IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received May 31, 2019, accepted June 27, 2019, date of publication July 15, 2019, date of current version August 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928559

Energy-Efficient Hybrid Precoding With Low Complexity for mmWave Massive MIMO Systems

YANG LIU[1], (Member, IEEE), QINGXIA FENG[1], QIONG WU[2], YINGHUI ZHANG[1], MINGLU JIN[3], (Member, IEEE), AND TIANSHUANG QIU[3], (Member, IEEE)

[1]College of Electronic Information Engineering, Inner Mongolia University, Hohhot 010021, China
[2]School of Biological Science and Medical Engineering, Beihang University, Beijing 100083, China
[3]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China

Corresponding author: Yinghui Zhang (zhangyinghui_imu@163.com)

This work was supported by the National Science Foundation of China under Grant 61761033 and Grant 61461036.

ABSTRACT Millimeter-wave (mmWave) massive multiple-input multiple-output (MIMO) utilizes large antenna arrays and is considered a promising technology for fifth-generation (5G) and beyond wireless communication systems. However, the high-power consumption of the radio-frequency (RF) chains makes it infeasible. To solve this problem, hybrid precoding is proposed, which is a combination of analog and digital precoding. The fully connected architecture hybrid precoding still requires a large number of phase shifters (PSs). The sub-connected architecture can greatly reduce the required power consumption, and however, it cannot obtain a satisfactory achievable rate. To avoid the high energy consumption and obtain a high resolution, we propose a novel partly connected architecture in this paper. In addition, we propose an energy-efficient successive interference cancelation (SIC) hybrid precoding based on the partly connected architecture, which transforms the problem of maximizing the total achievable rate with non-convex constraints into a series of sub-rate optimization problems. Furthermore, a low-complexity energy-efficient SIC hybrid precoding based on the partly connected architecture is developed, which uses the partial singular value decomposition (SVD) to realize the sub-rate optimization and significantly reduce the complexity. Theoretical analysis demonstrates the superiority of the proposed hybrid precoding in terms of complexity. The simulation results indicate that the proposed hybrid precoding algorithms enjoy better energy efficiency and achievable rate performance than some recently proposed hybrid precoding algorithms.

INDEX TERMS Millimeter wave communication, MIMO, energy efficiency, complexity theory, hybrid precoding.

I. INTRODUCTION

The growth of traffic in mobile communications, which is rapidly increasing with the popularity of mobile devices, has been drawn a great amount of important attention in recent years [1]. Millimeter wave (mmWave) communication with broad unlicensed bandwidth is a promising technology for future wireless communication systems [2]. The mmWave signal attenuation is severe because the free space propagation pathloss is inversely proportional to the wavelength [3]. However, the mmWave massive multiple input multiple output (MIMO) with shorter wavelength is able to make dozens or hundreds of antennas to be packed into a small size, which can compensate for the severe pathloss by sufficient beamforming. Moreover, it can utilize precoding to concentrate the signal in a specific direction [4]. Therefore, considerable attention has been paid to mmWave massive MIMO technology in wireless communications [5], [6].

Fully analog beamforming operates the phase of the signal sent by each antenna via analog phase shifters (PSs) [7]–[9]. Although the fully analog precoding is simply implemented, it cannot provide multiplexing gains for transmitting parallel data streams. For the MIMO systems in conventional frequency band, the classical fully digital precoding is usually used, which can obtain the optimal multiplexing gain [10]. Since the number of radio frequency (RF) chains equals that

The associate editor coordinating the review of this manuscript and approving it for publication was Ning Zhang.

of the antennas for traditional fully digital precoding, it is too costly for mmWave massive MIMO systems. Therefore, novel precoding structures are required for mmWave systems that are quite different from those of conventional communication systems [12]–[15].

The hybrid precoding architecture has drawn considerable attention for mmWave massive MIMO systems, since the number of RF chains is less than that of transmit/receive antennas. Its key idea is to realize precoding in high-dimensional analog and low-dimensional digital domains, where the analog beamforming is implemented by analog circuit, while the digital precoding requires a small number of RF chains [16]. The fully-connected and sub-connected architectures are employed by the existing hybrid precoding. For the fully-connected architecture, there are two ways to design the hybrid precoding. In the first way, a codebook is required to design for improving the performance when optimizing the hybrid precoding [17]–[19]. A codebook was designed in [18], in which wider beams were generated by turning off some antennas. For a multiuser mmWave system, a low-complexity codebook-based RF-baseband hybrid precoder was proposed in [19]. Some codewords in these methods require an analog switch for each antenna element path, which will result in additional costs and power consumption. In the second way, optimization problem is formulated as a sparsity signal reconstruction problem for hybrid precoding [20]–[22]. The hybrid precoding is transformed into a sparse digital precoding optimization problem [23]. A new hybrid precoding algorithm from the perspective of geometric construction was proposed in [24]. Although these algorithms do not consider the designing of codewords, they are all designed based on fully-connected architecture, which involves complicated phase shifter networks and high complexity.

To avoid the problem of fully-connected schemes, hybrid precoding based on sub-connected architecture has been attracted more attention [12], [25], [26]. Compared with fully-connected architecture, it can realize a great reduction in terms of hardware and improve the energy efficiency. However, this architecture leads to an unsatisfing achievable rate performance, which is significantly important in wireless communication systems with the rapid development of mobile communications and the ever growing demand for data rate.

In this paper, we propose a new partly-connected architecture for hybrid precoding. This architecture selects two RF chains to control one sub-antenna array and each antenna is controlled by two phase shifters connected by two RF chains. The partly-connected architecture can achieve a great increase in the achievable rate at the cost of increasing a small number of PSs. In addition, we propose an energy-efficient successive interference cancelation (SIC) hybrid precoding based on partly-connected architecture, which is called partly-connected successive interference cancelation (PC-SIC) precoding. For the proposed PC-SIC hybrid precoding, the total achievable rate optimization problem with non-convex constraints is transformed into a series of simple sub-rate optimization problems. Furthermore, to reduce the complexity of the proposed PC-SIC precoding, we further propose a low-complexity energy-efficient partly-connected SIC hybrid precoding (LPC-SIC) algorithm. The proposed LPC-SIC precoding uses partial singular value decomposition (SVD) in the sub-rate optimization, which can avoid the computational complexity of solving unrelated singular value vectors. Simulation results indicate that the proposed hybrid precoding algorithms can obtain satisfactory performance in terms of the achievable rate and energy efficiency for mmWave massive MIMO systems.

The rest of the paper is organized as follows. The channel model and power model are introduced in Section II. The proposed partly-connected architecture, PC-SIC precoding, and LPC-SIC precoding are described in Section III. In Section IV, the simulation results of the achievable rate and energy efficiency are provided. Section V concludes this paper.

Notation: In this paper, lower-case and upper-case boldface letters denote vectors and matrices, respectively. $(\cdot)^T$, $(\cdot)^{-1}$, $(\cdot)^H$, and $\|\cdot\|_F$ denote the transpose, inverse, conjugate transpose, and Frobenius norm of a matrix, respectively. $\mathbf{I}_N$ is the $N \times N$ identity matrix. $\mathbb{E}(\cdot)$ denotes the expectation. $\mathbb{C}^{m\times n}$ denotes an $m \times n$ dimensional complex space.

II. SYSTEM MODEL

A. CHANNEL MODEL

It is known that the mmWave channel no longer obeys the conventional Rayleigh fading and has different propagation characteristics compared to lower-frequency channels [27]. In this paper, the geometric Saleh-Valenzuela channel model is used for mmWave communications [28], [29]. Based on the Saleh-Valenzuela model, the channel matrix $\mathbf{H}$ can be represented as

$$\mathbf{H} = \tau \sum_{l=1}^{L} \alpha_l \Lambda_r(\phi_l^r, \theta_l^r) \Lambda_t(\phi_l^t, \theta_l^t) \mathbf{a}_r(\phi_l^r, \theta_l^r) \mathbf{a}_t^H(\phi_l^t, \theta_l^t), \tag{1}$$

where $\tau = \sqrt{N^2M/L}$ is a normalization factor, M is the average number of antennas connected to one RF chain, NM is the number of transmit antennas, L is the number of effective channel paths corresponding to limited scatters, and N is the number of RF chains, which usually requires $L \leq N$. α_l is the complex gain of the lth path that follows the Rayleigh distribution. $\phi_l^t(\theta_l^t)$ and $\phi_l^r(\theta_l^r)$ are the lth path's azimuth (elevation) angles of departure and arrival (AoD/AoA), respectively. $\Lambda_t(\phi_l^t, \theta_l^t)$ and $\Lambda_r(\phi_l^r, \theta_l^r)$ denote the transmit and receive antenna array gain at a specific AoD and AoA, respectively. $\mathbf{a}_t^H(\phi_l^t, \theta_l^t)$ and $\mathbf{a}_r(\phi_l^r, \theta_l^r)$ represent the normalized transmit and receive array response vectors based on the antenna array structures at the base station (BS) and the user, respectively. For the uniform linear arrays (ULAs), the array response vector can be presented as

$$\mathbf{a}_{ULA}(\phi) = \frac{1}{\sqrt{U}}\left[1, e^{j\frac{2\pi}{\lambda}d\sin(\phi)}, \cdots, e^{j(U-1)\frac{2\pi}{\lambda}d\sin(\phi)}\right]^T, \tag{2}$$

where λ is the signal wavelength, d is the distance between antenna elements, and U represents the elements for the ULAs.

