The Spectral vs. Energy Efficiency Trade-Off in Dynamic User Clustering Aided mmWave NOMA Networks

Sudhakar Rai, Ekant Sharma, *Senior Member, IEEE,* Aditya K. Jagannatham, *Senior Member, IEEE*, and Lajos Hanzo *Life Fellow, IEEE*

***Abstract*—The spectral efficiency (SE) and global energy efficiency (GEE) trade-off encountered in the design of millimeter-wave (mmWave)-based massive multi-input multi-output (MIMO) non-orthogonal multiple access (NOMA) networks is investigated with a particular focus on user clustering. By exploiting the similarity among user channels a pair of spectral and energy-efficient user clustering algorithms are proposed for dynamically selecting both the number of clusters and the number of users in each cluster. Subsequently, a joint analog precoder/combiner and user clustering technique is developed, followed by a multi-objective optimization (MOO) framework for flexibly balancing the GEE and SE objectives in a mmWave NOMA network subject to specific constraints. The MOO objective is initially transformed to a weighted sum rate maximization problem, followed by a quadratic-transform (QT)-based approach conceived for maximizing the non-convex objective by approximating it as a concave-convex function. Our simulation results demonstrate that the user clustering techniques designed attain a 85% performance gain over random clustering technique and demonstrating the benefits of the algorithm designed for mmWave NOMA networks.**

***Index Terms*—Hybrid precoding, User clustering, mmWave, MIMO, NOMA, spectral efficiency, energy efficiency, fractional programming.**

I. Introduction

Both millimeter-wave (mmWave) and non-orthogonal multiple access (NOMA) constitute widely recognized enabling technologies for next-generation (NG) networks, with great potential to address the diverse demands of emerging applications, including augmented reality, virtual reality, Internet of Things, and ultra-reliable low-latency communications [1]. The availability of wide blocks of spectrum in the mmWave frequency band spanning from 30 GHz to 300 GHz is a key enabler of high data rates and overall system capacity [2]. However, the mmWave signal experiences significant path loss. Thankfully, the short wavelength at mmWave frequencies allows implementing a large antenna array at the base station (BS) to compensate for the path loss via multiple-input multiple-output (MIMO) beamforming techniques [3]. Nevertheless, the prohibitively expensive hardware requirements and excessive energy consumption at high frequencies make it impractical to dedicate a separate radio frequency (RF) chain to each antenna element. To address the above challenge, hybrid mmWave systems significantly reduce the number of RF chains without an obvious performance loss by striking an attractable balance between performance and feasibility [4].

On the other hand, reducing the number of RF chains imposes an additional challenge as it limits the number of active users served simultaneously in mmWave networks [5]. This diminishes the potential of mmWave networks to fully exploit the benefits of multi-user (MU) MIMO technology by restricting its application to a scenario where massive connectivity is desirable. Integration of mmWave technology with the NOMA concept into massive MIMO-NOMA systems breaks the limits of the conventional mmWave network by allowing multiple users to be served using a single RF chain on the same time, frequency, code, and space resource block [6]. This fact motivates the use of NOMA in conjunction with mmWave technology to facilitate MU-MIMO transmission, hence further enhancing the spectral efficiency (SE), rate fairness, and connection density [7]. Compared to the traditional orthogonal multiple access (OMA) scheme, NOMA multiplexes the multiple user streams of a given resource block by distinguishing them in the power domain using superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver [8]. Although NOMA improves the massive connectivity, it is not advisable to directly employ NOMA in a massive user scenario. This is because the implementation of the SIC framework at the receiver is vulnerable to cross-user error propagation and leads to a significant decoding delay, when a large number of active users are present in the network. Thus, mmWave-based massive MIMO systems beneficially support the NOMA SIC framework of dense networks by grouping the users into multiple clusters. Each orthogonal beam in such a network only serves a single cluster, with NOMA being used within

An earlier part of this paper is accepted in part at the IEEE Asia-Pacific Conference on Communications (APCC) 2024.

L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council (EPSRC) projects under grant EP/Y037243/1, EP/W016605/1, EP/X01228X/1, EP/Y026721/1, EP/W032635/1 and EP/X04047X/1 as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028). The work of Aditya K. Jagannatham was supported in part by the Qualcomm Innovation Fellowship; in part by the Qualcomm 6G UR Gift; and in part by the Arun Kumar Chair Professorship. The work of Ekant Sharma was supported by IIITB COMET Foundation under the project "ITB-1983-ECD".

Sudhakar Rai and Aditya K. Jagannatham are with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur 208016, India (e-mail: sudhrai@iitk.ac.in; adityaj@iitk.ac.in).

Ekant Sharma is with the Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India (e-mail: ekant@ece.iitr.ac.in).

Lajos Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (email:lh@ecs.soton.ac.uk).

each cluster [9]. A brief review of the existing research in this area is presented next.

A. Review of existing contributions

The potential gains and challenges of mmWave-massive MIMO-based NOMA networks have been the focus of many research endeavors. Hybrid beamforming, user ordering, user clustering, and power allocation are the key design components of mmWave-based NOMA systems [9]–[14]. These components are inherently interlinked, and naturally, their optimization is intertwined. For example, the choice of user clustering affects the beamformer design, which, in turn, impacts both the power allocation and user ordering. In this context, the authors of [10] investigated a joint optimization problem of user clustering, beamforming, and power allocation in a mmWave MIMO NOMA downlink system. In their path-breaking work Pang *et al.,* [11], formulated a joint optimization problem for power allocation and hybrid beamforming to maximize the minimum user signal-to-leakage-plus-noise ratio, while ensuring rate fairness across the users. Similarly, in [9], a joint user grouping and power optimization framework is developed for secure mmWave NOMA systems. Similarly, Wang et al., [15] proposed a joint clustering and power optimization algorithm to maximize the sum rate by exploiting a Stackelberg game-based design.

Review of existing user clustering schemes:

The directional characteristics of the mmWave beams result in a stronger correlation between the user channels that lie within the main lobe of the beam, which is ideally suited for NOMA user clustering. Generally, user clustering is a combinatorial problem that has NP-hard time complexity. Therefore, various research efforts have been put forward to design the suboptimal low-complex user clustering for mmWave-based NOMA networks. Most of the existing literature in mmWave-based NOMA networks is focused on random user clustering [16], two-user pairing [17], a fixed number of users per cluster [14], [18], [19], and on a variable number of users per cluster [14], [15], [18], [20]–[22]. Moreover, the above treatises mainly consider the channel correlation-based [11], [12]; channel disparity-based [20], [23]; angle-based [7], [24]; joint channel correlation- and disparity-based [17], [21]; cluster head-based [25]; game theory-based [15]; graph theory-based [19]; agglomerative hierarchical [13], [26]; re-clustering-based [23]; beamwidth-based [24] and unsupervised machine learning-based [27] arrangements. These also include K-means-based [11], [12], [21], K-means variants-based [22], and EM algorithm-based [14] user clustering algorithms. By contrast, only a few studies [21], [27], [26] are available in the literature that proposed a user clustering scheme by dynamically updating both the cluster size and the number of clusters.

The widespread adoption of wireless devices and services requiring high data rates has triggered unprecedented the data traffic escalation, which has also been accompanied by an enormous increase in energy consumption [28]. At the same time, shrinking terminal size severely limits the battery capacity of such devices. Therefore, efficient utilization of the available spectrum and energy resources is of vital importance in these systems [29]. Numerous research efforts have focused on improving the SE in mmWave NOMA networks [7], [25], [30], [31]. An ingenious optimization problem conceived for power allocation and power splitting to maximize the SE in a simultaneous wireless information and power transfer-aided mmWave NOMA system has been investigated in [25]. Similarly, the authors of [30] formulated a joint beamforming and power allocation problem to maximize the sum-rate of a two-user scenario. Proceeding further, Shao *et al.,* [31] considered an angle-domain NOMA setting and improved both the user scheduling and precoder/decoder design strategies to maximize the sum-rate of a multi-cell mmWave system. Similarly, considering the angular orientation of the line of sight (LoS) path in a mmWave channel, the authors of the inspirational treatise [7] derived the closed-form sum-rate expression for an angle-domain mmWave NOMA system. Their analysis is both challenging and stimulating due to the incorporation of the angle estimation error. The energy-efficient design of wireless communication systems has become an increasingly prominent consideration due to their ecological and economic impacts. However, it is also important to note that prioritizing energy efficiency (EE) may inadvertently lead to under-utilization of the available spectral resources. Towards this, the authors of [17] put forward an excellent framework for energy-efficient power allocation design in a MU NOMA network for the mmWave band, subject to per-cluster power and per-user quality of service constraints. Along similar lines, the authors of [32] study the challenging issue of secure EE in these systems, considering the presence of an eavesdropper in the set of legitimate receivers. Next, the authors of [33] formulated an optimization problem to jointly optimize the hybrid precoding, power allocation, and bandwidth partitioning for maximizing the EE of mmWave NOMA HetNets. As a further advance, the authors of [34] adjusted the beamwidth of an analog beamformer to facilitate multiple NOMA users within each beam, and thereafter optimized the EE of the resultant beamwidth-controlled mmWave NOMA system.

It is widely exploited that there exists a fundamental trade-off between the EE and SE in a typical implementation. Common approaches to dealing with it are to maximize the EE subject to a constrained SE and vice versa. However, these approaches do not exploit the available degrees of freedom of the constrained objective [35]. Furthermore, introducing these constraints significantly adds to the complexity of obtaining a solution for the optimization problem as it restricts the feasible solution space [36]. Therefore, it is essential to strike a balance between these two factors, consequently, joint rather than constrained optimization of the EE-SE has gained significant research attention in recent years [37]. Only few works such as [38] have successfully investigated the beamforming design problem to find the optimal non-dominated SE-EE Pareto front in the context of mmWave systems for index modulated MIMO-orthogonal frequency division modulation (OFDM) systems. As for the NOMA systems, very few studies have focused on exploring the SE-EE trade-off. For instance, the exposition [39] proposed an imaginative multi-objective optimization (MOO) approach for beamformer design in a

multiple-input single-output-based NOMA system that strikes an excellent balance between the SE and EE. Along similar lines, a resource allocation scheme was proposed for hybrid time division multi-access (TDMA)-NOMA systems in [40].

