s}$, for $k = 0, \cdots, N-1$ with N denoting the number of subcarriers. Then, N_{RF}^t output streams are transformed into the time domain by the N-point inverse fast Fourier transform. After adding cyclic prefixes (CPs), the signals are further precoded by an analog precoder $\boldsymbol{F}_{\mathrm{RF}} \in \mathbb{C}^{N_t \times N_{\mathrm{RF}}^t}$ composed of a number of phase shifters. It is worth noting that in the HBF design for broadband systems, the digital beamformers can be optimized for different subcarriers, in contrast, the analog one is invariant for the whole frequency band and thus $\boldsymbol{F}_{\mathrm{RF}}$ is not related to the subcarrier index. It is also worth noting that $\boldsymbol{F}_{\mathrm{RF}}$ can represent different hybrid architectures. In particular, we define a connection matrix $\boldsymbol{U}_{\mathrm{p}} \in \mathbb{C}^{N_t \times N_{\mathrm{RF}}^t}, \left[\boldsymbol{U}_{\mathrm{p}}\right]_{ij} \in \{0,1\}$ to represent the connection strategy with any specific hybrid architecture, where $\left[\boldsymbol{U}_{\mathrm{p}}\right]_{i,j} = 1$ indicates that the j-th RF chain is connected to the i-th antenna.



▲Figure 1. Downlink single-user mmWave multiple-input multiple-output orthogonal frequency division multiplexing (MIMO-OFDM) system with hybrid beamforming (HBF)

The analog beamformer with any arbitrary fixed hybrid architecture can be represented by:

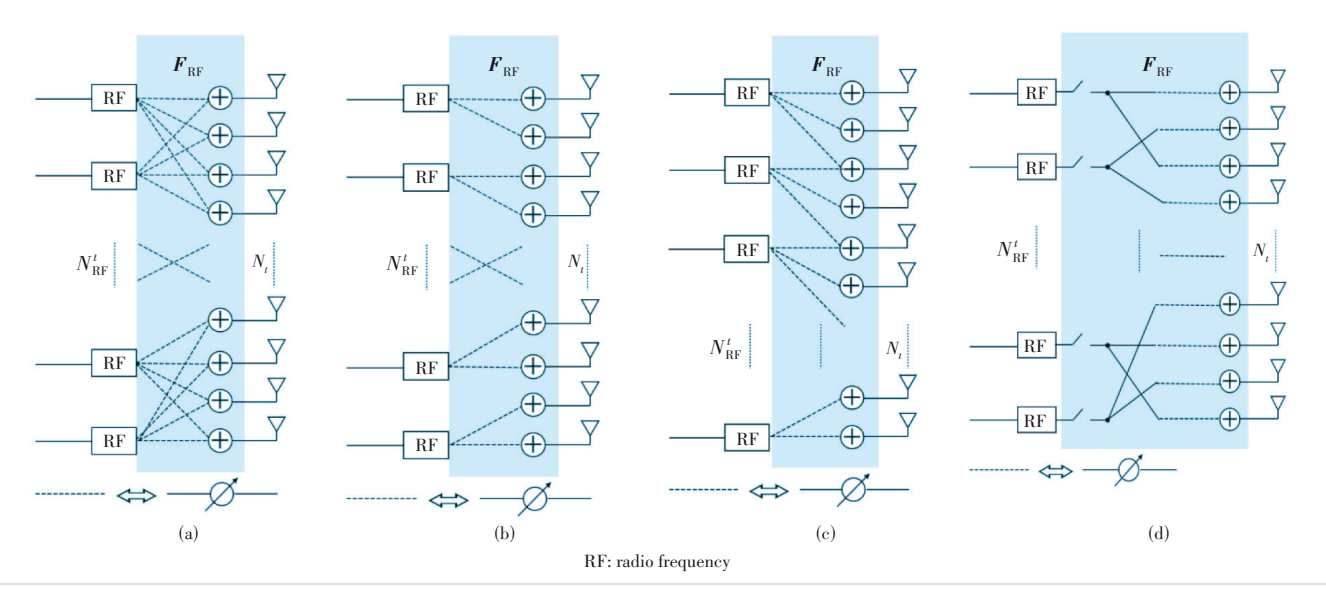
$$\boldsymbol{F}_{\mathrm{RF}} = \boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}} \odot \boldsymbol{U}_{\mathrm{p}}, \tag{1}$$

$$\boldsymbol{W}_{\mathrm{RF}} = \boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}} \odot \boldsymbol{U}_{\mathrm{c}}, \tag{2}$$

where $\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}$ and $\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}}$ represent the analog precoder and combiner with the FC architecture, respectively.

We consider four hybrid architectures for analog beamforming as depicted in Fig. 2. In the FC architecture shown in Fig. 2(a), each RF chain is connected to all antenna elements so that a total of $N_t N_{\mathrm{RF}}^t$ phase shifters are required. In the PC architecture shown in Fig. 2(b), each RF chain is connected to an antenna subarray while each antenna is connected to only one RF chain, so that a total number of N_t phase shifters are required. In the OL architecture shown in Fig. 2(c), the antenna subarrays connected to each RF chain can overlap, where the overlapped antennas are connected to multiple RF chains at the same time. The number of phase shifters required lies between $\left[N_t, N_t N_{\mathrm{RF}}^t\right]$. In the PC-dynamic architecture based on a switch network in Fig. 2(d), the partition of the antenna subarrays can be dynamically adjusted by turning on or off the switches according to the system state, and a total number of N_t phase shifters and $N_t N_{\mathrm{RF}}^t$ switches are required.

The transmitted signal at the k-th subcarrier via N_t antennas is represented by $\boldsymbol{x}_k = \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB},k} \boldsymbol{s}_k$, where $\boldsymbol{F}_{\mathrm{BB},k}$ and $\boldsymbol{F}_{\mathrm{RF}}$ satisfy the power constraint $\left\| \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB},k} \right\|_F^2 \leqslant 1$. After passing through the channel matrix at the k-th subcarrier $\boldsymbol{H}_k \in \mathbb{C}^{N_r \times N_t}$, the signals reach the receiver which is equipped with N_r antennas. The received signals are first processed by an analog combiner $\boldsymbol{W}_{\mathrm{RF}} \in \mathbb{C}^{N_r \times N_{\mathrm{RF}}^t}$, which is also shared by all subcarriers. Then, after removing CPs and performing the fast Fourier transform, a digital combiner $\boldsymbol{W}_{\mathrm{BB},k} \in \mathbb{C}^{N_{\mathrm{RF}}^t \times N_s}$ is deployed at



▲Figure 2. Diagram of four hybrid architectures

each subcarrier. Finally, the processed signal at the k-th subcarrier can be expressed as:

$$\boldsymbol{y}_k = \boldsymbol{W}_{\mathrm{BB},k}^H \boldsymbol{W}_{\mathrm{RF}}^H \boldsymbol{H}_k \boldsymbol{F}_{\mathrm{RF}} \boldsymbol{F}_{\mathrm{BB},k} \boldsymbol{s}_k + \boldsymbol{W}_{\mathrm{BB},k}^H \boldsymbol{W}_{\mathrm{RF}}^H \boldsymbol{n}_k, \quad \text{for } k = 0,\cdots,N-1, \tag{3}$$

where $\boldsymbol{n}_k \sim CN(0, {\sigma_k}^2 \boldsymbol{I}_{N_r})$ denotes the additive white Gaussian noise at the k-th subcarrier.

2.2 Channel Model

We consider the clustered delay line (CDL) model developed by 3GPP, which takes the mmWave propagation characteristics into account, and can better characterize spatial correlations for 3D channels. Besides the normalized delay and the power, the azimuth angle of departure (AOD), the azimuth angle of arrival (AOA), the zenith angle of departure (ZOD), and the zenith angle of arrival (ZOA) are also defined in the CDL model. Three types of the CDL model, i.e., CDL-A, CDL-B and CDL-C, are constructed to represent different channel profiles for the non-line of sight (NLOS) scenarios, while two types, i.e., CDL-D and CDL-E, are constructed for the line-of-sight scenarios[23]. The NLOS channel coefficient of the n-th cluster with M rays between the transmit and receive antennas (u and s respectively) at time instant t and delay τ is given by:

$$H_{u,s,n}^{\mathrm{NLOS}}(t) = \sqrt{\frac{P_n}{M}} \sum_{m=1}^{M} \begin{bmatrix} F_{\mathrm{rx},u,\theta}\left(\theta_{n,m,\mathrm{ZOA}}, \phi_{n,m,\mathrm{AOA}}\right) \\ F_{\mathrm{rx},u,\phi}\left(\theta_{n,m,\mathrm{ZOA}}, \phi_{n,m,AOA}\right) \end{bmatrix}^T \times$$
$$\begin{bmatrix} \exp\left(j\Phi_{n,m}^{\theta\theta}\right) & \sqrt{K_{n,m}^{-1}} \exp\left(j\Phi_{n,m}^{\theta\phi}\right) \\ \sqrt{K_{n,m}^{-1}} \exp\left(j\Phi_{n,m}^{\phi\theta}\right) & \exp\left(j\Phi_{n,m}^{\phi\phi}\right) \end{bmatrix} \begin{bmatrix} F_{\mathrm{tx},u,\theta}\left(\theta_{n,m,\mathrm{ZOD}}, \phi_{n,m,\mathrm{AOD}}\right) \\ F_{\mathrm{tx},u,\phi}\left(\theta_{n,m,\mathrm{ZOD}}, \phi_{n,m,\mathrm{AOD}}\right) \end{bmatrix} \times$$
$$\exp\left(j2\pi\lambda_0^{-1}\left(r_{\mathrm{rx},n,m}^T \dot{d}_{\mathrm{rx},s}\right)\right) \exp\left(j2\pi\lambda_0^{-1}\left(r_{\mathrm{tx},n,m}^T \cdot \dot{d}_{\mathrm{tx},s}\right)\right) \exp\left(j2\pi v_{n,m} t\right),$$

where the definitions of parameters are given in Table 1.