B. POWER MODEL

The total power consumption is usually modeled as

$$P_{total} = P_{tr} + NP_{RF} + N_{PS}P_{PS}, \tag{3}$$

where P_{tr} is the transmission power, P_{RF} is the power consumed by RF chain, and P_{PS} is the energy consumption of PS. N and N_{PS} are the numbers of RF chains and PSs, respectively.

The energy efficiency can be defined as the ratio of the achievable rate to the total power consumption, which is expressed as

$$\eta = \frac{R}{P_{total}} = \frac{R}{P_{tr} + NP_{RF} + N_{PS}P_{PS}}, \tag{4}$$

where R is the total achievable rate.

We aim to design the analog precoder $\boldsymbol{F}_A$ and digital precoder $\boldsymbol{F}_D$ to maximize the total achievable rate R, which can be expressed as

$$\begin{aligned}(\mathbf{F}_A^{opt}, \mathbf{F}_D^{opt}) = \underset{\mathbf{F}_A, \mathbf{F}_D}{\arg\max}\, R,\\ s.t.\ \mathbf{F}_A \in \mathcal{F},\\ ||\mathbf{F}_A\mathbf{F}_D||_F^2 \leq N,\end{aligned} \tag{5}$$

where $\mathcal{F}$ denotes the set of feasible analog precoders, N is the number of data streams.

III. PROPOSED ENERGY-EFFICIENT HYBRID PRECODING BASED ON PARTLY-CONNECTED ARCHITECTURE

In this section, we first develop a new partly-connected architecture and then propose an energy-efficient partly-connected hybrid precoding to achieve near-optimal performance. Furthermore, an energy-efficient low-complexity partly-connected hybrid precoding algorithm is proposed.

A. PARTLY-CONNECTED ARCHITECTURE FOR MMWAVE MASSIVE MIMO SYSTEM

A new type of partly connected architecture for hybrid precoding in mmWave massive MIMO systems is proposed, as shown in Fig. 1. The proposed partly connected architecture is a compromise proposal between the fully-connected architecture and the sub-connected architecture. In the proposed architecture, NM transmit antennas are equipped in the BS and N independent data streams are sent to K user antennas. Furthermore, the BS has N RF chains, which satisfy $N \leq NM$. Specifically, each two RF chains is connected to one sub-antenna array. Because one sub-antenna array has $2M$ antennas, one RF chain is connected to the sub-antenna array via $2M$ phase shifters and each of the two RF chains is connected to one sub-antenna array via $4M$ phase shifters. Since $2M$ phase shifters are required to transmit one data stream, to transmit N independent data streams the partly-connected architecture requires $2NM$ PSs, while the sub-connected architecture needs NM PSs, and the fully-connected architecture needs N^2M PSs [26].

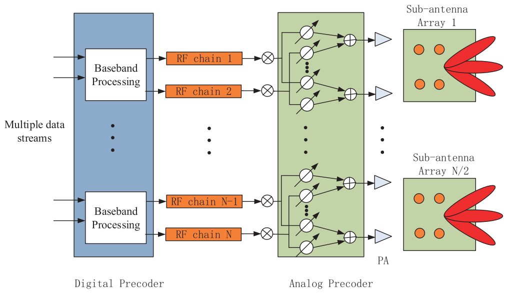

FIGURE 1. Partly-connected architecture.

It can be seen from Fig. 1, although the number of PSs for partly-connected architecture is twice that of the sub-connected architecture, the number of sub-antenna arrays is $N/2$ for partly-connected architecture, which is half of the sub-connected architecture. The hybrid precoder $\mathbf{F}$ at the BS is $NM \times N$. It should be formed from two parts: a high-dimensional analog precoder $\mathbf{F}_A \in \mathbb{C}^{NM\times N}$ and a low-dimensional digital precoder $\mathbf{F}_D \in \mathbb{C}^{N\times N}$, i.e., $\mathbf{F} = \mathbf{F}_A\mathbf{F}_D$ [28].

First, the N data streams in the baseband are precoded by digital precoder $\mathbf{F}_D$, and pass through N RF chains. $\mathbf{F}_D$ can be further specialized to be a subblock diagonal matrix as $\mathbf{F}_D = \text{diag}[\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{N/2}]$ based on the partly-connected architecture, where $\mathbf{d}_n \in \mathbb{C}^{2\times 2}$ for $n = 1, 2, \ldots N/2$. Afterwards, the N data streams are precoded again by an $NM \times N$ analog precoder, which is realized by $2NM$ PSs. $\mathbf{F}_A$ can also be specialized as a subblock diagonal matrix as $\mathbf{F}_A = \text{diag}[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{N/2}]$, where $\mathbf{a}_n \in \mathbb{C}^{2M\times 2}$ for $n = 1, 2, \ldots, N/2$. The elements of $\mathbf{F}_A$ have the same amplitude $1/\sqrt{2M}$, but different phases [20]. Therefore, for one sub-antenna array, two baseband data streams are precoded by $\mathbf{d}_n \in \mathbb{C}^{2\times 2}$, and then pass through two RF chains and are precoded by the analog weighting vector $\mathbf{a}_n \in \mathbb{C}^{2M\times 2}$, which is realized by $4M$ PSs. Therefore, the hybrid precoding for one sub-antenna array is $\mathbf{f}_n = \mathbf{a}_n\mathbf{d}_n$, which satisfies $\mathbf{f}_n \in \mathbb{C}^{2M\times 2}$.

Considering a block-fading propagation channel, the received $K \times 1$ signal vector $\mathbf{y} = [y_1, y_2, \ldots, y_K]^T$ at the user side can be expressed as

$$\mathbf{y} = \sqrt{\rho}\mathbf{H}\mathbf{F}_A\mathbf{F}_D\mathbf{s} + \mathbf{n} = \sqrt{\rho}\mathbf{H}\mathbf{F}\mathbf{s} + \mathbf{n}, \tag{6}$$

where ρ is the average received power, and $\mathbf{y}$ is the $K \times 1$ received vector. $\mathbf{H} \in \mathbb{C}^{K\times NM}$ denotes the channel matrix based on the Saleh-Valenzuela model between the BS and the user. $\mathbf{s} = [s_1, \ldots, s_N]^T$ presents the source signal vector in the baseband. $\mathbf{F} = \mathbf{F}_A\mathbf{F}_D$ is the hybrid precoding matrix

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{f}_{N/2} \end{bmatrix}. \tag{7}$$

The widely used Gaussian signal with a normalized signal power $\mathbb{E}(\mathbf{s}\mathbf{s}^H) = (1/N)\mathbf{I}_N$ is used [28]. To meet the constraint for the total transmit power, it is required that $\|\mathbf{F}\|_F \leq N$ [20]. The additive white Gaussian noise (AWGN) vector $\mathbf{n} = [n_1, n_2, \ldots, n_K]^T$ follows the independent and identically distribution (i.i.d.) $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_K)$.

The hybrid precoding matrix based on sub-connected architecture can be expressed as

$$\mathbf{F}_{sub} = \begin{bmatrix} \mathbf{f}_{sub(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_{sub(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{f}_{sub(N)} \end{bmatrix}, \tag{8}$$

where $\mathbf{f}_{sub(l)} \in \mathbb{C}^{M\times 1}$, for $l = 1, 2, \ldots, N$. It can be seen from the partly-connected architecture based hybrid precoding matrix (7), when the upper triangle and lower triangle of $\mathbf{f}_n$ are both zero matrices, $\mathbf{f}_n$ is equivalent to the rows from the $(2M(n-1)+1)$th one to the $(2M(n-1)+2M)$th one of $\mathbf{f}_{sub(2n-1)}$ and $\mathbf{f}_{sub(2n)}$, respectively, in which $n = 1, 2, \ldots, N/2$. For the partly-connected architecture, $\mathbf{f}_n$ in Equation (7) can be expressed as

$$\mathbf{f}_n = \begin{bmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{c} & \mathbf{d} \end{bmatrix}, \tag{9}$$

where $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$ are $M \times 1$ complex matrices. When the upper triangle $\mathbf{b}$ and lower triangle $\mathbf{c}$ are both zero matrices, $\mathbf{a}$ and $\mathbf{d}$ are the $(2M(n-1)+1)$th one to the $(2M(n-1)+2M)$th one of $\mathbf{f}_{sub(2n-1)}$ and $\mathbf{f}_{sub(2n)}$ $(n = 1, 2, \ldots, N/2)$, respectively. Since the block matrix of the partly-connected architecture can be a zero matrix, the hybrid precoding based on sub-connected architecture is a special case of the proposed partly-connected architecture based hybrid precoding when $\mathbf{b}$ and $\mathbf{c}$ are both zero matrices. Although the hybrid precoding based on sub-connected architecture enjoys better energy efficiency than the fully digital precoding and spatially sparse precoding based on fully-connected architecture [26], the partly-connected architecture can achieve a better performance of achievable rate than the sub-connected architecture. The number of PSs based on partly-connected architecture increases, however, the power consumption of PSs is only on the order of milliwatt. Therefore, the partly-connected architecture can obtain a great increasing of achievable rate at the cost of increasing a small number of PSs, and achieve a higher energy efficiency.