Against this backdrop, our contributions are boldly contrasted to the existing literature in Table I. Observe that there is only a single treatise in [41] that investigates a joint EE-SE objective-based design for mmWave NOMA systems in the presence of an intelligent reflecting surface between the BS and users. However, their model is limited to a two-user scenario and therefore it has limited applicability in dense multi-user networks. In addition, most of the existing literature in mmWave-NOMA exploits correlation-based user clustering, whose performance is sensitive to an empirically set correlation threshold [9], [17], [25], [33]. A significant drawback of these methodologies is that they lack a systematic framework to determine either the optimal value of the correlation threshold or the number of clusters to suppress the inter-cluster interference. Motivated by these shortcomings of the current state-of-the-art, this treatise aims for comprehensively designing the user clustering, hybrid beamforming, and power allocation schemes to achieve the optimal SE-EE trade-off in mmWave MU massive MIMO-NOMA systems. The key contributions of this treatise are listed below.

B. Contribution of this work

1) In contrast to [9], [10], [12], [17], [25], [30] and [24] that consider single-antenna downlink receivers, we introduce a general framework for an arbitrary number of antennas at the downlink receivers. To begin with, a novel algorithm is conceived for the dynamic selection of the cluster heads (CHs) by optimizing the number of clusters and maximizing the both SE as well as GEE of the system. In contrast to the clustering algorithm of [25] that iteratively increases the correlation threshold, potentially leading to increased inter-cluster interference, the clustering algorithm designed iteratively determines the optimal number of clusters to minimize the inter-cluster interference without necessitating an increase of the correlation threshold. Despite the need for the dynamic number of clusters [21], [27], [26], particularly for mmWave-based networks, a dearth of research investigations is available in these fields.

2) Exploiting the condition number as a correlation metric, more sophisticated cluster head selection and user clustering algorithms are proposed. In contrast to the existing clustering algorithms of [9], [17], [25], [33] that successively obtain suitable CHs, the proposed algorithm simultaneously obtains the CHs for each cluster, which is an efficient solution for user clustering as it allows for a more comprehensive evaluation of the correlation of users. Notably, this study is the, first, to explore using the condition number as a correlation metric for user clustering for the mmWave-based NOMA network. Furthermore, leveraging only the LoS path in the mmWave channel, a joint analog beamforming and user grouping algorithm is developed next for users that lie within the half power beam width (HPBW) of the beamformer main lobe, thereby ensuring that users are served by the most effective beamformer.

3) Following user grouping, and hybrid beamformer design, the power optimization problem is proposed next for striking a trade-off between the SE and global energy efficiency (GEE). Exploiting the MOO framework, an optimization problem is developed for jointly maximizing the SE and GEE of the network subject to total power, per user rate constraint, and successful SIC constraints. Unlike orthogonal multiple access (OMA) schemes, SIC constraint in NOMA allocates a larger portion of power to users with weaker channels to ensure successful SIC while optimizing either SE or GEE. Additionally, this work incorporates the SIC overhead at the users into the power consumption model which is also lacking in the existing literature of the energy-efficient mmWave-NOMA networks. A weighted-sum optimization problem is developed to solve the above MOO problem for the proposed system. The resultant optimization framework is general and exclusive SE optimization, as well as GEE optimization, constitutes its special cases.

4) Our simulation findings demonstrate the impact of the proposed user clustering schemes on the i) SE; ii) GEE; and iii) SE-GEE trade-off of the proposed mmWave-NOMA network by comparing them with random and other existing state-of-the-art clustering techniques [21], [25], [42]. The CN-based clustering algorithm outperforms the random and similarity-based clustering schemes under a moderate number of antennae and large user scenarios. However, its advantage diminishes as antenna numbers increase or user numbers decrease. The impact of the proposed user clustering scheme and hybrid precoding scheme on the SE-GEE trade-off performance was also studied by comparing it to the fully digital MIMO, and MIMO-OMA counterparts. Moreover, the effectiveness of the proposed optimization framework over the random power allocation is validated through simulations, demonstrating its ability to balance GEE and SE by fine-tuning the weights.

II. System Model

Consider the downlink of a single-cell MU mmWave MIMO-NOMA system where a BS simultaneously serves K users. The BS is equipped with N_B transmit antennas and N_{RF} RF chains, while each user in the network has N_U receive antennas and a single RF chain. In contrast to conventional mmWave transmission, where each RF chain at the BS supports only a single user, mmWave-based NOMA transmission allows multiple users to be served simultaneously using a single RF chain, i.e., $K > N_{RF}$. The K users are divided into G clusters by employing a suitable user clustering technique, where the number of clusters does not exceed the number of RF chains, i.e., $G \leq N_{RF}$.

A. Channel model

We harness the widely used narrowband Saleh-Valenzuela block-fading channel model for mmWave communication [4]. The downlink channel $\mathbf{H}_k \in \mathbb{C}^{N_U \times N_B}$ between the BS and kth user can be expressed as

$$\mathbf{H}_k = \sqrt{\frac{N_B N_U}{L_k + 1}} \left(\mathbf{H}_{k,0} + \sum_{l=1}^{L_k} \mathbf{H}_{k,l} \right), \tag{1}$$

Table I: Contrasting our solution to the literature of HP based mmWave MIMO-NOMA systems

	[25], [13], [33]	[8]	[34]	[31]	[35]	[36]	[39]	[38]	[41]	[21]	[26]	[27]	**Proposed**
mmWave-Based Massive MIMO	✓		✓	✓				✓	✓		✓	✓	✓
NOMA	✓	✓	✓	✓			✓		✓	✓	✓	✓	✓
Multi-Antenna Users		✓	✓	✓	✓								✓
Dynamic User Clustering										✓	✓	✓	✓
Joint User Clustering And Beamforming												✓	✓
Power Optimization	✓			✓	✓	✓	✓	✓	✓	✓		✓	✓
Balancing GEE-SE Trade Off					✓	✓	✓	✓	✓				✓

where $\mathbf{H}_{k,0} = \nu_{k,0}\mathbf{a}_R(\theta_{k,0})\mathbf{a}_T^H(\phi_{k,0})$ and $\mathbf{H}_{k,l} = \nu_{k,l}\mathbf{a}_R(\theta_{k,l})\mathbf{a}_T^H(\phi_{k,l})$ denote the channel matrices for the LoS and the lth non-line of sight (NLoS) paths, respectively, between the BS and kth user. Due to the high free-space path-loss and limited scattering in mmWave communication, the number of distinguishable paths $L_k + 1$ is small, where L_k denotes the number of NLoS paths. The parameter $\nu_{k,l}$ denotes the complex gain of the lth path and it is modeled as an independent random variable obeying the distribution $\mathcal{CN}\big(0, \sigma_{k,l}^2\big)$. Upon considering uniform linear arrays (ULAs) at both the transmitter and the receiver, the vectors $\mathbf{a}_T(\phi) = \frac{1}{\sqrt{N_B}}[1, e^{j\frac{2\pi}{\lambda}d\sin\phi}, \cdots, e^{j\frac{2\pi}{\lambda}d(N_B-1)\sin\phi}]^T \in \mathbb{C}^{N_B\times 1}$ and $\mathbf{a}_R(\theta) = \frac{1}{\sqrt{N_U}}[1, e^{j\frac{2\pi}{\lambda}d\sin\theta}, \cdots, e^{j\frac{2\pi}{\lambda}d(N_U-1)\sin\theta}]^T \in \mathbb{C}^{N_U\times 1}$ denote the normalized transmit and receive array steering vectors for the angular directions of ϕ and θ, respectively. In particular, θ and ϕ denote the azimuth angle of arrival (AoA) and azimuth angle of departure (AoD), respectively, which are uniformly distributed within the range of $[0, \pi]$. The parameters λ and $d = \frac{\lambda}{2}$ denote the carrier wavelength and antenna spacing, respectively.

B. Downlink data transmission

The BS is assumed to transmit G streams in order to serve the K users in the G clusters. Let $\mathcal{S} = \{1, 2, \cdots, G\}$ denote the set containing the cluster indices, while the set of user indices in the ith cluster is denoted as $\mathcal{U}_i = \{1, 2, \cdots, |\mathcal{U}_i|\}$. Here $|\mathcal{U}_i|$ represents the number of users in the ith cluster with $|\mathcal{U}_i| \neq 0$ for $\forall i$. For NOMA superposition coding and mmWave hybrid precoding, the transmitted signal of the BS can be expressed as

$$\mathbf{x} = \mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{s}, \tag{2}$$

where the vector $\mathbf{s} = \big[s_1, s_2, \cdots, s_G\big]^T \in \mathbb{C}^{G\times 1}$ denotes the symbols corresponding to all the clusters. The scalar $s_i = \sum_{j=1}^{|\mathcal{U}_i|}\sqrt{\alpha_{i,j}P_{\max}}s_{i,j}$ represents the NOMA superposition coded symbol for all the users in the ith cluster, where $s_{i,j}$ denotes the information symbol of the jth user in the ith cluster which is assumed to be independent identically distributed (i.i.d) with an average power of $\mathbb{E}\big[|s_{i,j}|^2\big] = 1$. The parameters $\alpha_{i,j}$ and $P_{\max}$ represent the power allocation coefficients corresponding to symbol $s_{i,j}$ and the average transmit power at the BS, respectively. The transmit power constraint can be formulated as $\sum_{i=1}^{G}\sum_{j=1}^{|\mathcal{U}_i|}\alpha_{i,j} \leq 1$. The matrices $\mathbf{F}_{RF} = \big[\mathbf{f}_{RF}^1, \mathbf{f}_{RF}^2 \cdots, \mathbf{f}_{RF}^G\big] \in \mathbb{C}^{N_B\times G}$ and $\mathbf{F}_{BB} = \big[\mathbf{f}_{BB}^1, \mathbf{f}_{BB}^2, \cdots, \mathbf{f}_{BB}^G\big] \in \mathbb{C}^{G\times G}$ denote the analog and baseband transmit precoding (TPC) matrices at the BS, respectively. The choice of $\alpha_{i,j}$, $\mathbf{F}_{RF}$ and $\mathbf{F}_{BB}$ will be discussed later in the following sections.

