IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 12, 2019, accepted August 1, 2019, date of publication August 5, 2019, date of current version August 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933328

Power Allocation for Energy Efficiency Optimization in Multi-User mmWave-NOMA System With Hybrid Precoding

XIANGBIN YU[1,2,3], (Member, IEEE), FANGCHENG XU[1], KAI YU[1], AND XIAOYU DANG[1], (Member, IEEE)

[1]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2]Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[3]National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China

Corresponding author: Xiangbin Yu (yxbxwy@gmail.com)

This work was supported in part by the Natural Science Foundation of China under Grant 61571225 and Grant 61571224, in part by the Natural Science Foundation of Jiangsu Province in China under Grant BK20181289, in part by the Open Research Fund Key Laboratory of Wireless Sensor Network and Communication of Chinese Academy of Sciences under Grant 2017006, and in part by the Open Research Fund of National Mobile Communications Research Laboratory of Southeast University under Grant 2017D03.

ABSTRACT Recently, the non-orthogonal multiple access (NOMA) scheme has been applied in millimeter-wave (mmWave) communication system to support more users and further improve the performance. In this paper, an energy-efficient power allocation (PA) scheme is designed for a downlink multi-user mmWave-NOMA system with hybrid precoding (HP), where two typical HP architectures, namely the fully-connected HP architecture (FHPA) and the sub-connected HP architecture (SHPA), are both considered. Firstly, we pair users in term of their channel difference and correlation. Then, analog beamforming schemes are proposed for the system with fully-connected and sub-connected HP architectures, respectively. Based on this, a two-step HP design with the proposed analog beamforming and zero-forcing precoding is presented. With these results, the optimization problem is formulated to maximize the energy efficiency (EE) of the system. This is a non-convex optimization problem, and can be approximately decomposed into independent convex sub-problems by applying the fractional programming theory. Using the coordinate descent method, the closed-form solutions of each sub-problem are derived. On these basis, an effective iterative algorithm is proposed to obtain the suboptimal power allocation. Simulation results verify the effectiveness of the proposed PA scheme, it has the same EE performance as the existing scheme with relatively low complexity, and can obtain the EE close to the exhaustive search scheme as well as the chaotic accelerated particle swarm optimization scheme. Moreover, the proposed analog beamforming obviously outperforms the conventional finite resolution analog beamforming, and the system with SHPA has higher EE than that with FHPA.

INDEX TERMS Analog beamforming, energy efficiency, hybrid precoding, millimeter-wave communications, multi-user pairing, non-orthogonal multiple access, power allocation.

I. INTRODUCTION

the perspective of green communication, the fifth generation (5G) mobile communication will not only put forward higher requirements on the conventional performance indicators such as spectrum efficiency (SE), transmission rate and communication delay, but also require high energy efficiency (EE).

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

On the one hand, the existing inefficient orthogonal multiple access (OMA) techniques, e.g., time-division multiple access (TDMA), may not be able to meet the needs of a large number of users in the future communication scenario. Different from OMA, non-orthogonal multiple access (NOMA) allows multiple users to share the same resource block with different power levels via the successive interference cancellation (SIC) technology [1]. In recent years, NOMA has been applied to several fields. Particularly, two schemes are proposed in [2] for generating artificial jamming at the base station (BS) in order to guarantee the security of NOMA networks, while in [3], joint trajectory and precoding optimization problem is studied for NOMA networks with unmanned aerial vehicle (UAV). On the other hand, millimeter-wave (mmWave) communication has been considered as one of key technologies in 5G, because it can utilize a great deal of spare spectrum in the high frequency band ranging from 30 GHz to 300 GHz [4]. In general, the number of radio frequency (RF) chains in a mmWave device is less than that of antennas due to high energy consumption and hardware cost, which limits the number of users. Therefore, it is necessary to combine NOMA and mmWave communication, i.e., mmWave-NOMA, to support more users and further improves the system performance [5]–[13].

Up to present, the power allocation schemes for SE maximization in mmWave-NOMA systems have been studied in some literatures [6]–[11]. In [6], a user scheduling and power allocation algorithm for a mmWave-NOMA system based on random beamforming was proposed. For a downlink mmWave-NOMA system, a sub-optimal scheme is presented in [7] to solve the joint power allocation and beamforming optimization problem. Similar to [7], [8] presented the joint power allocation and beamforming scheme for the uplink case. However, the SE maximization schemes proposed in [7] and [8] are only applicable to two-user case with the ideal beamforming assumption and may be difficult to be extended multi-user cases. In [9], NOMA was firstly combined with beamspace multiple input multiple output (MIMO) in mmWave communications. The authors proposed a dynamic power allocation to solve the SE optimization problem. In [10], a multi-beam NOMA framework for a multiple RF chain mmWave system with hybrid precoding (HP) was firstly realized, which multiple analog beams can be formed for each NOMA group. The authors studied the resource allocation maximizing the system sum-rate and obtained a suboptimal two-stage resource allocation scheme. In [11], the integration of simultaneous wireless information and power transfer (SWIPT) in mmWave massive MIMO-NOMA systems was firstly investigated. An iterative optimization algorithm was developed to solve the SE optimiaztion problem by jointly optimizing power allocation for mmWave massive MIMO-NOMA and power splitting factors for SWIPT. In contrast, only a few works have addressed the design of the energy-efficient power allocation schemes for mmWave-NOMA systems [12]–[14]. In [12], a near-optimal SIC-based power allocation scheme was presented for a HP mmWave-NOMA system. The EE maximization problem in a HP mmWave-NOMA system was investigated in [13], but the fully-connected HP architecture, which may against the improvement of EE, was adopted in the paper. Besides, the proposed algorithm in [13] is based on the Lagrange dual method, and the resulting complexity is relatively higher. An optimal power allocation algorithm was proposed in [14] for maximizing the EE of NOMA systems, but it does not consider the superiority of mmWave-NOMA systems.

Motivated by the analysis above, we will study the power allocation for EE maximization in downlink multi-user mmWave-NOMA system with HP by considering two typical HP architectures. A low-complexity energy-efficient power allocation scheme is proposed for the system, and superior EE performance is achieved. The major contributions of this paper are summarized as follows.

1) The downlink multi-user mmWave-NOMA system with the fully-connected and sub-connected HP architectures are respectively presented. For this system model, we pair every two users to form a cluster according to their channel difference and correlation, and then a two-step hybrid precoding is designed, i.e., the analog precoding and the digital precoding. For the analog precoding, we solve the optimization problem maximizing the sum of array gains for each cluster, and obtain the closed-form optimal solutions for fully-connected and sub-connected HP architectures, respectively. For the digital precoding, the zero-forcing (ZF) precoding is performed to eliminate the inter-cluster interference for the strong users in all clusters. Since the two users in each cluster have high channel correlation, the inter-cluster interference for the weak users in all clusters can be minimized. Meanwhile, the intra-cluster interference can be canceled by performing the SIC.

2) According to the analysis of EE, we formulate the EE maximization problem for multi-user mmWave-NOMA system. With the proposed user pairing and HP design scheme, the non-convex EE optimization problem can be approximately decomposed into independent convex sub-problems by applying the fractional programming theory. Furthermore, using the coordinate descent (CD) method, we derive the closed-form solution of each sub-problem. Based on this, an effective iterative algorithm is proposed to obtain a sub-optimal solution of the original non-convex problem.

3) Simulation results show that the proposed power allocation scheme can achieve near-optimal EE performance with relatively low complexity by comparing with other benchmark schemes, namely the existing scheme in [13], the chaotic accelerated particle swarm optimization (CAPSO) scheme, and the exhaustive search scheme. Moreover, the proposed analog beamforming has a great improvement on the system performance compared with the conventional finite resolution analog beamforming (FRAB).

The rest of this paper is organized as follows. Section II introduces the system model. Sections III presents the user pairing and HP design scheme. In Section IV, the EE maximization problem is formulated and an iterative algorithm is proposed to solve the non-convex problem. Section V provides the simulation results. Finally, conclusions are drawn in Section VI.