B. ENERGY-EFFICIENT HYBRID PRECODING ALGORITHM BASED ON PARTLY-CONNECTED ARCHITECTURE

Inspired by the successive interference cancelation (SIC) algorithm, we propose an energy-efficient PC-SIC hybrid precoding. The aim of precoding is to maximize the total achievable rate. The achievable rate of mmWave massive MIMO systems can be expressed as

$$R = \log_2\left(\left|\mathbf{I}_k + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{F}\mathbf{F}^H\mathbf{H}^H\right|\right). \tag{10}$$

According to the system model in Section II, the hybrid precoding matrix $\mathbf{F}$ can be represented as $\mathbf{F} = \mathbf{F}_A\mathbf{F}_D = \mathrm{diag}[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{N/2}] * \mathrm{diag}[\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{N/2}]$. There are three constraints on $\mathbf{F}$:

1) $\mathbf{F}$ should be a block diagonal matrix which is shown in (7), i.e., $\mathbf{F} = \mathrm{diag}[\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{N/2}]$, where $\mathbf{f}_n = \mathbf{a}_n\mathbf{d}_n$ is the $2M \times 2$ non-zero vector of the $(N-2n+1)$th and $(N-2n+2)$th columns $(n = 1, 2, \ldots N/2)$.
2) The non-zero elements of each column of $\mathbf{F}_A$ should have the same amplitude.
3) To fulfill the requirements of the total transmit power constraint, the Frobenius norm of $\mathbf{F}$ should meet $\|\mathbf{F}\|_F \leq N_s$.

Unfortunately, the total achievable rate R is difficult to be optimized because of the non-convex constraints on $\mathbf{F}$. Motivated by [30], we can decompose the complicated optimization problem (10) into a series of sub-rate optimization problems based on the partly-connected architecture, which is much easier to solve.

The precoding on different sub-antenna arrays is independent based on the block diagonal structure of the hybrid precoding matrix $\mathbf{F}$. Therefore, $\mathbf{F}$ can be divided into three parts as follows $\mathbf{F} = [\mathbf{F}_{1:N-2}\ \mathbf{f}_{N-1}\ \mathbf{f}_N]$, where $\mathbf{F}_{1:N-2}$ is an $NM \times (N-2)$ matrix containing the first $(N-2)$ columns of $\mathbf{F}$, $\mathbf{f}_{N-1}$ and $\mathbf{f}_N$ are the $(N-1)$th and the Nth column of $\mathbf{F}$, respectively. Then, the total achievable rate R in (10) can be presented as

$$\begin{aligned} R &= \log_2\left(\left|\mathbf{I}_k + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{F}\mathbf{F}^H\mathbf{H}^H\right|\right) \\ &= \log_2\left(\left|\mathbf{I}_k + \frac{\rho}{N\sigma^2}\mathbf{H}[\mathbf{F}_{1:N-2}\ \mathbf{f}_{N-1}\ \mathbf{f}_N]\right.\right. \\ &\qquad \left.\left.[\mathbf{F}_{1:N-2}\ \mathbf{f}_{N-1}\ \mathbf{f}_N]^H\mathbf{H}^H\right|\right) \\ &= \log_2\left(\left|\mathbf{I}_k + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{F}_{1:N-2}\mathbf{F}_{1:N-2}^H\mathbf{H}^H\right.\right. \\ &\qquad \left.\left.+ \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{f}_{N-1}\mathbf{f}_{N-1}^H\mathbf{H}^H + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{f}_N\mathbf{f}_N^H\mathbf{H}^H\right|\right). \end{aligned} \tag{11}$$

The auxiliary matrix $\mathbf{T}_{N-2} = \mathbf{I}_k + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{F}_{1:N-2}\mathbf{F}_{1:N-2}^H\mathbf{H}^H$ is defined, and then R can be stated in a closed-form as follows

$$\begin{aligned} R &= \log_2\left(|\mathbf{T}_{N-2}|\right) \\ &\quad + \log_2\left(\left|\mathbf{I}_k + \frac{\rho}{N\sigma^2}\mathbf{T}_{N-2}^{-1}\mathbf{H}\left(\mathbf{f}_{N-1}\mathbf{f}_{N-1}^H + \mathbf{f}_N\mathbf{f}_N^H\right)\mathbf{H}^H\right|\right) \\ &= \log_2\left(|\mathbf{T}_{N-2}|\right) \\ &\quad + \log_2\left(\left|\mathbf{I}_k + \frac{\rho}{N\sigma^2}\left(\mathbf{f}_{N-1}^H\mathbf{G}_{N-2}\mathbf{f}_{N-1} + \mathbf{f}_N^H\mathbf{G}_{N-2}\mathbf{f}_N\right)\right|\right), \end{aligned} \tag{12}$$

where $\mathbf{G}_{N-2} = \mathbf{H}^H\mathbf{T}_{N-2}^{-1}\mathbf{H}$. The first term $\log_2(|\mathbf{T}_{N-2}|)$ of (12) has the same form as (10). Hence, $\log_2(|\mathbf{T}_{N-2}|)$ can be further decomposed by using the similar method in (11), and it can be given by

$$\begin{aligned}\log_2(|\mathbf{T}_{N-2}|) = \log_2(|\mathbf{T}_{N-4}|) \\ + \log_2\left(\left|\mathbf{I}_k + \frac{\rho}{N\sigma^2}(\mathbf{f}_{N-3}^H\mathbf{G}_{N-4}\mathbf{f}_{N-3}\right.\right. \\ \left.\left. + \mathbf{f}_{N-2}^H\mathbf{G}_{N-4}\mathbf{f}_{N-2})\right|\right),\end{aligned} \tag{13}$$

where $\mathbf{G}_{N-4} = \mathbf{H}^H\mathbf{T}_{N-4}^{-1}\mathbf{H}$, $\mathbf{T}_{N-4} = \mathbf{I}_k + \frac{\rho}{N\sigma^2}\mathbf{H}\mathbf{F}_{1:N-4}$ $\mathbf{F}_{1:N-4}^H\mathbf{H}^H$.

After $N/2$ decompositions, the total achievable rate R in (12) can be expressed as

$$R = \sum_{n=1}^{N/2} \log_2\left(1 + M_{N-2n+1} + M_{N-2n+2}\right), \tag{14}$$

where $\mathbf{G}_{N-2n} = \mathbf{H}^H\mathbf{T}_{N-2n}^{-1}\mathbf{H}$, $M_{N-2n+1} = \frac{\rho}{N\sigma^2}\mathbf{f}_{N-2n+1}^H$ $\mathbf{G}_{N-2n}\mathbf{f}_{N-2n+1}$, and $M_{N-2n+2} = \frac{\rho}{N\sigma^2}\mathbf{f}_{N-2n+2}^H\mathbf{G}_{N-2n}\mathbf{f}_{N-2n+2}$.

The problem of maximizing the total achievable rate can be transformed into a series of sub-rate optimization problems of sub-antenna arrays based on (14). This can significantly simplify a complex problem that has non-convex constraints. First, because each antenna array is connected to two RF chains independently, the achievable capacity of the first antenna array can be optimized by assuming that all the other antenna arrays are closed. Afterwards, by excluding the contribution of the first antenna array from (10), the achievable rate of the second antenna array can be optimized. A similar procedure is executed until the last antenna array is considered.

According to the analysis above, the sub-rate optimization problem of the nth sub-antenna array can be given by

$$\begin{aligned}\mathbf{f}_{N-2n+1}^{opt}, \mathbf{f}_{N-2n+2}^{opt} &= \underset{\mathbf{p}_{N-2n+1}^{opt},\mathbf{p}_{N-2n+2}^{opt}\in F}{\arg\max} \log_2(1 \\ &\quad + M_{N-2n+1} + M_{N-2n+2}) \\ &\overset{(a)}{=} \underset{\mathbf{p}_{N-2n+1}^{opt},\mathbf{p}_{N-2n+2}^{opt}\in F}{\arg\max} \log_2(1 \\ &\quad + \frac{\rho}{N\sigma^2}(\mathbf{f}_{N-2n+1}^H\mathbf{G}_{N-2n}\mathbf{f}_{N-2n+1} \\ &\quad + \mathbf{f}_{N-2n+2}^H\mathbf{G}_{N-2n}\mathbf{f}_{N-2n+2})),\end{aligned} \tag{15}$$

where (a) is obtained by defining the auxiliary matrix $\mathbf{G}_{N-2n}$, $\mathcal{F}$ is the set of all the feasible vectors satisfying the three constraints proposed in Section III-A. The $(N-2n+1)$th and $(N-2n+2)$th precoding vectors $\mathbf{f}_{N-2n+1}^{opt}$ and $\mathbf{f}_{N-2n+2}^{opt}$ only have $2M$ non-zero elements from the $(2M(n-1)+1)$th one to the $(2M(n-1)+2M)$th one, in which $n = 1, 2, \ldots, N/2$. Therefore, the sub-rate optimization problem (15) becomes