For a given user clustering, hybrid precoding, and power allocation, the signal $\mathbf{r}_{i,j} \in \mathbb{C}^{N_U\times 1}$ received at the jth user of the ith cluster can be expressed as

$$\mathbf{r}_{i,j} = \mathbf{H}_{i,j}\mathbf{F}_{RF}\sum_{g=1}^{G}\mathbf{f}_{BB}^g\sum_{m=1}^{|\mathcal{U}_g|}\sqrt{\alpha_{g,m}P_{\max}}s_{g,m} + \boldsymbol{\eta}_{i,j}, \tag{3}$$

where the matrix $\mathbf{H}_{i,j}$ is the mmWave channel from the BS to the jth user in the ith cluster, furthermore, the vector $\boldsymbol{\eta}_{i,j} \in \mathbb{C}^{N_U\times 1}$, which follows the distribution $\mathcal{CN}\big(0, \sigma_\eta^2\mathbf{I}_{N_U}\big)$, denotes the additive white Gaussian noise (AWGN) added at the jth user in the ith cluster.

Owing to the fact that there is only a single RF chain at the users, signal received at the jth user in the ith cluster after applying the analog combiner $\mathbf{w}_{i,j} \in \mathbb{C}^{N_U\times 1}$ is written as

$$\mathbf{w}_{i,j}^H\mathbf{r}_{i,j} = \mathbf{w}_{i,j}^H\mathbf{H}_{i,j}\mathbf{F}_{RF}\sum_{g=1}^{G}\mathbf{f}_{BB}^g\sum_{m=1}^{|\mathcal{U}_g|}\sqrt{\alpha_{g,m}P_{\max}}s_{g,m} + \mathbf{w}_{i,j}^H\boldsymbol{\eta}_{i,j}. \tag{4}$$

The next section discusses the digital TPC design, the NOMA SIC framework and the achievable rate analysis for the desired user.

III. Achievable Rate Analysis

The received signal in (4) contains the desired signal, intra-cluster interference, and inter-cluster interference, which can be expressed as

$$\mathbf{y}_{i,j} = \underbrace{\mathbf{h}_{i,j}^H\mathbf{f}_{BB}^i\sqrt{\alpha_{i,j}P_{\max}}s_{i,j}}_{\text{desired signal}} + \underbrace{\mathbf{h}_{i,j}^H\mathbf{f}_{BB}^i\sum_{m\neq j}^{|\mathcal{U}_i|}\sqrt{\alpha_{i,m}P_{\max}}s_{i,m}}_{\text{intra-cluster interference}}$$
$$+ \underbrace{\mathbf{h}_{i,j}^H\sum_{g\neq i}^{G}\mathbf{f}_{BB}^g\sum_{m=1}^{|\mathcal{U}_g|}\sqrt{\alpha_{g,m}P_{\max}}s_{g,m}}_{\text{inter-cluster interference}} + \mathbf{w}_{i,j}^H\boldsymbol{\eta}_{i,j}, \tag{5}$$

where $\mathbf{h}_{i,j}^H = \mathbf{w}_{i,j}^H\mathbf{H}_{i,j}\mathbf{F}_{RF} \in \mathbb{C}^{1\times G}$ denotes the effective channel spanning from the BS to the jth user in the ith cluster. The BS minimizes the inter-cluster interference by designing the digital TPC for each cluster, whereas, the intra-cluster interference is minimized at each user by employing SIC.

Since the number of users exceeds the number of beams, a digital TPC can be designed for a single user in each cluster, which is shared among all the remaining users within the cluster. Therefore, it becomes impossible to completely suppress the inter-cluster interference. It is desirable to design the digital

TPC for the strongest channel user in each cluster in order to maximize the network throughput. By concatenating the effective channel vectors of the strongest channel users from each cluster as $\tilde{\mathbf{H}} = \left[\mathbf{h}_{(1)}, \mathbf{h}_{(2)}, \cdots, \mathbf{h}_{(G)}\right] \in \mathbb{C}^{G\times G}$ and applying the conventional zero forcing algorithm, the digital TPC matrix can be expressed as $\mathbf{F}_{BB} = \tilde{\mathbf{H}}\left(\tilde{\mathbf{H}}^H\tilde{\mathbf{H}}\right)^{-1}$. To satisfy the unit power constraint for the hybrid TPC, each column of $\mathbf{F}_{BB}$ is further normalized as $\mathbf{f}_{BB}^g = \frac{\mathbf{f}_{BB}^g}{||\mathbf{F}_{RF}\mathbf{f}_{BB}^g||}$, for $g = 1, 2, \cdots, G$.

On the other hand, for SIC, the decoding order of users within the ith cluster is set in the ascending order of their effective channel gains as follows

$$|\mathbf{h}_{i,1}^H\mathbf{f}_{BB}^i|^2 \leq |\mathbf{h}_{i,2}^H\mathbf{f}_{BB}^i|^2 \leq \cdots \leq |\mathbf{h}_{i,|\mathcal{U}_i|}^H\mathbf{f}_{BB}^i|^2. \quad (6)$$

The NOMA power allocation coefficients corresponding to users in the ith cluster are ordered in descending order, i.e., $1 \geq \alpha_{i,1} \geq \alpha_{i,2} \geq \cdots \geq \alpha_{i,|\mathcal{U}_i|} \geq 0$.

Following the SIC detection order in (6), the first user in the ith cluster detects the desired signal by treating the signals of all other users in the same cluster as interference. In particular, the jth user in the ith cluster employs SIC to detect all the previous user signals, and then progressively subtracts the interference before decoding its own information signal. Therefore, the signal-to-interference-plus-noise (SINR) expression at the jth user for detecting the kth user's signal in the ith cluster, along with $k \leq j$ can be expressed as

$$\text{SINR}_{k\leftarrow j}^i = \frac{|\mathbf{h}_{i,j}^H\mathbf{f}_{BB}^i|^2\alpha_{i,k}P_{\max}}{\left\{\begin{array}{l}|\mathbf{h}_{i,j}^H\mathbf{f}_{BB}^i|^2\sum\limits_{m=k+1}^{|\mathcal{U}_i|}\alpha_{i,m}P_{\max} \\ +\sum\limits_{g\neq i}^{G}|\mathbf{h}_{i,j}^H\mathbf{f}_{BB}^g|^2\sum\limits_{m=1}^{|\mathcal{U}_g|}\alpha_{g,m}P_{\max}+|\mathbf{w}_{i,j}^H\boldsymbol{\eta}_{i,j}|^2\end{array}\right\}}, \quad (7)$$

where $j \leq |\mathcal{U}_i|$, $1 \leq k \leq j$. It follows from (6) and (7) that if the kth user having a weaker effective channel can detect its message signal, then the jth user associated with a stronger effective channel is also guaranteed to be able to detect the message of the kth user having $\text{SINR}_{k\leftarrow j}^i \geq \text{SINR}_{k\leftarrow k}^i$ for $\forall$ $k \leq j$. This condition guarantees successful SIC operation at each user and it is always satisfied, once the detection order is set at the BS. Note that the above inequality derived for successful SIC turns out to be a non-convex constraint under the optimization framework of Section V. Therefore, (6) provides a more convenient and tractable alternative for ensuring successful SIC, when framing an optimization problem associated with NOMA [43]. The achievable rate at the jth user in the ith cluster upon detecting its own message signal is given as

$$R_{i,j} = \log_2\left(1 + \text{SINR}_{j\leftarrow j}^i\right). \quad (8)$$

In contrast to [4], which focused on developing the joint analog TPC and receiver combining (RC) for each user in the mmWave network, joint design is impossible for mmWave NOMA networks since users in a particular cluster share a common TPC designed for the CH. In the next section, the analog RCs are designed for each user by employing the instantaneous channel matrices. Subsequently, user grouping and analog precoding are designed for the proposed mmWave NOMA system.

Section IV-A: Analog-only combining

↓ Exploiting the beamforming codebook, analog-only combiners $\mathbf{w}_k$, $\forall k$ is designed for users in (10).

Section IV-B1: Cluster Head Selection Techniques

↓ a) *Similarity-based Cluster Head Selection Technique:* Dynamically selects the CHs for each cluster by investigating the correlation between user channels in Algorithm 2.
b) *Condition Number-Based Cluster Head Selection Technique:* Exploits the condition number as correlation index to determine to CHs for the all clusters in Algortihm 2.

Section IV-B2: Analog precoding

↓ Based on the equivalent channel vectors of CHs, analog precoding matrix $\mathbf{F}_{RF}$ is designed in (13).

Section IV-B3: Assignment of remaining users

↓ Similar to cluster head selection technique, the condition number is used to identify the remaining users in all the clusters.

Section IV-C: Joint analog precoding, analog combining and user grouping

↓ Exploiting the half-power beamwidth (HPBW) of the main lobe of the beam, a joint analog precoding, analog combining, and user clustering are discussed in Algorithm 3.

Section III: Digital precoding

Fig. 1: Hybrid precoding, combining and user clustering

IV. Analog Precoding, Analog Combining and User Clustering

Recall that, as the number of RF chains N_{RF} available at the BS is lower than the number of active users K, it is not possible to design the analog TPC $\mathbf{F}_{RF}$ and the baseband TPC $\mathbf{F}_{BB}$ for each user at the BS. Therefore, it becomes essential to choose a CH in each cluster and to design the RF and baseband TPC matrices for these. The remaining users within the cluster share the same TPC designed for the CH. In contrast to [14], [25], [44] and [13], where users equipped with a single antenna are considered, choosing a CH for each cluster in the MU-MIMO system under consideration in this work is a challenging task. To address this, the analog RC is specifically designed for each user, first, employing the user channels. Subsequently, a cluster head selection algorithm is proposed to determine the CH for each cluster, utilizing the equivalent channels of the users. The systematic flow of steps in the proposed hybrid TPC, RC and user clustering schemes are summarized in Fig. 1.

Remark: We recommend commencing with user clustering, then proceeding with analog and digital precoding, and finally power optimization. However, depending on the performance needs, other sequences may also considered. For example, in [18] beamforming has been performed first, followed by user clustering and power optimization.

A. Analog receiver combining

The analog-only RC at each user can be viewed as a typical single-user problem [4, Section IV]. Therefore, it can be designed using only point-to-point channel knowledge obtained by exploiting the codebook-based beam-training techniques developed in [45], which do not require explicit channel estimation and have a low training overhead. This allows each user to choose their respective analog-only RC vector from the pre-defined RF codebook, defined as follows [4]

$$\mathcal{W}_c = \big[\mathbf{a}_R(\theta_0), \mathbf{a}_R(\theta_1), \cdots, \mathbf{a}_R(\theta_{C_W-1})\big], \tag{9}$$

where C_W is the size of the codebook and $\{\theta_i\}_{i=0}^{G_{w-1}}$ are selected from the set $\mathbf{\Theta} = \{\theta_i; \theta_i \in \frac{2\pi i}{C_W}, \forall i\}$, which is the set of feasible quantized angular grid points.

