Proportional Fair Scheduling for Downlink mmWave Multi-User MISO-NOMA Systems

Mingshan Zhang, Yongna Guo, Lou Salaün, Chi Wan Sung, Chung Shue Chen

Proportional Fair Scheduling for Downlink mmWave Multi-User MISO-NOMA Systems

Mingshan Zhang, Yongna Guo, *Student Member, IEEE*, Lou Salaün, *Member, IEEE*, Chi Wan Sung, *Senior Member, IEEE*, and Chung Shue Chen, *Senior Member, IEEE*

***Abstract*—In this paper, we study non-orthogonal multiple access (NOMA) user scheduling and resource allocation problem for a generic downlink single-cell multiple input and single output (MISO) millimeter wave (mmWave) system. The larger number of packed antennas and the highly directional property of mmWave communications enable directional beamforming to achieve spatial diversity. Toward this end, we consider two different hybrid precoding schemes which are based on orthogonal matching pursuit (OMP). Users are assigned into different clusters and the base station (BS) transmits superposed signals that share the same precoding vector. Moreover, both fixed number of users per cluster and dynamic number of users per cluster are investigated. We aim to jointly optimize the user clustering, service scheduling, and power allocation strategy, in maximizing the proportional fairness (PF) among the users and exploring the multiuser diversity and multiplexing gain. Since the formulated joint user clustering, scheduling and power allocation problem is a mixed integer non-convex optimization problem, we propose a two-fold methodology. First, we apply a hybrid precoding and user clustering scheme, where the hybrid precoder is constructed by singular vector division (SVD) or minimum mean square error (MMSE). Then, with the obtained result, we approximate the proportional fairness power allocation problem by a sequence of Geometric Programming (GP) problems which are solved iteratively. The proposed scheme strikes a balance between the spectral efficiency and service fairness. Results show that the proposed MISO-NOMA scheme which is based on MMSE hybrid precoder and the proposed user scheduling and power allocation strategy under proportional fairness metric can outperform various conventional MISO schemes. Furthermore, our proposed dynamic number of users per cluster scheme outperforms the fixed scheme and can be considered as an upper bound in several aspects, including spectral efficiency and fairness.**

***Index Terms*—Millimeter wave, opportunistic hybrid beamforming, multiple input single output (MISO), non-orthogonal multiple access (NOMA), proportional fair scheduling.**

I. Introduction

With the explosive growth of mobile data services, the demand for higher network capacity and more diversified mobile network applications imposes challenges for 5G and beyond 5G (B5G) networks. Millimeter wave (mmWave) communication has been considered as a promising technology to meet the fast growing data rate demand [1], [2]. The large available spectrum bandwidth in mmWave (3-300 GHz) provides a potential solution to tackle the bandwidth shortage problem. Many countries and regions such as USA, Europe, Korea, Japan and China have announced their 5G spectrum strategies and roadmaps for the frequency spectrum 24.25-29.5 GHz as a key frequency band to deploy commercial systems for new radio (NR) [3]. Different from the characteristics of sub-6 GHz wireless communication, the propagation of mmWave is highly directional with severe path loss, low penetration and high signal attenuation [4]. Besides, the short wavelength of mmWave allows a large amount of antennas to be packed in a small form, to achieve spatial diversity through highly directional beamforming [5]. Thus, multiple-input multiple-output (MIMO) processing can in turn help reap the gains offered by mmWave.

In traditional MIMO, the beamforming for baseband signal is always done digitally, controlling both the signal's phase and amplitude, where each antenna element is controlled by dedicated baseband and radio frequency (RF) hardware. However, mmWave systems usually contain a large number of antennas, which would increase the cost and power consumption if full digital precoder is applied. Moreover, mmWave channels may be sparser such that fewer spatial degrees of freedom are available [6]. Indeed, the sparsity can be exploited for optimizing channel estimation and beam training. To address this issue, the idea of hybrid precoding was proposed in [7] and [8], where the conventional digital precoder is divided into two parts: a low-dimensional digital baseband precoder, and a high-dimensional RF precoder which is implemented via analog phase shifters. Thus, the hybrid precoding problem was formulated as a sparse approximation problem and solved by the proposed spatially sparse precoding (SSP) algorithm [9].

Non-orthogonal multiple access (NOMA) is another key technology in 5G/B5G, and has attracted much attention due to its excellent overload performance compared to traditional orthogonal multiple access (OMA) method [10]. Compared with OMA transmission, the complexity of the NOMA receiver is higher, but a higher spectrum efficiency can be obtained. For example, in [11], the authors investigate and reveal the ergodic sum-rate gain (ESG) of NOMA over OMA in uplink cellular communication systems. The application of NOMA in MIMO systems can further improve the system capacity. Regarding MIMO-NOMA, it was shown that the application of MIMO in NOMA is able to achieve an enhanced system performance compared to pure NOMA and pure MIMO [12]. Since beamforming plays an important role in MIMO systems, the authors in [13] investigate the beamforming problem in MISO-NOMA

M. Zhang is with Network and Edge Group, Intel Research Center, China. She was with Nokia Bell Labs, France. Email: mingshan.zhang@intel.com

Y. Guo and C. W. Sung are with the Dept. of Electrical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR. Email: yongnaguo2-c@my.cityu.edu.hk, albert.sung@cityu.edu.hk

L. Salaün and C. S. Chen are with Nokia Bell Labs, Paris-Saclay, Nozay 91620, France. Email: {lou.salaun, chung_shue.chen}@nokia-bell-labs.com

(M. Zhang and Y. Guo contributed equally to this work.)

A part of the work was carried out at the Laboratory for Information, Networking and Communication Sciences (www.lincs.fr).

system, and they further study beamforming design with the aid of reconfigurable intelligent surface (RIS) recently [14]. In [15], the user clustering, beamforming (BF), and power allocation problem for sum rate maximization of a single-cell MIMO-NOMA system is investigated, in which the number of antennas per user is more than that of the base stations. In [16], the authors investigate the power allocation problem of NOMA in a virtual MIMO system for IoT communications. However, it only considers the problem in one time slot but not the long-time scheduling problem. A k-means based machine-learning algorithm for user clustering in a mmWave system is proposed in [17], which can be performed online with low computational complexity. Moreover, unmanned aerial vehicle (UAV) can also be exploited with NOMA. In [18], the authors propose a UAV-assisted NOMA network, in which the UAV and BS cooperate with each other to serve ground users simultaneously. And the UAV-assisted NOMA transmission system in mmWave is further investigated in [19].

For multi-user scenario, a base station (BS) can track the channel condition of each user and schedule transmissions among users according to their instantaneous channel quality [20], so that the multi-user diversity benefits can be extracted. There is usually a trade-off between the user fairness and cell throughput. A proportional fairness (PF) scheduler can be used to strike a balance of them while harnessing multiuser diversity [21]. The user fairness of NOMA has been investigated in [22] and [23] where BS transmits to several users in each time slot. A proportional fair sub-carrier allocation scheduling scheme for downlink NOMA has been proposed in [24]. Besides, an opportunistic hybrid user scheduling strategy, called memory-based greedy user scheduling, was proposed in [25], aiming at maximizing the sum PF metric among users, while assuming that each user communicates with the BS via only a single stream.

To the best of our knowledge, the existing literature and user scheduling studies for mmWave systems usually consider simple user clustering, scheduling or power allocation problem, which may either transmit superposed signals of multiple users (i.e., NOMA) or transmit via different data streams for different users in each time slot (i.e., OMA). Besides, the joint user clustering, scheduling and power allocation optimization problem for generic multi-beam MISO-NOMA in mmWave systems has not been addressed yet, which needs a careful study on how to fully explore the multiuser diversity and multiplexing gain.

In this paper, we explore the potential of MISO-NOMA user clustering, scheduling, and power allocation joint optimization in downlink mmWave systems. The main contributions of the paper can be summarized as follows:

1) We model and formulate a joint user clustering, service scheduling and power allocation problem for a downlink mmWave system to maximize the proportional fairness metric.
2) To handle the joint mixed integer non-convex optimization problem and deal with its high complexity, we decouple the problem and tackle in a two-step strategy: a heuristic user clustering and an iterative power allocation scheme in solving a sequence of geometric programming (GP) problems.
3) Given the hybrid precoding matrix and user clustering strategy, we satisfy the proportional fairness metric through the proposed time scheduling and show that during each iteration and based on a fixed value of interference term, the subproblem can be solved by GP method. According to the simulation results, the algorithm always converges after a limited number of iterations and the final power allocation solution can be obtained.
4) The performance of the proposed schemes is evaluated by extensive simulations. Results show that our proposed MISO-NOMA user clustering and power allocation scheme built on proportional fairness metric and MMSE precoder can achieve a good tradeoff between user throughput and fairness. Besides, we study the impact of the maximum allowable number of users per cluster and propose an interesting strategy that dynamically decides the number of users for each cluster. We show that this dynamic scheme outperforms the standard fixed approach.

Note that in our previous work [26], a generic power minimization problem for a general downlink MISO-NOMA system was studied, which however did not consider user scheduling and explicit user fairness. Besides, here we focus on the mmWave system and problem, which is different and of great potential for 5G/B5G networks. Most importantly, we consider a thorough user clustering, service time scheduling and power allocation joint optimization problem in order to achieve enhanced spectrum utilization efficiency, user fairness, and number of supported users.

Notation: $(\cdot)^T$ is used to denote the transpose of a matrix while $(\cdot)^H$ is used to denote the conjugate transpose of a matrix. $|\cdot|$ denotes the magnitude of a scalar while $\|\cdot\|_F$ represents the Frobenius norm. Furthermore, $\mathcal{CN}(\cdot,\cdot)$ denotes a complex Gaussian distribution, and $\mathbf{I}_N$ indicates an $N \times N$ identity matrix. Furthermore, $\mathbb{E}[\cdot]$ is used to denote statistical expectation of random variable.

II. System Model and Problem Formulation

In this section, we define the system model and notations used throughout the paper.

A. System Model

In this work, we consider a downlink single-cell MISO-NOMA system, where a base station (BS) serves K users. The set of users is denoted by $\mathcal{K}$. It is assumed that each user experiences frequency-flat block fading channel, for which the channel condition remains constant within the frequency band and within a time slot. The transmitter consists of hybrid RF and baseband precoders to exploit the sparse-scattering nature of mmWave channels.

For massive MIMO hybrid beamforming, the transmitter architecture can be categorized as full connection architecture and sub-array partition architecture. For the fully connected

TABLE I
SYSTEM MODEL PARAMETERS

Notations	Description
N_s	The number of data streams
N_{RF}	The number of RF chains
N_{TX}	The number of physical transmit antennas
$\mathbf{F}_{\text{BB}}$	$N_{\text{RF}} \times N_s$ baseband precoder
$\mathbf{F}_{\text{RF}}$	$N_{\text{TX}} \times N_{\text{RF}}$ RF precoder

structure, each antenna unit is associated with the weighted sum of all outputs from each RF chain. Although its complexity is higher than that of the sub-array model, the full connection architecture provides better suppression of the sidelobe of its radiation pattern.