2.3 Problem Formulation

We jointly optimize the hybrid architecture and the hybrid beamformers to maximize the spectral efficiency over N subcarriers subject to the constant modulus constraint of the analog beamformers and the power constraint of the transmitter. The problem can be formulated as follows:

▼ Table 1. Definitions of some parameters in the clustered delay line (CDL) channel model

Parameter	Definition
P_n	Power of the n-th cluster
$F_{\mathrm{rx},u}, F_{\mathrm{tx},s}$	Radiation patterns of the receiving and the transmitting antennas
$\Phi_{n,m}$	Random initial phases of different polarization combinations
$K_{n,m}$	Cross polarization power ratio for the m-th ray in the n-th cluster
λ_0	Carrier wavelength
$r_{\mathrm{rx},n,m}, r_{\mathrm{tx},n,m}$	Spherical unit vectors of the receiving and the transmitting antennas
$\boldsymbol{v}_{n,m}$	Velocity vector

$$\begin{aligned} \underset{\boldsymbol{F}_{\mathrm{BB},k}, \boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}, \boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}}, \boldsymbol{W}_{\mathrm{BB},k}, \boldsymbol{U}_p, \boldsymbol{U}_c}{\text{maximize}} \quad & \frac{1}{N} \sum_{k=1}^{N} R_k \\ \text{subject to} \quad & \left\| \boldsymbol{F}_k \right\|_{\mathrm{F}}^2 \leqslant 1, \forall k; \\ & \left| \left[\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}} \right]_{ij} \right| = 1, \left| \left[\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}} \right]_{ij} \right| = 1, \forall i,j; \\ & \left[\boldsymbol{U}_{\mathrm{p}} \right]_{ij}, \left[\boldsymbol{U}_{\mathrm{c}} \right]_{ij} \in 0,1, \forall i,j; \\ & \left\| \boldsymbol{U}_p \right\|_1 = N_{p1}, \left(\boldsymbol{U}_p \right)_2 = N_{p2}, \end{aligned} \tag{4}$$

where $R_k = \log \left| \boldsymbol{I}_{N_s} + \sigma_k^2 \left(\boldsymbol{W}_k{}^H \boldsymbol{W}_k \right)^{-1} \boldsymbol{W}_k{}^H \boldsymbol{H}_k \boldsymbol{F}_k \boldsymbol{F}_k^H \boldsymbol{H}_k^H \boldsymbol{W}_k \right|$ is the achievable spectral efficiency at each subcarrier, and $\boldsymbol{F}_k = \left(\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}} \odot \boldsymbol{U}_{\mathrm{p}} \right) \boldsymbol{F}_{\mathrm{BB},k}, \boldsymbol{W}_k = \left(\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}} \odot \boldsymbol{U}_{\mathrm{c}} \right) \boldsymbol{W}_{\mathrm{BB},k}$. N_{p1} and N_{p2} are the predetermined numbers of phase shifters used at the transmitter and the receiver, respectively.

It has been proved in Ref. [13] that the SEM problem can be transformed into an equivalent WMMSE problem, which is more tractable. The modified mean square error (MSE) is defined as:

$$\boldsymbol{E} \triangleq \mathbb{E}\left[\left(\beta^{-1} \boldsymbol{y} - \boldsymbol{s} \right) \left(\beta^{-1} \boldsymbol{y} - \boldsymbol{s} \right)^H \right], \tag{5}$$

where β is a scaling factor to be jointly optimized with the hybrid beamformers. The WMMSE problem in the broadband scenario can be formulated as:

$$\begin{aligned} \underset{\boldsymbol{F}_{\mathrm{U},k}, \boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}, \boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}}, \boldsymbol{W}_{\mathrm{BB},k}, \boldsymbol{U}_{\mathrm{p}}, \boldsymbol{U}_{\mathrm{c}}, \beta_k, \boldsymbol{T}_k}{\text{minimize}} \quad & \frac{1}{N} \sum_{k=1}^{N} \left(\boldsymbol{T}_k \boldsymbol{E}_k \right) - \log \left| \boldsymbol{T}_k \right| \\ \text{subject to} \quad & \left(\boldsymbol{F}_k \right)_{\mathrm{F}}^2 \leqslant 1, \forall k; \\ & \left| \left[\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}} \right]_{ij} \right| = 1, \left| \left[\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}} \right]_{ij} \right| = 1, \forall i,j; \\ & \left[\boldsymbol{U}_{\mathrm{p}} \right]_{ij}, \left[\boldsymbol{U}_{\mathrm{c}} \right]_{ij} \in \{0,1\}, \forall i,j; \\ & \left(\boldsymbol{U}_{\mathrm{p}} \right)_1 = N_{\mathrm{p1}}, \left(\boldsymbol{U}_{\mathrm{c}} \right)_2 = N_{\mathrm{p2}}, \end{aligned} \tag{6}$$

where $\boldsymbol{T}_k$ and $\boldsymbol{E}_k = \boldsymbol{I}_{N_s} - \beta_k^{-1} \boldsymbol{F}_k^H \boldsymbol{H}_k^H \boldsymbol{W}_k - \beta_k^{-1} \boldsymbol{W}_k^H \boldsymbol{H}_k \boldsymbol{F}_k + \beta_k^{-2} \sigma_k^2 \boldsymbol{W}_k^H \boldsymbol{W}_k + \beta_k^{-2} \boldsymbol{W}_k^H \boldsymbol{H}_k \boldsymbol{F}_k \boldsymbol{F}_k^H \boldsymbol{H}_k^H \boldsymbol{W}_k$ are respectively the weight matrix and the MSE matrix for the k-th subcarrier, and $\boldsymbol{F}_{\mathrm{U},k} = \beta_k^{-1} \boldsymbol{F}_{\mathrm{BB},k}$. Since the joint optimization of the hybrid beamforming and architecture is hard to solve, we decompose the problem into two subproblems: the HBF optimization problem with a fixed architecture and the architecture optimization problem.

3 HBF Optimization with Fixed Architecture

In this section, we apply the alternating minimization method and the MO method to optimize the hybrid beamformers with a fixed hybrid architecture. The WMMSE problem is formulated as:

$$\begin{aligned}\underset{\boldsymbol{F}_{\mathrm{U},k},\boldsymbol{F}_{\mathrm{RF}},\boldsymbol{W}_{\mathrm{RF}},\boldsymbol{W}_{\mathrm{BB},k}\beta_k,\boldsymbol{T}_k}{\text{minimize}} \quad & \frac{1}{N}\sum_{k=1}^{N}\left(\boldsymbol{T}_k\boldsymbol{E}_k\right)-\log\left|\boldsymbol{T}_k\right| \\ \text{subject to} \quad & \left(\boldsymbol{F}_{\mathrm{U},k}\boldsymbol{F}_{\mathrm{RF}}\right)_{\mathrm{F}}^{2}\leqslant\beta_k^{-2},\forall k; \\ & \left|\left[\boldsymbol{F}_{\mathrm{RF}}\right]_{ij}\right|=\begin{cases}1,i,j\in\left\{i,j\middle|\left|\left[\boldsymbol{U}_{\mathrm{p}}\right]_{ij}\right|=1\right\}\\ 0,\text{ else}\end{cases}; \\ & \left|\left[\boldsymbol{W}_{\mathrm{RF}}\right]_{ij}\right|=\begin{cases}1,i,j\in\left\{i,j\middle|\left|\left[\boldsymbol{U}_{\mathrm{c}}\right]_{ij}\right|=1\right\},\forall i,j\\ 0,\text{ else}\end{cases}.\end{aligned} \tag{7}$$

Since the joint optimization problem of the five variables in Eq. (7) is hard to solve, we adopt the alternating minimization method to decouple the optimization of the transmitter and receiver and solve the two subproblems separately.

3.1 Transmitter Design

In this subsection, we fix the hybrid combiner $\boldsymbol{W}_k$ and optimize the hybrid precoder. Firstly, the closed form solution of $\boldsymbol{T}_k$ can be obtained as follows by differentiating the objective function with respect to $\boldsymbol{T}_k$

$$\boldsymbol{T}_k=\boldsymbol{E}_k^{-1}. \tag{8}$$

Secondly, the optimal digital precoder $\boldsymbol{F}_{\mathrm{BB},k}$ and the scaling factor β_k at each subcarrier can be derived with fixed $\boldsymbol{F}_{\mathrm{RF}}$. Considering the power constraint, it can be proved that the optimal β_k can only be achieved with the maximum transmit power, and the optimal β_k is given by:

$$\beta_k=1/\sqrt{\left\|\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{U},k}\boldsymbol{F}_{\mathrm{U},k}^{H}\boldsymbol{F}_{\mathrm{RF}}^{H}\right\|_F^2}. \tag{9}$$

According to the Karush-Kuhn-Tucker (KKT) conditions, $\boldsymbol{F}_{U,k}$ has a closed-form solution as follows:

$$\boldsymbol{F}_{U,k}=\left(\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{G}_k\boldsymbol{G}_k^{H}\boldsymbol{F}_{\mathrm{RF}}+\xi_k\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{F}_{\mathrm{RF}}\right)^{-1}\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{G}_k, \tag{10}$$

where $\xi_k=(\sigma_k^2\mathrm{tr}(\boldsymbol{T}_k\boldsymbol{W}_k^{H}\boldsymbol{W}_k))^{-1}$ and $\boldsymbol{G}_k=\boldsymbol{H}_k^{H}\boldsymbol{W}_k$.