Notations: Matrices and vectors are denoted by the upper-case and lower-case boldface letters, respectively. $(\cdot)^T$, $(\cdot)^H$ and $(\cdot)^{-1}$ denote the transpose, Hermitian transpose and matrix inversion, respectively. $\mathbb{C}^{m\times n}$ denotes the space of $m \times n$ complex matrices. $\|\cdot\|_2$ indicates the Euclidean norm of a vector, and $|\cdot|$ denotes the absolute value of a complex scalar. $[\mathbf{a}]_i$ denotes the i-th entry of $\mathbf{a}$. $\vec{\lambda}_{\max}(\mathbf{A})$ denotes the eigenvector of the maximum eigenvalue of $\mathbf{A} \in \mathbb{C}^{n\times n}$. $\min\{\cdot\}$ and $\max\{\cdot\}$ represent the minimum and the maximum of two real scalars, respectively. $\angle\{\cdot\}$ stands for the angle of a complex scalar or the angles of a vector. The complex Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathcal{CN}\left(\mu,\sigma^2\right)$. $\exp(\cdot)$ denotes the exponential function.

II. SYSTEM MODEL

Consider a downlink multi-user mmWave-NOMA system with hybrid precoding, where one BS equipped with N_{RF} RF chains and N antennas serves K single-antenna users simultaneously. To effectively reduce the inter-interference among different clusters, the number of clusters G can not exceed the number of RF chains, i.e., $G \le N_{\mathrm{RF}}$. For simplicity, we assume that $G = N_{\mathrm{RF}}$, and the same assumption can also be found in [9]–[13]. Besides, each cluster consists of two users, i.e., $K = 2G$.

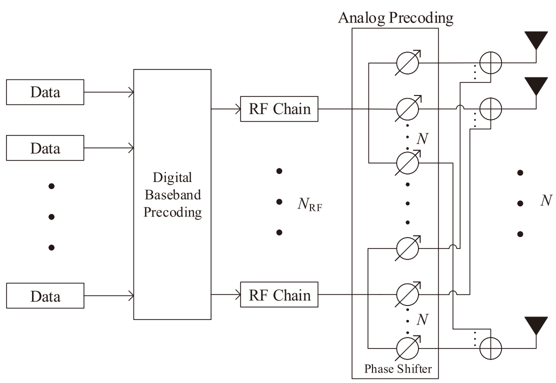


FIGURE 1. Fully-connected HP architecture.

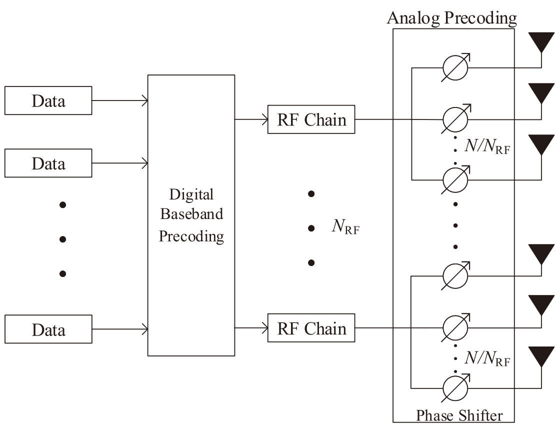


FIGURE 2. Sub-connected HP architecture.

In the conventional fully-connected HP architecture [15], as shown in Fig. 1, each RF chain is connected to all N antennas by finite resolution phase shifters with $N_{\mathrm{PS}} = NN_{\mathrm{RF}}$, where N_{PS} is the number of phase shifters. This architecture can bring not only full array gain but also high energy consumption. By contrast, in the sub-connected HP architecture [16], as shown in Fig. 2, each RF chain is connected to only a subset of N antennas, which means that only $N_{\mathrm{PS}} = N$ phase shifters are required.

Let $\mathbf{A} \in \mathbb{C}^{N\times N_{\mathrm{RF}}}$ be the analog precoding matrix. For the fully-connected HP architecture, the analog precoding matrix $\mathbf{A}^{(\mathrm{full})}$ can be expressed as

$$\mathbf{A}^{(\mathrm{full})} = \left[\mathbf{a}_1^{(\mathrm{full})}, \mathbf{a}_2^{(\mathrm{full})}, \cdots, \mathbf{a}_{N_{\mathrm{RF}}}^{(\mathrm{full})}\right], \tag{1}$$

where the elements of $\mathbf{a}_n^{(\mathrm{full})} \in \mathbb{C}^{N\times 1}$ for $n = 1, 2, \cdots, N_{\mathrm{RF}}$ have the same amplitude $1/\sqrt{N}$ but different phases [15].

For the sub-connected HP architecture, the analog precoding matrix $\mathbf{A}^{(\mathrm{sub})}$ can be expressed as

$$\mathbf{A}^{(\mathrm{sub})} = \begin{bmatrix} \mathbf{a}_1^{(\mathrm{sub})} & \mathbf{0}_{M\times 1} & \cdots & \mathbf{0}_{M\times 1} \\ \mathbf{0}_{M\times 1} & \mathbf{a}_2^{(\mathrm{sub})} & \cdots & \mathbf{0}_{M\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{M\times 1} & \mathbf{0}_{M\times 1} & \cdots & \mathbf{a}_{N_{\mathrm{RF}}}^{(\mathrm{sub})} \end{bmatrix}, \tag{2}$$

where $M = N/N_{\mathrm{RF}}$ is a positive integer, and the elements of $\mathbf{a}_n^{(\mathrm{sub})} \in \mathbb{C}^{M\times 1}$ for $n = 1, 2, \cdots, N_{\mathrm{RF}}$ have the same amplitude $1/\sqrt{M}$ but different phases [16].

Let $\mathbf{d}_g \in \mathbb{C}^{N_{RF}\times 1}$ $(g = 1, \ldots, G)$ be the digital precoding vector for the g-th cluster, then the signal transmitted by BS can expressed as

$$s = \mathbf{A}\sum_{g=1}^{G}\sum_{k=1}^{2}\mathbf{d}_g\sqrt{p_{g,k}}x_{g,k}, \tag{3}$$

where $x_{g,k}$ $(g = 1, \ldots, G, k = 1, 2)$ is the transmitted signal for the k-th user in the g-th cluster with $\mathbb{E}\{|x_{g,k}|^2\} = 1$, $p_{g,k}$ is the corresponding transmitted power, which is limited to the maximum transmission power $P_{\max}$, i.e., $p_{g,1} + p_{g,2} \le P_{\max}$ $(g = 1, \ldots, G)$.

Thus, the received signal of the k-th user in the g-th cluster can be expressed as

$$\begin{aligned} y_{g,k} &= \mathbf{h}_{g,k}^{\mathrm{H}}\mathbf{A}\sum_{i=1}^{G}\sum_{j=1}^{2}\mathbf{d}_i\sqrt{p_{i,j}}x_{i,j} + n_{g,k} \\ &= \underbrace{\mathbf{h}_{g,k}^{\mathrm{H}}\mathbf{A}\mathbf{d}_g\sqrt{p_{g,k}}x_{g,k}}_{\text{desired signal}} + \underbrace{\mathbf{h}_{g,k}^{\mathrm{H}}\mathbf{A}\mathbf{d}_g\sqrt{p_{g,m}}x_{g,m}}_{\text{intra-cluster interference}} \\ &\quad \underbrace{+\,\mathbf{h}_{g,k}^{\mathrm{H}}\mathbf{A}\sum_{i=1,i\neq g}^{G}\mathbf{d}_i\left(\sqrt{p_{i,1}}x_{i,1} + \sqrt{p_{i,2}}x_{i,2}\right)}_{\text{inter-cluster interference}} + n_{g,k}, \end{aligned} \tag{4}$$

where $g = 1, 2, \ldots, G$, $k = 1, 2$, $m \in \{1, 2\}$, $m \neq k$, $n_{g,k}$ denotes the noise following the distribution $\mathcal{CN}\left(0,\sigma^2\right)$, $\mathbf{h}_{g,k} \in \mathbb{C}^{N\times 1}$ is the mmWave channel vector of the k-th user in g-th cluster. Assuming that a uniform linear array (ULA) structure with a half-wavelength antenna space is adopted at the BS, the mmWave channel can be modeled as [5]–[18]

$$\tilde{\mathbf{h}}_{g,k} = \sqrt{\frac{N}{L_{g,k}}}\sum_{l=1}^{L_{g,k}}\lambda_{g,k}^{(l)}\mathbf{a}\left(N, \theta_{g,k}^{(l)}\right), \tag{5}$$

where $L_{g,k}$ is the number of the multi-path components (MPCs) for the k-th user in the g-th cluster, $\lambda_{g,k}^{(l)}$ and $\theta_{g,k}^{(l)}$ are the complex gain and angle of departure (AoD) of the l-th MPC, respectively. $\mathbf{a}(\cdot)$ represents the $N \times 1$ array steering vector defined by

$$\mathbf{a}(N,\theta) = \frac{1}{\sqrt{N}}\left[e^{j\pi 0\cos\theta}, e^{j\pi 1\cos\theta}, \ldots, e^{j\pi(N-1)\cos\theta}\right]^T. \quad (6)$$