A downlink multi-user mmWave system is considered. The BS is equipped with N_{TX} transmit antennas and N_{RF} RF chains, to transmit N_s ($\leq N_{\text{RF}}$) data streams to K users. Each user is equipped with a single receive antenna. In practical systems, the number of users K may be much larger than N_s, and the channel condition of each user may vary during each scheduling time slot, some users will be scheduled to transmit in each cluster during each time slot according to the scheduling scheme. Suppose that for each time slot the BS schedules transmissions to a certain number of users, and we denote the index set of scheduled users for time slot t by $\mathcal{U}(t)$. Denote by $\mathcal{J}$ the index set of data streams of the downlink multi-user transmissions, where $\mathcal{J} \triangleq \{1, 2, \ldots, N_s\}$. Each data stream $j \in \mathcal{J}$ serves one cluster with L_j users performing NOMA. Besides, denote by $\mathcal{U}_j(t)$ the index set of users who are assigned to the j-th cluster in time slot t, where $\mathcal{U}(t) = \mathcal{U}_1(t) \bigcup \mathcal{U}_2(t) \bigcup \ldots \bigcup \mathcal{U}_{N_s}(t)$. It is assumed that users in the same cluster communicate with the BS via only a single stream, and the receiver performs successive interference cancellation (SIC) to decode the superposed signal of users in the same cluster. Note that the time slot variable t can be omitted for simplicity unless it is needed for clarity.

The hardware block diagram of mmWave transmitter is shown in Fig. 1. The transmitter consists of a full-connected hybrid beamforming architecture, where each RF chain is connected to all N_{TX} transmit antennas. The hardware architecture of the transmitter consists of an $N_{\text{RF}} \times N_s$ baseband precoder $\mathbf{F}_{\text{BB}}$, followed by an $N_{\text{TX}} \times N_{\text{RF}}$ RF precoder $\mathbf{F}_{\text{RF}}$. The major notations of the system model are illustrated in Table I. Note that N_s, N_{RF}, N_{TX} are related to the transmitter physical settings, and are treated as constants in the optimization problem.

The transmitted signal is thus expressible as:

$$\boldsymbol{x} = \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\boldsymbol{s} = \mathbf{F}\boldsymbol{s} = \sum_{j=1}^{N_s} \mathbf{f}_j s_j,$$

where $\boldsymbol{s} \triangleq [s_1, s_2, \ldots, s_{N_s}]^T$ is used to denote the transmitted data stream vector of dimension $N_s \times 1$, and $\mathbf{F} \triangleq \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}$ which consists of column vector $\mathbf{f}_j$ of dimension $N_{\text{TX}} \times 1$ with $j = 1, 2, \ldots, N_s$. We assume that the channel state information can be obtained in the transmitter side by some channel estimation methods in mmWave channels such as using multiple frequency tones [27] or deep learning methods [28]. It is assumed that data stream s_j represents the superposed signal of NOMA users in cluster j, i.e., $s_j \triangleq \sum_{k \in \mathcal{U}_j} \sqrt{p_k} b_k$, where b_k is the desired data of user k, and p_k is the transmit power for user k. Let $\boldsymbol{p} \triangleq (p_1, p_2, \ldots, p_K)$ be the power vector of all users in the cell. Further, let $q_j \triangleq \sum_{k \in \mathcal{U}_j} p_k$ be the sum transmit power of users in the j-th cluster and we use power vector $\boldsymbol{q} \triangleq (q_1, q_2, \ldots, q_{N_s})$ to denote the sum transmit power of users of all clusters.

It should be noted that RF precoder is implemented using analog phase shifters, thus all entries of $\mathbf{F}_{\text{RF}}$ are of constant modulus, i.e., $|\mathbf{F}_{\text{RF}}^{m,n}| = 1/\sqrt{N_{\text{TX}}}$. It is also assumed that the phase of each entry in $\mathbf{F}_{\text{RF}}$ is quantized to $\mathcal{Q}$ bits, thus $\mathbf{F}_{\text{RF}}^{m,n} = \frac{1}{\sqrt{N_{\text{TX}}}} e^{j\varphi_{m,n}}$, where $\varphi_{m,n} \in \{0, \frac{2\pi}{2^{\mathcal{Q}}}, \ldots, \frac{2(2^{\mathcal{Q}}-1)\pi}{2^{\mathcal{Q}}}\}$.

In narrowband block-fading channels, the received signal of user k is expressible as:

$$y_k = \mathbf{h}_k^T \boldsymbol{x} + n_k = \mathbf{h}_k^T \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\boldsymbol{s} + n_k = \mathbf{h}_k^T \mathbf{f}_j s_j + \mathbf{h}_k^T \sum_{j' \neq j}^{N_s} \mathbf{f}_{j'} s_{j'} + n_k, \tag{1}$$

where y_k is the received signal of user k who is assigned to cluster j, $\mathbf{h}_k$ is the $N_{\text{TX}} \times 1$ channel vector such that $E[\|\mathbf{h}_k\|_F^2] = N_{\text{TX}}$, $\mathbf{n} \triangleq [n_1, \ldots, n_K]^T$ is the vector of i.i.d. $\mathcal{CN}(0, \sigma^2)$ noise, and $\|\cdot\|_F$ refers to Frobenius norm.

Let $\mathbf{f}_{\text{BB},j}$ be each column vector of $\mathbf{F}_{\text{BB}}$, where $j = 1, 2, \ldots, N_s$. The inter-cluster interference plus noise of user k who is assigned to cluster j can be obtained by (1) as follows:

$$\hat{I}_k \triangleq \sum_{j' \neq j}^{N_s} q_{j'} |\mathbf{h}_k^T \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j'}|^2 + \sigma^2, \tag{2}$$

where the first term is the inter-cluster interference and the second term is the noise power at user k. To simplify the notation, we use I_k to represent the normalized inter-cluster interference plus noise value of user k, i.e.,

$$I_k \triangleq \frac{\hat{I}_k}{|\mathbf{h}_k^T \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j}|^2}, \tag{3}$$

where $\hat{I}_k$ is given by (2).

In each cluster, the user signals are superposed for transmission, and decoded by SIC at the reception of users. In general downlink communication scenarios, the users with better channel condition are allocated with lower power. The optimal SIC decoding for downlink is in the decreasing order of normalized interference plus noise value [29]–[31], which depends on a user's channel gain and also its experienced interference and noise level. We define power vector $\boldsymbol{q}_{-j} \triangleq (q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_{N_s})$. The decoding order of users in each cluster should be decided by their normalized interference plus noise value, which is determined by the power allocation for other clusters, i.e., $\boldsymbol{q}_{-j}$. The decoding order of users in cluster j can be thus denoted by:

$$\boldsymbol{\pi}_j(\boldsymbol{q}_{-j}) \triangleq (\pi_j(1), \pi_j(2), \ldots, \pi_j(L_j)), \tag{4}$$

where $\pi_j(l)$, $l = 1, 2, \ldots, L_j$, being the l-th element of $\boldsymbol{\pi}_j$, means that user $\pi_j(l)$ will first decode the signals of $\pi_j(1)$ to $\pi_j(l-1)$ and then subtract these signals to decode its own signal while treating the signals of $\pi_j(l+1)$ to $\pi_j(L)$ as

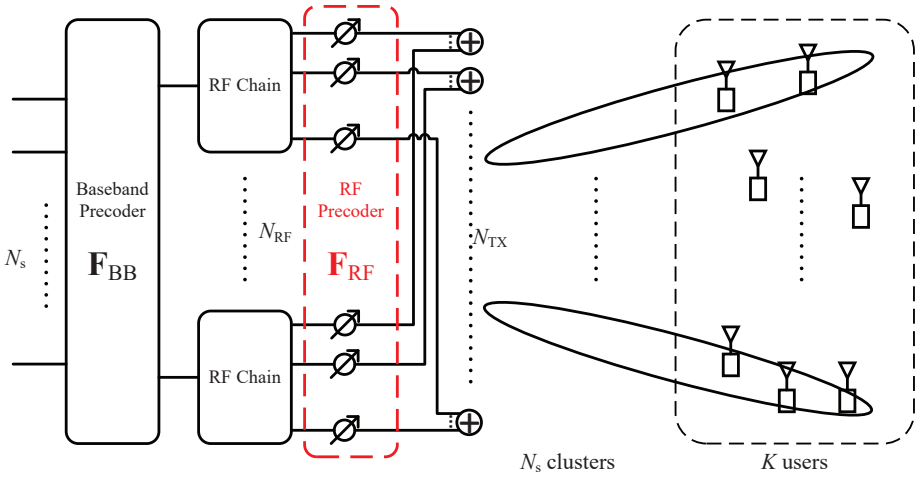


Fig. 1. Simplified hardware block diagram of mmWave transmitter with digital baseband precoding followed by constrained radio frequency precoding implemented using RF phase shifters.

noise according to optimal decoding order and SIC principle [32]–[34]. We have the following permutation of users in each cluster accordingly:

1) Users are sorted in the descending order of their interference plus noise values such that $I_{\pi_j(1)} \geq I_{\pi_j(2)} \geq \cdots \geq I_{\pi_j(L_j)}$.
2) If two users have the same interference plus noise value, i.e., $I_{\pi_j(l)} = I_{\pi_j(l')}$ for $l < l'$, then $\pi_j(l) < \pi_j(l')$.

For convenience, if user k is the l-th decoded user of cluster j, i.e., $\pi_j(l) = k$, we define $p_{j,l}$ as the power of user k, which is equivalent to p_k. Suppose that the decoding order of users in cluster j is $\hat{\boldsymbol{\pi}}_j$, thus the signal to interference plus noise ratio (SINR) of user k can be calculated based on the received signal in (1), which depends on the mmWave precoder and also the power allocation strategy $\boldsymbol{p}$:

$$\begin{aligned}
&\text{SINR}_k(\boldsymbol{p}, \mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}) \\
=&\frac{p_{j,\hat{\pi}_j^{(-1)}(k)} |\mathbf{h}_k^T \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j}|^2}{\left(\sum_{l'=\hat{\pi}_j^{(-1)}(k)+1}^{L_j} p_{j,l'} |\mathbf{h}_k^T \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j}|^2 \; + \right.} \\
&\left. \sum_{j' \neq j}^{N_s} \sum_{l=1}^{L_{j'}} p_{j',l} |\mathbf{h}_k^T \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j'}|^2 + \sigma^2 \right) \\
=&\frac{p_{j,k}}{\sum_{k'=\hat{\pi}_j^{(-1)}(k)+1}^{L_j} p_{j,k'} + I_k},
\end{aligned} \tag{5}$$

where I_k is given by (3) and we define $\hat{\pi}_j^{(-1)}(k)$ to denote the decoding order of user k in cluster j, according to $\hat{\boldsymbol{\pi}}_j$, such that $\hat{\pi}_j^{(-1)}(k) = l$ for $\hat{\pi}_j(l) = k$. The achievable data rate of user k assigned to stream j is thus given by:

$$R_k(\boldsymbol{p}, \mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}) = W \log_2(1 + \text{SINR}_k(\boldsymbol{p}, \mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}})), \tag{6}$$

where $k \in \mathcal{K}$ and W is the system bandwidth. Note that users in the same cluster communicate with the BS via only a single stream.