Once the codebook $\mathcal{W}_c$ is designed, each user exhaustively sweeps over $\mathcal{W}_c$ to determine the RF-only RC. Therefore, the RC vector $\mathbf{w}_k$ at the kth user can be selected as follows

$$\mathbf{w}_k^H = \arg\max_{\forall \mathbf{w}_i \in \mathcal{W}_c} \|\mathbf{w}_i^H \mathbf{H}_k\| \text{ , for } \forall k. \tag{10}$$

It can be observed from (10) that for a given channel matrix $\mathbf{H}_k$, the kth user selects the analog RC $\mathbf{w}_k$ that maximizes the strength of the equivalent channel vector $\bar{\mathbf{h}}_k^H = \mathbf{w}_k^H \mathbf{H}_k \in \mathbb{C}^{1\times N_B}$.

Algorithm 1: Similarity-based CH selection algorithm

Input: Given a correlation threshold, $\epsilon = 0.1$; Number of users, K with their channel vectors, $\{\bar{\mathbf{h}}_i\}_{i=1}^K$; Number of clusters, G; $g = 2$

Output: CH set, Ω^*; Remaining user set, Ω^c; Number of clusters, G

$\mathbf{A} = \{\bar{\mathbf{a}}_i\}_{i=1}^K$ where $\bar{\mathbf{a}}_i = \bar{\mathbf{h}}_i / \|\bar{\mathbf{h}}_i\|$;
$[\sim, \Omega^c] = \text{sort}\big(\{\|\bar{\mathbf{h}}_1\|, \|\bar{\mathbf{h}}_2\|, \cdots, \|\bar{\mathbf{h}}_K\|\}, \text{'descend'}\big)$;
$\Omega^* = \Omega^c(1)$; $\Omega^c = \Omega^c \setminus \Omega^*$;
while $g \le G$ **do**
  $\mathcal{S} = \{j \in \Omega^c; |\bar{\mathbf{a}}_i^H \bar{\mathbf{a}}_j| < \epsilon, \forall i \in \Omega^*, \bar{\mathbf{a}}_i, \bar{\mathbf{a}}_j \in \mathbf{A}\}$;
  $\boldsymbol{\rho} = \{|\bar{\mathbf{a}}_i^H \bar{\mathbf{a}}_j| < \epsilon, \forall i \in \Omega^*, j \in \mathcal{S}\}$;
  if $\boldsymbol{\rho} \neq \Phi$ **then**
    $j^* = \arg\min_{j\in\mathcal{S}} \boldsymbol{\rho}$;
    $\Omega^* = \Omega^* \cup \Omega^c(j^*)$; $\Omega^c = \Omega^c \setminus \Omega^*$;
  else
    $G = g - 1$;
  $g = g + 1$;

B. User clustering

User clustering can be leveraged to minimize both the inter-cluster and the intra-cluster interference by efficiently classifying the users within each cluster. Intuitively, user channels within the same NOMA cluster are expected to be highly correlated so that they can be aligned within the same beam, while ensuring successful SIC operation. In addition, the correlation between the channel gains of the users in different NOMA clusters should be minimized for reducing the inter-cluster interference. Therefore, user clustering can be performed, first, by selecting the CH for each cluster based on the user's equivalent channel vectors $\{\bar{\mathbf{h}}_i\}_{i=1}^K$ and then, associating the remaining users with the selected CHs.

Algorithm 2: Condition number-based CH selection algorithm

Input: Given a CN threshold, $\epsilon = 10$dB; Number of users, K with their channel vectors, $\{\bar{\mathbf{h}}_i\}_{i=1}^K$; User indices set, $\Omega^c = \{1, 2, \cdots, K\}$; Number of clusters, G; flag = true

Output: Cluster head set Ω^*; Remaining user set Ω^c; Number of clusters, G

while *flag* = *true* **do**
  Possible number of choices for cluster head set, $\pi = \binom{K}{G}$;
  Set of user indices corresponds to the ith choice of cluster head set, Ω_{π_i} with $\Omega_{\pi_i} = \{\Omega_{\pi_i}^1, \Omega_{\pi_i}^2, \cdots, \Omega_{\pi_i}^G\}$, for $\Omega_{\pi_i}^k \in \{1, 2, \cdots, K\}$;
  $\mathbf{\Gamma}_i = \big[\bar{\mathbf{h}}_{\Omega_{\pi_i}^1}, \bar{\mathbf{h}}_{\Omega_{\pi_i}^2} \cdots, \bar{\mathbf{h}}_{\Omega_{\pi_i}^G}\big]$, for $i = 1, 2, \cdots, \pi$;
  $\big[\backsim, \mathbf{\Lambda}_i, \backsim\big] = \text{SVD}\big(\mathbf{\Gamma}_i\big), \forall\, i$;
  $\sigma_{i_{\max}} = \max\big(\text{diag}\big(\mathbf{\Lambda}_i\big)\big)$; $\sigma_{i_{\min}} = \min\big(\text{diag}\big(\mathbf{\Lambda}_i\big)\big), \forall\, i$;
  $\kappa_i = 10\log_{10}\left(\frac{\sigma_{i_{\max}}^2}{\sigma_{i_{\min}}^2}\right)$, $\forall\ i$;
  $C_i = \log_2\big(|\mathbf{I}_{N_T} + \frac{P_{\max}}{G}\mathbf{\Gamma}_i\mathbf{\Gamma}_i^H|\big)$;
  $\mathcal{S} = \{i; \kappa_i \le \epsilon, \forall i\}$; $\Delta = \{C_j; j \in \mathcal{S}\}$;
  $j^* = \arg\max_{j\in\mathcal{S}} \Delta$;
  if $\Delta == \Phi$ **then**
    $G = G - 1$; flag = true;
  else
    $\Omega^* = \Omega_{\pi_{j^*}}$; $\Omega^c = \Omega^c \setminus \Omega^*$; flag = false;
  if $G == 1$ **then**
    $\Omega^* = [\ ]$; $\Omega^c = U$;

1) Cluster head selection techniques

This section proposes a pair of spectral- and energy-efficient CH selection algorithms by dynamically selecting the number of clusters. These are unlike the existing clustering algorithms of [12], [17], [25], where the number of clusters is fixed. Following the development of a CH selection technique in Algorithm 1, which successively obtains the CHs for all the clusters, the subsequent CH selection technique of Algorithm 2 reflects a novel approach to simultaneously obtain the CHs.

I) Similarity-based cluster head selection: The proposed CH selection Algorithm 1 identifies the CH for each cluster by comparing the degree of correlation between the equivalent user channels to a correlation threshold ϵ. Specifically, users whose channel correlation is below the correlation threshold can be treated as uncorrelated users in **Step 6** and are identified to be the CHs in **Step 9**. It is worth noting here that the proposed algorithm dynamically updates the number of clusters in **Step 11** when there exists no potential candidate for the CH with a correlation below the specified threshold. unlike [25], which iteratively updates the correlation threshold by increasing its value until there is no potential candidate to be set as the CH, the proposed clustering algorithm complies with stringent correlation threshold criterion.

II) Condition number[1]*-based cluster head selection*: Exploiting the combinatorial nature of user clustering, a more efficient CH selection technique is proposed in Algorithm 2 for simultaneously determining the CHs for all clusters. This

[1]Condition number specifies the extent to which a matrix is ill-conditioned. Mathematically, it is expressed as the ratio of maximum singular value to the minimum singular value of a matrix i.e., $\kappa(\mathbf{A}) = 10\log_{10}\frac{\sigma_{max}^2(\mathbf{A})}{\sigma_{min}^2(\mathbf{A})}$.

is unlike the CH selection algorithm discussed in the previous sub-section, where the selection of each CH depends on the previously selected CHs. To minimize inter-cluster interference, it is crucial to select cluster heads (CHs) with mutually uncorrelated channels. Leveraging the condition number (CN) as a correlation index, the proposed scheme identifies the potential CHs from the set of all K users. The condition number of a matrix indicates the degree to which the matrix is ill-posed.

Given the equivalent channel vectors $\{\bar{\mathbf{h}}_k\}_{k=1}^{K}$ for all users, Algorithm 2 first identifies all possible combinations of cluster heads π in **Step 2**. For instance, the legitimate selection of 4 cluster heads out of 8 users in the network yielding $\pi = 70$ combinations. The index set Ω_{π_i}, for $i = 1, 2, \cdots, \pi$, in **Step 3** contains the user indices corresponding to each specific combination. Next, the cluster head matrix (CHM) is constructed by concatenating the equivalent channel vectors of users having indices given in the index set Ω_{π_i} i.e., $\mathbf{\Gamma}_i = [\bar{\mathbf{h}}_{\Omega^1_{\pi_i}}, \bar{\mathbf{h}}_{\Omega^2_{\pi_i}}, \cdots, \bar{\mathbf{h}}_{\Omega^G_{\pi_i}}]$, as shown in Steps 4 of Algortihm 2. Specifically, the matrix $\mathbf{\Gamma}_i \in \mathbb{C}^{N_B \times G}$ contains the equivalent channel vectors of the potential CH candidates for all G beams. Thereafter, the condition number of each CHM is calculated as follows

$$\kappa_i = 10\log_{10}\left(\frac{\sigma^2_{i_{\max}}}{\sigma^2_{i_{\min}}}\right),\ i = 1, 2, \cdots, \pi, \tag{11}$$

where $\sigma^2_{i_{\max}}$ and $\sigma^2_{i_{\min}}$ are the maximum and minimum singular values of the ith cluster head matrix, respectively. Subsequently, the capacity C_i delivered by each CHM is calculated in **Step 8**.

Following this, the index set $\mathcal{S}$ is formed by selecting the indices of only those CHMs whose CN κ_i meets the predefined CN threshold criteria in **Step 9**. Specifically, those CHMs that possess a condition number below the specified CN threshold ϵ are selected. This indicates that the equivalent channel vectors of the CHs in these CHMs are uncorrelated with each other. Finally, we select the desired CHM that maximizes the capacity in **Step 10**, and then we proceed to obtain the CH set by extracting the user indices corresponding to the CHM selected in **Step 14**. If no CHM exists that meets the CN threshold criterion in **Step 11**, this implies that there is no set of G users whose equivalent channels are uncorrelated. In that case, the proposed algorithm reduces the number of clusters by one in **Step 12**vand repeats with **Step 2** to identify the least-correlated cluster heads with a reduced number of clusters. Hence, the proposed algorithm dynamically obtains the number of clusters in **Step 12** for which the CH matrix obtained satisfies the strict pre-defined CN threshold criterion.