$$\begin{aligned}\bar{\mathbf{f}}_{N-2n+1}^{opt}, \bar{\mathbf{f}}_{N-2n+2}^{opt} &= \underset{\bar{\mathbf{p}}_{N-2n+1}^{opt},\bar{\mathbf{p}}_{N-2n+2}^{opt}\in\bar{F}}{\arg\max} \log_2(1 \\ &\quad + \frac{\rho}{N\sigma^2}(\bar{\mathbf{f}}_{N-2n+1}^H\bar{\mathbf{G}}_{N-2n}\bar{\mathbf{f}}_{N-2n+1} \\ &\quad + \bar{\mathbf{f}}_{N-2n+2}^H\bar{\mathbf{G}}_{N-2n}\bar{\mathbf{f}}_{N-2n+2})),\end{aligned} \tag{16}$$

where $\overline{\mathcal{F}}$ includes all the possible $2M \times 2$ vectors satisfying the last two constraints, $\overline{\mathbf{G}}_{N-2n}$ of size $2M \times 2M$ is the corresponding sub-matrix of $\mathbf{G}_{N-2n}$ by only keeping the rows and columns of $\mathbf{G}_{N-2n}$ from the $(2M(n-1)+1)$th one to the $(2M(n-1)+2M)$th one, which can be presented as

$$\overline{\mathbf{G}}_{N-2n} = \mathbf{R}\mathbf{G}_{N-2n}\mathbf{R}^H = \mathbf{R}\mathbf{H}^H\mathbf{T}_{N-2n}^{-1}\mathbf{H}\mathbf{R}^H, \tag{17}$$

where $\mathbf{R} = [\mathbf{0}_{2M\times 2M(n-1)}\ \mathbf{I}_{2M}\ \mathbf{0}_{2M\times 2M(N/2-n)}]$ is the corresponding selection matrix.

The SVD of $\overline{\mathbf{G}}_{N-2n}$ can be represented by $\overline{\mathbf{G}}_{N-2n} = \mathbf{V}\sum\mathbf{V}^H$, where $\mathbf{V}$ is a $2M \times 2M$ unitary matrix, $\mathbf{v}_1$ is the first column of $\mathbf{V}$ and $\mathbf{v}_2$ is the second column of $\mathbf{V}$. The $\sum$ is a $2M \times 2M$ diagonal matrix that contains the singular values of $\overline{\mathbf{G}}_{N-2n}$ in decreasing order. Since the elements of $\mathbf{v}_1$ and $\mathbf{v}_2$ do not obey the second constraint in Section III-A, the hybrid precoding vectors $\bar{\mathbf{f}}_{N-2n+1}^{opt}$ and $\bar{\mathbf{f}}_{N-2n+2}^{opt}$ cannot be directly chosen as $\mathbf{v}_1$ and $\mathbf{v}_2$, respectively. To satisfy the constraints, the following approach is proposed.

First, the analog precoding is given by

$$\bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt} = \frac{1}{\sqrt{2M}} e^{jangle(\mathbf{v}_{1:2})}. \tag{18}$$

Then, $\mathbf{v}_1$ and $\mathbf{v}_2$ are chosen as $\bar{\mathbf{f}}_{N-2n+1}^{opt}$ and $\bar{\mathbf{f}}_{N-2n+2}^{opt}$, respectively. To make $\bar{\mathbf{f}}_{N-2n+1}^{opt}$ and $\bar{\mathbf{f}}_{N-2n+2}^{opt}$ sufficiently close to $\mathbf{v}_1$ and $\mathbf{v}_2$, the digital precoding can be given by

$$\begin{aligned}&\bar{\mathbf{d}}_{N-2n+1:N-2n+2}^{opt} \\ &= \frac{\left(\mathbf{f}_{N-2n+1:N-2n+2}^{opt}\right)^T \bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt}}{2\left(\bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt}\right)^T \bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt}} \\ &\quad + \frac{\left(\bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt}\right)^T \mathbf{f}_{N-2n+1:N-2n+2}^{opt}}{2\left(\bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt}\right)^T \bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt}}.\end{aligned} \tag{19}$$

Finally, let

$$\bar{\mathbf{f}}_{N-2n+1:N-2n+2}^{opt} = \bar{\mathbf{a}}_{N-2n+1:N-2n+2}^{opt}\bar{\mathbf{d}}_{N-2n+1:N-2n+2}^{opt}. \tag{20}$$

Thus, $\bar{\mathbf{f}}_{N-2n+1}^{opt}$ and $\bar{\mathbf{f}}_{N-2n+2}^{opt}$ are updated. It is worth pointing out that $\mathbf{v}_1$ and $\mathbf{v}_2$ are columns of the unitary matrix $\mathbf{V}$, and each element v_i of $\mathbf{v}_1$ and $\mathbf{v}_2$ (for $i = 1, 2, \ldots, 2M$) has an amplitude less than one, then $\left\|\bar{\mathbf{f}}_{N-2n+1:N-2n+2}^{opt}\right\|_2^2 \leq 2$. Because the optimal solution $\bar{\mathbf{f}}_{N-2n+1:N-2n+2}^{opt}$ for $n = 1, 2, \ldots, N/2$ has a similar form, we can conclude that $\left\|\mathbf{f}^{opt}\right\|_F^2 = \left\|\mathrm{diag}\left\{\bar{\mathbf{f}}_1^{opt}, \bar{\mathbf{f}}_2^{opt}, \ldots, \bar{\mathbf{f}}_{N/2}^{opt}\right\}\right\|_F^2 \leq N$, in which $\bar{\mathbf{f}}_n^{opt} \in \mathbb{C}^{2M\times 2}$.

As mentioned above, the proposed PC-SIC hybrid precoding algorithm, i.e., the following steps in Algorithm 1, can solve the sub-rate optimization problem of the nth sub-antenna array (for $n = 1, 2, \ldots N/2$).

C. LOW-COMPLEXITY HYBRID PRECODING ALGORITHM BASED ON PARTLY-CONNECTED ARCHITECTURE

To reduce the complexity of the proposed PC-SIC precoding, an energy-efficient low-complexity partly-connected architecture based successive interference cancelation hybrid precoding, i.e., LPC-SIC, is proposed in this section.

Algorithm 1 Proposed PC-SIC Hybrid Precoding Algorithm

```
Initialize: $\overline{\mathbf{G}}_{N-2}$;
for $n = 1, 2, \ldots, N/2$ do
    Obtain $\mathbf{v}_1$ and $\mathbf{v}_2$ by performing the SVD of $\overline{\mathbf{G}}_{N-2n}$;
    Obtain the value of $\overline{\mathbf{a}}^{opt}_{N-2n+1:N-2n+2} = \mathrm{A}$ (18);
    Let $\mathbf{v}_1$ and $\mathbf{v}_2$ as $\overline{\mathbf{f}}^{opt}_{N-2n+1}$ and $\overline{\mathbf{f}}^{opt}_{N-2n+2}$, respectively;
    Obtain the value of $\overline{\mathbf{d}}^{opt}_{N-2n+1:N-2n+2} = \mathrm{D}$ (19);
    Let $\overline{\mathbf{f}}^{opt}_{N-2n+1:N-2n+2} = \mathrm{A} \times \mathrm{D}$;
    Update matrices $\mathbf{T}_{N-2n}$ and $\overline{\mathbf{G}}_{N-2n}$ for the next $(m+1)$th sub-antenna array. (for $m = 1, 2, \ldots(N/2-1)$);
end for
```
Output: hybrid precoding matrix $\mathbf{F}$;

The SVD algorithm usually obtains all the singular values of the matrix. To obtain the complete SVD of the matrix, redundant calculations are inevitably generated in massive MIMO systems. When the number of singular values or singular vectors required is much smaller than the matrix dimension, the waste of computational is enormous. Fortunately, hybrid precoding based on partly-connected architecture does not need to obtain all the singular value distributions of the matrix and corresponding singular vectors. Only the singular vectors corresponding to the first two large singular values need to be obtained to further gain a hybrid precoding matrix. Therefore, a partial singular value decomposition method is proposed in this section, which only calculates the singular values and singular vectors required by the partly-connected architecture.

The partial singular value decomposition method mainly exploits the Givens transformation. This transformation can make the original matrix orthogonal, which can be expressed as

$$\mathbf{AV} = \mathbf{W}, \tag{21}$$

where $\mathbf{A} \in \mathbb{C}^{m \times n}$, $m > n$, $\mathbf{V}$ is an unitary matrix, $\mathbf{W}$ satisfies the column vectors are orthogonal to each other, which is given by

$$\mathbf{W} = \mathbf{U}\Lambda, \tag{22}$$

where $\mathbf{U}$ is the left singular vector of $\mathbf{A}$, and Λ is the singular value of $\mathbf{A}$.