B. mmWave Channel

Due to the limited scattering characteristic of mmWave channels, a mmWave channel can be modeled as the summation of M scattering paths as follows [4]: The channel vector $\mathbf{h}_k$ between the BS and user k at time slot t can be expressed as:

$$\mathbf{h}_k(t) = \sqrt{N_{\text{TX}} D(d_k)} \sum_{m=1}^{M} \alpha_{k,m}(t) \mathbf{a}_{\text{BS}}^H(\phi_{k,m}(t)),$$

where M is the number of propagation paths that can be observed by each user, $D(\cdot)$ is used to denote the path loss function which is given by $D(d_k) = \eta d_k^{-\alpha}$, where $\eta = (\frac{c}{4\pi f_c})^2$ is the frequency-dependent factor with $c = 3 \times 10^8$ m/s and f_c is the carrier frequency, $\alpha_{k,m}(t)$ is the complex link gain of the m-th path, and $\phi_{k,m}(t) \in [-\pi, \pi]$ is the angle of departure (AoD) of the m-th path. By assuming a uniform linear array (ULA) [35] case, we use $\mathbf{a}_{\text{BS}}(\cdot)$ to denote the array response vector, which is expressible as:

$$\begin{aligned}
&\mathbf{a}_{\text{BS}}(\phi_{k,m}(t)) = \\
&\frac{1}{\sqrt{N_{\text{TX}}}} \left[1, e^{j\frac{2\pi d}{\lambda}\sin(\phi_{k,m}(t))}, \ldots, e^{j(N_{\text{TX}}-1)\frac{2\pi d}{\lambda}\sin(\phi_{k,m}(t))}\right]^T,
\end{aligned} \tag{7}$$

where λ is the signal wavelength and d is the minimum distance between antenna elements.

Since the channels will evolve over time, it is considered that the link gains have time correlation according to a Gauss-Markov random process [36] and the AoD of each path experiences time evolution similar to [37], [38] as follows:

$$\alpha_{k,m}(t) = \rho \cdot \alpha_{k,m}(t-1) + \sqrt{1-\rho^2} \Delta\alpha_{k,m}(t), \tag{8}$$

$$\phi_{k,m}(t) = \phi_{k,m}(t-1) + \Delta\phi_{k,m}(t), \tag{9}$$

where $\Delta\alpha_{k,m}(t) \sim \mathcal{CN}(0,1)$, $\rho \in [0,1]$ is the correlation coefficient, and $\Delta\phi_{k,m}(t) \sim \mathcal{N}(0, \sigma_u^2)$ with variance $\sigma_u^2 = (\frac{\pi}{360})^2$, for $k = 1, 2, \ldots, K$ and $m = 1, 2, \ldots, M$.

III. Precoding Algorithms for mmWave Systems

It has been shown that the hybrid precoding approach proposed in [7] and [8] can achieve similar performance to full digital precoding but at lower system cost and power consumption for MIMO mmWave schemes. In [39] and in [40], the hybrid precoding is designed according to the clustering results. In our proposed scheme, the beamforming users of each cluster are selected firstly to design the hybrid precoding. Then, the other users are matched to each cluster according to their channel correlation values with the designed precoder of each cluster. In this section, we introduce two exemplary designs of hybrid precoder: one based on singular value decomposition (SVD) and the other based on minimum mean-squared error (MMSE).

A. Spatially Sparse Precoding Design via OMP

A spatially sparse precoding algorithm based on orthogonal matching pursuit (OMP) is proposed in [8], which exploits the sparse scattering structure of mmWave channels. Algorithms based on OMP are designed to approximate optimal unconstrained precoders and combiners such that sub-optimal performance can be achieved with low-cost RF hardware. The precoder design problem is formulated as to minimize the distance between $\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}$ and the channel's optimal unconstrained precoder $\mathbf{F}_{\text{opt}}$, see (10) below, which can be solved by singular value decomposition (SVD) of the channel matrix. Define the channel's ordered SVD to be $\mathbf{H} = \mathbf{U}\sum\mathbf{V}^H$, where $\mathbf{U}$ is an $N_s \times \text{rank}(\mathbf{H})$ unitary matrix, $\sum$ is a $\text{rank}(\mathbf{H}) \times \text{rank}(\mathbf{H})$ diagonal matrix of singular values ranging in decreasing order, and $\mathbf{V}$ is a $N_{\text{TX}} \times \text{rank}(\mathbf{H})$ unitary matrix while $\mathbf{V}^H$ is the conjugate transpose of $\mathbf{V}$. The optimal unconstrained unitary precoder for $\mathbf{H}$ is thus given by $\mathbf{F}_{\text{opt}} = \mathbf{V}$. Given $\mathbf{F}_{\text{opt}}$, the precoder optimization problem can be stated as:

$$(\mathbf{F}_{\text{RF}}^{\text{opt}}, \mathbf{F}_{\text{BB}}^{\text{opt}}) = \underset{\mathbf{F}_{\text{RF}}, \mathbf{F}_{\text{BB}}}{\arg\min} \|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F, \tag{10}$$

subject to

$$\begin{aligned} &\mathbf{F}_{\text{RF}} \in \mathcal{F}_{\text{RF}}, \\ &\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F^2 = N_s, \end{aligned} \tag{11}$$

where $\mathcal{F}_{\text{RF}}$ is the set of $N_{\text{TX}} \times N_{\text{RF}}$ matrices which are feasible RF precoders, that is, each column of $\mathcal{F}_{\text{RF}}$ is necessarily equal to one of the following response vectors $\mathbf{a}_{\text{BS}}(\cdot)$'s. We denote them by matrix $\mathbf{A}_{\text{BS}}$ to represent for the convenience of coming discussions:

$$\begin{aligned} &\mathbf{A}_{\text{BS}} \triangleq \\ &\quad [\mathbf{a}_{\text{BS}}(\phi_{1,1}), \ldots, \mathbf{a}_{\text{BS}}(\phi_{1,M}), \mathbf{a}_{\text{BS}}(\phi_{2,1}), \ldots, \mathbf{a}_{\text{BS}}(\phi_{K,M})], \end{aligned} \tag{12}$$

where $\mathbf{a}_{\text{BS}}(\cdot)$ is defined in (7).

For completeness, we describe the procedure to compute the hybrid spatially sparse precoder in Algorithm 1. In each iteration of the algorithm, a column vector along which the "residual precoding matrix" $\mathbf{F}_{\text{res}}$ has the largest projection is selected from $\mathbf{A}_{\text{BS}}$, to be appended to the RF precoder matrix $\mathbf{F}_{\text{RF}}$, as described by line 3, 4 and 5. Note that $\mathbf{F}_{\text{RF}} = [\mathbf{F}_{\text{RF}}|\mathbf{A}_{\text{BS}}^{(l^*)}]$ in line 5 indicates that the column vector $\mathbf{A}_{\text{BS}}^{(l^*)}$ is appended to the matrix $\mathbf{F}_{\text{RF}}$. Then, the solution of $\mathbf{F}_{\text{BB}}$ is calculated (see line 6), and the contribution of the selected vector on $\mathbf{F}_{\text{RF}}$ is removed (see line 7). The procedure continues for N_{RF} iterations, until all beamforming vectors have been selected. Finally, the baseband precoder is normalized (see line 9).

Algorithm 1 Spatially Sparse Precoding Design via OMP

Input: $\mathbf{F}_{\text{opt}}$, $\mathbf{A}_{\text{BS}}$
Output: $\mathbf{F}_{\text{RF}}$, $\mathbf{F}_{\text{BB}}$
1: Initialization: $\mathbf{F}_{\text{RF}} =$ empty matrix, $\mathbf{F}_{\text{res}} = \mathbf{F}_{\text{opt}}$
2: **for** $i \leq N_{\text{RF}}$ **do**
3: $\quad \mathbf{\Phi} = \mathbf{A}_{\text{BS}}^H \mathbf{F}_{\text{res}}$
4: $\quad l^* = \text{argmax}_{l=1,\ldots,KL_j} (\mathbf{\Phi}\mathbf{\Phi}^H)_{l,l}$
5: $\quad \mathbf{F}_{\text{RF}} = [\mathbf{F}_{\text{RF}}|\mathbf{A}_{\text{BS}}^{(l^*)}]$
6: $\quad \mathbf{F}_{\text{BB}} = (\mathbf{F}_{\text{RF}}^H \mathbf{F}_{\text{RF}})^{-1} \mathbf{F}_{\text{RF}}^H \mathbf{F}_{\text{opt}}$
7: $\quad \mathbf{F}_{\text{res}} = \frac{\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}}{\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F}$
8: **end for**
9: $\mathbf{F}_{\text{BB}} = \sqrt{N_s} \frac{\mathbf{F}_{\text{BB}}}{\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F}$

B. Hybrid MMSE Precoding via OMP

Another hybrid precoder is the minimum mean-squared error (MMSE) precoder proposed in [41], which can also be easily implemented using an orthogonal matching pursuit algorithm. The difference between the MMSE precoder and aforementioned SVD precoder is that the MMSE precoder jointly designs $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$ for minimizing the sum mean-square-error (MSE) of all streams and having the received signal as close as possible to the original signal, i.e., minimizing $\mathbb{E}[||\boldsymbol{s} - \hat{\boldsymbol{s}}||^2]$, where $\boldsymbol{s}$ is transmitted signal while $\hat{\boldsymbol{s}}$ is the received signal. The procedure to find the hybrid MMSE precoder is presented in Algorithm 2 with inputs $\mathbf{A}_{\text{BS}}$ and $\widetilde{\mathbf{H}}$, where $\widetilde{\mathbf{H}} \triangleq [\widetilde{\mathbf{h}}_1, \ldots, \widetilde{\mathbf{h}}_{N_s}]$ and $\widetilde{\mathbf{h}}_j, j = 1, \ldots, N_s$, denotes the highest channel gain for stream j.

The major difference between MMSE precoding and SVD precoding lies in the method of designing the baseband precoder $\mathbf{F}_{\text{BB}}$. By line 6 of Algorithm 1, the baseband precoder $\mathbf{F}_{\text{BB}}$ is calculated as the solution to the unconstrained least-square minimization of $\|\mathbf{F}_{\text{opt}} - \mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\|_F$. While in line 6 of Algorithm 2, the baseband precoder is derived as a closed-form MMSE solution. As is analyzed in [41], the two precoding designs yield the same total complexity order.

IV. Proportional Fair Scheduling Metric

In the following, we consider proportional fairness metric to strike a balance between cell throughput and user service fairness.