It is important to note that Algorithm 2 only maximizes the capacity for the cluster heads. This does not guarantee that the overall capacity of the mmWave NOMA network is maximized, as the remaining users are assigned to clusters later in Section IV-B.3.

2) Analog precoding

This subsection focuses on the analog TPC design at the BS using the equivalent channel vectors corresponding to the CHs $\{\bar{\mathbf{h}}_{\Omega^*(g)}\}_{g=1}^{G}$ obtained, where Ω^* is the set containing the user indices corresponding to the CHs. The beamforming codebook $\mathcal{F}_c \in \mathbb{C}^{N_B \times C_F}$ for the RF TPC can be designed, similar to (9), using transmit steering vectors $\mathbf{a}_T(\phi_i)$ with quantized phase angles and constant modulus entries as follows

$$\mathcal{F}_c = \left[\mathbf{a}_T(\phi_0), \mathbf{a}_T(\phi_1), \cdots, \mathbf{a}_T(\phi_{C_{F-1}})\right], \tag{12}$$

where C_F denotes the codebook size. By exploiting (12), the analog TPC can be designed for maximizing the array gain of the CHs as follows

$$\mathbf{f}^g_{RF} = \arg\max_{\forall \mathbf{f}_i \in \mathcal{F}_c} \|\bar{\mathbf{h}}^H_{\Omega^*(g)}\mathbf{f}_i, \| \text{ for } g = 1, 2, \cdots, G. \tag{13}$$

Applying the analog TPC, the effective channel vector $\tilde{\mathbf{h}}^H_k \in \mathbb{C}^{1\times G}$ at the kth user can be expressed as

$$\mathbf{h}^H_k = \bar{\mathbf{h}}^H_k \mathbf{F}_{RF},\ \text{for } k = 1, 2, \cdots, K. \tag{14}$$

It is noteworthy here that the analog TPC vectors are designed exclusively for maximizing the desired signal power of the CHs and are shared among the other users of the same cluster. To effectively utilize the TPC among the other users in the same cluster, user grouping has been investigated, next, for the proposed system.

3) Assignment of remaining users

Similar to the CH selection criterion in Algorithm 2, the condition number can be used as the correlation index to identify the remaining users in a given cluster. First, the matrix $\mathbf{G}^g_i \in \mathbb{C}^{G\times 2}$ is constructed by concatenating the effective channel vectors of the CH for the gth cluster and the ith user from the remaining user index set Ω^c as follows

$$\mathbf{G}^g_i = \left[\mathbf{h}_{i\in\Omega^c}\, \mathbf{h}_{\Omega^*(g)}\right],\ \text{for } g = 1, 2, \cdots, G. \tag{15}$$

There exist a set of G possible matrices for the ith user, denoted by $\{\mathbf{G}^g_i\}_{g=1}^{G}$. In the next step, the ith remaining user is assigned to the cluster $\hat{g}$, provided that the condition number of the matrix $\mathbf{G}^{\hat{g}}_i$ is maximized as follows

$$\hat{g} = \underset{g=1,2,\cdots,G}{\arg\max}\ 10\log_{10}\left(\frac{\sigma^g_{i_{\max}}}{\sigma^g_{i_{\min}}}\right)^2. \tag{16}$$

Here, $\sigma^g_{i_{\max}}$ and $\sigma^g_{i_{\min}}$ are the maximum and minimum singular values corresponding to the matrix $\mathbf{G}^g_i$, respectively. This process is repeated until all the users in the remaining user set Ω^c are assigned to their respective clusters.

To summarize, the proposed user clustering technique first identifies the CHs for each cluster by dynamically updating the number of clusters using Algorithms 1 and 2 in Subsection IV-B1 and proceeds to assign the remaining users to the CHs obtained, as discussed in Section IV-B3. Therefore, the performance of the above user clustering technique depends on the effectiveness of the CHs chosen. To further improve the performance, it is necessary to investigate a user clustering mechanism that jointly determines the CHs and assigns the remaining users for each cluster. Moreover, the user clustering problem has to be jointly investigated with the analog TPC and RC design problem to further improve the performance. Considering this, a joint analog TPC, analog RC, and user clustering problem is investigated in the next sub-section.

C. Joint analog precoding, combining and user grouping

Exploiting the half-power beamwidth (HPBW) of the main lobe of the beam, a more sophisticated user grouping technique

Algorithm 3: Joint analog precoding, analog combining, and user grouping

Input: Half-power beamwidth (HPBW) $= 0.891\frac{2\pi}{N_B}$; $\Delta\theta_g = \frac{\text{HPBW}}{2}$ for $g = 1, 2, \ldots, G$; Number of users, K, with their channel matrices, $\{\mathbf{H}_i\}_{i=1}^{K}$; Transmit/Receive Array steering vectors, $\{\mathbf{a}_R(\theta_{k,0})\}_{k=1}^{K}, \{\mathbf{a}_T^H(\phi_{k,0})\}_{k=1}^{K}$; Number of clusters, G; Intializations: Set of user indices in the g-th cluster, $\Omega_g = [\ ]$ for $g = 1, 2, \ldots, G$; Set containing remaining user indices, $\Omega^c = \{1, 2, \ldots, K\}$; $\epsilon = \frac{2\pi}{100}$; $g = 1$; flag=false; iter=0;

Output: Ω_g for $g = 1, 2, \ldots, G$

```
while flag=true do
    for g ← 1 to G do
        if iter==0 then
            [j*] = arg max_{j∈Ω^c} a_R^H(θ_{j,0}) H_j a_T(φ_{j,0}),
            w_{j*} = a_R(θ_{j*}); f_{RF}^g = a_T(φ_{j*});
            φ_g = ∠f_{RF}^g;
            Ω_g = Ω_g ∪ {j*}; Ω^c = Ω^c \ Ω_g;
        S = {l ∈ Ω^c; φ_l ∈ [φ_g − Δφ_g, φ_g + Δφ_g]};
        if S == φ then
            Δφ_g = Δφ_g + ε;
        else
            Ω_g = Ω_g ∪ S;
            w_{g,m} = a_R(θ_l) for ∀m, l ∈ Ω_g;
            Ω^c = Ω^c \ Ω_g;
        if Ω^c == φ then
            flag=false;
            break;
    iter=iter+1;
```

is now investigated by integrating analog precoding, analog combining, and user clustering, as formulated in Algorithm 3. In this approach, only those users located within the main lobe of the beam are allowed to form a NOMA cluster. Here, particularly, the joint design considers the availability of angular information of the LoS paths of the mmWave channels between all the users and the BS to design the analog TPC and RC vectors. First of all, the strongest user in the first cluster is identified by maximizing the product $\mathbf{a}_R^H(\theta_{j,0})\mathbf{H}_j\mathbf{a}_T(\phi_{j,0})$ for $j = 1, 2, \cdots, K$ in **Step 4**. Thus, the analog TPC $\mathbf{f}_{RF}^1$, analog RC $\mathbf{w}_{1,1}$ and the first user in the first cluster are jointly obtained. Next, by exploiting the angular information (AoD) ϕ_l corresponding to the remaining users $l \in \Omega^c$, the users are assigned to the first cluster. Specifically, those users are assigned to the first cluster that lie within the HPBW of the main lobe, as given in **Step 12**. This ensures that the users are efficiently allocated to clusters based on their angular positions relative to the main lobe direction. Subsequently, the analog RC for all the users in the first cluster are set as the respective receive array response vectors $\mathbf{a}_R(\theta_l)$ for $l \in \Omega_g$, as given in **Step 13**. Finally, the first cluster formulation is completed in **Step 14**. The same procedure is repeated for the formation of the remaining clusters. However, when no user is found who lies within the main beam, the algorithm expands its search to fill the desired cluster with the remaining users upon increasing the angular range by ϵ in **Step 10**.

V. Problem Formulation

The major challenge in future wireless networks is to maximize the system performance by identifying a suitable utility function that guarantees efficient utilization of the available network resources. In the context of mmWave-based NOMA systems, two widely adopted utility functions are the SE and GEE [7], [17], [25], [30]–[34].

The achievable SE of the system, measured in bits/s/Hz, is defined as the average number of bits transmitted per unit bandwidth, which can be expressed as follows

$$\text{SE}(\boldsymbol{\alpha}) = \sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} R_{g,m}. \tag{17}$$

The network SE can be written as $\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\log_2\left(1+\frac{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}\right)$. The terms $\mathcal{N}_{g,m}(\boldsymbol{\alpha})$ and $\mathcal{D}_{g,m}(\boldsymbol{\alpha})$ represent the numerator and denominator, respectively, of the downlink SINR expression derived in (7).

On the other hand, the GEE is defined as the ratio of network SE to the total power consumption, measured in bits/Joule/Hz, which can be written as

$$\text{GEE}(\boldsymbol{\alpha}) = \frac{\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\log_2\left(1+\text{SINR}_{m\leftarrow m}^{g}\right)}{\upsilon_{\text{PA}}^{-1}\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\alpha_{g,m}P_{\max}+P_c}. \tag{18}$$

Here, the quantity $\upsilon_{\text{PA}} \in (0, 1]$ denotes the power amplifier efficiency at the transmitter. The term P_c denotes the circuit power dissipation at the transceiver.