In general, due to the sparse characteristic of the mmWave channel, the original channel model in (5) can be simplified as [7], [8], [13], [17]

$$\mathbf{h}_{g,k} = \sqrt{\frac{N}{L_{g,k}}}\lambda_{g,k}\mathbf{a}\left(N,\theta_{g,k}\right), \quad (7)$$

where $\lambda_{g,k} = \lambda_{g,k}^{(m_{g,k})}$, $\theta_{g,k} = \theta_{g,k}^{(m_{g,k})}$, and $m_{g,k}$ is the index of the strongest MPC for the k-th user in the g-th cluster.

III. USER PAIRING AND HYBRID PRECODING DESIGN SCHEME

In this section, we will address the user pairing and HP scheme design for the multi-user mmWave-NOMA system. As for the user pairing, every two users are paired and form a cluster according to their channel difference and correlation. Afterwards, based on the two-step HP scheme in [19], we will give the analog precoding design firstly, and then low-complexity digital precoding is presented.

A. USER PAIRING

The channel difference and correlation between arbitrary two users can be defined as [13], [20], [21]

$$\begin{cases} \mathrm{Diff}_{(i,j)} = \left|10\lg\|\mathbf{h}_i\|_2^2 - 10\lg\|\mathbf{h}_j\|_2^2\right|, \\ \mathrm{Corr}_{(i,j)} = \dfrac{\|\mathbf{h}_i^{\mathrm{H}}\mathbf{h}_j\|}{\|\mathbf{h}_i\|_2\|\mathbf{h}_j\|_2}, \end{cases} \quad (8)$$

respectively, where $\mathbf{h}_i$ and $\mathbf{h}_j$ denote the $N \times 1$ channel vector of unpaired user-i and user-j, respectively.

For NOMA systems, one of the basic user pairing criteria is to select users with obvious channel difference to pair, which can help improve the successful rate of SIC [22]. In addition, it can be seen from (4) that two users in the same cluster are affected by the same precoding vectors for other clusters, so the higher is the channel correlation between two users in the same cluster, the more similar is the channel, and the better is the effect of inter-cluster interference cancellation. Thus, another user pairing criterion is to pair users with high channel correlation.

Specifically, we pair users by the following steps. Firstly, we find out all candidate user pairs, where the channel correlation of each user pair is greater than the predetermined threshold value ρ. This threshhold is a real constant with $\rho \in [0,1]$, and its value can be attained in terms of [13], [20], [21]. Then, the second step is to select the user pair with maximum channel difference to form a cluster and remove it from the candidate user pairs. Lastly, repeat the above-mentioned two steps until all users are paired. After the user pairing, the users in each cluster will be resorted such that $\|\mathbf{h}_{g,1}\|_2 \geq \|\mathbf{h}_{g,2}\|_2$ $(g = 1, 2, \ldots, G)$.

B. ANALOG PRECODING

One principle of the analog precoding design is to maximize the array gain for the strong user in the g-th cluster $(g = 1, 2, \ldots, G)$, i.e., $|\mathbf{h}_{g,1}^{\mathrm{H}}\mathbf{a}_g^{(\mathrm{full})}|^2$ for the fully-connected HP architecture and $|\mathbf{h}_{g,1}^{\mathrm{H}}\mathbf{a}_g^{(\mathrm{sub})}|^2$ for the sub-connected HP architecture [11].

With the help of the low-complexity FRAB [11], [18], let N_{C} be the number of candidate phases, then the i-th element of $\mathbf{a}_g^{(\mathrm{full})}$ is given by

$$\left[\mathbf{a}_g^{(\mathrm{full})}\right]_i = \frac{1}{\sqrt{N}}\exp\left(j2\pi\hat{n}_1/N_{\mathrm{C}}\right), \quad (9)$$

where $i = 1, \cdots, N$, and

$$\hat{n}_1 = \underset{n_1\in\{1,2,\ldots,N_{\mathrm{C}}\}}{\arg\min}\left|\angle\left(\left[\mathbf{h}_{g,1}\right]_i\right) - 2\pi n_1/N_{\mathrm{C}}\right|. \quad (10)$$

Similarly, the i-th element of $\mathbf{a}_g^{(\mathrm{sub})}$ is given by

$$\left[\mathbf{a}_g^{(\mathrm{sub})}\right]_i = \frac{1}{\sqrt{M}}\exp\left(j2\pi\hat{n}_2/N_{\mathrm{C}}\right), \quad (11)$$

where $i = (g-1)M+1, (g-1)M+2, \ldots, gM$, and

$$\hat{n}_2 = \underset{n_2\in\{1,2,\ldots,N_{\mathrm{C}}\}}{\arg\min}\left|\angle([\hat{\mathbf{h}}_{g,1}]_i) - 2\pi n_2/N_{\mathrm{C}}\right|, \quad (12)$$

in which $\hat{\mathbf{h}}_{g,k}$ corresponds to the $((g-1)M+1)$ th row to the gM th row of $\mathbf{h}_{g,k}$ $(g = 1, 2, \ldots, G, k = 1, 2)$.

Since one analog beamforming generated by one RF chain should support the two NOMA users in a cluster, another principle of the analog precoding is to maximize the sum of array gains for the two users in the g-th cluster $(g = 1, 2, \ldots, G)$. Based on this, effective analog precodings are presented for the system.

For the fully-connected HP architecture, $\mathbf{a}_g^{(\mathrm{full})}$ is obtained by solving the following problem

$$\begin{aligned} &\max_{\mathbf{a}_g^{(\mathrm{full})}} \left|\mathbf{h}_{g,1}^{\mathrm{H}}\mathbf{a}_g^{(\mathrm{full})}\right|^2 + \left|\mathbf{h}_{g,2}^{\mathrm{H}}\mathbf{a}_g^{(\mathrm{full})}\right|^2 \\ &\text{s.t.} \left|\left[\mathbf{a}_g^{(\mathrm{full})}\right]_i\right| = \frac{1}{\sqrt{N}}, \quad i \in \{1, 2, \ldots, N\}. \end{aligned} \quad (13)$$

From (13), it can be derived that the optimal solution is

$$\mathbf{a}_g^{(\mathrm{full})} = \frac{1}{\sqrt{N}}\exp\left(j\angle\left(\vec{\lambda}_{\max}\left(\mathbf{h}_{g,1}\mathbf{h}_{g,1}^{\mathrm{H}} + \mathbf{h}_{g,2}\mathbf{h}_{g,2}^{\mathrm{H}}\right)\right)\right). \quad (14)$$

Similarly, the optimal solution for the sub-connected HP architecture is

$$\mathbf{a}_g^{(\mathrm{sub})} = \frac{1}{\sqrt{M}}\exp\left(j\angle\left(\vec{\lambda}_{\max}\left(\hat{\mathbf{h}}_{g,1}\hat{\mathbf{h}}_{g,1}^{\mathrm{H}} + \hat{\mathbf{h}}_{g,2}\hat{\mathbf{h}}_{g,2}^{\mathrm{H}}\right)\right)\right). \quad (15)$$

After the analog precoding, the users in each cluster will be resorted such that $\|\mathbf{h}_{g,1}^{\mathrm{H}}\mathbf{A}\|_2 \geq \|\mathbf{h}_{g,2}^{\mathrm{H}}\mathbf{A}\|_2$ $(g = 1, 2, \ldots, G)$.