Thirdly, by substituting $\boldsymbol{T}_k$,β_k and $\boldsymbol{F}_{\mathrm{BB},k}$ back into the original objective function, the optimization problem of $\boldsymbol{F}_{\mathrm{RF}}$ can be obtained as follows:

$$\begin{aligned}\underset{\boldsymbol{F}_{\mathrm{RF}}}{\text{minimize}} \quad & f\left(\boldsymbol{F}_{\mathrm{RF}}\right) \\ \text{subject to} \quad & \left|\left[\boldsymbol{F}_{\mathrm{RF}}\right]_{ij}\right|=\begin{cases}1,i,j\in\left\{i,j\middle|\left|\left[\boldsymbol{U}_{p}\right]_{ij}\right|=1\right\},\\ 0,\text{ else}\end{cases}\end{aligned} \tag{11}$$

where $f(\boldsymbol{F}_{\mathrm{RF}})=\frac{1}{N}\sum_{k=1}^{N}\mathrm{tr}\left(\left(\boldsymbol{T}_k^{-1}+\xi_k\boldsymbol{G}_k^{H}\boldsymbol{F}_{\mathrm{RF}}(\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{F}_{\mathrm{RF}})^{-1}\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{G}_k\right)^{-1}\right)$.

Next, we use the MO method to design $\boldsymbol{F}_{\mathrm{RF}}$. The basic idea is to define a Riemannian manifold considering the constant modulus constraint, and iteratively update $\boldsymbol{F}_{\mathrm{RF}}$ along the direction of the Riemann gradient in a way similar to the conventional Euclidean gradient descent algorithm[12]. The key is to derive the Euclidean conjugate gradient of $f\left(\boldsymbol{F}_{\mathrm{RF}}\right)$ with the FC architecture, which is given by:

$$\begin{aligned}\nabla_{\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}}f\left(\boldsymbol{F}_{\mathrm{RF}}\right)=&\frac{1}{N}\sum_{k=1}^{N}\xi_k\left(\boldsymbol{F}_{\mathrm{RF}}\left(\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{F}_{\mathrm{RF}}\right)^{-1}\boldsymbol{F}_{\mathrm{RF}}^{H}-\boldsymbol{I}_{N_t}\right)\\ &\boldsymbol{G}_k\boldsymbol{\Omega}_k^{-2}\boldsymbol{G}_k^{H}\boldsymbol{F}_{\mathrm{RF}}\left(\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{F}_{\mathrm{RF}}\right)^{-1},\end{aligned} \tag{12}$$

where $\boldsymbol{\Omega}_k\triangleq\boldsymbol{T}_k^{-1}+\xi_k\boldsymbol{G}_k^{H}\boldsymbol{F}_{\mathrm{RF}}\left(\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{F}_{\mathrm{RF}}\right)^{-1}\boldsymbol{F}_{\mathrm{RF}}^{H}\boldsymbol{G}_k$. Since $f\left(\boldsymbol{F}_{\mathrm{RF}}\right)$ is only related to the antennas that are connected to each RF chain with any specified hybrid architecture and calculating the gradient involves the derivative with respect to each entry of $\boldsymbol{F}_{\mathrm{RF}}$[13], with $\boldsymbol{F}_{\mathrm{RF}}=\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}\odot\boldsymbol{U}_{\mathrm{p}}$, it can be shown that:

$$\nabla f\left(\boldsymbol{F}_{\mathrm{RF}}\right)=\nabla_{\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}}f\left(\boldsymbol{F}_{\mathrm{RF}}\right)\odot\boldsymbol{U}_{\mathrm{p}}. \tag{13}$$

Then, we can obtain the Riemannian gradient by projecting the Euclidean gradient $\nabla f\left(\boldsymbol{F}_{\mathrm{RF}}\right)$ onto the tangent space, and update $\boldsymbol{F}_{\mathrm{RF}}$ with a proper step size determined by the well-known Armijo backtracking algorithm. Finally, the retraction operation is applied to make the result satisfy the constant modulus constraint[11] as follows:

$$\mu\boldsymbol{d}\mapsto\mathrm{Retr}_{\boldsymbol{x}}\left(\mu\boldsymbol{d}\right)=\mathrm{vec}\left[\frac{\left(\boldsymbol{x}+\mu\boldsymbol{d}\right)_i}{\left|\left(\boldsymbol{x}+\mu\boldsymbol{d}\right)_i\right|}\right]. \tag{14}$$

It is worth noting that with Eqs. (12) and (13), the above algorithm based on the MO method can be adopted in the HBF design with arbitrary hybrid architectures as there is no specific requirement to the connection matrix $\boldsymbol{U}_{\mathrm{p}}$. Finally, the precoder design with arbitrary fixed hybrid architectures is summarized in Algorithm 1.

Algorithm 1: Hybrid precoder design based on the MO method

Input: $\xi_k,\boldsymbol{G}_k,\boldsymbol{T}_k,\boldsymbol{U}_{\mathrm{p}}$
1: Initialize $\boldsymbol{F}_{\mathrm{RF},0}$ with random phases, $i=0$
2: **repeat**
3: Select the step size μ
4: Update $\mathrm{vec}\left(\boldsymbol{F}_{\mathrm{RF},i+1}\right)$ according to Eq. (14)
5: Update the Riemannian gradient $\boldsymbol{g}_i=\nabla f\left(\boldsymbol{F}_{\mathrm{RF},i+1}\right)$ according to Eqs. (12) and (13)
6: Calculate $\boldsymbol{g}_i^{+},\boldsymbol{d}_i^{+}$ from $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+1}$
7: Select Polak-Ribiere parameter η_{i+1}
8: Calculate the conjugate direction $\boldsymbol{d}_{i+1}=-\boldsymbol{g}_{i+1}+\eta_{i+1}\boldsymbol{d}_i^{+}$
9: Update $i\leftarrow i+1$
10: **until** a stopping condition is satisfied
Output: $\boldsymbol{F}_{\mathrm{RF}}$

3.2 Receiver Design

With the fixed hybrid precoder, we can get the optimization problem for the hybrid combiner. By differentiating Eq. (7) with respect to $\boldsymbol{W}_{\mathrm{BB},k}$, the closed-form solution of $\boldsymbol{W}_{\mathrm{BB},k}$ is given by:

$$\boldsymbol{W}_{\mathrm{BB},k} = \left(\boldsymbol{W}_{\mathrm{RF}}^H \tilde{\boldsymbol{G}}_k \tilde{\boldsymbol{G}}_k^H \boldsymbol{W}_{\mathrm{RF}} + \tilde{\xi}_k \boldsymbol{W}_{\mathrm{RF}}^H \boldsymbol{W}_{\mathrm{RF}}\right)^{-1} \boldsymbol{W}_{\mathrm{RF}}^H \tilde{\boldsymbol{G}}_k, \tag{15}$$

where $\tilde{\boldsymbol{G}}_k = \beta_k^{-1} \boldsymbol{H}_k \boldsymbol{F}_k$, $\tilde{\xi}_k = \sigma_k^2 \beta_k^{-2}$. By substituting Eq. (15) back into Eq. (7), we can get the optimization problem of $\boldsymbol{W}_{\mathrm{RF}}$ as follows:

$$\begin{aligned} & g\left(\boldsymbol{W}_{\mathrm{RF}}\right) = \\ \underset{\boldsymbol{W}_{\mathrm{RF}}}{\text{minimize}} \quad & \frac{1}{N}\sum_{k=1}^{N} \mathrm{tr}\left(\boldsymbol{T}_k\left(\boldsymbol{I}_{N_s} + \tilde{\xi}_k^{-1} \tilde{\boldsymbol{G}}_k^H \boldsymbol{W}_{\mathrm{RF}}\left(\boldsymbol{W}_{\mathrm{RF}}^H \boldsymbol{W}_{\mathrm{RF}}\right)^{-1} \boldsymbol{W}_{\mathrm{RF}}^H \tilde{\boldsymbol{G}}_k\right)^{-1}\right) \\ \text{subject to} \quad & \left|\left[\boldsymbol{W}_{\mathrm{RF}}\right]_{ij}\right| = \begin{cases} 1, i, j \in \left\{i, j \middle| \left|\left[\mathbf{U}_c\right]_{ij}\right| = 1\right\} \\ 0, \text{else} \end{cases}. \end{aligned} \tag{16}$$

This problem is difficult to tackle due to the non-convex constraint. However, since it has a similar form to the design problem of the analog precoder in Eq. (11), we can also adopt the MO method to optimize the analog combiner in the same way as we optimize the analog precoder. Firstly, the key step is to derive the Euclidean gradient of $g\left(\boldsymbol{W}_{\mathrm{RF}}\right)$ with the FC architecture, which is given by:

$$\begin{aligned} g\left(\boldsymbol{W}_{\mathrm{RF}}\right) = & \frac{1}{N}\sum_{k=1}^{N} \tilde{\xi}_k \Big(\boldsymbol{W}_{\mathrm{RF}} \left(\boldsymbol{W}_{\mathrm{RF}}^H \boldsymbol{W}_{\mathrm{RF}}\right)^{-1} \boldsymbol{W}_{\mathrm{RF}}^H - \\ & \boldsymbol{I}_{N_t}\Big) \tilde{\boldsymbol{G}}_k \boldsymbol{T}_k \boldsymbol{J}_k^{-2} \tilde{\boldsymbol{G}}_k^H \boldsymbol{W}_{\mathrm{RF}} \left(\boldsymbol{W}_{\mathrm{RF}}^H \boldsymbol{W}_{\mathrm{RF}}\right)^{-1}, \end{aligned} \tag{17}$$

where $\boldsymbol{J}_k = \boldsymbol{T}_k\left(\boldsymbol{I}_{N_s} + \tilde{\xi}_k \tilde{\boldsymbol{G}}_k^H \boldsymbol{W}_{\mathrm{RF}}\left(\boldsymbol{W}_{\mathrm{RF}}^H \boldsymbol{W}_{\mathrm{RF}}\right) \boldsymbol{W}_{\mathrm{RF}}^H \tilde{\boldsymbol{G}}_k\right)$. Secondly, we can use the formula $\nabla g\left(\boldsymbol{W}_{\mathrm{RF}}\right) = \nabla_{\boldsymbol{W}_{\mathrm{RF}}^{\mathrm{FC}}} g\left(\boldsymbol{W}_{\mathrm{RF}}\right) \odot \mathbf{U}_c$ to obtain the Euclidean gradient of $g\left(\boldsymbol{W}_{\mathrm{RF}}\right)$ with any specified architecture. Finally, with the derived gradient, we can optimize $\boldsymbol{W}_{\mathrm{RF}}$ by applying a procedure similar to that in Algorithm 1.