Power consumption: The term P_c constitutes the power consumed by the analog and digital circuitry of the fully connected hybrid transceiver and it is calculated as $P_c = P_{\text{FIX}}+P_{\text{TX}}+KP_{\text{RX}}+P_{\text{SIC}}\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}(m-1)$. The first three terms of P_c constitute the power consumed of the transceiver architecture for mmWave systems [3]. Recall that NOMA harness SIC at the users to detect the transmitted signal. The number of SIC operations at each user depends upon the size of each NOMA cluster and the user index in the SIC detection order in (6). Driven by the aforementioned property, this work also accounts for the SIC overhead at the users in the last term of P_c [46]. The first term P_{FIX} denotes the fixed component of the circuit power required for baseband signal processing, controlling, and site-cooling. Next, the term P_{TX} denotes the static power consumption at the transmitter RF front-end in the hybrid architecture and it is expressed as $P_{\text{TX}} = N_{RF}(2P_{\text{DAC}} + P_{\text{RF}} + N_BP_{\text{PS}}) + N_BP_{\text{PA}}$. Here, the terms P_{DAC}, P_{RF}, P_{PS} and P_{PA} denote the power dissipated at the digital-to-analog converter (DAC) for each I/Q channel, RF chain, analog phase shifter, and power amplifier, respectively. Similarly, the static power consumption at the RF front-end of each user can be expressed as $P_{\text{RX}} = 2P_{\text{ADC}} + P_{\text{RF}} + N_UP_{\text{PS}} + N_UP_{\text{LNA}}$. Here, the terms P_{ADC} and P_{LNA} denote the power dissipated at the analog-to-digital converter (ADC)

at the receiver and low-noise amplifier, respectively. Further, the quantity P_{RF} is computed as $P_{\text{RF}} = P_{\text{LO}} + 2P_M + 2P_{\text{LPF}}$, where P_M, P_{LO}, and P_{LPF} denote the power dissipated by a mixer, a local oscillator, and a low-pass filter, respectively. The insertion losses offered by passive devices, which include power splitters, power combiners and analog phase shifters, are considered to be negligible in the power consumption model. The term P_{SIC} denotes the average power dissipation during each layer of SIC decoding.

There exists a fundamental trade-off between the GEE and SE that prevents them from being simultaneously maximized. To address the challenge of providing a high throughput with limited power resources, it is imperative to simultaneously maximize both the GEE and SE. To this end, a multi-objective optimization (MOO) framework is proposed next to jointly design the GEE and the SE of mwWave NOMA systems.

A. Spectral efficiency and global energy efficiency trade-off

Due to the inability of conventional resource allocation schemes to simultaneously maximize the SE and the GEE, in this section, we conceive a MOO framework, for jointly optimizing the SE and the GEE. Considering both SE and GEE as primary objectives, a constrained bi-objective problem can be formulated as

$$\mathbf{P1}: \underset{\{\alpha_{g,m}\}_{m=1\,g=1}^{|\mathcal{U}_g|\;G}}{\text{maximize}} \quad \Big(\text{SE}\left(\boldsymbol{\alpha}\right), \text{GEE}\left(\boldsymbol{\alpha}\right)\Big) \tag{19a}$$

$$\text{subject to} \quad \sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} \alpha_{g,m} \leq 1, \tag{19b}$$

$$\alpha_{g,m} \geq 0, \quad \forall g, m \tag{19c}$$

$$\alpha_{g,i} \geq \sum_{j=i+1}^{|\mathcal{U}_g|} \alpha_{g,j}, \quad \forall g, i \tag{19d}$$

$$R_{g,m} \geq R_{g,m}^{\text{QoS}}, \quad \forall g, m\,, \tag{19e}$$

where the vector $\boldsymbol{\alpha}$ denotes the set of transmit power coefficients $\{\alpha_{g,m}\}_{m=1\,g=1}^{|\mathcal{U}_g|\;G}$. The constraint in (19b) limits the transmit power at the BS to $P_{\max}$, i.e. we have $\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} \alpha_{g,m} P_{\max} \leq P_{\max}$, and the constraint in (19c) ensures non-negative power coefficients. The constraint (19d) guarantees successful SIC operation at each user, whereas (19e) represents the rate constraint for each user, with $R_{g,m}^{\text{QoS}}$ as the minimum rate requirement of the mth user in the gth cluster. It is noteworthy here that in the absence of (19d) and (19e), the above problem can be solved by allocating a large fraction of the available power to the stronger channel users in each cluster, as has been routinely done in conventional OMA systems. This causes the remaining users in each cluster to be treated unfairly. The constraint in (19d) guarantees successful SIC operation at the users by allocating a large fraction of the power to the weaker channel users in each cluster. Together with (19d), (19e) constrains the weaker channel users to be treated fairly, while maximizing the network's spectral efficiency. Due to the existence of inter-cluster and intra-cluster interference, the objectives as well as the constraint (19e) in $\mathbf{P1}$ are non-convex, which renders solving the above maximization problem challenging.

It is important to note here that the SE in the above MOO problem is a monotonically increasing function of the transmit power and it is maximized by completely exhausting the available transmit power budget $P_{\max}$. On the other hand, GEE is maximized by using only a small portion of the available power budget, also known as green power, which prevents the SE from increasing further. Therefore, the solution space of the above MOO problem is classified as trivial or non-trivial depending on the available transmit power budget.

Trivial solution: Let $\boldsymbol{\alpha}_{\text{SE}}^*$ and $\boldsymbol{\alpha}_{\text{GEE}}^*$ be the unique unconstrained maximizers of the SE and GEE maximization problems such that $\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} \alpha_{g,m}^{\text{GEE}*} = \sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} \alpha_{g,m}^{\text{SE}*}$ [47]. This implies that the solution space of the constrained GEE maximization problem lies in the low transmit power region, where both the SE and GEE are increasing functions of the transmit power [47], [48]. This is because the SE is maximized by using all the available power, i.e., $\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} \alpha_{g,m}^{\text{SE}*} = 1$, whereas in the low transmit power regime, the GEE is maximized by maximizing the SE in the numerator of (18), rather than reducing the available power. Therefore, there exists a unique maximizer for the problem $\mathbf{P1}$ that simultaneously maximizes both the SE and the GEE.

Non-trivial solution: If $\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} \alpha_{g,m}^{\text{GEE}*} < \sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|} \alpha_{g,m}^{\text{SE}*}$ and $R_{g,m}\left(\alpha_{g,m}^{\text{GEE}*}\right) \geq R_{g,m}^{\text{QoS}}$ for $\forall g, m$, the solution space of the problem $\mathbf{P1}$ lies in the moderate and high transmit power region where the SE and GEE compete with each other [48]. In such a scenario, the problem $\mathbf{P1}$ does not possess a unique solution, but rather a set of solutions that lie on the SE-GEE trade-off region. These solutions are commonly referred to as Pareto-optimal solutions. Typically, the Pareto optimal solutions are obtained by employing a dominance test over the solution space of problem $\mathbf{P1}$. A feasible solution $\boldsymbol{\alpha}_u$ from the solution space is said to be Pareto dominant over another feasible solution $\boldsymbol{\alpha}_v$, if and only if [49]

$$\begin{aligned}&\text{SE}\left(\boldsymbol{\alpha}_u\right) \geq \text{SE}\left(\boldsymbol{\alpha}_v\right), \text{GEE}\left(\boldsymbol{\alpha}_u\right) > \text{GEE}\left(\boldsymbol{\alpha}_v\right), \text{or}\\ &\text{SE}\left(\boldsymbol{\alpha}_u\right) > \text{SE}\left(\boldsymbol{\alpha}_v\right), \text{GEE}\left(\boldsymbol{\alpha}_u\right) \geq \text{GEE}\left(\boldsymbol{\alpha}_v\right).\end{aligned} \tag{21}$$

If there exists no other solution that dominates $\boldsymbol{\alpha}_u$, then the vector $\boldsymbol{\alpha}_u$ is a member of the non-dominated set. The set is constructed by all such non-dominant solutions over the entire feasible space, termed the Pareto optimal set, and the boundary defined by mapping them onto the corresponding objective space is termed the Pareto Front. Specifically, the Pareto front is constitute of the collection of all optimal points, where neither the GEE nor the SE may be improved without degrading the other.

It is crucial to note here that the SE and the GEE are both non-convex, since the SE is the sum of the log-ratios that are rational functions of the optimization variables, whereas the GEE is a single-ratio fractional program. Therefore, solving a bi-objective problem in the face of non-convex conflicting objectives over a high-dimensional solution space necessitates significant computational resources. Consequently, it becomes extremely challenging to solve $\mathbf{P1}$. To obtain the unique globally optimal solution, the original multi-objective problem is first transformed into a single-objective problem via a

weighted sum approach [50]. Such an approach, also termed as scalarization, is a popular method of reformulating the MOO problem by transforming it into a single objective optimization (SOO) problem by assigning each objective the weighting factor $w_i \in [0,1]$ for $i = 1, 2$, so that $w_1 + w_2 = 1$. The SOO problem, therefore, can be expressed as follows

$$\mathbf{P2}: \underset{\{\alpha_{g,m}\}_{g=1\,m=1}^{G\quad|\mathcal{U}|_g}}{\text{maximize}} \quad w_1 \text{SE}(\boldsymbol{\alpha}) + w_2 \text{GEE}(\boldsymbol{\alpha}) \tag{22a}$$

$$\text{subject to} \quad (19b), (19c), (19d), (19e). \tag{22b}$$

The weights w_1 and w_2 are adjusted in accordance with the relative importance of the two objectives. It can be readily observed that the above maximization problem simplifies to the SE maximization framework by setting $w_1 = 1$ and the GEE maximization problem for $w_2 = 1$. Furthermore, a unique solution is obtained by setting different values for the weighting factors over problem $\mathbf{P2}$ and the set of all the unique solutions obtained by choosing different possible values for weighting factors constitutes the Pareto-optimal set of the original MOO problem.

To construct a weighted sum of the SE and the GEE in the above SOO problem due to their different units as well as scales, it is required to normalize the first term of (22a) with P_{sum} as follows

$$\mathbf{P3}: \underset{\{\alpha_{g,m}\}_{g=1\,m=1}^{G\quad|\mathcal{U}|_g}}{\text{maximize}} \quad \text{GEE}(\boldsymbol{\alpha}) + \beta \frac{\text{SE}(\boldsymbol{\alpha})}{P_{\text{sum}}} \tag{23a}$$

$$\text{subject to} \quad (19b), (19c), (19d), (19e). \tag{23b}$$

Here the constant $P_{\text{sum}} = \upsilon_{\text{PA}}^{-1} P_{\max} + P_c$, is defined similarly to the total network power consumption, so that both the terms in (23a) are on the same scale. Note that the problem $\mathbf{P3}$ is equivalent to $\mathbf{P2}$ with $\frac{w_1}{w_2} = \frac{\beta}{P_{\text{sum}}}$.

The problem $\mathbf{P3}$ is still challenging to solve because both the GEE and SE objectives are non-convex. Exploiting the quadratic transform (QT) technique, an iterative algorithm is developed next to attain the optimal solution of the SE-GEE trade-off problem [51].