A. Proportional Fairness for Single User Scheduling

In MISO-NOMA system, more than one user can be scheduled for transmissions simultaneously. Always serving the users with the best channel conditions can maximize the total throughput, but the fairness among users would be poor unless their channels are symmetric and have the same fading statistics. In reality, the users' channels are not the same and

Algorithm 2 Hybrid MMSE Precoding via OMP

Input: $\widetilde{\mathbf{H}}$, $\mathbf{A}_{\text{BS}}$
Output: $\mathbf{F}_{\text{RF}}$, $\mathbf{F}_{\text{BB}}$
1: Initialization: $\mathbf{F}_{\text{RF}}$ = empty matrix, $\mathbf{V}_{\text{res}} = \mathbf{I}_{N_s}$, where $\mathbf{I}_{N_s}$ is an identity matrix
2: **for** $i \leq N_{\text{RF}}$ **do**
3: $\quad \mathbf{\Phi} = \mathbf{A}_{\text{BS}}^H \widetilde{\mathbf{H}}^H \mathbf{V}_{\text{res}}$
4: $\quad l^* = \text{argmax}_{l=1,\ldots,KL_j} (\mathbf{\Phi}\mathbf{\Phi}^H)_{l,l}$
5: $\quad \mathbf{F}_{\text{RF}} = [\mathbf{F}_{\text{RF}} | \mathbf{A}_{\text{BS}}^{(l^*)}]$
6: $\quad \mathbf{V}_{\text{BB}} = (\mathbf{F}_{\text{RF}}^H \widetilde{\mathbf{H}}^H \widetilde{\mathbf{H}} \mathbf{F}_{\text{RF}} + \sigma^2 \mathbf{F}_{\text{RF}}^H \mathbf{F}_{\text{RF}})^{-1} \mathbf{F}_{\text{RF}}^H \widetilde{\mathbf{H}}^H$
7: $\quad \mathbf{V}_{\text{res}} = \frac{\mathbf{I}_{N_s} - \mathbf{F}_{\text{RF}} \mathbf{V}_{\text{BB}}}{||\mathbf{I}_{N_s} - \mathbf{F}_{\text{RF}} \mathbf{V}_{\text{BB}}||_F}$
8: **end for**
9: $\gamma = \frac{||\mathbf{F}_{\text{RF}} \mathbf{V}_{\text{BB}}||_F^2}{N_s}$
10: Normalize $\mathbf{F}_{\text{BB}} = \sqrt{1/\gamma} \mathbf{V}_{\text{BB}}$

a sum-rate maximization scheduling may degrade the service quality of cell edge users. A proper time scheduling strategy is relevant and can significantly affect the system throughput and user fairness.

In proportional fair (PF) scheduling, the target is to maximize the long-term average throughput of users. In delay-tolerant packet data systems, multi-user diversity can be achieved by tracking the channel fluctuations of the users and scheduling service time to users whose instantaneous channel condition is relatively good compared to the time average in the past. This strategy has been widely used in wireless communications as it balances system sum rate and user fairness and can be defined as follows [42]:

$$k^* = \underset{k}{\text{argmax}} \frac{R_k(t)}{T_k(t)},$$

where $R_k(t)$ is the instantaneous throughput of user k at time t and $T_k(t)$ is its average throughput achieved at time t, which is updated according to:

$$T_k(t+1) = \begin{cases} (1 - \frac{1}{t_c}) T_k(t) + \frac{1}{t_c} R_k(t), & \text{if } k = k^*, \\ (1 - \frac{1}{t_c}) T_k(t), & \text{if } k \neq k^*, \end{cases}$$

where t_c is the duration of considered past time window or the throughput averaging parameter. Note that t_c is considered as a constant in the setting and does not change with time slot t. It is known that the above PF scheduling is also equivalent to maximizing the sum of the logarithms of the user average throughput. It can be observed that the larger t_c, the less important the fairness constraint, and thus the above PF scheduling tends to a maximum throughput scheduling when $t_c \to \infty$. Note that in the case of large t_c, a long delay may appear between the successive transmissions of a user.

B. Proportional Fairness for Multiple User Scheduling

Recall that $\mathcal{U}(t)$ is the set of scheduled users over each time slot t. The average throughput of user k is then updated by:

$$T_k(t+1) = \begin{cases} (1 - \frac{1}{t_c}) T_k(t) + \frac{1}{t_c} R_k(t), & \text{if } k \in \mathcal{U}(t), \\ (1 - \frac{1}{t_c}) T_k(t), & \text{if } k \notin \mathcal{U}(t). \end{cases}$$

Proposition 1. *The optimal power allocation strategy $\boldsymbol{p}^*$ that maximizes $\sum_{k \in \mathcal{U}(t)} \frac{R_k(t)}{T_k(t)}$ for long-term time average is proportional fair (PF), where $\mathcal{U}(t)$ represents the set of scheduled users for each time slot.*

Proof. Suppose that the scheduled users interfere each other, it is known [43], [44] that the following power allocation policy maximizes the PF metric:

$$\underset{\boldsymbol{p}}{\arg\max} \prod_{k \in \mathcal{U}(t)} (1 + \frac{R_k(t)}{(t_c - 1) T_k(t)}). \tag{13}$$

See also for example [23], [45], the use of (13) for NOMA user scheduling problems. By expansion, we can rewrite (13) as:

$$\prod_{k \in \mathcal{U}(t)} (1 + \frac{R_k(t)}{(t_c - 1) T_k(t)}) = 1 + \frac{1}{t_c - 1} \left(\sum_{k \in \mathcal{U}(t)} \frac{R_k(t)}{T_k(t)} \right) + (\frac{1}{t_c - 1})^2 \left(\sum_{\substack{\{k_1, k_2\} \in \mathcal{U}(t), \\ k_1 \neq k_2}} \frac{R_{k_1}(t)}{T_{k_1}(t)} \frac{R_{k_2}(t)}{T_{k_2}(t)} \right) + \cdots .$$

When t_c tends to infinity, (13) is asymptotically equivalent to:

$$\underset{\boldsymbol{p}}{\arg\max} \left(1 + \frac{1}{t_c - 1} \sum_{k \in \mathcal{U}(t)} \frac{R_k(t)}{T_k(t)} \right). \tag{14}$$

Hence, for $t_c \gg 1$, maximizing (14) is PF, which is indeed equivalent to maximizing the value of $\sum_{k \in \mathcal{U}(t)} \frac{R_k(t)}{T_k(t)}$ for each time slot. This completes the proof. □

By the result of Proposition 1, the PF scheduling problem in (13) can be approximated and formalized as the following weighted sum-rate (WSR) maximization problem to address:

$$\underset{b_{j,k}, p_k}{\arg\max} \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} b_{j,k} \frac{R_k}{T_k} \tag{15}$$

$$\boldsymbol{s.t.} \quad C1: b_{j,k} \in \{0, 1\}, \forall j \in \mathcal{J}, \forall k \in \mathcal{K}, \tag{15a}$$

$$C2: \sum_{k \in \mathcal{K}} b_{j,k} \leq L_j, \forall j \in \mathcal{J}, \tag{15b}$$

$$C3: \sum_{j \in \mathcal{J}} b_{j,k} \leq 1, \forall k \in \mathcal{K}, \tag{15c}$$

$$C4: \sum_{k=1}^{K} p_k \leq P_{\max}, \tag{15d}$$

$$C5: p_k \geq 0, \forall k \in \mathcal{K}, \tag{15e}$$

where we omit t for the sake of notational simplicity and the binary variable $b_{j,k}$ decides whether user k is scheduled in cluster j. Constraints $C2$ and $C3$ ensure that at most

L_j users can be scheduled in each cluster and each user receives a single beam at most. Constraint $C4$ represents the limited transmit power $P_{\max}$ available at the BS. Constraint $C5$ refers to the non-negativity of transmit power. It can be observed that (15) is a mixed integer and non-convex optimization problem because of constraints $C1$–$C3$ (where $b_{j,k}$'s are binary decision variables) and the interference power term in the objective function.

V. User Clustering and Scheduling

In order to reduce the computational complexity, we decouple the optimization problem (15) into two sub-problems such that the user clustering and power allocation are performed separately and by iterative algorithms. In the first step, the BF matrix is constructed by OMP, based on SVD or MMSE precoding. Then, the users are clustered according to their channel conditions using the same principle of the Channel Condition based User Clustering (CCUC) method in [26], since CCUC can be conducted independently of the power control part and has low computational complexity. Finally, given the user clustering, the power allocation problem is solved by using geometric programming (GP) techniques [46].

A. Heuristic User Clustering Method

When the hybrid beamforming precoder is obtained, the user selection and clustering is computed based on the channel conditions such that users that are allocated to cluster j are highly correlated with the hybrid precoder for cluster j, i.e., $\mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB},j}$. The CCUC algorithm proposed in [26] adopts zero forcing for constructing the precoding matrix, thus the channel condition metric is the correlation value between user channels. However, in this work, since the precoding matrix is constructed based on OMP method, thus, the channel condition metric will be the correlation value between user channels and the precoding matrix of each BF user. The maximum number of users per cluster is limited to L_j, as a constraint due to receiver decoding complexity and error propagation issues in practical SIC implementations [21].

The proposed heuristic user clustering is summarized in Algorithm 3. For the simplicity of notation, we use $\mathcal{U}_j$ to denote the user set of cluster j, while $\mathcal{J}_{\text{full}}$ denotes the set of clusters that are full (i.e., any cluster j which has already L_j users). The correlation value between each user k and the hybrid precoder of each cluster j is determined and there are in total $K \times N_s$ correlation values. At each iteration (see line 2–8), a user k^* is assigned to a cluster j^* according to the channel quality based metric (see line 3) such that k^* and j^* have the highest correlation value given that cluster j^* is not full. The algorithm terminates when either all the users are selected or all the clusters are full.