3.3 Alternating Optimization

We develop a joint hybrid precoding and combining optimization algorithm based on the WMMSE criterion by iteratively and alternatively using Algorithm 1. During each iteration, with the fixed hybrid combiner $\boldsymbol{W}_k$ and the weight matrix $\boldsymbol{T}_k$, we first optimize $\boldsymbol{F}_{\mathrm{RF}}$,$\boldsymbol{F}_{\mathrm{BB},k}$ according to Algorithm 1. Then, with the fixed hybrid precoder, we optimize $\boldsymbol{W}_{\mathrm{RF}}$,$\boldsymbol{W}_{\mathrm{BB},k}$. Finally, we update $\boldsymbol{T}_k$ according to Eq. (8). These steps are repeated until the stopping condition is satisfied. The stopping condition could be set as a maximum number of iterations or it depends on whether the relative difference between the objective function values of two consecutive iterations is smaller than a specific value. The HBF algorithm with a fixed hybrid architecture is summarized in Algorithm 2, which is referred to as the hybrid beamforming algorithm using manifold optimization under the WMMSE criterion (HBF-WMO algorithm).

Algorithm 2: The HBF-WMO algorithm

Input: $\sigma_k, \boldsymbol{H}_k, \forall k \in \{1,\cdots,N\}, \boldsymbol{U}_{\mathrm{p}}, \boldsymbol{U}_{\mathrm{c}}$
1: Initialize $\boldsymbol{W}_{\mathrm{RF},0}, \boldsymbol{F}_{\mathrm{RF},0}, \boldsymbol{T}_{k,0}, \boldsymbol{W}_{\mathrm{BB},k,0}, i = 0$
2: **repeat:**
3: Compute $\boldsymbol{F}_{\mathrm{RF},i}$ according to Algorithm 1
4: Compute $\beta_{k,i}, \boldsymbol{F}_{\mathrm{U},k,i}$ according to Eqs. (9) and (10)
5: Compute $\boldsymbol{W}_{\mathrm{RF},i}$ according to Algorithm 1
6: Compute $\boldsymbol{W}_{\mathrm{BB},k,i}$ according to Eq. (15)
7: Compute $\boldsymbol{T}_{k,i} = \boldsymbol{E}_{k,i}^{-1}$
8: $i \leftarrow i + 1$
9: **until** a stopping condition is satisfied
10: $\boldsymbol{F}_{\mathrm{BB},k} = \beta_k \boldsymbol{F}_{\mathrm{U},k}$
Output: $\boldsymbol{F}_{\mathrm{RF}}, \boldsymbol{F}_{\mathrm{BB},k}, \boldsymbol{W}_{\mathrm{RF}}, \boldsymbol{W}_{\mathrm{BB},k}$

4 Subarray Partition Optimization with the PC-Dynamic Architecture

In this section, we propose an algorithm to dynamically allocate antennas to each RF chain through a switch network in accordance with the channel state variation, and under the constraint that each antenna element is allocated only once. Since the size of the receive antenna array is relatively small compared with that of the transmitter, we only consider the architecture optimization of the transmitter here. Based on the derived objective function in Eq. (11), the WMMSE problem for the optimization of the subarray partition with the PC-dynamic architecture is given by

$$\begin{aligned} \underset{\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}, \boldsymbol{U}_{\mathrm{p}}}{\text{minimize}} \quad & f\left(\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}} \odot \boldsymbol{U}_{\mathrm{p}}\right) \\ \text{subject to} \quad & \left|\left[\boldsymbol{F}_{\mathrm{RF}}^{\mathrm{FC}}\right]_{ij}\right| = 1; \\ & \left[\boldsymbol{U}_{\mathrm{p}}\right]_{ij} \in \{0,1\}; \\ & \left(\boldsymbol{U}_{\mathrm{p}i\cdot}\right)_1 = 1, \forall i,j, \end{aligned} \tag{18}$$

where $\boldsymbol{U}_{\mathrm{p}i\cdot}$ dedicates the i-th row of $\boldsymbol{U}_{\mathrm{p}}$. The original problem is difficult to solve directly, so we transform it into a more tractable one. First, let S_r denote the antenna subset connected to the r-th RF chain. We partition N_t antennas into N_{RF}^t subsets as:

$$\bigcup_{r=1}^{N_{\mathrm{RF}}^t} S_r = S_{\mathrm{ant}}, \left|S_r\right| > 0, S_i \cap S_j = \varnothing, \forall i,j \in \left\{1,\cdots,N_{\mathrm{RF}}^t\right\}, i \neq j, \tag{19}$$

where $\left|S_r\right|$ dedicates the number of elements in S_r and $S_{\mathrm{ant}} = \left\{1,\cdots,N_{\mathrm{t}}\right\}$. The optimization problem can be formulated as:

$$\begin{aligned}&\underset{\{S_r\}_{r=1}^{N_{\mathrm{RF}}^t}}{\text{minimize}} \quad \frac{1}{N}\sum_{k=1}^{N}\mathrm{tr}\left(\left(\boldsymbol{T}_k^{-1}+\xi_k\boldsymbol{G}_k^H\boldsymbol{F}_{\mathrm{RF}}\left(\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{F}_{\mathrm{RF}}\right)^{-1}\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{G}_k\right)^{-1}\right)\\&\text{subject to} \quad \left|\left[\boldsymbol{F}_{\mathrm{RF}}\right]_{ij}\right|=\begin{cases}1, i\in S_j\\0, \text{else}\end{cases};\\&\bigcup_{r=1}^{N_{\mathrm{RF}}^t}S_r=S_{\mathrm{ant}}, |S_r|>0, S_i\cap S_j=\varnothing, i\neq j.\end{aligned} \tag{20}$$

It can be shown that the analog precoder with the conventional fixed PC architecture satisfies $\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{F}_{\mathrm{RF}}=\frac{N_t}{N_{\mathrm{RF}}^t}\boldsymbol{I}_{N_{\mathrm{RF}}^t}$, since the number of antennas connected to each RF chain is assumed to be the same. The equality does not hold in general with the PC-dynamic architecture since the number of antennas connected to each RF chain is not the same. However, as all the RF chains are treated equally and the number of RF chains is assumed to be much less than the number of antennas, it is very likely that the number of antennas connected to different RF chains tends to be close to each other, i. e., $\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{F}_{\mathrm{RF}}\approx\frac{N_t}{N_{\mathrm{RF}}^t}\boldsymbol{I}_{N_{\mathrm{RF}}^t}$. According to the simulation results, the number of antennas in each subarray varies little with the optimized partition. Thus, by using this approximation, the objective function in Eq. (20) can be written as:

$$J(\boldsymbol{F}_{\mathrm{RF}})=\frac{1}{N}\sum_{k=1}^{N}\mathrm{tr}\left(\left(\frac{\xi_k}{N_{\mathrm{t}}}\boldsymbol{G}_k^H\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{G}_k\right)^{-1}\right). \tag{21}$$

Inspired by Ref. [12], where the authors derived a lower bound of the original MMSE problem and proposed the EVD algorithm with the FC architecture, we derive the lower bound of $J(\boldsymbol{F}_{\mathrm{RF}})$ as:

$$\begin{aligned}&\sum_{k=1}^{N}t\left(\boldsymbol{M}_k^{-1}\right)=\sum_{k=1}^{N}\sum_{s=1}^{N_s}\lambda_s^{-1}\left(\boldsymbol{M}_k\right)\geqslant\\&\sum_{k=1}^{N}N_{\mathrm{s}}^2\left(\sum_{s=1}^{N_s}\lambda_s\left(\boldsymbol{M}_k\right)\right)^{-1}\geqslant N^2N_{\mathrm{s}}^2\left(\sum_{k=1}^{N}\mathrm{tr}\left(\boldsymbol{M}_k\right)\right)^{-1},\end{aligned} \tag{22}$$

where $\boldsymbol{M}_k\triangleq\boldsymbol{T}_k^{-1}+\frac{\xi_k}{N_{\mathrm{t}}}\boldsymbol{G}_k^H\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{G}_k$, and $\lambda_s(\cdot)$ dedicates the eigenvalues of a matrix. The equality holds only if the values of $\mathrm{tr}\left(\boldsymbol{M}_k\right)$ at all subcarriers are the same. By taking the lower bound as the objective function and omitting the constants, the problem becomes:

$$\begin{aligned}&\underset{\{S_r\}_{r=1}^{N_{\mathrm{RF}}^t}}{\text{maximize}} \quad \mathrm{tr}\left(\frac{1}{N}\sum_{k=1}^{N}\xi_k\boldsymbol{G}_k^H\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{G}_k\right)\\&\text{subject to} \quad \left|\left[\boldsymbol{F}_{\mathrm{RF}}\right]_{ij}\right|=\begin{cases}1, i\in S_j\\0, \text{else}\end{cases};\\&\bigcup_{r=1}^{N_{\mathrm{RF}}^t}S_r=S_{\mathrm{ant}}, |S_r|>0, S_i\cap S_j=\varnothing, i\neq j.\end{aligned} \tag{23}$$