The QT framework is adopted for transforming the non-convex objective in problem $\mathbf{P3}$ to a quadratic concave function. Using the QT framework, the numerator and denominator of the GEE in $\mathbf{P3}$ are decoupled by introducing the auxiliary variable y. The problem $\mathbf{P3}$ can therefore be recast as follows

$$\mathbf{P4}: \underset{\boldsymbol{\alpha}}{\text{maximize}} \; 2y\sqrt{\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\log_2\left(1+\frac{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}\right)} - y^2\mathcal{P}_{\text{tot}}(\boldsymbol{\alpha}) + \frac{\beta}{P_{\text{sum}}}\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\log_2\left(1+\frac{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}\right) \tag{24a}$$

$$\text{subject to} \quad \sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\alpha_{g,m} \leq 1, \tag{23b}$$

$$\alpha_{g,m} \geq 0, \quad \forall g, m \tag{23c}$$

$$\alpha_{g,i} \geq \sum_{j=1}^{i-1}\alpha_{g,j}, \; \forall g, i \tag{23d}$$

$$\log_2\left(1+\frac{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}\right) \geq R_{g,m}^{\text{QoS}}, \; \forall g, m. \tag{23e}$$

Since the optimization variables are coupled in the numerator and denominator of the term $\frac{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}$, the optimization problem $\mathbf{P4}$ is still non-convex. Therefore, QT can be applied once again to transform the problem $\mathbf{P4}$ into the equivalent problem $\mathbf{P5}$ by decoupling each of the ratios inside the logarithms as follows

$$\mathbf{P5}: \max_{\boldsymbol{\alpha}} \; 2y\sqrt{\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\log_2\left(1+2z_{g,m}\sqrt{\mathcal{N}_{g,m}(\boldsymbol{\alpha})} - z_{g,m}^2\mathcal{D}_{g,m}(\boldsymbol{\alpha})\right)} - y^2\mathcal{P}_{\text{tot}}(\boldsymbol{\alpha}) + \frac{\beta}{P_{\text{sum}}}\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\log_2\left(1+2z_{g,m}\sqrt{\mathcal{N}_{g,m}(\boldsymbol{\alpha})} - z_{g,m}^2\mathcal{D}_{g,m}(\boldsymbol{\alpha})\right) \tag{26a}$$

$$\text{subject to} \quad (23b), (23c), (23d) \tag{26b}$$

$$\log_2\left(1+2z_{g,m}\sqrt{\mathcal{N}_{g,m}(\boldsymbol{\alpha})} - z_{g,m}^2\mathcal{D}_{g,m}(\boldsymbol{\alpha})\right) \geq R_{g,m}^{\text{QoS}}, \quad \forall g, m. \tag{26c}$$

The auxiliary variables $\mathbf{z} = \{z_{g,m}\}_{m=1\,g=1}^{|\mathcal{U}_g|\quad G}$ decouple the ratios $\frac{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}$ in (24a) and (23e). For the given auxiliary variables $\{y, \mathbf{z}\}$, the problem $\mathbf{P5}$ is concave over $\boldsymbol{\alpha}$ and can be solved by using a standard toolbox such as CVX [52]. For a fixed value of $\boldsymbol{\alpha}$, the optimal values of the auxiliary variables $\{y, \mathbf{z}\}$ are computed in the closed form below [53]

$$y^* = \frac{\sqrt{\sum_{g=1}^{G}\sum_{m=1}^{|\mathcal{U}_g|}\log_2\left(1+\frac{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}\right)}}{\mathcal{P}_{\text{tot}}(\boldsymbol{\alpha})}, \text{ and} \tag{27}$$

$$z_{g,m}^* = \frac{\sqrt{\mathcal{N}_{g,m}(\boldsymbol{\alpha})}}{\mathcal{D}_{g,m}(\boldsymbol{\alpha})}. \tag{28}$$

Next, the power allocation coefficients $\boldsymbol{\alpha}$ are iteratively optimized by solving problem $\mathbf{P5}$, while updating the auxiliary variables $\{y^*, \mathbf{z}^*\}$. This process is repeated until a stationary point of the original optimization problem is obtained.

To summarize, the following steps are carried out in sequence to balance the SE-GEE trade-off: i) a bi-objective problem $\mathbf{P1}$ is developed using the MOO framework; ii) the bi-objective problem is then transformed into a single-objective optimization problem $\mathbf{P2}$ by using the weighted-sum scalarization technique; iii) next, the objectives in the $\mathbf{P2}$ optimization problem are normalized in $\mathbf{P3}$ to make them consistent with units as well as range; and iv) finally, the SE-GEE trade-off optimization problem is convexified by adopting the QT framework in $\mathbf{P5}$, which iteratively updates the auxiliary and power variables. The various steps of the QT approach for SE-GEE trade-off maximization are summarized

Table II: SYSTEM PARAMETERS

Parameter	Specification	Parameter	Specification
Noise variance, σ_η^2	1	Number of BS antennas, N_B	64
Number of user antennas, N_U	2	Number of RF chains, N_{RF}	4
Number of users, K	$\{8, 12\}$	Number of clusters, G	N_{RF}
Size of transmit beamforming codebook, C_F	$2N_B$	Size of receive beamforming codebook, C_W	32
User's minimum rate constraint, $R_{g,m}^{\text{QoS},k}$, $\forall k$	0.01 bps/Hz	Weight factor, β	1
Tolerance of Algorithm 4 termination, ϵ	10^{-3}	BS power amplifier efficiency, υ_{PA}	1
Average channel gain of LoS component, $\sigma_{k,0}^2$, for $\forall k$	$\mathcal{U}_{[1,10]}$	Average channel gain of NLoS components, $\sigma_{k,l}^2$, for $\forall k$	10^{-1}

in Algorithm 1, which computes the auxiliary variables $\{y, \mathbf{z}\}$ in **Step 3** and solves **P5** in **Step 4**.

Remarks: This treatise assumes having block fading channels between the BS and users in (1), which remain static within the coherence interval. Therefore, the near-perfect channel estimates can be obtained using sufficiently long training sequences by employing any of the existing channel estimation techniques discussed in [54]. Therefore, the proposed user clustering, beamforming, and SE-GEE optimization techniques can be effectively adapted to scenarios with imperfect CSI at a modest performance degradation.

Algorithm 4: SE-GEE trade-off maximization using QT

Input: Initialize $\boldsymbol{\alpha}^{(0)} = \{\alpha_{g,m}\}_{m=1\,g=1}^{|\mathcal{U}_g|\quad G}$ with random entries for each cluster arranged in decreasing order. Set tolerance $\epsilon > 0$ and the maximum number of iterations L.

Output: $\boldsymbol{\alpha}$ as the solution.

for $t \leftarrow 1$ **to** L **do**
- Given the latest updates of the optimization variable $\boldsymbol{\alpha}^{(t-1)}$, use (27) and (28) to calculate the auxiliary variables y^t and $\mathbf{z}^t$.
- Solve **P5** to obtain $\boldsymbol{\alpha}^{(t)}$.
- **if** $\|\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}^{(t-1)}\|^2 < \epsilon$ **then**
  - $\boldsymbol{\alpha}^* \leftarrow \boldsymbol{\alpha}^{(t)}$ and break the loop
- **else**
  - Repeat steps 2 and 3.

B. Computational complexity of the proposed algorithms

- *Complexity of similarity-based CHS scheme in Algorithm 1:* The computational complexity of Algorithm 1 is primarily dominated by **Step 5**, which involves calculating the correlation between elements of the cluster head set Ω^* and the remaining user set Ω^c. For each cluster index g, the number of correlation operations is $(K-g)g$. Consequently, the total number of correlation operations required is given by the summation $\sum_{g=1}^{G}(K-g)g$. This summation can be simplified as $K \cdot \frac{G(G+1)}{2} - \frac{G(G+1)(2G+1)}{6}$. Thus, for $K >> G$, the time complexity of **Step 5** can be expressed as $\mathcal{O}(KG^2)$, indicating that the complexity grows with the number of users K and the number of clusters G.
- *Complexity of CN-based CHS scheme in Algorithm 2:* The time complexity of Algorithm 2 is primarily determined by the number of possible cluster head choices in **Step 2** and the complexity of the SVD operation in **Step 5**. Specifically, the algorithm involves evaluating $\binom{K}{G}$ possible choices for the cluster head set with complexity $\mathcal{O}(K^G)$. For each choice, the SVD is computed, which has a time complexity of $\mathcal{O}(N_B G^2)$. Thus, the overall time complexity of the algorithm is $\mathcal{O}\left(K^G \cdot N_B G^2\right)$. This escalates for large values of K and G.
- *Complexity of joint analog precoding, digital precoding and clustering in Algorithm 3:* The complexity of Algorithm 3 is primarily driven by the user selection and updating processes. Specifically, the selection of the user having the strongest beamforming gain in each cluster involves evaluating the beamforming vectors for $K-g$ users, which incurs a complexity of $\mathcal{O}(N_B G)$ per user. Given that this process is repeated for each of the G clusters, the total complexity for user selection is $\mathcal{O}(GKN_B)$. Additionally, updating the analog precoding vector and checking the user angles contributes complexity order of $\mathcal{O}(K)$ and $\mathcal{O}(N_B G)$, respectively, per iteration. Considering the iterative nature of the algorithm, the overall complexity is $\mathcal{O}(K^2 G N_B)$.
- *Complexity of SE-GEE trade-off optimization in Algorithm 4:* The per iteration complexity of Algorithm 4 is calculated by computing the complexity of auxiliary variables $\left(y, \{z_{g,m}\}_{m=1\,g=1}^{|\mathcal{U}_g|\quad G}\right)$ in **Step 2** and solving **P5** in **Step 3**. Algorithm 4 solves the SE-GEE optimization problem involving $2K+1$ real variables and $3K+1$ constraints and has overall computational complexity of order $\mathcal{O}\left[\mu_\pi\left((5K+2)^{3.5}(2K+1)^2\right)\right]$, where μ_π is the number of iterations required by Algorithm 4 to converge.