Algorithm 3 The proposed user clustering algorithm

Input: $\mathbf{F}_{\text{RF}}$, $\mathbf{F}_{\text{BB}}$, and $\mathbf{H}$
Output: $\mathcal{U}_j$ for $j \in \mathcal{J}$
1: **Initialization:** $\mathcal{U}_j = \varnothing$ for $j \in \mathcal{J}$, $\mathcal{J}_{\text{full}} = \varnothing$, $\mathcal{K}' \triangleq \mathcal{U}_1 \bigcup \ldots \bigcup \mathcal{U}_{N_s}$
2: **while** $\mathcal{K}' \neq \mathcal{K}$ and $\mathcal{J}_{\text{full}} \neq \mathcal{J}$ **do**
3: Match a new user to a cluster such that they have the highest correlation value, i.e., $(k^*, j^*) \triangleq \underset{(k,j)\,:\,k\in\mathcal{K}\setminus\mathcal{K}',\, j\in\mathcal{J}\setminus\mathcal{J}_{\text{full}}}{\arg\max} \frac{|\mathbf{h}_k^T \mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB},j}|}{\|\mathbf{h}_k\|\cdot\|\mathbf{F}_{\text{RF}}\mathbf{f}_{\text{BB},j}\|}$
4: $\mathcal{U}_{j^*} \leftarrow \mathcal{U}_{j^*} \bigcup \{k^*\}$
5: **if** $|\mathcal{U}_{j^*}| = L_j$ **then**
6: $\mathcal{J}_{\text{full}} \leftarrow \mathcal{J}_{\text{full}} \bigcup j^*$
7: **end if**
8: **end while**

B. Geometric Programming for Power Allocation

Suppose that the user clustering strategy is settled and the mmWave precoders $\mathbf{F}_{\text{RF}}$ and $\mathbf{F}_{\text{BB}}$ are given based on the channel conditions of BF users, we now have the power allocation problem below according to (15):

$$\underset{\boldsymbol{p}}{\arg\max} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{U}_j} \frac{R_k}{T_k} \tag{16}$$

$$\boldsymbol{s.t.} \quad C1: \sum_{k\in\mathcal{U}} p_k \leq P_{\max}, \tag{16a}$$

$$C2: p_k \geq 0, k \in \mathcal{U}. \tag{16b}$$

Note that $b_{j,k}$'s in (15) are now removed, since the user clustering has already been done previously. Problem (16) can be converted to the linear fractional programming problem and then solved optimally by the monotonic optimization approach using the polyblock algorithm [47]. The fractional programming problem can be derived as follows:

$$\underset{\boldsymbol{p}}{\arg\max} \prod_{j\in\mathcal{J}} \prod_{k\in\mathcal{U}_j} \left[\frac{f_k(\boldsymbol{p})}{g_k(\boldsymbol{p})}\right]^{\frac{1}{T_k}} \tag{17}$$

$$\boldsymbol{s.t.} \quad C1: \sum_{k\in\mathcal{U}} p_k \leq P_{\max}, \tag{17a}$$

$$C2: p_k \geq 0, k \in \mathcal{U}, \tag{17b}$$

where $f_k(\boldsymbol{p}) = p_{j,k} + \sum_{k'=\hat{\pi}_j^{(-1)}(k)+1}^{L_j} p_{j,k'} + I_k$ and $g_k(\boldsymbol{p}) = \sum_{k'=\hat{\pi}_j^{(-1)}(k)+1}^{L_j} p_{j,k'} + I_k$. The problem dimension of the fractional programming is denoted by $D = \sum_{j\in\mathcal{J}} |\mathcal{U}_j|$. Note that the polyblock algorithm has worst-case exponential-time complexity in D, which makes it not practical when $D \geq 10$ [48]–[50]. Moreover, it requires a large number of iterations when $D \geq 5$. Therefore, we propose a low-complexity iterative algorithm by geometric programming (GP) techniques to solve the power allocation problem. Note that the problem (16) can be re-written as follows.

Proposition 2. *Problem* (16) *can be converted into the following optimization problem with* $\boldsymbol{R} \triangleq (R_k)_{k\in\mathcal{U}}$ *as the*

optimization variable.

$$\underset{\boldsymbol{R}}{\operatorname{argmin}} \log e^{-\sum_{k\in\mathcal{U}} \frac{R_k}{T_k}} \tag{18}$$

$$\boldsymbol{s.t.}$$

$$C1: \log e^{-R_k} \leq 0, \forall k \in \mathcal{U}, \tag{18a}$$

$$C2: \sum_{j=1}^{N_s}\sum_{l=1}^{L_j} \frac{(I_{\pi_j(l)} - I_{\pi_j(l+1)})}{P_{\max} + \sum_{j'=1}^{N_s} I_{\pi_{j'}(1)}} e^{\sum_{l'=1}^{l} R_{\pi_j(l')} \frac{\ln 2}{W}} \leq 1. \tag{18b}$$

Proof. See Appendix A. □

It can be seen that the objective function of (18) is linear and the constraint $C1$ is convex in $\boldsymbol{R}$. However, the constraint $C2$ is not convex, since the interference term $I_{\pi_j(l)}$, $j = 1, \ldots, N_s$ and $l = 1, \ldots, L$, depends on the power allocation strategy and is thus a function of the data rates of the other users. According to geometric programming (GP), the objective function of a problem needs to be linear while the constraints should be convex. As a consequence, we design an iterative power control algorithm. At each iteration, the interference terms are assumed to be fixed values, independent of each user's data rate, such that the constraint $C2$ becomes convex in $\boldsymbol{R}$, and the problem (18) can be solved by GP method. At each iteration, the interference of each user is calculated based on the power allocation strategy of the former iteration and the data rates of the other users. Here, the GP problem can be solved by using interior point methods, which is a very efficient algorithm, with the worst-case polynomial-time complexity [46], [51]. Then, given the data rate, the power allocation strategy can be determined and be used to calculate the interference in the next iteration.

Let $I_k^{(n)}$ and $R_k^{(n)}$ be the normalized interference plus noise value and the achievable data rate of user k at iteration n, respectively. In addition, we define vector $\boldsymbol{R}^{(n)} \triangleq (R_k^{(n)})_{k\in\mathcal{U}}$ to represent the data rates of users at the n-th iteration. Note that the iterations are in the same time slot, and thus the average throughput of each user T_k does not change. We solve the following problem at each iteration for problem (18):

$$\underset{\boldsymbol{R}^{(n)}}{\operatorname{argmin}} \log e^{-\sum_{k\in\mathcal{U}} \frac{R_k^{(n)}}{T_k}} \tag{19}$$

$$\boldsymbol{s.t.}$$

$$C1: \log e^{-R_k^{(n)}} \leq 0, \forall k \in \mathcal{U}, \tag{19a}$$

$$C2: \log \sum_{j=1}^{N_s}\sum_{l=1}^{L_j} \frac{(I_{\pi_j(l)}^{(n-1)} - I_{\pi_j(l+1)}^{(n-1)})}{P_{\max} + \sum_{j'=1}^{N_s} I_{\pi_{j'}(1)}^{(n-1)}} e^{\sum_{l'=1}^{l} R_{\pi_j(l')}^{(n)} \frac{\ln 2}{W}} \leq 0. \tag{19b}$$

At each iteration n, since $I_{\pi_j(l)}^{(n-1)}$'s, for $j = 1, \ldots, N_s$ and $l = 1, \ldots, L_j$, are given fixed values, the problem (19) is exactly in the form of GP and we can solve it by using interior point methods at each iteration.

Theorem 3. *Given the data rates of all users $\boldsymbol{R}$, the power consumption of cluster $j, j \in \mathcal{J}$, with the decoding order $\boldsymbol{\pi}_j$ can be calculated as:*

$$q_j(\boldsymbol{\pi}_j, \boldsymbol{q}_{-j}) = \sum_{l=1}^{L_j} \chi_l(\boldsymbol{\pi}_j) I_{\hat{\pi}_j(l)}, \tag{20}$$

where

$$\chi_l(\boldsymbol{\pi}_j) \triangleq \Big[\prod_{l'<l} (\gamma_{\pi_j(l')} + 1)\Big] \gamma_{\hat{\pi}_j(l)}, \tag{21}$$

and

$$\gamma_k \triangleq 2^{\frac{R_k}{W}} - 1, k \in \mathcal{K}. \tag{22}$$

Proof. See Appendix B. □

We describe our iterative power allocation algorithm for solving (16) by Algorithm 4. In each iteration, the interference terms are regarded as fixed values such that the power allocation problem can be converted as a GP problem with data rate vector as the optimization variable. We solve the GP problem in the first place to obtain the data rate allocation strategy $\boldsymbol{R}^{(n)}(t)$. Then, the power vector $\boldsymbol{p}^{(n)}$ can be uniquely determined accordingly. We conduct exhaustive computer simulations and find that the algorithm always converges after a limited number of iterations and the final data rate allocation strategy $\boldsymbol{R}^*$ can be obtained.

VI. Simulation Results

In this section, we evaluate the performance of the proposed user clustering and power allocation strategy under different scheduling metrics and precoding schemes. In OMA scheme, only one user is assigned to each cluster. However, in NOMA scheme, the maximum number of users per cluster can be larger than one ($L_j \geq 1$). We will consider both proportional fairness (PF) and sum rate (SR) maximization metrics for comparison. The SR maximization metric is adopted by setting $T_k = 1, \forall k \in \mathcal{K}$ in (15). Note that we implement the SVD and MMSE precoding schemes under the assumption of perfect CSI at the BS.

Following 3GPP specifications [52], we set the power limit of BS, $P_{\max}$, as 46 dBm and the noise spectral density at -174 dBm/Hz. The carrier frequency of the mmWave system is set at 28 GHz while the bandwidth is equal to 100 MHz. We simulate for each network topology a duration of 10^4 time slots, in which the BS is located at the center of a cell and the users are uniformly generated within the cell with channel conditions set according to 3GPP and standard mmWave radio propagation model (see Section II). The system parameters are summarized in Table I. It is worth noting that the plots of simulation results are obtained by simulating 10^4 instances of randomly generated network realization while each point is the obtained average result.

First, we evaluate the sum rate performance of different precoding, user scheduling and power allocation schemes. We set $L_j = L, \forall j$, for simplicity. Fig. 2(a) and Fig. 2(b) illustrate the sum spectral efficiency against the number of users under different schemes. Here, we assume that each cluster contains two users (i.e., $L_j = L = 2, \forall j$). It can be seen that the sum spectral efficiency increases generally with the number of users due to multi-user diversity. From Fig. 2(a), we observe that NOMA-SR-MMSE, which uses SR maximization as the

Algorithm 4 The iterative power allocation with GP for one time slot

Input: $\mathcal{U}_j$ for $j \in \mathcal{J}$, T_k for $k \in \mathcal{U}$, and $I_k^{(n-1)}$ for $k \in \mathcal{U}$
Output: $R_k^{(n)}$ for $k \in \mathcal{U}$, $q_j^{(n)}$ for $j \in \mathcal{J}$, and $\boldsymbol{p}_j^{(n)}$ for $j \in \mathcal{J}$