Because not all the singular values are required in the system, only r singular values are required ($r < n$),

$$\mathbf{AV} = (\mathbf{W}_1, \mathbf{W}_2). \tag{23}$$

Let

$$\mathbf{W}_1{=}\mathbf{U}_1\Lambda_1, \tag{24}$$

where

$$\mathbf{U}_1(:,i){=}\frac{\mathbf{W}_1(:,i)}{\|\mathbf{W}_1(:,i)\|_2}. \tag{25}$$

It is satisfied that $0 < i < r$. $\mathbf{W}_1$ and $\mathbf{W}_2$ satisfy

$$\mathbf{W}_1^H\mathbf{W}_1 = \Lambda_1^2, \tag{26}$$

$$\mathbf{W}_1^H\mathbf{W}_2 = \mathbf{0}_{r\times(n-r)}. \tag{27}$$

Taking Equation (24) into Equation (26) and Equation (27), we can obtain

$$\mathbf{U}_1^H\mathbf{U}_1 = \mathbf{I}_r, \tag{28}$$

$$\mathbf{U}_1^H\mathbf{W}_2 = \mathbf{0}_{r\times(n-r)}. \tag{29}$$

And then

$$\mathbf{AV} = (\mathbf{U}_1\Lambda_1, \mathbf{W}_2). \tag{30}$$

Multiply $\mathbf{U}_1^H$ on both sides of the above equation, and we can get

$$\mathbf{U}_1^H\mathbf{AV} = (\mathbf{U}_1^H\mathbf{U}_1\Lambda_1, \mathbf{U}_1^H\mathbf{W}_2) = (\Lambda_1, \mathbf{0}). \tag{31}$$

Then,

$$\begin{aligned}\mathbf{U}_1^H\mathbf{AV}\left(\mathbf{U}_1^H\mathbf{AV}\right)^H &= \mathbf{U}_1^H\mathbf{AVV}^H\mathbf{A}^H\mathbf{U}_1\\ &= \mathbf{U}_1^H\mathbf{AA}^H\mathbf{U}_1 = \Lambda_1^2.\end{aligned} \tag{32}$$

Thus,

$$\mathbf{AA}^H\mathbf{U}_1 = \mathbf{U}_1\Lambda_1^2. \tag{33}$$

Therefore, Λ_1^2 is the singular value of $\mathbf{AA}^H$. In addition, $\mathbf{U}_1$ is the singular vector corresponding to a singular value.

As mentioned above, the key idea of the partial SVD algorithm is to make better use of the orthogonal mapping, which can make the two vectors orthogonal. Because only the partial singular values and corresponding singular vectors are required by the partly-connected architecture, the partial column vectors are orthogonal to each other and the partial column vectors and the remaining column vectors are orthogonal to each other. Therefore, the partial SVD method eliminates the orthogonal work of the unrelated column vectors, which can significantly reduce the computational load for each round.

Based on partly-connected architecture, the singular values of $\overline{\mathbf{G}}_{N-2n}$ are required. They are arranged in descending order. r (for a partly-connected architecture, $r = 2$) singular values and r singular vectors are required by the system. We assume that the ith singular value is σ_i, in which, $0 \leq i < r$.

First, the 2-norm of the column vector of matrix $\overline{\mathbf{G}}_{N-2n}$, i.e., $\overline{\mathbf{G}}_{N-2n}(:,i)$ and $\overline{\mathbf{G}}_{N-2n}(:,j)$ are calculated, in which $i < j \leq r$. If $\left\|\overline{\mathbf{G}}_{N-2n}(:,j)\right\|_2 < \left\|\overline{\mathbf{G}}_{N-2n}(:,i)\right\|_2$, directly perform the orthogonal transformation. Otherwise, $\overline{\mathbf{G}}_{N-2n}(:,j)$ and $\overline{\mathbf{G}}_{N-2n}(:,i)$ are changed, and the orthogonal transformation is then performed. If $i > r$, end the loop. According to Equation (25), after several iteration cycles, the singular vectors can be calculated and expressed as follows

$$\mathbf{v}_i = \frac{\widetilde{\mathbf{G}}_{N-2n}(:,i)}{\left\|\widetilde{\mathbf{G}}_{N-2n}(:,i)\right\|_2}, \tag{34}$$

where $\widetilde{\mathbf{G}}_{N-2n}(:,i)$ is the updated version of $\overline{\mathbf{G}}_{N-2n}(:,i)$. A detailed description of the proposed partial SVD algorithm is summarized in the following.

In conclusion, the proposed LPC-SIC algorithm uses partial SVD in the first step of the sub-rate optimization problem

TABLE 1. Computational complexity analysis.

	Number of Multiplications	Number of Divisions
Spatially sparse precoding[20]	$\mathcal{O}\left(N^4M+N^2L^2+N^2M^2L\right)$	$\mathcal{O}\left(2N^3\right)$
SIC-based hybrid precoding[26]	$\mathcal{O}\left(M^2\left(NS+K\right)\right)$	$\mathcal{O}\left(2NS\right)$
Proposed PC-SIC precoding	$\mathcal{O}\left(4M^3N+16MN\right)$	$\mathcal{O}\left(4N\right)$
Proposed LPC-SIC precoding	$\mathcal{O}\left(4M^2N+14MN\right)$	$\mathcal{O}\left(4N\right)$

```
Algorithm 2 Partial Singular Value Decomposition
Input: $\overline{\mathbf{G}}_{N-2n}$ and $r$;
Initialize: $i = 1$ and $j = 2$;
Calculate 2-norm of column vector of matrix $\overline{\mathbf{G}}_{N-2n}$;
while $j \leq r$ do
    if $\left\|\overline{\mathbf{G}}_{N-2n}(:,j)\right\|_2 < \left\|\overline{\mathbf{G}}_{N-2n}(:,i)\right\|_2$ then
        Orth $\left\|\overline{\mathbf{G}}_{N-2n}(:,i)\right\|_2, \left\|\overline{\mathbf{G}}_{N-2n}(:,j)\right\|_2$;
    else
        Change and orth $\left\|\overline{\mathbf{G}}_{N-2n}(:,i)\right\|_2, \left\|\overline{\mathbf{G}}_{N-2n}(:,j)\right\|_2$;
    end if
    $i++, j++$;
end while
Output: $\mathbf{v}_i = \widetilde{\mathbf{G}}_{N-2n}(:,i) / \left\|\widetilde{\mathbf{G}}_{N-2n}(:,i)\right\|_2$ (34);
```

to replace the SVD algorithm, which can significantly reduce the complexity.

D. COMPUTATIONAL COMPLEXITY EVALUATION

The computational complexity of the proposed partly-connected architecture based hybrid precoding algorithms is mainly composed of the following three parts.

1) The first part is from the SVD of matrix $\overline{\mathbf{G}}_{N-2n}$. The dimension of matrix $\overline{\mathbf{G}}_{N-2n}$ is $2M \times 2M$. The computational complexity of directly implementing SVD is $\mathcal{O}\left(4M^3N\right)$. For the proposed LPC-SIC hybrid precoding algorithm which requires the first two column vectors and uses a partial SVD algorithm, the computational complexity is $\mathcal{O}\left(4M^2N-2MN\right)$.

2) The second part from obtaining a digital precoding for solving Equation (19), which needs $\mathcal{O}\left(12MN\right)$ multiplications and $\mathcal{O}\left(4N\right)$ divisions.

3) The third part is from obtaining a hybrid precoding for solving Equation (20), which needs $\mathcal{O}\left(4MN\right)$ multiplications.

According to the above analysis, the overall complexity of the proposed PC-SIC hybrid precoding and the proposed LPC-SIC hybrid precoding is $\mathcal{O}\left(4M^3N+16MN+4N\right)$ and $\mathcal{O}\left(4M^2N+14MN+4N\right)$, respectively.

The complexity comparison of the spatially sparse precoding [20], SIC-based hybrid precoding [26], and the proposed PC-SIC and LPC-SIC precodings is shown in Table 1. It is assumed that $N = 8$, $M = 8$, $K = 16$, $L = 3$, and $S = 5$, which are the typical values in mmWave MIMO systems [20]. Note that the computational complexity of the spatially sparse precoding is approximately 5×10^4 multiplications and 10^3 divisions. The complexity of the SIC-based hybrid precoding requires approximately 4×10^3 multiplications and 10^2 divisions. The multiplications and divisions times of the proposed PC-SIC precoding are approximately 2×10^4 and 3×10, respectively. While the number of multiplications and divisions of the proposed LPC-SIC precoding are approximately 3×10^3 and 3×10, respectively. Therefore, the proposed LPC-SIC precoding enjoys significantly lower complexity, which is only approximately 7% of the spatially sparse precoding complexity and approximately 18% of the proposed PC-SIC precoding complexity. In addition, the computational complexity of the proposed LPC-SIC precoding is approximately 82% of SIC-based precoding complexity.