We can write $\boldsymbol{G}_k$ as a combination of N_{RF}^t block matrixes:

$$\boldsymbol{G}_k^H=\left[\boldsymbol{G}_{k,S_1}^H\ \boldsymbol{G}_{k,S_2}^H\ \cdots\ \boldsymbol{G}_{k,S_{N_{\mathrm{RF}}^t}}^H\right], \tag{24}$$

where $\boldsymbol{G}_{k,S_r}^H=\boldsymbol{G}_k^H(:,S_r)$, so the objective function in Eq. (23) can be written as:

$$\begin{aligned}&\mathrm{tr}\left(\frac{1}{N}\sum_{k=1}^{N}\xi_k\boldsymbol{G}_k^H\boldsymbol{F}_{\mathrm{RF}}\boldsymbol{F}_{\mathrm{RF}}^H\boldsymbol{G}_k\right)=\\&\frac{1}{N}\sum_{k=1}^{N}\xi_k\left(\boldsymbol{G}_k^H\boldsymbol{F}_{\mathrm{RF}}\right)_F^2=\\&\frac{1}{N}\sum_{k=1}^{N}\xi_k\left(\left[\boldsymbol{G}_{k,S_1}^H\boldsymbol{f}_{\mathrm{RF},S_1}\cdots\boldsymbol{G}_{k,S_{N_{\mathrm{RF}}^t}}^H\boldsymbol{f}_{\mathrm{RF},S_{N_{\mathrm{RF}}^t}}\right]\right)_F^2=\\&\sum_{r=1}^{N_{\mathrm{RF}}^t}\left(\boldsymbol{f}_{\mathrm{RF},S_r}^H\boldsymbol{R}_{S_r}\boldsymbol{f}_{\mathrm{RF},S_r}\right),\end{aligned} \tag{25}$$

where $\boldsymbol{R}_{S_r}=\frac{1}{N}\sum_{k=1}^{N}\xi_k\boldsymbol{G}_{k,S_r}^H\boldsymbol{G}_{k,S_r}\boldsymbol{f}_{\mathrm{RF},S_r}\in\mathbb{C}^{N_t\times 1}$, $\left[\left|\boldsymbol{f}_{\mathrm{RF},S_r}\right|\right]_i=\begin{cases}1, i\in S_j\\0, i\notin S_j\end{cases}$. Without consideration of the constant modulus constraint, the maximum value of Eq. (25) is given by $\sum_{r=1}^{N_{\mathrm{RF}}^t}\lambda_1\left(\boldsymbol{R}_{S_r}\right)$, where $\lambda_1(\cdot)$ is the maximum eigenvalue of a matrix. Therefore, we can solve Eq. (23) by maximizing the sum of the maximum eigenvalues of $\boldsymbol{R}_{S_r}$ corresponding to each subarray. The optimization problem of the subarray partition can be formulated as:

$$\begin{aligned}&\underset{\{S_r\}_{r=1}^{N_{\mathrm{RF}}^t}}{\text{maximize}} \quad \sum_{r=1}^{N_{\mathrm{RF}}^t}\lambda_1\left(\boldsymbol{R}_{S_r}\right)\\&\text{subject to} \quad \bigcup_{r=1}^{N_{\mathrm{RF}}^t}S_r=S_{ant}, |S_r|>0, S_i\cap S_j=\varnothing, i\neq j.\end{aligned} \tag{26}$$

The optimal solution of Eq. (26) is a complex combinatorial optimization problem and calculating the eigenvalues leads to relatively high computational cost. Therefore, two operations are adopted here to further simplify the optimization problem. Firstly, according to Ref. [20], the maximum eigenvalue can be approximated with the l_1 norm of a matrix as follows:

$$\hat{\lambda}_1\left(\boldsymbol{R}_{S_r}\right)\triangleq\frac{1}{|S_r|}\sum_{i,j\in S_r}\left|\left[\boldsymbol{R}\right]_{i,j}\right|, \tag{27}$$

where $\boldsymbol{R}=\frac{1}{N}\sum_{k=1}^{N}\xi_k\boldsymbol{G}_{k,S_r}^H\boldsymbol{G}_{k,S_r}$ is the average channel covariance matrix of all subcarriers. Then, the problem in Eq. (26) becomes[1]:

$$\begin{aligned}&\underset{\{S_r\}_{r=1}^{N_{\mathrm{RF}}^t}}{\text{maximize}} \quad \frac{1}{|S_r|}\sum_{i,j\in S_r}\left|\left[\mathbf{R}\right]_{i,j}\right|\\&\text{subject to} \quad \bigcup_{r=1}^{N_{\mathrm{RF}}^t}S_r=S_{\mathrm{ant}}, |S_r|>0, S_i\cap S_j=\varnothing, i\neq j.\end{aligned} \tag{28}$$

Secondly, a greedy algorithm is proposed here to solve the

1 It is worth noting that compared with the original optimization objective function in Eq. (20), the omission of the constant modulus constraint and the use of several approximations above will bring in some performance loss, which is of interest for further investigation in future work.

problem. Since $\boldsymbol{R}$ is a Hermitian matrix, we only need to consider the upper triangular part when calculating the objective function in Eq. (28). The main idea of the greedy algorithm is that the objective function does not decrease after adding $\left|[\boldsymbol{R}]_{ij}\right|, i<j$ into or removing it out of the subset S_r. The algorithm consists of two steps. The first one is to define an initial full antenna set $S_0=\{1,\cdots,N_t\}$ and sort all the elements $\left|[\boldsymbol{R}]_{ij}\right|, i<j$ in descending order. To ensure that the constraint $|S_r|>0$ is satisfied, two antennas i and j corresponding to $[\boldsymbol{R}]_{ij}$ are partitioned into S_r in order. The second step is to consider different cases of partitioning antenna i,j into different subsets and choose the one that maximizes the objective function. In this process, if two antennas i and j belong to two different subsets, only four cases related to two subsets are considered. Otherwise, N_{RF}^t cases are considered if both antennas belong to S_0. For simplicity of description, we define a function $f_{\boldsymbol{R}_{S_r}}$ as follows:

$$f_{\boldsymbol{R}_{S_r}}(S_r,n_{sel},r)=\begin{cases}0, \ |S_r|=0 \text{ or } \{n_{\mathrm{sel}}=N_{\mathrm{RF}}^t, \text{and } r=0\}\\ \frac{1}{|S_r|}\left(\sum_{i,j\in S_r}\left|[\boldsymbol{R}_{\mathrm{up}}]_{i,j}\right|+\frac{1}{2}\sum_{i\in S_r}\left|[\boldsymbol{R}]_{i,i}\right|\right), \text{otherwise,}\end{cases} \tag{29}$$

where n_{sel} denotes the number of subsets that already have elements, and $\boldsymbol{R}_{\mathrm{up}}$ is the upper triangular part of $\boldsymbol{R}$. The partition optimization algorithm is summarized in Algorithm 3. With $\{S_r\}_{r=1}^{N_{\mathrm{RF}}^t}$, we can easily get $\boldsymbol{U}_{\mathrm{p}}$, and then use the HBF-WMO algorithm proposed in Section 3 to optimize $\boldsymbol{F}_{\mathrm{RF}}$ with the PC-dynamic architecture.

Algorithm 3: Dynamic subarrays partition optimization

Input: $\boldsymbol{R}, N_{\mathrm{RF}}^t, S_0, n_{\mathrm{sel}}=0$

1: Sort $\boldsymbol{R}_{\mathrm{up}}$ in descending order:
$\left|[\boldsymbol{R}]_{i,j}\right|\geqslant\left|[\boldsymbol{R}]_{i_kj_k}\right|\geqslant\cdots\geqslant\left|[\boldsymbol{R}]_{i_Kj_K}\right|, 1\leqslant i_k<j_k\leqslant N_t, K=N_t(N_t-1)/2$

2: For $k=1{:}K$ **repeat**:

3: If $i_k,j_k\in S_0$:

4: If $n_{\mathrm{sel}}<N_{\mathrm{RF}}^t$:
$S_{n_{\mathrm{sel}}}\leftarrow\{i_k,j_k\}, S_0\backslash\{i_k,j_k\}, n_{\mathrm{sel}}\leftarrow n_{\mathrm{sel}}+1$

5: Else: $r_{\max}=\underset{r}{\mathrm{argmax}}\ (f_{\boldsymbol{R}_{S_r}}(S_r\cup\{i_k,j_k\},n_{\mathrm{sel}},r))$
$S_{r_{\max}}\leftarrow\{i_k,j_k\}, S_0\backslash\{i_k,j_k\}$