Solve GP problem (19) to obtain the optimal data rates of scheduled users R_k for $k \in \mathcal{U}$.
for $j := 1, 2, \ldots, N_s$, **do**
  Calculate the optimal decoding order, $\boldsymbol{\pi}_j$, based on (4), with inputs $I_k^{(n-1)}$ for $k \in \mathcal{U}$
  for $l := 1, 2, \ldots, L_j$, **do**
    Calculate $\gamma_{j,\pi_j(l)}$ by (22), i.e.,

$$\gamma_{j,\pi_j(l)} := 2^{\frac{r_{j,\pi_j(l)}}{W}} - 1$$

    Calculate χ_l by (21), i.e.,

$$\chi_l := \Big[\prod_{l'<l}(\gamma_{\pi_j(l')}+1)\Big]\gamma_{\pi_j(l),m}$$

  end for
  Determine the least required transmit power of cluster j as follows:

$$q_j^{(n)} := \sum_{l=1}^{L_j} \chi_l I_{\pi_j(l)}^{(n-1)}$$

  Determine the power allocation for user $\pi_j(L_j)$, which is quoted below:

$$p_{\pi_j(L)}^{(n)} := \gamma_{\pi_j(L_j)} I_{\pi_j(L)}^{(n)}$$

  for $l := L_j - 1, L_j - 2, \ldots, 1$, **do**
    Calculate the allocated power to user $\pi_j(l)$ as follows:

$$p_{\pi_j(l)}^{(n)} := \gamma_{\pi_j(l)} \left(\sum_{l'>l} p_{\pi_j(l')}^{(n)} + I_{\pi_j(l)}^{(n-1)} \right)$$

  end for
end for
return $R_k^{(n)}$ for $k \in \mathcal{U}$, $q_j^{(n)}$ for $j \in \mathcal{J}$, and $\boldsymbol{p}_j^{(n)}$ for $j \in \mathcal{J}$

user scheduling metric, achieves the highest spectral efficiency. On the other hand, NOMA-PF-MMSE scheme, which uses PF maximization as the user scheduling metric, obtains a slightly lower spectral efficiency compared to NOMA-SR-MMSE. It is because NOMA-PF-MMSE scheme takes user fairness into concern and some spectral efficiency performance are sacrificed for the tradeoff. We also simulate for OMA schemes. It is shown in Fig. 2(a) that NOMA-PF-MMSE outperforms OMA-PF-MMSE generally. OMA-SR-MMSE has the lowest spectral efficiency performance compared to other MMSE schemes and does not increase a lot along with the increase of user number, it is because the maximum scheduled user number of OMA is limited as the number of antennas and less user diversity can be achieved. As expected, we can see from Fig. 2(b), the spectral efficiency of NOMA schemes is clearly higher than that of OMA, similar to that in Fig. 2(a). Another observation is that MMSE precoder has higher spectral efficiency than SVD precoder generally. It is because MMSE is able to form very narrow beams by RF antenna arrays with high number of transmit and receive antennas. It can also be seen in Fig. 2(b) that the spectral efficiency of all the schemes for 16 users per cell is higher than that of 20 users per cell, and the curves for NOMA-PF-SVD and the OMA-PF-SVD are not too smooth. The reason is that there is tradeoff between the spectral efficiency and user fairness, which may sacrifice the spectral efficiency for the user fairness.

TABLE II
SIMULATION PARAMETERS

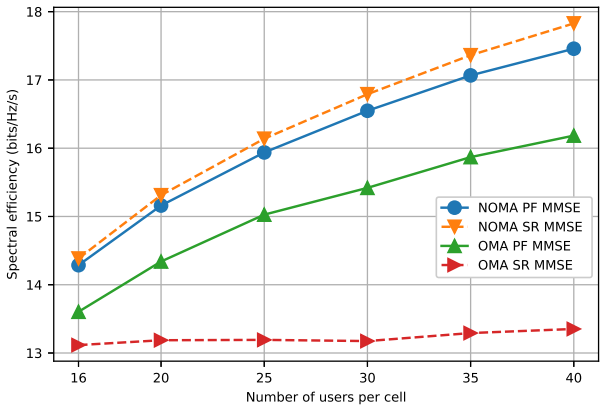
Parameters	Values
Cell radius (m)	10
Channel correlation coefficient, ρ	0.9924
Noise power spectral density	-174 dBm/Hz
Transmit power budget of a BS, $P_{\max}$	46 dBm
Throughput calculation	Shannon's capacity formula
Carrier frequency, f_c	28 GHz
System bandwidth, W	100 MHz
Maximum number of iterations	100
Number of transmit antennas N_{TX}	32
Number of RF chains N_{RF}	8
Number of data streams N_s	8
No. of simulation instances for each case	10^4

To evaluate the user fairness, we use the Jain's fairness index [53], which is given by:

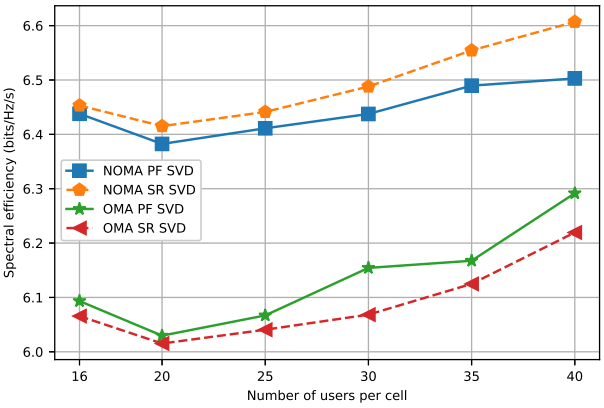
$$\text{JFI} = \frac{\left(\sum_{k=1}^{K} \bar{R}_k\right)^2}{K \sum_{k=1}^{K} \bar{R}_k^2}, \tag{23}$$

where we have $\bar{R}_k$ to denote the average throughput of user k during the simulation period while the throughput of user is measured over a sliding window of 30 slots. Fig. 3 shows the user fairness against the number of users under the different schemes. It can be seen that the user fairness decreases with the increase of the number of users per cell. This is because the users compete to access and the sum transmitted power is limited, which leads to the unfairness among users. As expected, NOMA-SR-MMSE scheme has the lowest fairness since its target is to maximize the sum rate instead of user fairness. On the other hand, the NOMA-PF-SVD scheme can achieve the highest user fairness at the expense of its low spectral efficiency (see Fig. 2(b)). Further, it can also be seen that with a same precoding strategy, NOMA can achieve higher user fairness than OMA. Taking into account both the sum-rate and user fairness performance, we see that overall NOMA-PF-MMSE can achieve a good balance between the spectral efficiency and user fairness, and is the most interesting scheme.

Consider that the maximum number of users per cluster is at $L = 2$ and the number of data streams is at $N_s = 8$, we assume 16 users to be clustered and scheduled for transmission. It should be noted that the actual number of users that are allocated with transmit power in the optimization solution can be less than or equal to 16 since some users could be allocated with zero power. Let $K^+(t)$ be the number of users that are allocated with positive power in time slot t. Fig. 4 plots the average value of $K^+(t)$ over t, say the average number



(a) MMSE precoder



(b) SVD precoder

Fig. 2. Sum spectral efficiency vs. the number of users per cell, for $L = 2$.

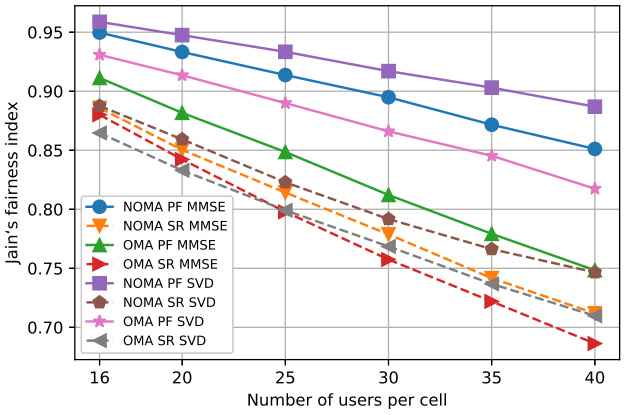


Fig. 3. Jain's fairness index vs. the number of users per cell, for $L = 2$.

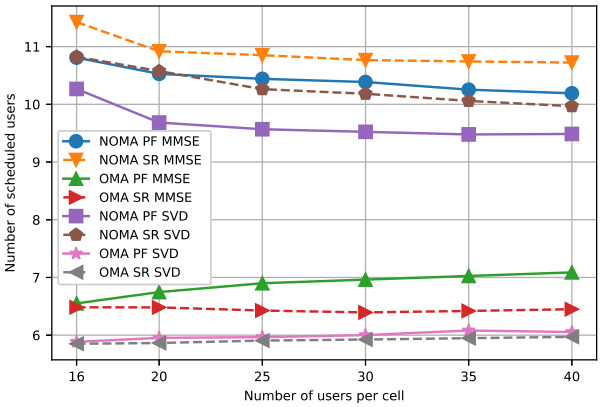


Fig. 4. The average value of $K^+(t)$ over t, say the average number of users allocated with positive power, vs. the number of users per cell, for $L = 2$.

of scheduled users, against the total number of users in the system, where $L = 2$, under the different schemes. It can be seen that NOMA-SR-MMSE scheme can accommodate the highest number of scheduled users. NOMA-PF-MMSE has very close performance compared to NOMA-SR-MMSE. On the other hand, the OMA scheme has the lowest number of scheduled users on average. It is because only N_s users are selected in the user clustering procedure and then considered during power allocation.

Moreover, we evaluate the performance of NOMA-PF-MMSE with different number of antennas, N_{TX}, and RF chains, N_{RF}, respectively. The number of users is fixed to be $K = 40$. Fig. 5 shows the sum spectral efficiency and Jain's fairness index performance of NOMA-PF-MMSE versus the number of antennas. We can see that the spectral efficiency grows with the number of antennas and the curve for Jain's fairness index is flat since a larger number of antennas can provide a higher degree of freedom. Fig. 6 shows the performance of NOMA-PF-MMSE versus the number of RF chains. It can be seen from Fig. 6 that the spectral efficiency decreases slowly with the number of RF chains and the Jain's fairness index remains stable. The reason is that when the number of RF chains increases, it will bring some interference.

We also investigate the influence of the maximum number of users per cluster, denoted by L, on the system's performance. We consider two strategies of setting L, namely *fixed-L* strategy and *dynamic-L* strategy, to compare. In *fixed-L* strategy, the maximum number of users of each cluster is fixed and equal to L (such that the number of scheduled users in each time slot is equal to N_sL), whereas in *dynamic-L* strategy, all users are clustered during the user clustering step and $L = \lceil K/N_s \rceil$, where $\lceil \cdot \rceil$ represents rounding up to an integer. Note that when $L = 1$, it is equivalent to OMA, where each cluster contains only one user.

Fig. 7 and Fig. 8 show the sum spectral efficiency and user fairness performance for different L, respectively, considering NOMA-MMSE-PF scheme especially. It can be seen that both the spectral efficiency and user fairness are improved with the increase of L under *fixed-L* schemes. On the other hand, the *dynamic-L* scheme indeed performs better than the *fixed-L* schemes. Fig. 9 shows the average number of scheduled users per time slot increases with the number of users per

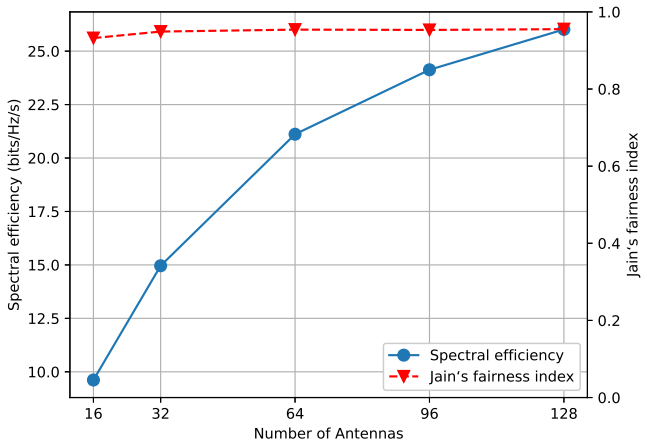


Fig. 5. Sum spectral efficiency and Jain's fairness index of NOMA-PF-MMSE vs. the number of antennas ($K = 40$).