IV. SIMULATION RESULTS

In this section, we provide simulation results to demonstrate the performance advantages of the proposed algorithms. We compare our hybrid precoding with the fully-connected architecture based optimal unconstrained precoding and spatially sparse precoding. The performance of the SIC-based hybrid precoding with the sub-connected architecture and the conventional analog precoding with sub-connected architecture is also presented. The mmWave channel model in Equation (1) is used where the number of scatters is set as $L = 3$ [20], [31]. The ULAs with antenna spacing of $d = \lambda/2$ are employed for both transmit and receive antenna arrays. The AoA and AoD of each element are uniformly distributed in $[-\pi, \pi]$ and $\left[-\pi/6, \pi/6\right]$, respectively. The carrier frequency is 28 GHz. The perfect channel state information (CSI) scenario is considered. The SNR is defined as $\mathrm{SNR} = \frac{\rho}{\sigma^2}$, where the noise variance is $\sigma^2 = 1$.

A. THE PERFORMANCE OF ENERGY EFFICIENCY

We consider the scenario of a small cell transmission and set the parameters as follows, $\mathrm{P_{RF}}$= 250 mW, $\mathrm{P_{PS}}$= 1 mW, and $\mathrm{P_t}$= 1 W [26]. Fig. 2 shows the energy efficiency against the number of RF chains N, where SNR $=$ 0 dB, $NM = K = 128$ ($N = 8, 16, \ldots, 128$ to ensure that M is an integer). The number of RF chains from 63 to 65 is shown in Fig. 2. It can be seen clearly that the proposed precoding algorithms are more energy efficient than the other algorithms when the number of RF chains is between 8 and 64. The simulation results also demonstrate that with an increasing number of RF chains, the SIC hybrid precoding based on

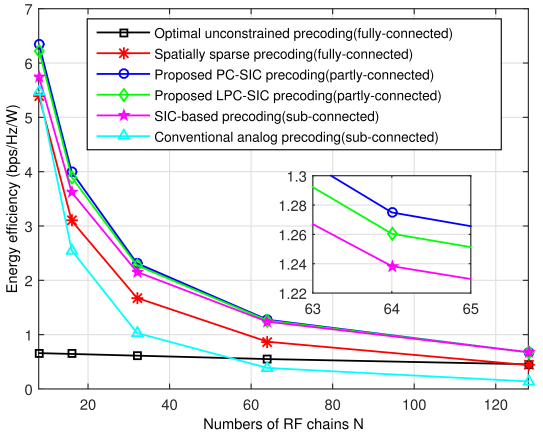

FIGURE 2. Energy efficiency comparison against the numbers of RF chains N, where $NM = K = 128$.

sub-connected architecture has a similar performance to those of the proposed PC-SIC and LPC-SIC hybrid precoding algorithms. It is also obvious that fully digital precoding, i.e., optimal unconstrained precoding, has low energy efficiency, since the corresponding large number of RF chains and PSs requires a high energy consumption. Therefore, the proposed partly-connected architecture can achieve satisfied energy efficiency.

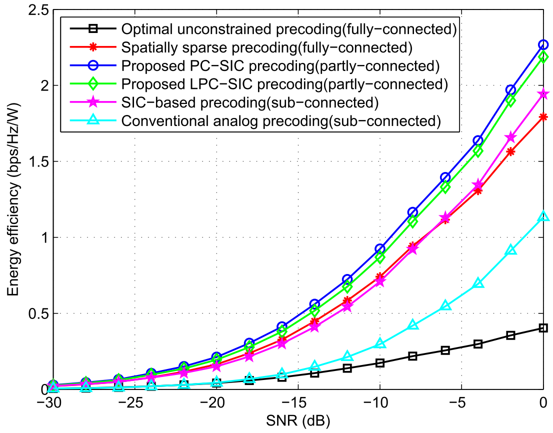

FIGURE 3. Energy efficiency comparison against SNR, where $NM = 128$ and $K = 16$.

Fig. 3 shows the energy efficiency of the proposed PC-SIC and LPC-SIC hybrid precoding algorithms in a 128 × 16 mmWave massive MIMO system, where $N = 16$. It can be observed that the proposed algorithms can achieve better energy efficiency performance than the hybrid precoding based on fully-connected architecture (including the optimal unconstrained precoding and spatially sparse precoding) and sub-connected architecture (including SIC-based precoding and conventional analog precoding). In addition, the energy efficiency gap between the proposed precoding algorithms and the other algorithms will increase as the SNR increases, which implies that the proposed algorithms have much better energy efficiency for high SNR.

B. EFFECT OF RF CHAINS

Two typical mmWave massive MIMO configurations with $NM \times K = 128 \times 16$ ($M = 8$) and $NM \times K = 128 \times 32$ ($M = 4$) are used to study the effect of the number of RF chains [30]. Fig. 4 and Fig. 5 show a comparison of achievable rate against SNR in the mmWave massive MIMO system. It is obvious that the achievable rates of the proposed PC-SIC and LPC-SIC hybrid precoding algorithms are close to those of the optimal unconstrained precoding and spatially sparse hybrid precoding using fully-connected architecture. The proposed PC-SIC and LPC-SIC algorithms outperform SIC-based hybrid precoding and conventional analog precoding with sub-connected architecture over the whole simulated SNR range.

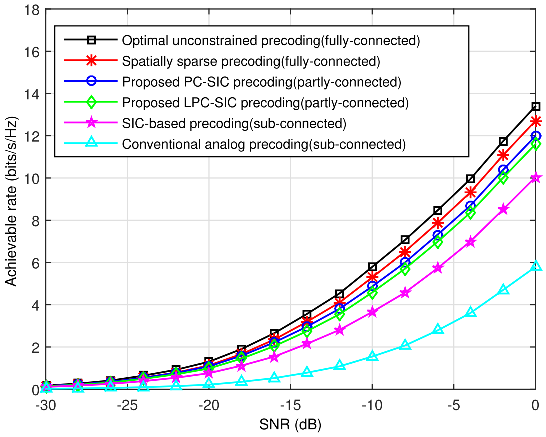

FIGURE 4. The achievable rate comparison against SNR for mmWave massive MIMO systems, where $NM \times K = 128 \times 16$ ($M = 8$).

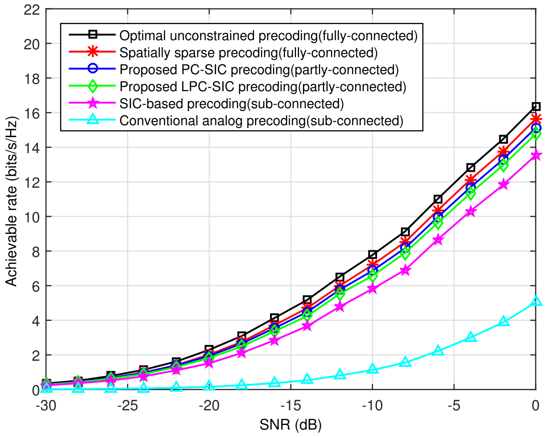

FIGURE 5. The achievable rate comparison against SNR for mmWave massive MIMO systems, where $NM \times K = 128 \times 32$ ($M = 4$).

In addition, it can be seen from Fig. 4 and Fig. 5 that the performance of hybrid precoding can be greatly improved by increasing the number of RF chains. For example, when SNR = 0 dB and the number of RF chains is 16, the achievable rate of the proposed LPC-SIC algorithm is 11.5995 bits/s/Hz. While the achievable rate of the proposed

LPC-SIC algorithm is 14.7650 bits/s/Hz when SNR = 0 dB and the number of RF chains is 32. Furthermore, the proposed LPC-SIC precoding achieves more than 97% of the achievable rate achieved by the proposed PC-SIC precoding in both of the simulated configurations. Considering the complexity analysis mentioned in Section III-D, we conclude that the proposed algorithms impose significantly lower complexity at the cost of a modest degradation in achievable rate. Moreover, the proposed LPC-SIC algorithm achieves the achievable rate performance close to that of the proposed PC-SIC algorithm, which indicates that the proposed partial SVD enjoys low complexity and satisfactory achievable rate.

C. EFFECT OF ANTENNAS

The performance of the achievable rate against the numbers of BS and user antennas is shown in Fig. 6. In this case, the number of base station antennas equals the number of user antennas. The number of RF chains is fixed at 8 and the SNR = 0 dB. We note that the performance of the proposed PC-SIC and LPC-SIC algorithms can be improved by increasing the number of BS antennas and user antennas. Since the energy consumption of RF chains is much higher than that of antennas, the energy consumption of increasing the user antennas can be ignored [32]. Therefore, compared to increasing the number of energy-intensive RF chains, the proposed hybrid precoding can significantly reduce the energy cost. In addition, it can be seen from Fig. 6 that the proposed algorithms are obviously superior to the algorithms based on sub-connected architecture and close to the algorithms based on fully-connected architecture.

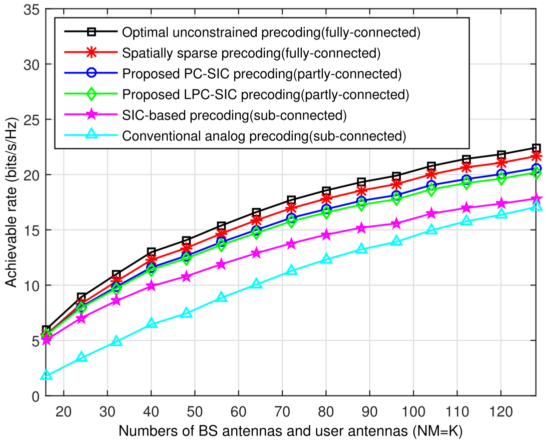

FIGURE 6. The achievable rate comparison against the numbers of BS antennas and user antennas *NM* = *K*, where *N* = 8 and SNR = 0 dB.