6: Else if $i_k\in S_m, j_k\in S_l, \forall m,l\in\{0,1,\cdots,n_{sel}\}, m\neq l$:
$u_{\mathrm{crt}}=f_{\boldsymbol{R}_{S_r}}(S_m,n_{sel},m)+f_{\boldsymbol{R}_{S_r}}(S_l,n_{\mathrm{sel}},l)$
$u_{\mathrm{newj}}=f_{\boldsymbol{R}_{S_r}}(S_m\cup\{j_k\},n_{\mathrm{sel}},m)+f_{\boldsymbol{R}_{S_r}}(S_l\backslash\{j_k\},n_{\mathrm{sel}},l)$
$u_{\mathrm{newi}}=f_{\boldsymbol{R}_{S_r}}(S_m\backslash\{i_k\},n_{sel},m)+f_{\boldsymbol{R}_{S_r}}(S_l\cup\{i_k\},n_{sel},l)$
$u_{\mathrm{newij}}=f_{\boldsymbol{R}_{S_r}}((S_m\backslash\{i_k\})\cup\{j_k\},n_{\mathrm{sel}},m)+f_{\boldsymbol{R}_{S_r}}((S_l\backslash\{j_k\})\cup\{i_k\},n_{\mathrm{sel}},l)$
$u=[u_{\mathrm{crt}},u_{\mathrm{newj}},u_{\mathrm{newi}},u_{\mathrm{newij}}]$
$\max(u)=u_{\mathrm{newj}}$ and $m\neq0$, $S_m\leftarrow\{j_k\}, S_l\backslash\{j_k\}$
$\max(u)=u_{\mathrm{newi}}$ and $l\neq0$, $S_m\backslash\{i_k\}, S_l\leftarrow\{i_k\}$
$\max(u)=u_{\mathrm{newij}}$ and $m,l\neq0$,
$(S_m\backslash\{i_k\})\leftarrow\{j_k\}, (S_l\backslash\{j_k\})\leftarrow\{i_k\}$

Output: $\{S_r\}_{r=1}^{N_{\mathrm{RF}}^t}$

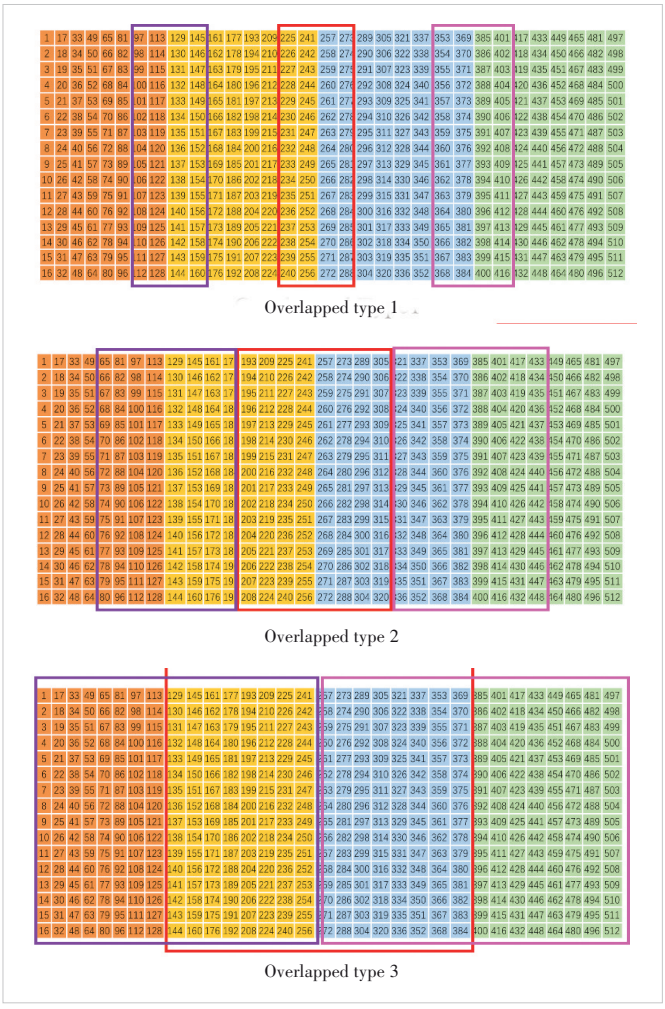
5 Simulation Results

In this section, we first provide some simulation results to show the spectral efficiency performance of the proposed HBF-WMO algorithm in Section 3 with fixed hybrid architectures, in comparison with the optimal fully-digital one. Then, we compare the spectral efficiency performance of fixed and dynamic partitions of antenna subarrays with the PC architecture.

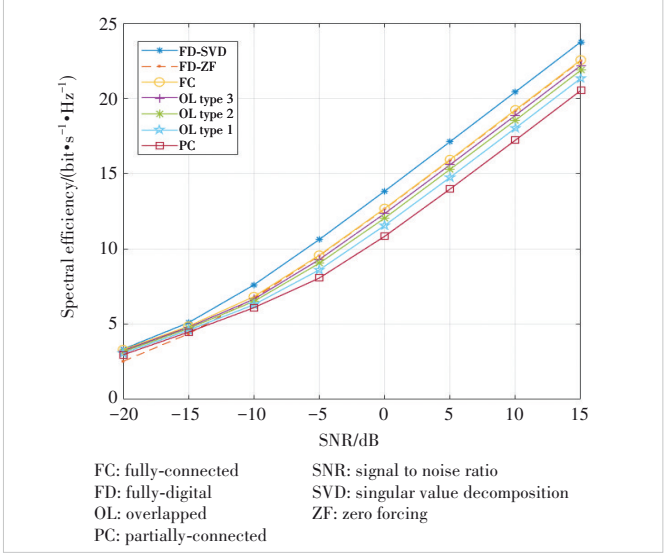
Considering an mmWave MIMO-OFDM system as that in Fig. 1, we assume that the transmitter takes a half-wavelength spaced UPA with $N_t=512$ antennas for the transmission of $N_s=2$ data streams. Five fixed hybrid architectures at the transmitter are evaluated, including the FC and PC architectures, and three types of the OL architecture. Considering the issue of practical implementation of the HBF architecture, we propose three specified OL architectures when four RF chains are employed with a 16×32 UPA at the transmitter, which are shown in Fig. 3. The numbers indicate the antenna indexes, and the units in the same color mean that the corresponding antenna elements are connected to the same RF chain. The antennas within framed squares are overlapped and connected to multiple RF chains. The receiver takes a UPA with $N_r=8$ antennas and $N_{\mathrm{RF}}^r=2$ RF chains with a fixed PC architecture. The number of subcarriers is set to $N=64$, the bandwidth is 100 MHz and the center frequency is 28 GHz. The signal-to-noise ratio (SNR) is defined as $\frac{1}{\sigma^2}$. We take CDL-A as the channel model to evaluate the system performance in a more practical NLOS scenario. In the simulation, the stopping condition is set as the relative difference between the objective function values of two consecutive iterations becomes smaller than $\delta=10^{-3}$.

5.1 Performance with Different Fixed Hybrid Architectures

Fig. 4 shows the performance of spectral efficiency as a function of SNR for the proposed HBF-WMO algorithm with different fixed architectures in CDL-A when four RF chains are equipped at the transmitter. The performance curves of the FC, PC, and OL architectures are labeled as "FC", "PC", and "OL", respectively. For comparison, two FDBF algorithms, namely the FDBF with singular value decomposition (SVD) on



▲ **Figure 3. Diagram of three types of the overlapped subarray-connected architecture at the transmitter with N_t=512, N_{RF}^t = 4**

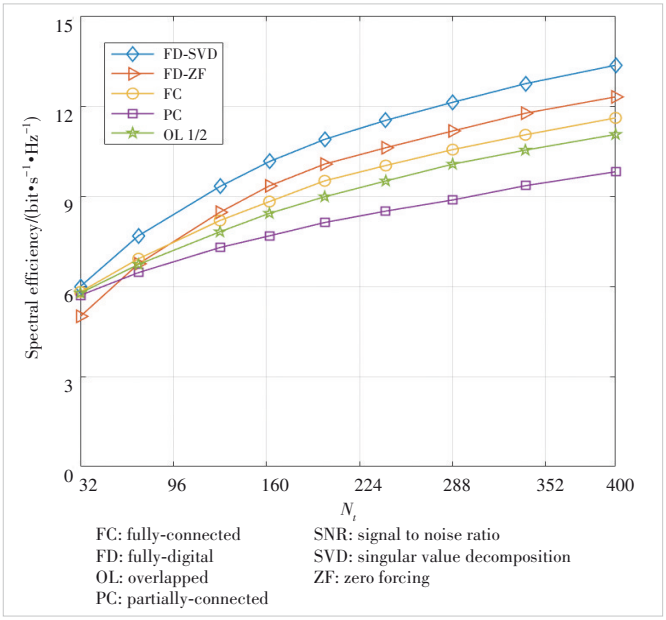


▲ **Figure 4. Spectral efficiency vs SNR for the hybrid beamforming (HBF-WMO) algorithm with different fixed hybrid architectures for a massive multiple-input multiple-output-orthogonal frequency division multiplexing (MIMO-OFDM) system with N=64, N_t=512, N_r=8, N_{RF}^t = 4, N_{RF}^r = 2, N_s=2**

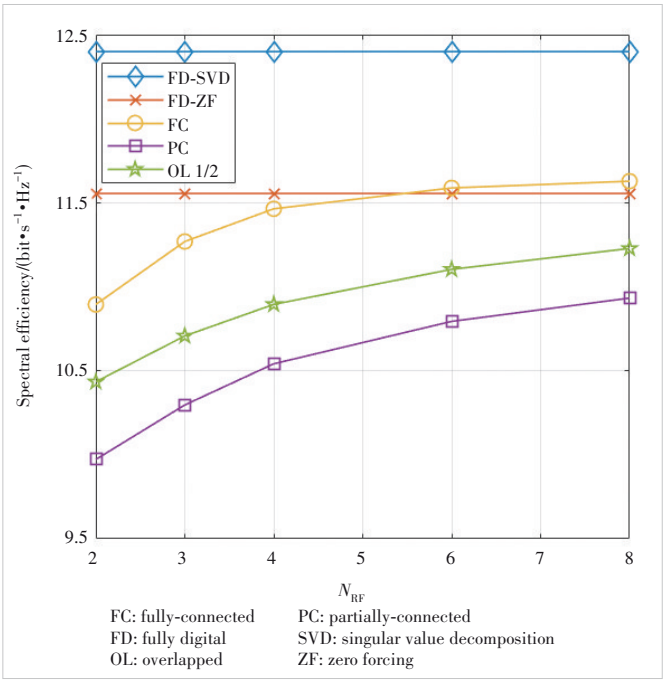
each subcarrier and the FDBF with zero-forcing (ZF), are adopted, and their performance curves are labeled as "FD-SVD" and "FD-ZF", respectively.