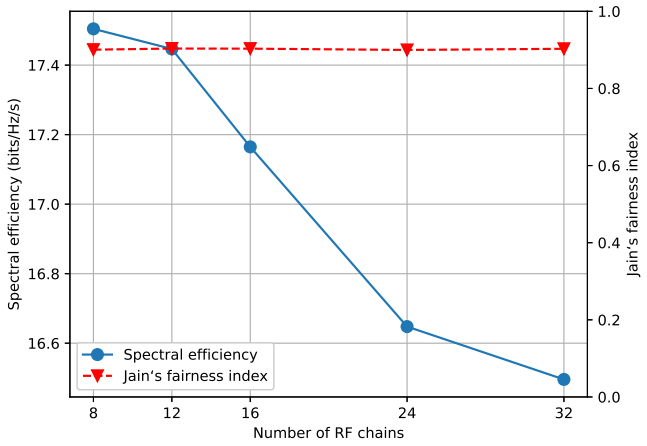


Fig. 6. Sum spectral efficiency and Jain's fairness index of NOMA-PF-MMSE vs. the number of RF chains ($K = 40$).

cell. We can see that *dynamic-L* scheme outperforms the *fixed-L* schemes and can be considered as an upper bound. Note that *dynamic-L* scheme does not discard users during the user clustering step and let the power allocation optimization determine during the user scheduling. Thus, *dynamic-L* scheme is more adaptive to different user density in a system. However, it should be noted that the computational complexity of SIC receiver grows with L and in practice there exists error propagation during SIC [21].

In Fig. 8, we can see that the user fairness performance is not monotone and it first increases and then decreases. This is because the user diversity is low when K is small (e.g., when $K = 8$, all of the users are served simultaneously by 8 beams) such that the channels of some users can be highly correlated, which would result in large inter-beam interference and poor achievable data rates at the users. Since the optimization targets to maximize the user fairness in the system, the solution will arrive to a result where a few users are served. This situation occurs more frequently when K is small and substantially degrade the system's overall performance and also the user fairness. The user fairness increase from $K = 8$ to $K = 11$. When K is larger, one can expect higher user diversity in the system. However, it should be noted that when the number of users grows, there would be an increasing number of users of low data rate although the PF scheduler will try to maintain user fairness, it would not support high throughput to users under poor channel conditions, due to the limited power and the interference limited system, resulting to lower service fairness in the system (from $K > 11$).

Furthermore, we compare the performance of the GP scheme with that of the polyblock solution. In our former simulation, the number of data streams $N_s = 8$ and $L = 2$, thus the problem dimension is $D = \sum_{j \in \mathcal{J}} |\mathcal{U}_j| = N_s \times L = 16$. For the comparison of the GP and polyblock algorithms and their practical implementation, we set the number of data streams $N_s = 3$ and $L = 2$, and thus $D = N_s \times L = 6$. We then compare the performance of the two algorithms under the NOMA-PF-MMSE scheme. Fig. 10 shows the spectral efficiency performance versus the maximum number of iterations of the polyblock algorithm. The number of users is 20. Since the polyblock algorithm will take a very long (computation) time for 10,000 time slots, each point is obtained by taking average of the sum rate of each time slot in 100 time slots for one randomly generated network realization. The red curve represents the spectral efficiency of the GP algorithm. We can see from Fig. 10 that when the number of iterations is small, the performance of the polyblock algorithm is worse than that of the GP algorithm. The spectral efficiency performance of the polyblock algorithm increases with the maximum number of iterations and the curve becomes flat when the number of iterations is larger than 400. When the number of maximum iterations is 500, the GP algorithm can still achieve 85.1% of the spectral efficiency of the polyblock algorithm. Fig. 11 shows the running time (in log scale) versus the maximum number of iterations of polyblock. It is shown that the running time of the polyblock algorithm increases very fast with the maximum number of iterations, which is much longer than that of the GP algorithm.

VII. Conclusion

In this paper, we study for MISO-NOMA user clustering, scheduling, and power allocation joint optimization in downlink mmWave communications. The mixed integer non-convex optimization problem is solved in a two-step scheme: a heuristic user clustering and precoding step, and an iterative GP power allocation step. Since the propagation of mmWave is highly directional with severe path loss, different hybrid precoding strategies are investigated and users are grouped into clusters for superposed transmissions. A heuristic user clustering algorithm is proposed, which takes the inter-correlation value between channels as criteria and is applicable for various number of users per cluster. To strikes a balance between the spectral efficiency and user fairness, we adopt PF metric for service scheduling and power allocation. One can see that the resource allocation problem can be expressed as a

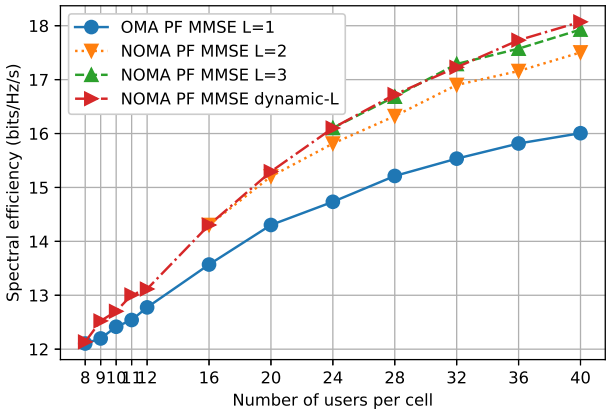


Fig. 7. Sum spectral efficiency vs. the number of users per cell.

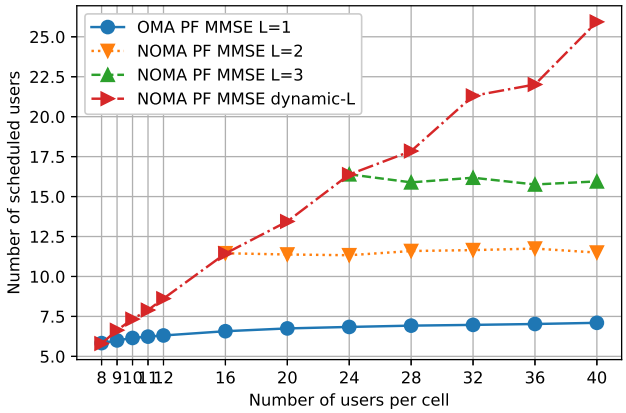


Fig. 9. The average value of $K^+(t)$ over t, say the average number of users allocated with positive power vs. the number of users per cell K.

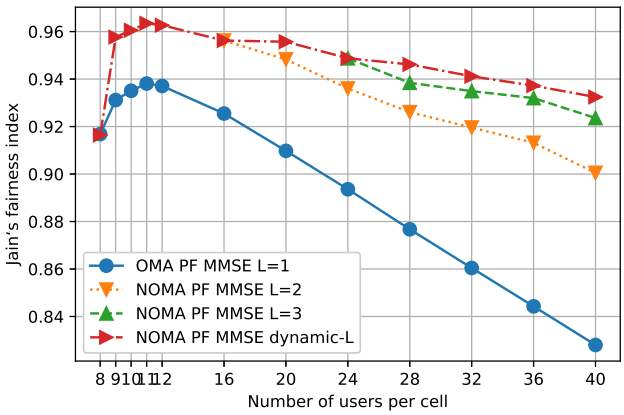


Fig. 8. Jain's fairness index vs. the number of users per cell.

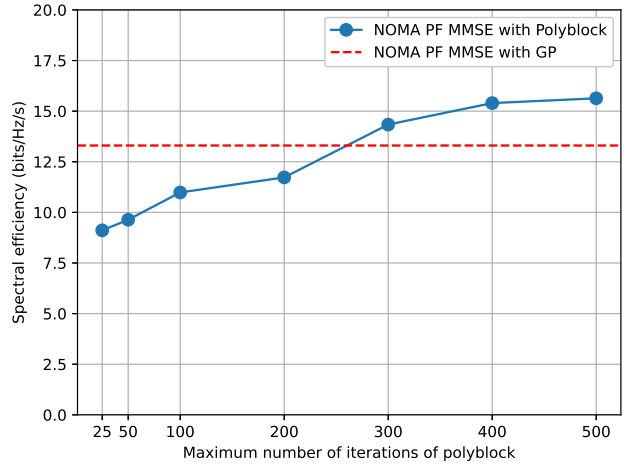


Fig. 10. Sum spectral efficiency vs. the maximum number of iterations of polyblock.

weighted sum rate maximization problem, for maximizing the PF metric. The non-convex optimization problem is finally solved in an iterative way, where in each iteration, the problem can be converted into a GP problem and be solved by standard interior point methods. Simulation results show that, NOMA-PF-MMSE can achieve an overall optimal balance between the sum spectral efficiency and user fairness. Note that although SVD precoder can achieve higher user fairness, it has quite poor sum spectral efficiency when compared to that of MMSE precoder. Besides, we investigate the influence of the maximum number of users per cluster, denoted by L. It is shown that the proposed *dynamic-L* scheme can achieve a strictly better performance than the standard fixed approach and can adapt to various user density and practical scenario. One may also consider in the future to use NOMA and mmWave techniques for mobile edge computing, fog radio access, mission-critical IoT, and massive machine type communications in 6G systems.

Appendix A
Proof of Proposition 2

Proof. We convert the optimization problem (16) into a GP problem by following the method in [54]. The achievable rate region of user k, who is assigned to cluster j, can be expressed as:

$$\boldsymbol{R}(I_k, p_k) = \left\{ R_{\pi_j(l)} : R_{\pi_j(l)} \leq W \log\left(1 + \frac{p_{\pi_j(l)}}{I_{\pi_j(l)} + \sum_{l'>l} p_{\pi_j(l')}}\right), \; l = 1, \ldots, L_j \right\}. \tag{24}$$

Assume that the data rate vector of all users $\{R_k\}$ reaches its boundary, the power consumption of each user can be calculated according to the data rate:

$$p_{\pi_j(l)} = \left(e^{R_{\pi_j(l)} \ln 2/W} - 1\right)\left(I_{\pi_j(l)} + \sum_{l'>l} p_{\pi_j(l')}\right). \tag{25}$$

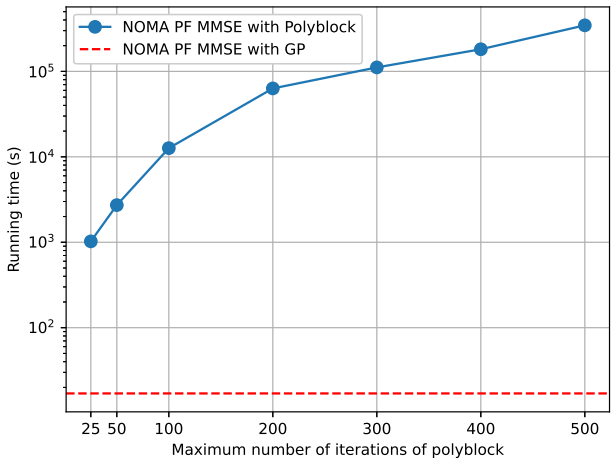


Fig. 11. Running time vs. the maximum number of iterations of polyblock.