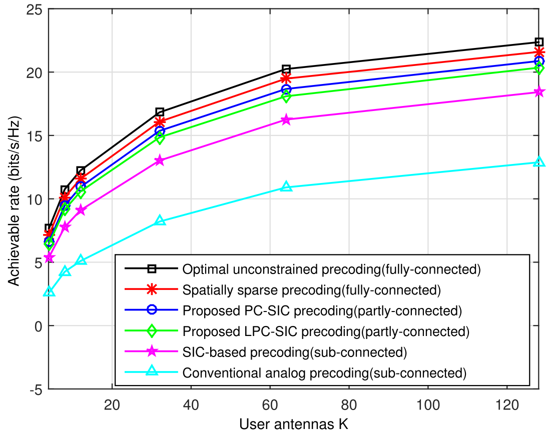

FIGURE 7. The achievable rate comparison against the numbers of user antennas when SNR = 0 dB and *NM* = 128.

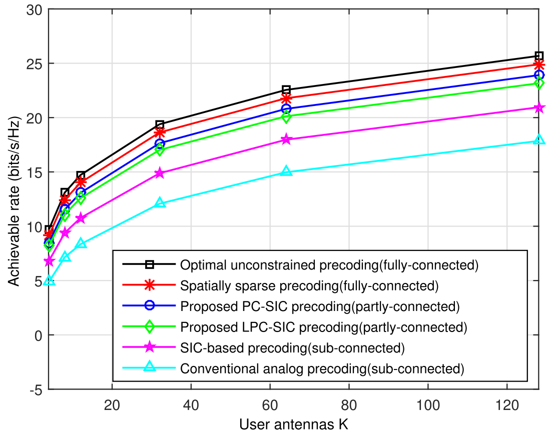

FIGURE 8. The achievable rate comparison against the numbers of user antennas when SNR = 0 dB and *NM* = 256.

D. EFFECT OF TRANSMIT AND USER ANTENNAS

The achievable rate comparison against the number of user antennas is shown in Fig. 7 and Fig. 8, where SNR = 0 dB, and $N = 16$. The number of user antennas is 4, 8, 16, 32, 64 and 128. First, it can be seen that the achievable rate will improve significantly when increasing the number of user antennas. The proposed algorithms are still obviously superior to the sub-connected architecture as the number of user antennas increases. Furthermore, it can be observed that increasing the number of user antennas can compensate for the performance loss of the proposed algorithms. For example, in Fig. 7 when the number of user antennas is 64, the achievable rate of optimal unconstrained precoding based on fully-connected architecture is 20.2409 bits/s/Hz. While the achievable rate of the proposed LPC-SIC precoding is 20.3474 bits/s/Hz when the number of user antennas is 128, which approximately equals the achievable rate of the optimal unconstrained precoding with 64 user antennas. From Fig. 8, the achievable rate of the optimal unconstrained precoding with 64 user antennas is similar to that of the proposed LPC-SIC precoding with 128 user antennas. Furthermore, the required number of PSs for the optimal unconstrained precoding is $N^2M = 4096$ in this environment. While the required number of PSs of the proposed PC-SIC and LPC-SIC precoding algorithms is $2NM = 512$. Therefore, the proposed architecture can effectively reduce the number of PSs and improve the energy efficiency.

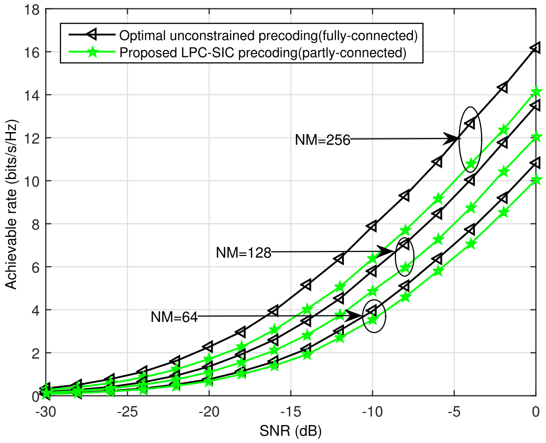

FIGURE 9. Achievable rate comparison under different transmit antennas for the mmWave massive MIMO system, where $K = 16$.

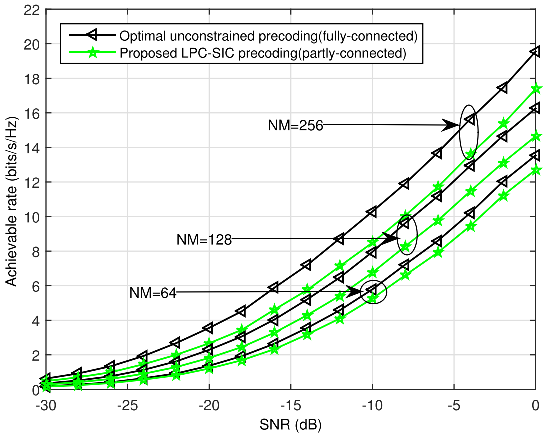

FIGURE 10. Achievable rate comparison under different transmit antennas for the mmWave massive MIMO system, where $K = 32$.

As mentioned above, the proposed LPC-SIC hybrid precoding can obtain a much better trade-off between the performance and computational complexity than the proposed PC-SIC hybrid precoding. Therefore, we mainly analyze the performance of the proposed LPC-SIC hybrid precoding. First, we analyze the relationship between the number of transmit antennas and the achievable rate. In this simulation, the SNR varies from -30 dB to 0 dB. The number of transmit antennas is 64, 128, and 256. The number of RF chains is 16. The impact of the number of transmit antennas on the achievable rate is shown in Fig. 9 and Fig. 10. It can be seen clearly from Fig. 9 that as the number of transmit antennas decreases, the achievable rate of the proposed LPC-SIC hybrid precoding gradually becomes close to that of the optimal unconstrained precoding. For example, when the number of transmit antennas is 256 and SNR $=$ 0 dB, the achievable rates of the optimal unconstrained hybrid precoding based on fully-connected architecture and the proposed LPC-SIC hybrid precoding are 16.1914 bits/s/Hz and 14.1543 bits/s/Hz, respectively. The proposed LPC-SIC hybrid precoding can obtain 87.42% of the achievable rate of the optimal unconstrained precoding. In contrast, the proposed LPC-SIC hybrid precoding can obtain 89.11% and 92.87% of the achievable rate of the optimal unconstrained precoding, when the numbers of transmit antennas are 128 and 64, respectively. In addition, the effect of the SNR is no longer obvious as the number of transmit antennas decreases. When increasing the SNR, the gap in the achievable rate between the optimal unconstrained and proposed LPC-SIC precoding will decrease as the number of transmit antennas NM decreases. Moreover, we can observe that the achievable rate can improve significantly as the number of user antennas increases. It can reduce the requirement for the SNR by increasing the number of user antennas. For example, when the number of transmit antennas is 256, the achievable rate of the proposed LPC-SIC precoding with $K = 16$ and SNR $= 0$ dB is approximately equal to that with $K = 32$ and SNR $= -4$ dB.

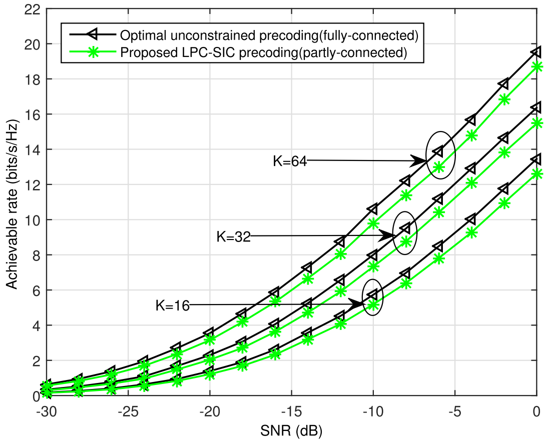

FIGURE 11. Achievable rate comparison under different user antennas for mmWave massive MIMO system, where $NM = 128$.

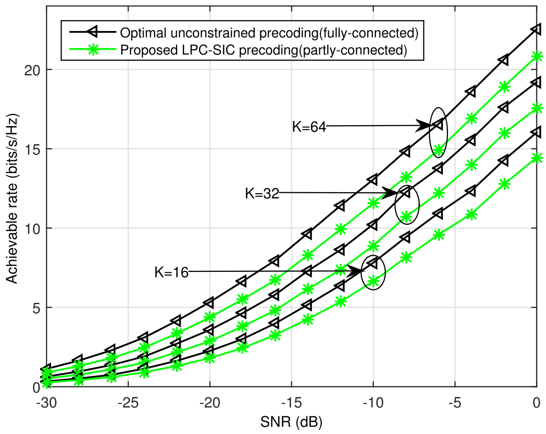

FIGURE 12. Achievable rate comparison under different user antennas for mmWave massive MIMO system, where $NM = 256$.