C. DIGITAL PRECODING

After the user pairing and analog precoding, the equivalent channel vector for the k-th user in the g-th cluster can be expressed as

$$\bar{\mathbf{h}}_{g,k}^{\mathrm{H}} = \mathbf{h}_{g,k}^{\mathrm{H}}\mathbf{A}, \tag{16}$$

where $g = 1, \ldots, G, k = 1, 2$. Then, the digital precoding is designed to eliminate the inter-cluster interference by the low-complexity ZF precoding [23]. Without loss of generality, we define the equivalent channel matrix as

$$\bar{\mathbf{H}} = \left[\bar{\mathbf{h}}_{1,1}, \bar{\mathbf{h}}_{2,1}, \ldots, \bar{\mathbf{h}}_{G,1}\right]. \tag{17}$$

According to the principle of ZF precoding, the digital precoding matrix can be formulated as

$$\bar{\mathbf{D}} = \left[\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_2, \ldots, \bar{\mathbf{d}}_G\right] = \bar{\mathbf{H}}\left(\bar{\mathbf{H}}^{\mathrm{H}}\bar{\mathbf{H}}\right)^{-1}. \tag{18}$$

Thus, the digital precoding vector for the g-th cluster can be obtained by normalizing $\bar{\mathbf{d}}_g$, i.e.,

$$\mathbf{d}_g = \frac{\bar{\mathbf{d}}_g}{\|\mathbf{A}\bar{\mathbf{d}}_g\|_2}. \tag{19}$$

After the digital precoding, we have:

$$\bar{\mathbf{h}}_{i,1}^{\mathrm{H}}\mathbf{d}_j = 0, \tag{20}$$

where $i, j = 1, \ldots, G$, and $i \neq j$. Therefore, the inter-cluster interference for the strong user in each cluster can be cancelled. Since the users in each cluster have high channel correlation after the user pairing, the inter-cluster interference for the weak user in the same cluster can be minimized.

Finally, the users in each cluster will be resorted such that $\|\bar{\mathbf{h}}_{g,1}^{\mathrm{H}}\mathbf{d}_g\|_2 \geq \|\bar{\mathbf{h}}_{g,2}^{\mathrm{H}}\mathbf{d}_g\|_2$ $(g = 1, 2, \ldots, G)$.

IV. ENERGY EFFICIENCY OPTIMIZATION

In this section, we will give the energy efficiency optimization scheme. By the user pairing and hybrid precoding, the remaining received signals of two users in the g-th cluster can be expressed as

$$\hat{y}_{g,1} = \bar{\mathbf{h}}_{g,1}^{\mathrm{H}}\mathbf{d}_g\sqrt{p_{g,1}}x_{g,1} + \bar{\mathbf{h}}_{g,1}^{\mathrm{H}}\mathbf{d}_g\sqrt{p_{g,2}}x_{g,2} + n_{g,1}, \tag{21}$$

and

$$\begin{aligned}\hat{y}_{g,2} = {} & \bar{\mathbf{h}}_{g,2}^{\mathrm{H}}\mathbf{d}_g\sqrt{p_{g,2}}x_{g,2} + \bar{\mathbf{h}}_{g,2}^{\mathrm{H}}\mathbf{d}_g\sqrt{p_{g,1}}x_{g,1} \\ & + \bar{\mathbf{h}}_{g,2}^{\mathrm{H}}\sum_{i=1,i\neq g}^{G}\mathbf{d}_i\left(\sqrt{p_{i,1}}x_{i,1} + \sqrt{p_{i,2}}x_{i,2}\right) + n_{g,2},\end{aligned} \tag{22}$$

respectively.

Accordingly, the signal-interference-noise-ratio (SINR) for the strong user to detect the weak user's signal can be expressed as

$$\mathrm{SINR}_{2\to1}^{g} = \frac{p_{g,2}\gamma_{g,1}^{g}}{p_{g,1}\gamma_{g,1}^{g} + 1}, \tag{23}$$

and the SINR for the weak user to detect its own signal can be expressed as

$$\begin{aligned}\mathrm{SINR}_{2\to2}^{g} &= \frac{p_{g,2}\gamma_{g,2}^{g}}{p_{g,1}\gamma_{g,2}^{g} + \sum_{i=1,i\neq g}^{G}\left(p_{i,1} + p_{i,2}\right)\gamma_{g,2}^{i} + 1} \\ &\leq \frac{p_{g,2}\gamma_{g,2}^{g}}{p_{g,1}\gamma_{g,2}^{g} + 1},\end{aligned} \tag{24}$$

where $\gamma_{g,k}^{g} = \|\bar{\mathbf{h}}_{g,k}^{\mathrm{H}}\mathbf{d}_g\|_2^2/\sigma^2$ $(g = 1, 2, \ldots, G, k = 1, 2)$, and $\gamma_{g,k}^{i} = \|\bar{\mathbf{h}}_{g,k}^{\mathrm{H}}\mathbf{d}_i\|_2^2/\sigma^2$ $(i = 1, 2, \ldots, G, i \neq g, k = 1, 2)$.

After sorting, $\gamma_{g,1}^{g} \geq \gamma_{g,2}^{g}$ $(g = 1, 2, \ldots, G)$, and thus we have $\mathrm{SINR}_{2\to1}^{g} \geq \mathrm{SINR}_{2\to2}^{g}$. For the sake of theoretical analysis, we assume that the perfect SIC can be carried out in all clusters.

By performing SIC, the received SINR of two users in the g-th cluster can be expressed as

$$\mathrm{SINR}_{1}^{g} = p_{g,1}\gamma_{g,1}^{g}, \tag{25}$$

and

$$\mathrm{SINR}_{2}^{g} = \frac{p_{g,2}\gamma_{g,2}^{g}}{p_{g,1}\gamma_{g,2}^{g} + \sum_{i=1,i\neq g}^{G}\left(p_{i,1} + p_{i,2}\right)\gamma_{g,2}^{i} + 1}, \tag{26}$$

respectively.

On these basis, the EE of the presented mmWave-NOMA with HP is given by

$$\eta_{\mathrm{EE}} = \frac{\sum_{g=1}^{G}\left(R_{g,1} + R_{g,2}\right)}{\sum_{g=1}^{G}\left(p_{g,1} + p_{g,2}\right) + P_{\mathrm{C}}}, \tag{27}$$

where $R_{g,k} = \log_2\left(1 + \mathrm{SINR}_{k}^{g}\right)$, $g = 1, 2, \ldots, G, k = 1, 2$, and $P_{\mathrm{C}} = N_{\mathrm{RF}}P_{\mathrm{RF}} + N_{\mathrm{PS}}P_{\mathrm{PS}} + P_{\mathrm{BB}}$, in which P_{RF}, P_{PS}, and P_{BB} denote the power consumption of the RF chain, the phase shifter and the baseband, respectively [12], [13].

Thus, the EE maximization problem can be formulated as

$$\begin{aligned}&\max_{\{p_{g,1},p_{g,2}\}} \quad \eta_{\mathrm{EE}} \\ &\text{s.t. } \mathcal{C}_1 : R_{g,k} \geq R_{\min}, \quad k \in \{1, 2\}, \ g \in \{1, \ldots, G\}, \\ &\qquad \mathcal{C}_2 : p_{g,1} + p_{g,2} \leq P_{\max}, \quad g \in \{1, \ldots, G\},\end{aligned} \tag{28}$$

where $\mathcal{C}_1$ denotes the minimum rate constraint with $R_{\min}$ for the k-th user in the g-th cluster, and $\mathcal{C}_2$ denotes the maximum power constraint with $P_{\max}$ for the g-th cluster.

According to the fractional programming theory [24], we can transform (28) into the following problem formulated as

$$\begin{aligned}\max_{\{p_{g,1},p_{g,2}\}} \ T(\omega) = {} & \sum_{g=1}^{G}\left(R_{g,1} + R_{g,2}\right) \\ & - \omega\left[\sum_{g=1}^{G}\left(p_{g,1} + p_{g,2}\right) + P_{\mathrm{C}}\right] \\ \text{s.t. } \mathcal{C}_1, \mathcal{C}_2, &\end{aligned} \tag{29}$$

where ω is a non-negative factor. Based on the Dinkelbach's algorithm, ω is updated in the iterative procedure, where the optimal ω (denoted by ω^*) is achieved if and only if $T(\omega^*) = 0$, then we can obtain the optimal value of (28) as $\eta_{\mathrm{EE}}^* = \omega^*$ [24]. In addition, it is noteworthy that if we set $\omega = 0$ in (29), the corresponding optimization problem is equivalent to maximizing the system SE.