The following equation stands for the sum power of all users assigned to cluster j:

$$\sum_{l=1}^{L_j} p_{\pi_j(l)} = \sum_{l=1}^{L_j} e^{\sum_{l'=1}^{l} R_{\pi_j(l')} \ln 2/W} (I_{\pi_j(l)} - I_{\pi_j(l+1)}) - I_{\pi_j(1)}. \tag{26}$$

And the sum power consumption of BS can be calculated by summing up the power of each cluster:

$$\sum_{j=1}^{N_s} \sum_{l=1}^{L_j} p_{\pi_j(l)} = \sum_{j=1}^{N_s} \left(\sum_{l=1}^{L_j} e^{\sum_{l'=1}^{l} R_{\pi_j(l')} \ln 2/W} (I_{\pi_j(l)} - I_{\pi_j(l+1)}) - I_{\pi_j(1)} \right). \tag{27}$$

Intuitively, the optimal power allocation strategy of the weighted sum rate maximization can be obtained when the sum power is maximum. Thus, the sum power of the BS should be less than or equal than $P_{\max}$, and (27) can be converted to:

$$\sum_{j=1}^{N_s} \sum_{l=1}^{L_j} \frac{(I_{\pi_j(l)} - I_{\pi_j(l+1)})}{P_{\max} + \sum_{j'=1}^{N_s} I_{\pi_j(1)}} e^{\sum_{l'=1}^{l} R_{\pi_j(l')} \ln 2/W} \leq 1. \tag{28}$$

Therefore, problem (16) can be converted to (18) with data rate vector $\{R_k\}$ as the optimization variable, and the proof is completed. □

Appendix B
Proof of Theorem 3

Proof. According to (6), the allocated power to user $\pi_j(l)$ is as follows:

$$p^{(t)}_{\pi_j(l)} := \gamma_{\pi_j(l)} \left(\sum_{l'>l} p^{(t)}_{\pi_j(l')} + I^{(t-1)}_{\pi_j(l)} \right). \tag{29}$$

The power of user $\pi_j(L)$ and $\pi_j(L-1)$ in cluster j can be calculated as follows:

$$p^{(t)}_{\pi_j(L)} := \gamma_{\pi_j(L)} I^{(t)}_{\pi_j(L)}, \tag{30}$$

$$p^{(t)}_{\pi_j(L-1)} := \gamma_{\pi_j(L-1)} \times (p^{(t)}_{\pi_j(L)} + I^{(t)}_{\pi_j(L-1)}). \tag{31}$$

Based on (30) and (31), we obtain that

$$\sum_{l=L-1}^{L} p_{\pi_j(l)} = \gamma_{\pi_j(L-1)} I_{\pi_j(L-1)} + (\gamma_{\pi_j(L-1)} + 1)\gamma_{\pi_j(L)} I_{\pi_j(L)}. \tag{32}$$

Then, we calculate the power of user $\pi_j(L-2)$, which is as follows:

$$p_{\pi_j(L-2)} = \gamma_{\pi_j(L-2)} \times (p_{\pi_j(L-1)} + p_{\pi_j(L)} + I_{\pi_j(L-2)}). \tag{33}$$

Based on (32) and (33), the required power of the last three users is obtained. We repeat the aforementioned steps, the summation of all the $L_{j,m}$ users' transmit power is obtained, i.e.,

$$\sum_{l=1}^{L} p_{\pi_j(l)} = \sum_{l=1}^{L} \left[\prod_{l'<l} (\gamma_{\pi_j(l')} + 1) \right] \times \gamma_{\pi_j(l)} I_{\pi_j(l)}, \tag{34}$$

which completes the proof. □

References

[1] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, C. I, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1909–1935, 2017.

[2] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter wave communication: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1616–1653, 2018.

[3] 3GPP Radio Access Network Working Group and others, "New frequency range for NR (24.25-29.5 GHz) (Release 15)," 3GPP TR 38.815, Tech. Rep., 2018.

[4] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1502–1517, 2017.

[5] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, 2016.

[6] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Communications magazine*, vol. 55, no. 9, pp. 134–141, 2017.

[7] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *Information Theory and Applications Workshop (ITA)*, 2013.

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

[9] S. Han, I. Chih-Lin, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, 2015.

[10] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Non-orthogonal multiple access for 5G and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.

[11] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, 2020.

[12] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 537–552, 2015.

[13] Y. Li, M. Jiang, Q. Zhang, Q. Li, and J. Qin, "Secure beamforming in downlink MISO nonorthogonal multiple access systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7563–7567, 2017.

[14] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Joint beamforming design in multi-cluster MISO NOMA reconfigurable intelligent surface-aided downlink communication networks," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 664–674, 2020.

[15] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: user clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, 2016.

[16] Z. Shi, H. Wang, Y. Fu, G. Yang, S. Ma, F. Hou, and T. A. Tsiftsis, "Zero-forcing-based downlink virtual MIMO-NOMA communications in IoT networks," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2716–2737, 2019.

[17] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7425–7440, 2018.

[18] 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, 2019.

[19] X. Pang, J. Tang, N. Zhao, X. Zhang, and Y. Qian, "Energy-efficient design for mmWave-enabled NOMA-UAV networks," *Science China Information Sciences*, vol. 64, no. 4, Apr. 2021, Art. no. 140303.

[20] Y. Liu, C. S. Chen, C. W. Sung, and C. Singh, "A game theoretic distributed algorithm for FeICIC optimization in LTE-A HetNets," *IEEE/ACM Transactions on Networking*, vol. 25, no. 6, pp. 3500–3513, 2017.

[21] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[22] J. Umehara, Y. Kishiyama, and K. Higuchi, "Enhancing user fairness in non-orthogonal access with successive interference cancellation for cellular downlink," in *IEEE International Conference on Communication Systems*, 2012, pp. 324–328.

[23] F. Liu, P. Mähönen, and M. Petrova, "Proportional fairness-based user pairing and power allocation for non-orthogonal multiple access," in *IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2015, pp. 1127–1131.

[24] E. Okamoto, "An improved proportional fair scheduling in downlink non-orthogonal multiple access system," in *IEEE Vehicular Technology Conference (VTC)*, 2015.

[25] T. Van Le and K. Lee, "Opportunistic hybrid beamforming based on adaptive perturbation for mmwave multi-user MIMO systems," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2020.

[26] Y. Fu, M. Zhang, L. Salaün, C. W. Sung, and C. S. Chen, "Zero-forcing oriented power minimization for multi-cell MISO-NOMA systems: A joint user grouping, beamforming and power control perspective," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1925–1940, 2020.

[27] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1576–1590, 2017.

[28] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmwave massive MIMO systems," *IEEE Wireless Communications Letters*, vol. 7, no. 5, pp. 852–855, 2018.

[29] A. Li, A. Harada, and H. Kayama, "A novel low computational complexity power assignment method for non-orthogonal multiple access systems," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 97, no. 1, pp. 57–68, 2014.

[30] Y. Fu, Y. Chen, and C. W. Sung, "Distributed power control for the downlink of multi-cell NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6207–6220, 2017.

[31] L. Salaün, M. Coupechoux, and C. S. Chen, "Weighted sum-rate maximization in multi-carrier NOMA with cellular power constraint," in *IEEE Conference on Computer Communications (INFOCOM)*, 2019, pp. 451–459.

[32] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *IEEE PIMRC*, 2013, pp. 611–615.

[33] Y. Fu, L. Salaün, C. W. Sung, and C. S. Chen, "Subcarrier and power allocation for the downlink of multicarrier NOMA systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 833–11 847, 2018.

[34] L. Salaün, M. Coupechoux, and C. S. Chen, "Joint subcarrier and power allocation in NOMA: Optimal and approximate algorithms," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2215–2230, 2020.

[35] T. A. Thomas, H. C. Nguyen, G. R. MacCartney, and T. S. Rappaport, "3D mmWave channel model proposal," in *IEEE 80th Vehicular Technology Conference*, 2014, pp. 1–6.

[36] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electronics letters*, vol. 27, no. 23, pp. 2145–2146, 1991.

[37] J. He, T. Kim, H. Ghauch, K. Liu, and G. Wang, "Millimeter wave MIMO channel tracking systems," in *IEEE Globecom Workshop*, 2015.

[38] C. Zhang, D. Guo, and P. Fan, "Tracking angles of departure and arrival in a mobile millimeter wave channel," in *IEEE International Conference on Communications (ICC)*, 2016.

[39] 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, 2019.

[40] 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, 2017.

[41] D. H. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, "Hybrid MMSE precoding and combining designs for mmWave multiuser systems," *IEEE Access*, vol. 5, pp. 19 167–19 181, 2017.

[42] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, 2002.

[43] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Communications Letters*, vol. 9, no. 3, pp. 210–212, 2005.

[44] M. Kountouris and D. Gesbert, "Memory-based opportunistic multi-user beamforming," in *IEEE International Symposium on Information Theory (ISIT)*, 2005, pp. 1426–1430.

[45] M.-R. Hojeij, C. Abdel Nour, J. Farah, and C. Douillard, "Weighted proportional fair scheduling for downlink nonorthogonal multiple access," *Wireless Communications and Mobile Computing*, 2018.

[46] S. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi, "A tutorial on geometric programming," *Optimization & Engineering*, vol. 8, no. 1, p. 67, 2007.

[47] N. T. H. Phuong and H. Tuy, "A unified monotonic approach to generalized linear fractional programming," *Journal of Global Optimization*, vol. 26, no. 3, p. 229, 2003.

[48] H. Tuy, F. Al-Khayyal, and P. T. Thach, "Monotonic optimization: Branch and cut methods," in *Essays and Surveys in Global Optimization*. Springer, 2005, pp. 39–78.

[49] A. Zappone, E. Björnson, L. Sanguinetti, and E. Jorswieck, "Globally optimal energy-efficient power control and receiver design in wireless networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 11, pp. 2844–2859, 2017.

[50] C. S. Chen, K. W. Shum, and C. W. Sung, "Round-robin power control for the weighted sum rate maximisation of wireless networks over multiple interfering links," *European Transactions on Telecommunications*, vol. 22, no. 8, pp. 458–470, 2011.

[51] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

[52] 3GPP Radio Access Network Working Group and others, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 16)," 3GPP TR 38.901, Tech. Rep., 2019.

[53] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 1984.

[54] K. Seong, "Cross-layer resource allocation for multi-user communication systems," Ph.D. dissertation, Stanford University, 2008.