It can be seen from this figure that the proposed HBF-WMO algorithm with the FC architecture performs well with far fewer RF chains than antenna elements, and its performance gain over the PC architecture is about 4 dB in CDL-A. The reason is that there are much fewer entries that can be optimized in the HBF matrix of the PC architecture than the FC architecture, since the PC architecture employs far fewer phase shifters. Results also show that the OL architecture achieves the compromise between the system performance and hardware costs, when compared with the FC and PC architectures. In particular, it achieves higher spectral efficiency than the PC architecture at the cost of higher power consumption and implementation complexity introduced by more phase shifters. For example, with 192 more phased shifters employed, the first type of OL architecture has a performance gain of about 1 dB over the PC architecture. According to Ref. [24], based on the state-of-the-art technique, the power consumed by each phase shifter is about 10 mW. With the size of overlapped antennas among different subarrays growing, the performance improvement of the OL architecture over the PC architecture gets bigger. For example, with the third type of OL architecture, the performance gain is about 3 dB over the PC one.

To verify the generality of the proposed HBF-WMO algorithm, we provide in Figs. 5 and 6 the spectral efficiency performance with different numbers of transmit antennas and RF chains under fixed hybrid architectures. The label "OL 1/2"



▲ **Figure 5. Spectral efficiency vs number of transmit antennas for the HBF-WMO algorithm with different fixed hybrid architectures for a massive multiple-input multiple-output-orthogonal frequency division multiplexing (MIMO-OFDM) system with SNR=0 dB, N=64, N_r=8, N_{RF}^t = N_{RF}^r = 2, N_s=2**



▲ Figure 6. Spectral efficiency vs number of transmit RF chains for the HBF-WMO algorithm with different fixed hybrid architectures for a massive multiple-input multiple-output-orthogonal frequency division multiplexing (MIMO-OFDM) system with SNR=0 dB, N=64, N_t=512, N_r=8, N_{RF}^t = 2, N_s=2

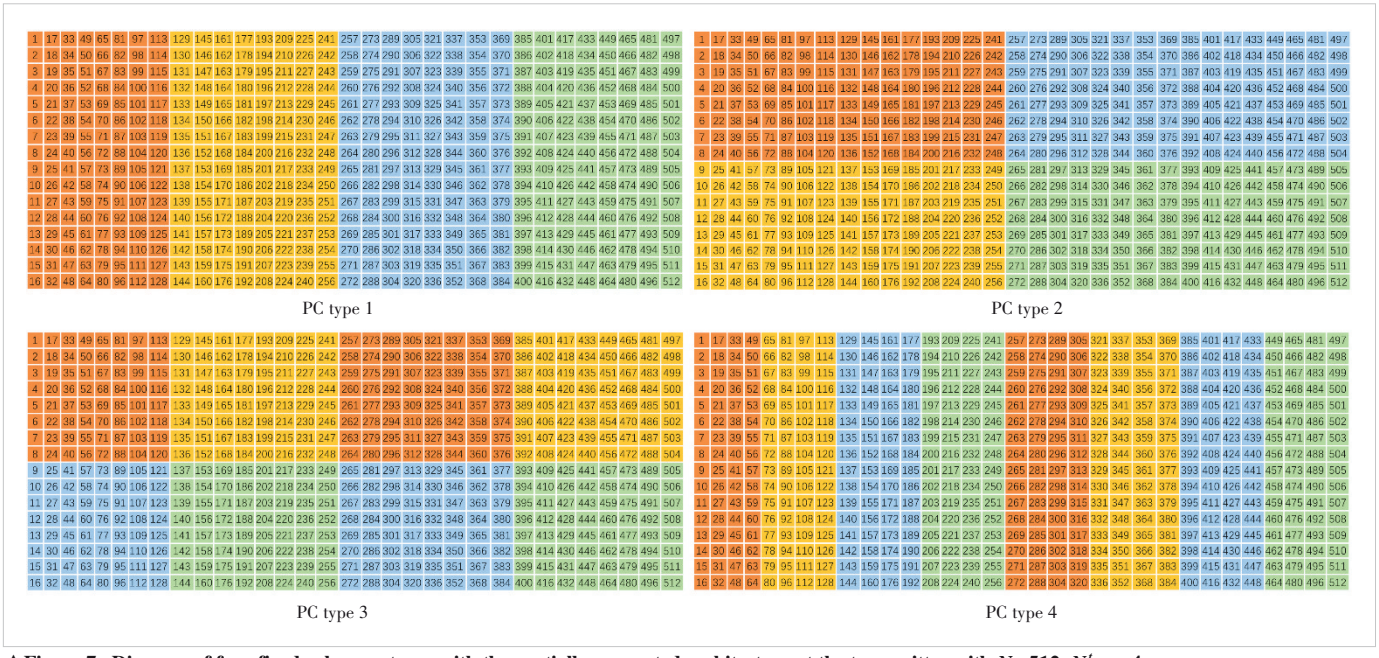
refers to the case where the number of overlapped antennas equals half the number of transmit antennas in the OL architecture. As shown in Fig. 5, the performance of the HBF-WMO algorithm improves with more transmit antennas. Fig. 6 also shows that the gap between the performance of the HBF-WMO algorithm and the optimal FDBF algorithm narrows with more RF chains equipped at the transmitter.

5.2 Performance with Different Partitions of Antenna Subarrays

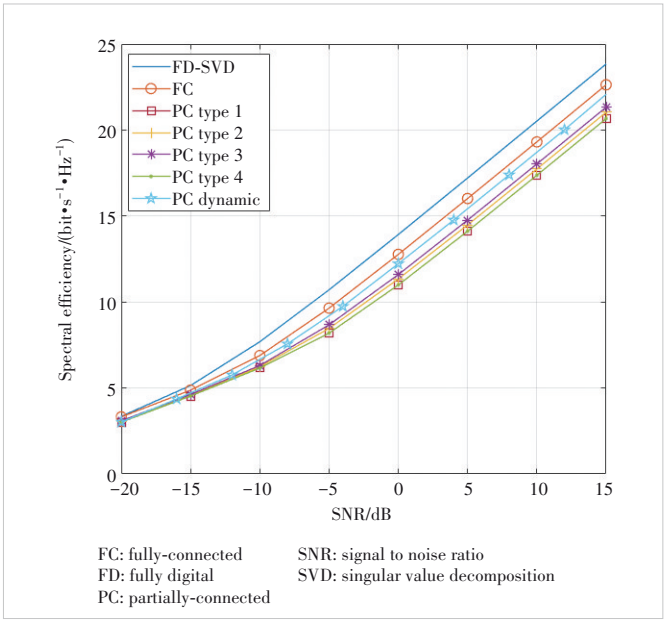
Next, we compare the performance of fixed and dynamic partitions of antenna subarrays with the PC architecture. Four types of partitions under the fixed PC architecture are considered at the transmitter, as shown in Fig. 7. For the HBF system with the PC-dynamic architecture, the antenna subarrays are first dynamically partitioned based on the algorithm proposed in Section 4 and then the HBF matrices are optimized based on the HBF-WMO algorithm in Section 3. Fig. 8 shows the average spectral efficiency performance as a function of SNR with fixed and dynamic antenna subarrays. It can be seen that of the four fixed partition types, the third one achieves the best performance. This is mainly due to the balanced horizontal and vertical angle resolution of the subarray in the third type and the original distribution of angles in CDL-A. It is also shown that the dynamic partition outperforms the third fixed partition type with about 1 dB at the cost of more power consumption and the associated complex circuit brought by N_t = 512 more switches employed in the switch network. According to Ref. [24], the power consumed by each switch is about 5 mW.

6 Conclusions

With relatively small hardware costs and performance loss compared with FDBF, HBF for mmWave communication systems has attracted much attention. Meanwhile, the design of hybrid architecture has also become a research hotspot considering the practical implementation complexity. We have inves-



▲Figure 7. Diagram of four fixed subarray types with the partially-connected architecture at the transmitter with N_t=512, N_{RF}^t = 4



▲Figure 8. Spectral efficiency vs SNR for the HBF-WMO algorithm with different partitions of subarrays for a massive multiple-input multiple-output-orthogonal frequency division multiplexing (MIMO-OFDM) system with N=64, N_t=512, N_r=8, N_{RF}^t = 4, N_{RF}^r = 2, N_s=2

tigated HBF with different hybrid architectures in this paper. After transforming the original SEM problem to a more tractable equivalent WMMSE problem, we propose the HBF-WMO algorithm with different fixed architectures. Simulation results have shown that the OL architecture achieves a compromise between the hardware costs and system performance compared with the conventional fixed architectures. We have also proposed a low-complexity subarray partition optimization algorithm based on the maximum eigenvalue approximation with the PC-dynamic architecture and combined it with the HBF-WMO algorithm. Simulation results show that the PC-Dynamic architecture achieves some performance gain over the fixed PC architecture.