Furthermore, we analyze the relationship between the number of user antennas and the achievable rate for the proposed LPC-SIC precoding. In this simulation, the SNR varies from -30 dB to 0 dB. The number of user antennas is $K = 16, 32, 64$. Fig. 11 and Fig. 12 depict the achievable rate of the optimal unconstrained precoding based on fully-connected architecture and the proposed LPC-SIC precoding against the SNR under different numbers of user antennas. It can be seen that the achievable rate will increase as the number of user antennas increases when employing the same number of transmit antennas. In addition, the system can reduce the requirement for the SNR by employing more antennas at the BS. For example, when the number of user antennas is $K = 64$, the achievable rate of the proposed LPC-SIC precoding with $NM = 128$ and SNR $= -10$ dB is approximately equal to that with $NM = 256$ and SNR $= -12$ dB.

V. CONCLUSION

To improve the energy efficiency of mmWave massive MIMO systems, we propose a new partly-connected architecture and two hybrid precoding algorithms based on partly-connected architecture and SIC in this paper. The partly-connected architecture selects two RF chains to control one sub-antenna array, which can effectively reduce the required PSs and the energy consumption. Then, we propose the PC-SIC and LPC-SIC hybrid precoding algorithms based on the partly-connected architecture to avoid high energy consumption for mmWave massive MIMO systems. The PC-SIC hybrid precoding decomposes the total achievable rate into a series of sub-rate optimization problems based on sub-antenna arrays of partly-connected architecture. Furthermore, to reduce the complexity of the PC-SIC precoding, we further propose a LPC-SIC hybrid precoding that uses the partial SVD algorithm to replace the SVD in the sub-rate optimization process. The complexity analysis shows that the proposed LPC-SIC hybrid precoding can significantly reduce the computational complexity. The simulation results demonstrate that the proposed algorithms can provide better achievable rate and energy efficiency.

ACKNOWLEDGMENT

The authors would like to thank the support of National Science Foundation of China.

REFERENCES

[1] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband system," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[2] S. Sun, T. S. Rappaport, R. W. Heath, Jr., A. Nix, and S. Rangan, "MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both?" *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 110–121, Dec. 2014.

[3] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[4] B. Biglarbegian, M. Fakharzadeh, D. Busuioc, M.-R. Nezhad-Ahmadi, and S. Safavi-Naeini, "Optimized microstrip antenna arrays for emerging millimeter-wave wireless applications," *IEEE Trans. Antennas Propag.*, vol. 59, no. 5, pp. 1742–1747, May 2011.

[5] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[6] M. J. Azizipour, K. Mohamed-Pour, and A. Lee Swindlehurst, "A burst-form CSI estimation approach for FDD massive MIMO systems," *Signal Process.*, vol. 162, pp. 106–114, Sep. 2019.

[7] V. Venkateswaran and A. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4131–4143, Aug. 2010.

[8] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 949–973, 2nd Quart., 2016.

[9] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[10] J. Joung and A. H. Sayed, "Multiuser two-way amplify-and-forward relay processing and power control methods for beamforming systems," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1833–1846, Mar. 2010.

[11] A. Azizzadeh, R. Mohammadkhani, S. V. A.-D. Makki, and E. Björnson, "BER performance analysis of coarsely quantized uplink massive MIMO," *Signal Process.*, vol. 161, pp. 259–267, Aug. 2019.

[12] J. Zhang, Y. Huang, T. Yu, J. Wang, and M. Xiao, "Hybrid precoding for multi-subarray millimeter-wave communication systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 440–443, Jun. 2018.

[13] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Select. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[14] K. Song, B. Ji, Y. Huang, M. Xiao, and L. Yang, "Performance analysis of heterogeneous networks with interference cancellation," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 6969–6981, Aug. 2017.

[15] C. Zhang, Y. Huang, Y. Jing, S. Jin, and L. Yang, "Sum-rate analysis for massive MIMO downlink with joint statistical beamforming and user scheduling," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2181–2194, Apr. 2017.

[16] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[17] J. Tan, L. Dai, J. Li, and S. Jin, "Angle-based codebook for low-resolution hybrid precoding in millimeter-wave massive MIMO systems," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Oct. 2017, pp. 1–5.

[18] T. He and Z. Xiao, "Suboptimal beam search algorithm and codebook design for millimeter-wave communications," *Mobile Netw. Appl.*, vol. 20, no. 1, pp. 86–97, Feb. 2015.

[19] A. Alkhateeb, G. Leus, and R. W. Heath, Jr., "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.

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

[21] Y.-Y. Lee, C.-H. Wang, and Y.-H. Huang, "A hybrid RF/baseband precoding processor based on parallel-index-selection matrix-inversion-bypass simultaneous orthogonal matching pursuit for millimeter wave MIMO systems," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 305–317, Jan. 2015.

[22] D. H. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, Jr., "Hybrid MMSE precoding and combining designs for mmWave multiuser systems," *IEEE Access*, vol. 5, pp. 19167–19181, 2017.

[23] C. Xu, R. Ye, Y. Huang, S. He, and C. Zhang, "Hybrid precoding for broadband millimeter-wave communication systems with partial CSI," *IEEE Access*, vol. 6, pp. 50891–50900, 2018.

[24] M. Su, Y. Huang, C. Zhang, J. Zhang, and Y. Li, "Hybrid precoder design for millimeter wave systems based on geometric construction," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[25] S. He, C. Qi, Y. Wu, and Y. Huang, "Energy-efficient transceiver design for hybrid sub-array architecture MIMO systems," *IEEE Access*, vol. 4, pp. 9895–9905, 2017.

[26] X. Gao, L. Dai, S. Han, I. Chih-Lin, and R. W. Heath, Jr., "Energy-efficient hybrid analog and digital precoding for MmWave MIMO systems with large antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 998–1009, Apr. 2016.

[27] Z. Gao, L. Dai, and Z. Wang, "Structured compressive sensing based superimposed pilot design in downlink large-scale MIMO systems," *Electron. Lett.*, vol. 50, no. 12, pp. 896–898, Jun. 2014.

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

[29] A. Hu, "User scheduling for capacity-Jain's fairness tradeoff in millimeter-wave MIMO systems," *Signal Process.*, vol. 158, pp. 141–149, May 2019.

[30] L. Dai, X. Gao, J. Quan, S. Han, and C.-L. I, "Near-optimal hybrid analog and digital precoding for downlink mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 1334–1339.

[31] T. Xie, L. Dai, X. Gao, M. Z. Shakir, and J. Li, "Geometric mean decomposition based hybrid precoding for millimeter-wave massive MIMO," *China Commun.*, vol. 15, no. 5, pp. 229–238, 2018.

[32] C. A. Balanis, *Antenna Theory and Design*. Hoboken, NJ, USA: Wiley, 2012.

YANG LIU received the B.Eng. degree in electronic engineering from Inner Mongolia University and the Ph.D. degree in electronic engineering from the Dalian University of Technology, in 2003 and 2012, respectively. Since 2017, he has been a Senior Research Scholar with the Department of Electronic Engineering, Tsinghua University. His research interests include array signal processing, non-Gaussian signal processing, and wireless communications with a focus on millimeter-wave multi-antenna techniques.

QINGXIA FENG received the B.S. degree in electronic engineering from Anhui University, Anhui, China, in 2018. She is currently pursuing the M.S. degree in electronic engineering with Inner Mongolia University, Hohhot, China. Her research interests include massive MIMO, millimeter-wave systems, channel estimation, and hybrid precoding.

QIONG WU received the B.Eng. degree from Tianjin Polytechnic University, in 2017. She is currently pursuing the Ph.D. degree with Beihang University, Beijing, China. Her main research interests include massive MIMO, millimeter-wave systems, and hybrid precoding.

YINGHUI ZHANG received the B.Eng. and M.S. degrees from Xidian University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015. She is currently an Associate Professor with the College of Electronic Information Engineering, Inner Mongolia University. Her research interests include 5G Technologies, millimeter-wave communications, cooperative communications, and relay networks. She serves as a member of the Inner Mongolia Communications Association.

MINGLU JIN received the B.S. degree from the University of Science and Technology, in 1982, and the M.S. and Ph.D. degrees from Beihang University, in 1984 and 1995, respectively. He was a Visiting Scholar with the Arimoto Laboratory, Osaka University, Osaka, Japan, from 1987 to 1988. He was a Research Fellow with the Radio and Broadcasting Research Laboratory, Electronics Telecommunications Research Institute, South Korea, from 2001 to 2004. He is currently a Professor with the Dalian University of Technology. His research interests include wireless communications, wireless sensor networks, and signal processing for wireless communication systems.

TIANSHUANG QIU received the B.S. degree from Tianjin University, Tianjin, China, in 1983, the M.S. degree from the Dalian University of Technology, Dalian, Liaoning, China, in 1993, and the Ph.D. degree from Southeastern University, Nanjing, China, in 1996, all in electrical engineering. He was a Postdoctoral Fellow with the Department of Electrical Engineering, Northern Illinois University, USA, from 1996 to 2000. He is currently a Professor with the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interests include wireless signal processing, biomedical signal processing, and non-Gaussian signal processing.