For the given ω, (29) can be simplified as

$$\begin{aligned}\max_{\{p_{g,1},p_{g,2}\}} & \sum_{g=1}^{G}\left[R_{g,1}+R_{g,2}-\omega\left(p_{g,1}+p_{g,2}\right)\right] \\ \text{s.t.} & \ \mathcal{C}_1,\ \mathcal{C}_2,\end{aligned} \tag{30}$$

However, the transformed non-convex problem is still difficult to solve directly. Since the inter-cluster interference for the weak user in each cluster can be minimized by the proposed user pairing and HP design scheme, we can adopt the CD method to obtain a sub-optimal solution of (30). Firstly, we set an initial power allocation $\tilde{\mathbf{p}} = \left[\tilde{p}_{1,1}, \tilde{p}_{1,2}, \ldots, \tilde{p}_{G,1}, \tilde{p}_{G,2}\right]$, then (30) can be approximately decomposed into G independent sub-problems. After that, we solve them one by one, in which the g-th ($g = 1, 2, \cdots, G$) sub-problem is formulated as

$$\begin{aligned}\max_{p_{g,1},p_{g,2}} & \ \tilde{R}_{g,1}+\tilde{R}_{g,2}-\omega\left(p_{g,1}+p_{g,2}\right) \\ \text{s.t.} & \ \mathcal{C}_3: p_{g,1}+p_{g,2}\le P_g^{\mathrm{UB}}, \\ & \ \mathcal{C}_4: \tilde{R}_{g,1}\ge R_{\min}, \\ & \ \mathcal{C}_5: \tilde{R}_{g,2}\ge R_{\min},\end{aligned} \tag{31}$$

where $\tilde{R}_{g,1} = R_{g,1}$, $\tilde{R}_{g,2} = \log_2\left(1+\frac{p_{g,2}\gamma_{g,2}^g}{p_{g,1}\gamma_{g,2}^g+\tilde{\alpha}}\right)$, $\tilde{\alpha} = 1+\sum_{i=1,i\neq g}^{G}\tilde{P}_i\gamma_{g,2}^i$, $\tilde{P}_i = \tilde{p}_{i,1}+\tilde{p}_{i,2}$ ($i = 1, 2, \cdots, G$), and $P_g^{\mathrm{UB}} = \min\left\{P_{\max}, \left\{P_{g,i}^{\mathrm{UB}}\right\}\right\}$ ($i = 1, \ldots, G, i \neq g$), in which $P_{g,i}^{\mathrm{UB}}$ represents the upper bound of the total power allocation for the g-th cluster in order to satisfy the minimum rate constraint $R_{\min}$ for the i-th cluster when the total power allocation of the i-th cluster is constant during the iteration of the CD algorithm. The detailed derivation for $P_{g,i}^{\mathrm{UB}}$ will be given below.

By introducing $P_g = p_{g,1}+p_{g,2}$ ($g = 1, 2, \ldots, G$), (31) can be equivalently transformed into the following problem:

$$\begin{aligned}\max_{P_g}\max_{p_{g,1}} & \left\{f_1\left(p_{g,1}\right)\right\}+\log_2\left(P_g\gamma_{g,2}^g+\tilde{\alpha}\right)-\omega P_g \\ \text{s.t.} & \ \mathcal{C}_4,\ \mathcal{C}_5, \\ & \ \mathcal{C}_6: 0\le p_{g,1}\le P_g; \\ & \ \mathcal{C}_7: 0\le P_g\le P_g^{\mathrm{UB}},\end{aligned} \tag{32}$$

where $f_1\left(p_{g,1}\right)$ is the objective function of the inner sub-problem in (32), which is formulated as:

$$\begin{aligned}\max_{p_{g,1}} & \ f_1\left(p_{g,1}\right) = \log_2\left(\frac{p_{g,1}\gamma_{g,1}^g+1}{p_{g,1}\gamma_{g,2}^g+\tilde{\alpha}}\right) \\ \text{s.t.} & \ \mathcal{C}_4,\ \mathcal{C}_5,\ \mathcal{C}_6.\end{aligned} \tag{33}$$

The derivative of $f_1\left(p_{g,1}\right)$ is given by

$$\frac{\partial f_1\left(p_{g,1}\right)}{\partial p_{g,1}} = \frac{\tilde{\alpha}\gamma_{g,1}^g-\gamma_{g,2}^g}{\left(1+p_{g,1}\gamma_{g,1}^g\right)\left(p_{g,1}\gamma_{g,2}^g+\tilde{\alpha}\right)\ln 2}. \tag{34}$$

From (34), we can further obtain $\partial f\left(p_{g,1}\right)/\partial p_{g,1} > 0$ because we have $\gamma_{g,1}^g \ge \gamma_{g,2}^g$ after sorting for the g-th cluster, which means that the optimal $p_{g,1}$ is obtained by its upper bound introduced by the constraint $\mathcal{C}_4$, i.e.,

$$p_{g,1}^* = \frac{P_g}{2^{R_{\min}}}+\frac{2^{-R_{\min}}-1}{\gamma_{g,2}^g}\tilde{\alpha}. \tag{35}$$

Moreover, when solving the power allocation problem for the g-th cluster, the optimal power allocation of the strong user in the i-th cluster is expressed as

$$p_{i,1}^* = \frac{\tilde{P}_i}{2^{R_{\min}}}+\frac{2^{-R_{\min}}-1}{\gamma_{i,2}^i}\left(\tilde{\alpha}-\tilde{P}_i\gamma_{i,2}^i+P_g\gamma_{i,2}^g\right), \tag{36}$$

where, $\gamma_{i,k}^i = \|\bar{\mathbf{h}}_{i,k}^{\mathrm{H}}\mathbf{d}_i\|_2^2/\sigma^2$, and $\gamma_{i,k}^g = \|\bar{\mathbf{h}}_{i,k}^{\mathrm{H}}\mathbf{d}_g\|_2^2/\sigma^2$, and $i = 1, 2, \cdots, G$, $g = 1, 2, \ldots, G$, $i \neq g$, $k = 1, 2$.

With (36), for the i-th cluster, the minimum rate constraint for the weak user can be satisfied, but the minimum rate constraint for the strong user may not be satisfied, thus we have $\log_2\left(1+p_{i,1}^*\gamma_{i,1}^i\right) \ge R_{\min}$. As a result, $P_{g,i}^{\mathrm{UB}}$ in (31) is given by

$$P_{g,i}^{\mathrm{UB}} = \frac{1}{\gamma_{i,2}^g}\left[\left(\frac{\tilde{P}_i}{2^{R_{\min}}-1}-\frac{1}{\gamma_{i,1}^i}\right)2^{R_{\min}}\gamma_{i,2}^i-\tilde{\alpha}\right]. \tag{37}$$

With (35) and (37), the problem in (31) can be rewritten as

$$\begin{aligned}\max_{P_g} & \ f_2\left(P_g\right) = \log_2\left[1+\left(\frac{P_g}{2^{R_{\min}}}+\tilde{\beta}\right)\gamma_{g,1}^g\right] \\ & \qquad +R_{\min}-\omega P_g \\ \text{s.t.} & \ P_g^{\mathrm{LB}}\le P_g\le P_g^{\mathrm{UB}},\end{aligned} \tag{38}$$

where $P_g^{\mathrm{LB}} = 2^{R_{\min}}\left(2^{R_{\min}}-1-\tilde{\beta}\gamma_{g,1}^g\right)/\gamma_{g,1}^g$ denotes the lower bound of P_g induced by the constraint $\mathcal{C}_4$, in which $\tilde{\beta} = \left(2^{-R_{\min}}-1\right)\tilde{\alpha}/\gamma_{g,2}^g$.

Clearly, (38) is a standard convex optimization problem. The corresponding Lagrangian function is given by

$$L\left(P_g, u_1, u_2\right) = f_2\left(P_g\right)+u_1\left(P_g-P_g^{\mathrm{LB}}\right)+u_2\left(P_g^{\mathrm{UB}}-P_g\right), \tag{39}$$

where u_1 and u_2 denote the Lagrange multipliers.