In this paper, certain problems in the practical implementation of the proposed architecture and HBF-WMO algorithm under massive MIMO-OFDM systems have not been investigated. On one hand, the dynamic architecture and the OL architecture would lead to more power consumption and insertion power loss with more required phase shifters, splitters, combiners and switches than the conventional PC architecture, and thus the energy efficiency could be considered for HBF optimization. On the other hand, some studies have made efforts to alleviate the effect of beam squint while developing hybrid precoding schemes for massive MIMO-OFDM systems, such as carrying out a phase compensation operation at each subcarrier[25] or projecting all frequencies to the central frequency[26]. In future work, we will make efforts to study the energy efficiency performance of different HBF architectures and extend our investigation to the scenario when the system is subject to the beam squint effect.

Reference

[1] PI Z Y, KHAN F. An introduction to millimeter-wave mobile broadband systems [J]. IEEE communications magazine, 2011, 49(6): 101 – 107. DOI: 10.1109/MCOM.2011.5783993
[2] ROH W, SEOL J Y, PARK J, et al. Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results [J]. IEEE communications magazine, 2014, 52(2): 106 – 113. DOI: 10.1109/MCOM.2014.6736750
[3] RANGAN S, RAPPAPORT T S, ERKIP E. Millimeter-wave cellular wireless networks: potentials and challenges [J]. Proceedings of the IEEE, 2014, 102(3): 366 – 385. DOI: 10.1109/JPROC.2014.2299397
[4] PAULRAJ A J, GORE D A, NABAR R U, et al. An overview of MIMO communications: a key to gigabit wireless [J]. Proceedings of the IEEE, 2004, 92(2): 198 – 218. DOI: 10.1109/JPROC.2003.821915
[5] AKDENIZ M R, LIU Y P, SAMIMI M K, et al. Millimeter wave channel modeling and cellular capacity evaluation [J]. IEEE journal on selected areas in communications, 2014, 32(6): 1164 – 1179. DOI: 10.1109/JSAC.2014.2328154
[6] ZHANG J, YU X H, LETAIEF K B. Hybrid beamforming for 5G and beyond millimeter-wave systems: a holistic view [J]. IEEE open journal of the communications society, 2019, 1: 77 – 91. DOI: 10.1109/OJCOMS.2019.2959595
[7] LARSSON E G, EDFORS O, TUFVESSON F, et al. Massive MIMO for next generation wireless systems [J]. IEEE communications magazine, 2014, 52(2): 186 – 195. DOI: 10.1109/MCOM.2014.6736761
[8] MOLISCH A F, RATNAM V V, HAN S Q, et al. Hybrid beamforming for massive MIMO: a survey [J]. IEEE communications magazine, 2017, 55(9): 134 – 141. DOI: 10.1109/MCOM.2017.1600400
[9] AYACH O E, RAJAGOPAL S, ABU-SURRA S, et al. Spatially sparse precoding in millimeter wave MIMO systems [J]. IEEE transactions on wireless communications, 2014, 13(3): 1499 – 1513. DOI: 10.1109/TWC.2014.011714.130846
[10] SOHRABI F, YU W. Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays [J]. IEEE journal on selected areas in communications, 2017, 35(7): 1432 – 1443. DOI: 10.1109/JSAC.2017.2698958
[11] YU X H, SHEN J C, ZHANG J, et al. Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems [J]. IEEE journal of selected topics in signal processing, 2016, 10(3): 485 – 500. DOI: 10.1109/JSTSP.2016.2523903
[12] LIN T, CONG J Q, ZHU Y, et al. Hybrid beamforming for millimeter wave systems using the MMSE criterion [J]. IEEE transactions on communications, 2019, 67(5): 3693 – 3708. DOI: 10.1109/TCOMM.2019.2893632
[13] ZHAO X Y, LIN T, ZHU Y, et al. Partially-connected hybrid beamforming for spectral efficiency maximization via a weighted MMSE equivalence [J]. IEEE transactions on wireless communications, 2021, 20(12): 8218 – 8232. DOI: 10.1109/TWC.2021.3091524
[14] ZHANG D D, WANG Y F, LI X H, et al. Hybridly connected structure for hybrid beamforming in mmWave massive MIMO systems [J]. IEEE transactions on communications, 2018, 66(2): 662 – 674. DOI: 10.1109/TCOMM.2017.2756882
[15] GAUTAM P R, ZHANG L. Hybrid precoding for partial-full mixed connection mmWave MIMO [C]//The IEEE Statistical Signal Processing Workshop (SSP). IEEE, 2021: 271 – 275. DOI: 10.1109/SSP49050.2021.9513847
[16] SONG N, YANG T, SUN H. Overlapped subarray based hybrid beamforming for millimeter wave multiuser massive MIMO [J]. IEEE signal processing letters, 2017, 24(5): 550 – 554. DOI: 10.1109/LSP.2017.2681689
[17] CHEN Y, CHEN D, JIANG T, et al. Millimeter-wave massive MIMO systems relying on generalized sub-array-connected hybrid precoding [J]. IEEE transactions on vehicular technology, 2019, 68(9): 8940 – 8950. DOI: 10.1109/TVT.2019.2930639
[18] MÉNDEZ-RIAL R, RUSU C, GONZÁLEZ-PRELCIC N, et al. Hybrid MIMO architectures for millimeter wave communications: phase shifters or switches? [J]. IEEE access, 2016, 4: 247 – 267. DOI: 10.1109/ACCESS.2015.2514261
[19] ALKHATEEB A, NAM Y H, ZHANG J Z, et al. Massive MIMO combining with switches [J]. IEEE wireless communications letters, 2016, 5(3): 232 – 235. DOI: 10.1109/LWC.2016.2522963
[20] PARK S, ALKHATEEB A, HEATH R W. Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems [J]. IEEE transactions on wire-

less communications, 2017, 16(5): 2907 – 2920. DOI: 10.1109/TWC.2017.2671869
[21] JIN J N, XIAO C S, CHEN W, et al. Channel-statistics-based hybrid precoding for millimeter-wave MIMO systems with dynamic subarrays [J]. IEEE transactions on communications, 2019, 67(6): 3991 – 4003. DOI: 10.1109/TCOMM.2019.2899628
[22] YAN L F, HAN C, YUAN J H. A dynamic array-of-subarrays architecture and hybrid precoding algorithms for terahertz wireless communications [J]. IEEE journal on selected areas in communications, 2020, 38(9): 2041 – 2056. DOI: 10.1109/JSAC.2020.3000876
[23] 3GPP. Study on channel model for frequencies from 0.5 to 100 GHz: TR 38.901 V16.2.0 [S]. 2020
[24] LI H Y, LI M, LIU Q. Hybrid beamforming with dynamic subarrays and low-resolution PSs for mmWave MU-MISO systems [J]. IEEE transactions on communications, 2020, 68(1): 602 – 614. DOI: 10.1109/TCOMM.2019.2950905
[25] CHEN Y, CHEN D, JIANG T, et al. Channel-covariance and angle-of-departure aided hybrid precoding for wideband multiuser millimeter wave MIMO systems [J]. IEEE transactions on communications, 2019, 67(12): 8315 – 8328. DOI: 10.1109/TCOMM.2019.2942307
[26] CHEN Y, XIONG Y F, CHEN D, et al. Hybrid precoding for wideband millimeter wave MIMO systems in the face of beam squint [J]. IEEE transactions on wireless communications, 2021, 20(3): 1847 – 1860. DOI: 10.1109/TWC.2020.3036945

Biographies

TANG Yuanqi received her BS degree in communication science and engineering from Fudan University, China in 2022, where she is currently pursuing her MS degree. Her research interests include hybrid beamforming for massive MIMO systems, millimeter wave signal processing and reconfigurable intelligent surface.

ZHANG Huimin received her BS degree in communication science and engineering from Fudan University, China in 2021, where she is currently pursuing her MS degree. Her current research interests include hybrid beamforming for massive MIMO systems and energy efficiency in intelligent reflecting surface-aided systems.

ZHENG Zheng received his BS and PhD degrees in information science and electronic engineering from Zhejiang University, China in 2013 and 2019, respectively. He is currently a senior algorithm engineer working on physical layer algorithms in ZTE Corporation. His research interests include wireless communications, array signal processing and artificial intelligence algorithms.

LI Ping received her MS degree in communication and information engineering from Xi'an Jiaotong University, China in 2004. She is currently a senior algorithm system engineer at ZTE Corporation, responsible for national key projects. Her research interests include digital signal processing, multiple antenna, system performance optimization, reconfigurable intelligent surface, networking technology, network planning, integrated sensing and communications (ISAC), and key technologies in 5G-A. She has applied for nearly 100 patents and published over 10 papers in various journals and conferences.

ZHU Yu (zhuyu@fudan.edu.cn) received his BE degree (Hons.) in electronics engineering and ME degree (Hons.) in communication and information engineering from the University of Science and Technology of China in 1999 and 2002, respectively, and got his PhD degree in electrical and electronic engineering from the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, China in 2007. Since 2008, he has been with Fudan University, China, where he is currently a professor with the School of Information Science and Technology. His current research interests include broadband wireless communication systems and networks and signal processing for communications. He has served as an editor for the *IEEE Wireless Communications Letters*, and as an editor of the *Journal of Communications and Information Networks*.