According to the Karush-Kuhn-Tucker (KKT) conditions [25], the optimal solution $\{P_g^*, u_1^*, u_2^*\}$ of (39) should satisfy the following equations.

$$\begin{aligned}& \partial f_2(P_g)/\partial P_g|_{P_g^*}+u_1^*-u_2^* = 0, \\ & u_1^*\left(P_g^*-P_g^{\mathrm{LB}}\right) = u_2^*\left(P_g^{\mathrm{UB}}-P_g^*\right) = 0, \\ & P_g^{\mathrm{LB}}\le P_g^*\le P_g^{\mathrm{UB}}, \quad u_1^*\ge 0, u_2^*\ge 0,\end{aligned} \tag{40}$$

where $\partial f_2(P_g)/\partial P_g$ is given by

$$\frac{\partial f_2\left(P_g\right)}{\partial P_g}=\frac{\gamma_{g,1}^{g}}{\left[2^{R_{\min}}+\left(P_g+2^{R_{\min}}\tilde{\beta}\right)\gamma_{g,1}^{g}\right]\ln 2}-\omega, \quad (41)$$

and $\partial f_2(P_g)/\partial P_g|_{P_g^*}$ denotes the value of $\partial f_2\left(P_g\right)/\partial P_g$ when $P_g=P_g^*$.

From (41), the zero point of $\partial f_2\left(P_g\right)/\partial P_g$ can be derived as

$$\hat{P}_g=\frac{1}{\omega\ln 2}-2^{R_{\min}}\left(\tilde{\beta}+\frac{1}{\gamma_{g,1}^{g}}\right). \quad (42)$$

If $\hat{P}_g<P_g^{\mathrm{LB}}$, i.e., $\partial f_2(P_g)/\partial P_g<0$ for $P_g\in[P_g^{\mathrm{LB}},P_g^{\mathrm{UB}}]$, $u_1^*>0$ can be obtained by (40), which means $P_g^*=P_g^{\mathrm{LB}}$;

If $\hat{P}_g>P_g^{\mathrm{UB}}$, i.e., $\partial f_2(P_g)/\partial P_g>0$ for $P_g\in[P_g^{\mathrm{LB}},P_g^{\mathrm{UB}}]$, $u_2^*>0$ can be obtained by (40), which means $P_g^*=P_g^{\mathrm{UB}}$;

If $P_g^{\mathrm{LB}}\le\hat{P}_g\le P_g^{\mathrm{UB}}$, $P_g^*=\hat{P}_g$ can be derived by $u_1^*=u_2^*=0$.

In conclusion, P_g^* can be expressed as

$$P_g^*=\min\left\{P_g^{\mathrm{UB}},\max\left\{P_g^{\mathrm{LB}},\hat{P}_g\right\}\right\}. \quad (43)$$

Based on the above analysis, the energy-efficient power allocation algorithm for the presented mmWave-NOMA with HP can be summarized as Algorithm 1.

Algorithm 1 Energy-Efficient Power Allocation Algorithm

Initialize tolerate $\varepsilon>0$, $\omega=0$
repeat
  Initialize power allocation $\tilde{\mathbf{p}}$
  repeat
    for $i=1$ **to** G **do**
      Calculate P_g^* by (43)
      Calculate $p_{g,1}^*$ by (35)
      $p_{g,2}^*=P_g^*-p_{g,1}^*$
    end for
  until $\tilde{\mathbf{p}}$ converge
  $\delta=\sum_{g=1}^{G}\left[R_{g,1}+R_{g,2}-\omega\left(p_{g,1}+p_{g,2}\right)\right]-\omega P_C$
  $\omega=\sum_{g=1}^{G}\left(R_{g,1}+R_{g,2}\right)/\left[\sum_{g=1}^{G}\left(p_{g,1}+p_{g,2}\right)+P_C\right]$
until $\delta\le\varepsilon$

In what follows, we will analyze the complexity of Algorithm 1. Let I_1 and I_2 denote the number of outer and inner iterations, respectively. For each inner iteration, the closed-form solution of each sub-problem can be attained. Therefore, the complexity of Algorithm 1 is $\mathcal{O}\left(GI_1I_2\right)$. In contrast, the energy-efficient power allocation algorithm proposed in [13] does not take into account the influence of minimum rate constraints for other clusters within the iteration of each sub-problem, and use the Lagrange dual method to obtain the solution iteratively. The complexity is $\mathcal{O}\left(D^2GI_1I_2\right)$, where $D=3$ denotes the number of dual variables induced by sub-gradient updated-based Lagrange dual method. Therefore, Algorithm 1 has lower complexity than the one proposed in [13]. When solving the SE maximization problem for the proposed mmWave-NOMA system, we just need the inner iteration of Algorithm 1, namely Algorithm 2, whose complexity is $\mathcal{O}\left(GI_2\right)$.

For a given $\{P_1,P_2,\ldots,P_G\}$, the optimal power allocation $\{p_{g,1}^*,p_{g,2}^*\}$ for the g-th cluster ($g=1,2,\ldots,G$) can be attained by (35) and $p_{g,2}^*=P_g^*-p_{g,1}^*$. Hence, the global optimal power allocation $\{P_g^*\}$ ($g=1,2,\ldots,G$) can be found by the exhaustive search method. Namely, they can be directly searched over $(0,P_{\max}]$ for all clusters while satisfying the minimum rate constraints. However, the complexity of the exhaustive search method is extremely higher, i.e., $\mathcal{O}((\frac{1}{\epsilon})^G)$, where ϵ is the search precision. Alternatively, we can use the chaotic accelerated particle swarm optimization (CAPSO) method in [26] to obtain the global optimal power allocation with high probability. For the CAPSO method, the position of each particle is a G-dimension vector, and thus its complexity is $\mathcal{O}(C_1C_2G)$, where C_1 denotes the number of iterations, and C_2 is the number of particles [26]. By comparison, the complexity of the CAPSO method linearly increases as G, while the complexity of the exhaustive search method exponentially increases as G.

To evaluate the EE performance of the proposed scheme with Algorithm 1, the EE comparison of the proposed scheme, the existing scheme in [13], the CAPSO scheme, and the high-complexity exhaustive search scheme will be given in the next section.

TABLE 1. The system simulation parameters.

Parameters	Default
HP architecture	Fully-connected
Number of antennas	$N=16$
Number of clusters	$G=4$
Number of RF chains	$N_{\mathrm{RF}}=4$
Number of optional phases	$N_{\mathrm{C}}=16$
Power of noise	$\sigma^2=1\mathrm{mW}$
Power consumption of RF chain	$P_{\mathrm{RF}}=300\mathrm{mW}$
Power consumption of phase shifter	$P_{\mathrm{PS}}=20\mathrm{mW}$
Power consumption of baseband	$P_{\mathrm{BB}}=200\mathrm{mW}$
Threshold of channel correlation	$\rho=0.8$
Minimum rate constraint	$R_{\min}=1\mathrm{bit/s/Hz}$

V. SIMULATION RESULTS

In this section, simulation results are provided to testify the effectiveness of the proposed energy-efficient power allocation scheme in the multiuser mmWave-NOMA system, where the Monte-Carlo method is employed for simulation. We consider a single-cell with one BS transmission scenario and $K(K=2G)$ users and assume that there are enough users to form G clusters [13]. The mmWave channel is modelled as (7). It is assumed that $\lambda_{g,k}$ follows the distribution $\mathcal{CN}(0,1)$, $\theta_{g,k}$ is uniformly distributed over $[0,2\pi]$, and $L_{g,k}=1\,(g=1,\ldots,G,k=1,2)$. Unless otherwise stated, we adopt the proposed analog beamforming in (14) for the fully-connected HP and (15) for the sub-connected HP, respectively, to perform the corresponding analog precoding, and the main simulation parameters are listed in Table 1.

Fig.3 illustrates the comparisons of EE performance of the system with different schemes, i.e., the proposed scheme,

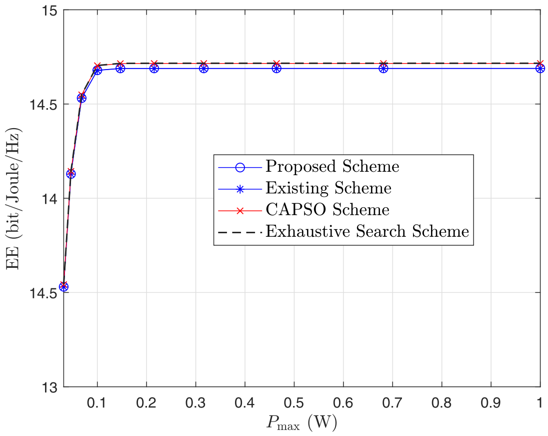


FIGURE 3. Comparison of EE performances of the system with different schemes.

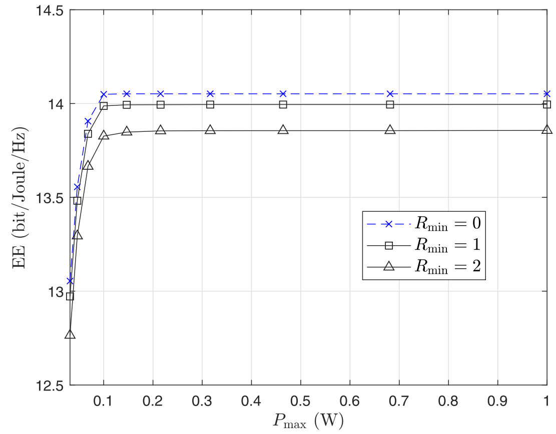


FIGURE 5. Comparison of EE performances of the system with different values of R_{min}.

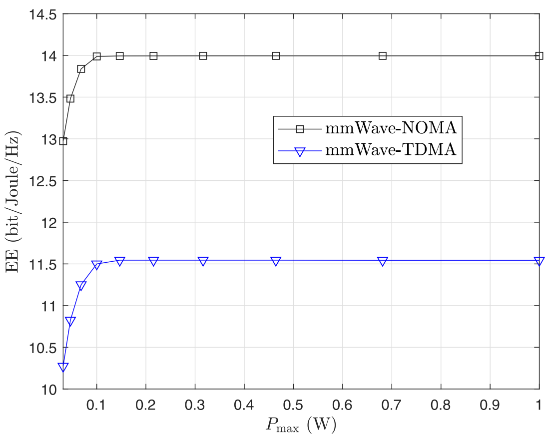


FIGURE 4. Comparison of EE performances between mmWave-NOMA and mmWave-TDMA.

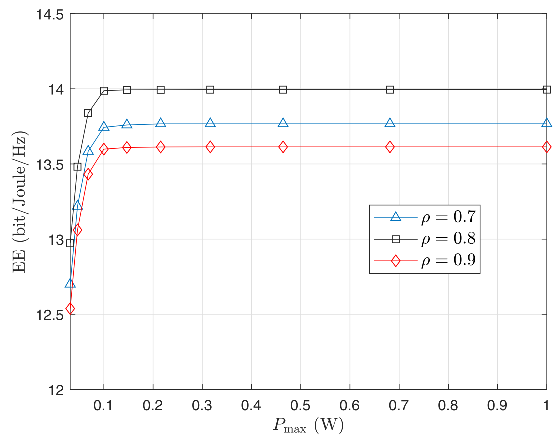


FIGURE 6. Comparison of EE performances of the system with different values of ρ.

the existing scheme in [13], the CAPSO scheme, and the exhaustive search scheme are compared, where we set $N = 8$ and $G = N_{RF} = 2$ for the convenience of comparison. From Fig.3, it is observed that the proposed scheme can obtain the EE identical to that of the existing scheme, and has almost the same EE performance as the exhaustive search scheme as well as the CAPSO scheme, but our scheme has the lowest complexity among them, which can be seen from the complexity analysis in above section. The results show the effectiveness of the proposed scheme. Meanwhile, the CAPSO scheme has the same EE performance as the exhaustive search scheme, which proves that the CAPSO scheme can effectively converge to the global optimal power allocation. This result indicates that the superiority of the CAPSO scheme over the exhaustive search scheme, especially for large G. It can be seen that when $P_{\max}$ is small, the EEs of the considered four schemes increase with the increase of $P_{\max}$. After $P_{\max}$ reaches a certain value, the optimal total power allocation for each cluster is no longer changed and less than $P_{\max}$. Thus, the EE performances tend to be stable.

Fig.4 shows the EE performances of the mmWave system with NOMA and TDMA, namely "mmWave-NOMA" and "mmWave-TDMA", respectively. For the "mmWave-TDMA", each user in the same cluster will be allocated with equal time slots. We assume that the interference can be eliminated by performing the corresponding HP scheme in each time slot. We can see from Fig.4 that the EE performance of "mmWave-NOMA" is significantly better than that of "mmWave-TDMA", which means that the application of NOMA in the mmWave communication system can effectively enhance the EE performance of the system.

Fig. 5 gives the comparison of EE performances of the system in the presence of different rate constraints, where the rate constraints $\{R_{\min}\}$ are set equal to 0, 1, 2 bit/s/Hz, respectively. It can be observed that with the increase of $R_{\min}$, the EE performance of the system decreases gradually. This is because the rate needs to be increased when $R_{\min}$ increases, which will consume more transmit powers. Thus, the possibility of the system achieving higher EE performance will be greatly lowered. Especially, the system without rate constraint (i.e., $R_{\min} = 0$) will obtain higher EE than that with rate constraint (i.e., $R_{\min} > 0$), as expected.

In Fig.6, we plot the EE performances of the system with different values of the user pairing threshold ρ, where ρ is equal to 0.7, 0.8, 0.9, respectively. It can be observed that

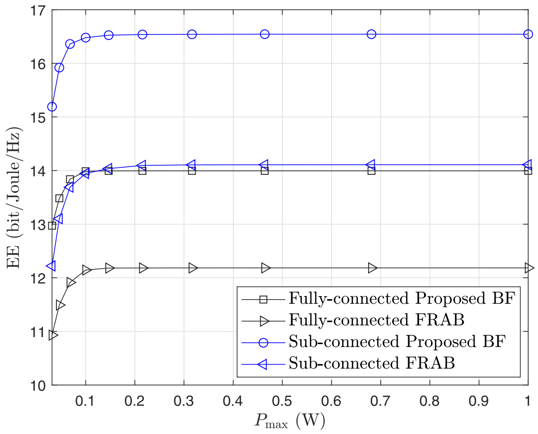


FIGURE 7. EE comparison of the system with two HP architectures for different analog precoding schemes.

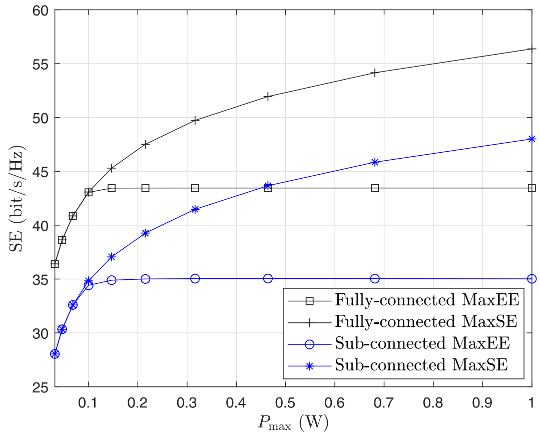


FIGURE 8. SE comparison of the system with two HP architectures for the "MaxEE" scheme and the "MaxSE" scheme.

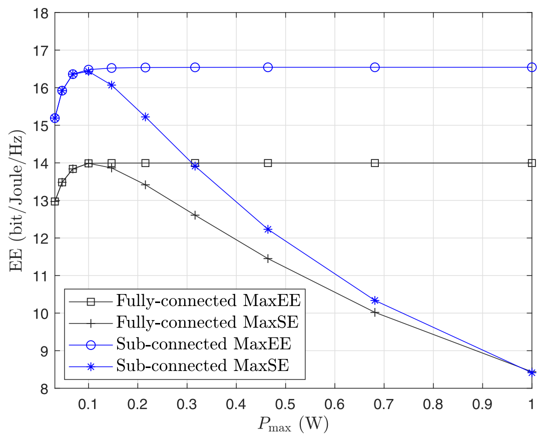


FIGURE 9. EE comparison of the system with two HP architectures for the "MaxEE" scheme and the "MaxSE" scheme.

the user pairing threshold ρ has a great influence on the EE performance of the system, and the EE performance is not linearly increasing or decreasing with respect to ρ. The reason is that as ρ increases, the higher the correlation between these two users in each cluster is, the better the inter-cluster interference cancellation of the weak users can be attained, but the intra-cluster channel difference may be decreased at the same time, which will bring small sum of rate gains for all clusters by performing SIC.

In Fig.7, we present the EEs of both fully-connected HP and sub-connected HP systems with different analog precoding schemes, where the legends "FRAB" denote the analog beamforming in (9) for the fully-connected and (11) for the sub-connected, respectively, and the legends "Proposed BF" denote the proposed analog beamforming in (14) for the fully-connected and (15) for the sub-connected, respectively. We can observe that the proposed analog beamforming obviously outperforms the conventional FRAB with regard to the system EE for both the fully-connected and sub-connected HP architectures, which means that the proposed analog precoding can be employed to further improve the system performance.

In Fig. 8 and Fig. 9, we plot the SE and EE of the system with two typical HP architectures, respectively, where the fully-connected HP architecture and the sub-connected HP architecture are both employed, "MaxEE" denotes the proposed EE maximization scheme, and "MaxSE" represents the SE maximization scheme, which refers to $\omega = 0$ in (29). From Fig. 8, it is found that with the increase of $P_{\max}$, the SE of the "MaxEE" scheme tends to be stable gradually, while the SE of the "MaxSE" scheme is still increasing. When $P_{\max}$ is smaller, the "MaxSE" scheme has the same SE as the "MaxEE" scheme, which means that the PA strategies of the two schemes are identical under this case. This is because the obtained optimal power allocation is beyond $P_{\max}$ under this case, and resultant same power allocations are achieved due to the the limitation of maximum power. However, when $P_{\max}$ is larger, the sum power P_g will become large as well, and correspondingly, "MaxSE" scheme will consume more inefficient power to pursue higher SE because this scheme do not care how much power is consumed, and the "MaxEE" scheme will terminate the increase of SE brought by the inefficient power consumption which will result in the reduction of EE. As a result, "MaxSE" scheme will exhibit high SE, and 'MaxEE" scheme has the stable SE. Besides, it is obvious that the system with fully-connected HP architecture outperforms than that with sub-connected counterpart in terms of SE. The reason is that the RF chains of the former can make full use of array gains.

From Fig. 9, we can see that with the increase of $P_{\max}$, the EE of "MaxEE" scheme tends to be stable, while the EE of "MaxSE" scheme begins to decline gradually. This is because for "MaxSE" scheme, its SE improvement is obtained by sacrificing more transmit power, which will result in the reduction of EE. When $P_{\max}$ is small, these two schemes have almost the same EE performances due to the reason analyzed in Fig. 8. Moreover, the sub-connected HP architecture can achieve higher EE performance than the fully-connected HP architecture. The reason for this is that although the SE of the former is less than that of the latter, the power consumption of the former is much smaller than that of the latter.

VI. CONCLUSION

We have investigated the EE optimization in a downlink multi-user mmWave-NOMA system with hybrid precoding by considering the fully-connected and sub-connected HP architectures, and a suboptimal energy-efficient power allocation scheme with low complexity is proposed for the system. Based on the analysis of EE, the user pairing scheme and two-step HP design are firstly presented. Namely, the analog beamforming is presented to improve the performance, and the digital ZF precoding is designed to eliminate the inter-cluster interference for the strong users in all clusters. With these results, the EE maximization problem is then formulated, and can be divided into independent convex sub-problems by using the fractional programming theory. Moreover, for these sub-problems, we have derived the closed-form solutions by means of the CD method. On these basis, a low-complexity iterative algorithm is proposed to find the suboptimal solution for the original optimization problem. Computer simulation indicates that the proposed scheme can achieve superior EE performance with low complexity and the proposed analog precoding enjoys higher performance compared with the FRAB.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the Editor for their valuable comments which improve the quality of this paper greatly.

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf.*, Berlin, Germany, Jun. 2013, pp. 1–5.

[2] N. Zhao, W. Wang, J. Wang, Y. Chen, Y. Lin, Z. Ding, and N. C. Beaulieu, "Joint beamforming and jamming optimization for secure transmission in MISO-NOMA networks," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2294–2305, Mar. 2019.

[3] N. Zhao, X. Pang, Z. Li, Y. Chen, F. Li, Z. Ding, and M.-S. Alouini, "Joint trajectory and precoding optimization for UAV-assisted NOMA networks," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3723–3735, May 2019.

[4] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[5] Z. Xiao, P. Xia, and X.-G. Xia, "Enabling UAV cellular with millimeter-wave communication: Potentials and approaches," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 66–73, May 2016.

[6] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.

[7] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X.-G. Xia, "Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2961–2974, May 2018.

[8] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Joint power control and beamforming for uplink non-orthogonal multiple access in 5G millimeter-wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6177–6189, Sep. 2018.

[9] 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.

[10] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "Multi-beam NOMA for hybrid mmWave systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1705–1719, Feb. 2019.

[11] 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. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[12] 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.

[13] W. Hao, M. Zeng, Z. Chu, and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec. 2017.

[14] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[15] 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.

[16] 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.

[17] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[18] Z. Ding, L. Dai, R. Schober, and H. V. Poor, "NOMA meets finite resolution analog beamforming in massive MIMO and millimeter-wave networks," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1879–1882, Aug. 2017.

[19] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 653–656, Dec. 2014.

[20] B. Kimy, S. Lim, H. Kim, S. Suh, J. Kwun, S. Choi, C. Lee, S. Lee, and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2013, pp. 1278–1283.

[21] Y. Zhao, W. Xu, and S. Jin, "An minorization-maximization based hybrid precoding in NOMA-mMIMO," in *Proc. 9th Int. Conf. Wireless Commun. Signal Process.*, Nanjing, China, Oct. 2017, pp. 1–6.

[22] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

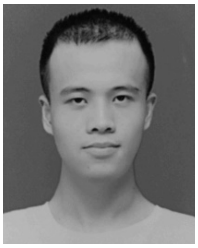
[23] A. Wiesel, Y. C. Eldar, and S. Shamai (Shitz), "Zero-forcing precoding and generalized inverses," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, Sep. 2008.

[24] W. Dinkelbach, "On nonlinear fractional programming," *Manage. Sci.*, vol. 13, no. 7, pp. 492–498, 1967.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge U.K.: Cambridge Univ. Press, 2004.

[26] A. H. Gandomi, G. J. Yun, X.-S. Yang, and S. Talatahari, "Chaos-enhanced accelerated particle swarm optimization," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 18, no. 2, pp. 327–340, Feb. 2013.

XIANGBIN YU received the Ph.D. degree in communication and information systems from the National Mobile Communications Research Laboratory, Southeast University, China, in 2004. From 2010 to 2011, he was a Research Fellow with the Department of Electronic Engineering, City University of Hong Kong. From 2014 to 2015, he was a Visiting Scholar with the Electrical and Computer Engineering, University of Delaware, USA. He is currently a Full Professor with the Nanjing University of Aeronautics and Astronautics, China. His research interests include distributed MIMO, NOMA, precoding design, mmWave communication, and green communication. He has been a member of the IEEE ComSoc Radio Communications Committee (RCC), since 2007, and a Senior Member of the Chinese Institute of Electronics, since 2012. He has served as a Technical Program Committee Member of the 2006 and 2017 IEEE Global Telecommunications Conference, the 2011 and 2017 Wireless Communications and Signal Processing, and the 2015 and 2018 IEEE International Conference on Communications.



FANGCHENG XU received the B.S. degree in information engineering from the Nanjing University of Aeronautics and Astronautics, where he is currently pursuing the M.Sc. degree.

KAI YU received the M.S. degree in communication and information systems from Dalian Maritime University. He is currently pursuing the Ph.D. degree with the Nanjing University of Aeronautics and Astronautics.

XIAOYU DANG (M'09) received the Ph.D. degree in electrical engineering from Brigham Young University, Provo, UT, USA, in 2009. From 2016 to 2017, he was a Visiting Scholar with The University of Tennessee, Knoxville, TN, USA. He is currently a Full Professor with the College of Electronics and Information Engineering, Nanjing University of Aeronautics and Astronautics, Jiangsu, China. His technical interests include coding and modulation, diversity techniques, and reliable transmission of signals in deep space environments.