VI. Simulation Results

This section validates the effectiveness of the proposed clustering algorithms for the mmWave-based NOMA downlink system and investigates the fundamental trade-off between SE and GEE for the proposed system. The performance of the proposed optimization Algorithm 4 is shown to strike a flexible SE-GEE trade-off. The transmit signal-to-noise ratio (SNR) at the BS is defined as $10\log_{10}\frac{P_{\max}}{\sigma_\eta^2}$ dB. The wireless channel matrix between the BS and kth user is modeled based on (1) with $L_k = 3$ NLoS components. Furthermore, equal power allocation is used among the clusters, i.e., $\frac{P_{\max}}{G}$ for all the clusters, while random power allocation (RPA) satisfying the SIC decoding condition in (6) is adopted for the users within

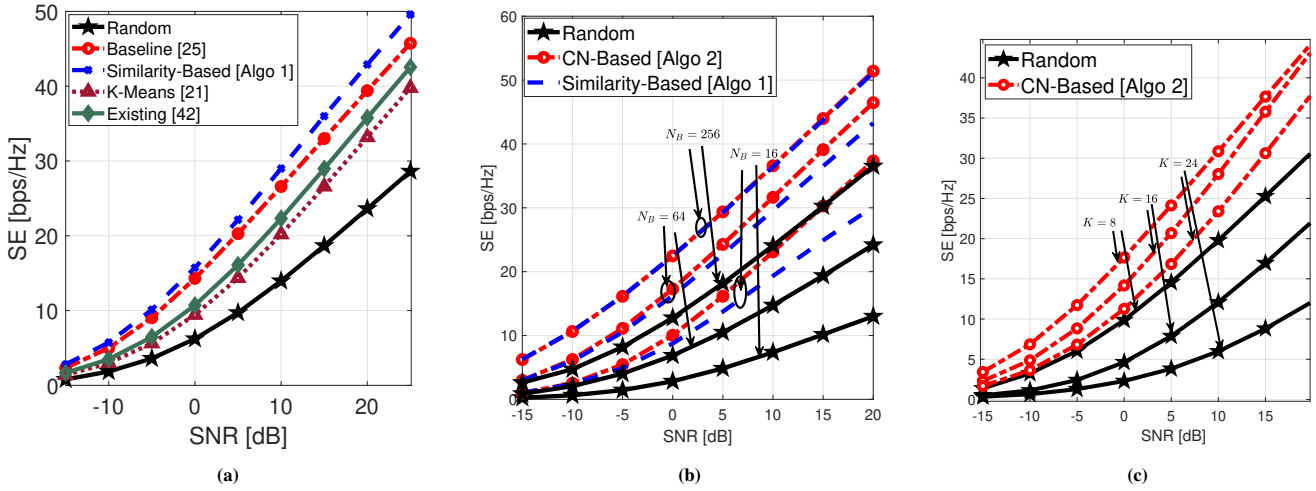


Fig. 2: (a) SE versus SNR performance of similarity-based clustering scheme in Algorithm 1; (b) SE versus SNR performance of CN-based clustering scheme in Algorithm 2; (c) SE performance of CN-based clustering scheme with different number of users K.

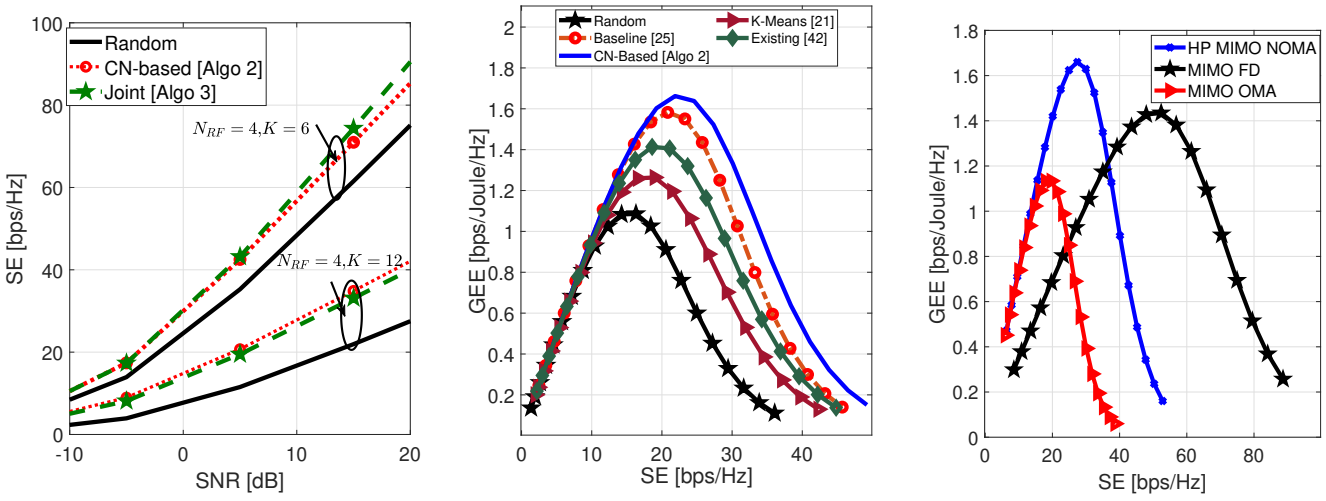


Fig. 3: (a) Comparison of CN-based scheme with joint clustering for different values of $K - N_{RF}$; (b) Comparison of GEE versus SE trade-off performance for the different state-of-the-art clustering schemes with $N_B = 64$, $K = 10$ and $N_{RF} = 4$; (c) Comparison of GEE versus SE trade-off performance for the different state-of-the-art beamforming techniques with $N_B = 64$, $K = 10$ and $N_{RF} = 4$.

each cluster. The parameters P_{FIX}, P_{RF}, P_{PS}, P_{LNA} and P_{SIC} of circuit power consumption are set as 5 W, 31.6 mW, 2 mW, 39 mW and 0.2 W, respectively. The remaining components of the power consumption model are set as given in [3]. All the simulation parameters are summarized at a glance in Table II, unless stated otherwise.

A. Clustering performance

The efficacy of the proposed user clustering in Section IV-B along with the cluster head selection techniques proposed in Algorithms 1 and 2, in terms of SE, are validated in this subsection by comparing them to both the random clustering, K-means [21] and the other state-of-the-art clustering techniques of [25], [42].

Fig. 2a illustrates the SE versus SNR performance of the proposed similarity-based dynamic cluster head selection Algorithm 1 by comparing it to the i) random; ii) K-means [21]; and iii) existing state-of-the-art user clustering schemes of [25], [42]. For this study, the number of users K and the number of RF chains N_{RF} are set as 12 and 4, respectively. It is noted that the proposed Algorithm 1 shows a 85% performance gain compared to the random clustering technique, while the gain over the baseline technique in [25] is 10%. It is also observed that the proposed clustering scheme surpasses both the K-means [21] and the existing user clustering techniques [25], [42]. The SE improvement can be justified by the fact that when no potential user is found to be uncorrelated, the existing clustering schemes assign the correlated users as CHs for different clusters, whereas, the proposed clustering scheme guarantees the uncorrelated users to be chosen as CHs by dynamically adjusting the number of clusters.

Fig. 2b compares the SE performance of the CN-based clustering technique proposed in Algorithm 2 to the similarity-based clustering scheme of Algorithm 1 and to random clustering. The impact of the number of transmit antennas N_B is also investigated. Similar to the previous studies, the number of users, K, and the number of RF chains, N_{RF}

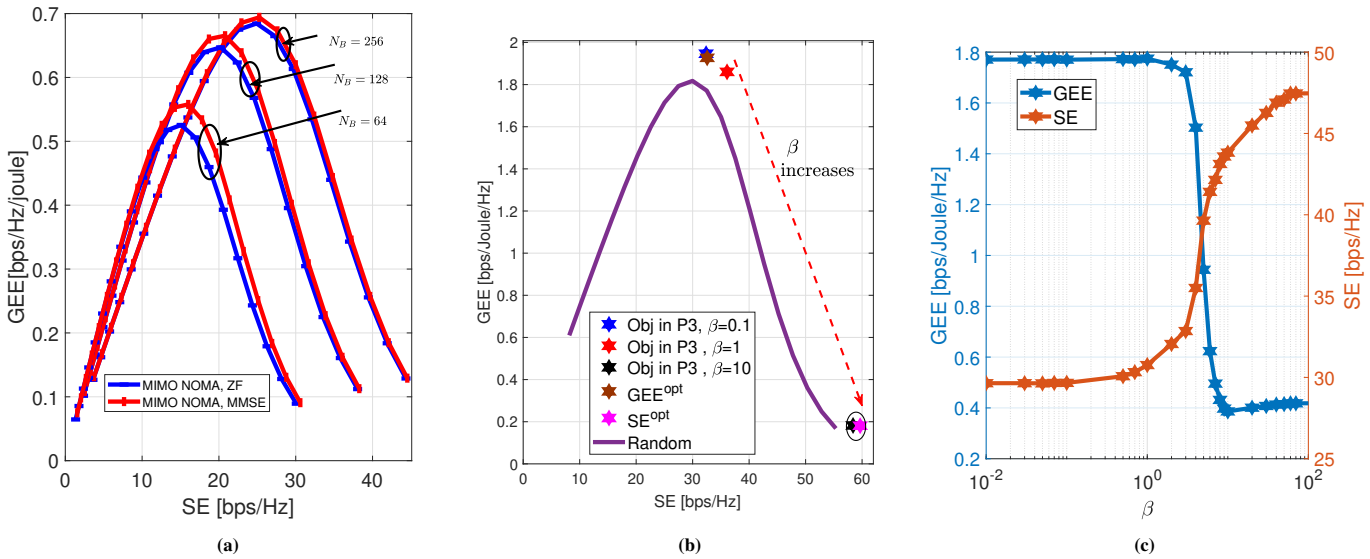


Fig. 4: (a) Comparison of GEE versus SE trade-off performance for the ZF-based and MMSE-based digital precoding techniques with $N_B = \{64, 128, 256\}$, $K = 16$, and $N_{RF} = 4$; (b) GEE-SE trade-off performance of mmWave-based NOMA systems with $K = 8$, $N_{RF} = 4$, and $N_B = 64$ for the proposed OPA scheme in Algorithm 4; (c) GEE and SE performance of Algorithm 4 by varying β for the fixed transmit SNR $P_{\max} = 15$ dB with $K = 8$, $N_{RF} = 4$, and $N_B = 64$.

are set as 12 and 4, respectively. It can be seen that for $N_B = 16$, the relative performance of the proposed CN-based clustering technique is better compared to both the similarity-based and to the random clustering techniques. On the other hand, as N_B increases, both the CN-based and the similarity-based clustering techniques perform similarly. This is because uncorrelated users of the network are easily obtained, as N_B increases. Consequently, both the proposed clustering techniques effectively identify the uncorrelated users for different clusters.

Fig. 2c demonstrates the impact of the number of users K by comparing the proposed CN-based clustering in Algortihm 2 to random clustering. The number of BS antennae N_B and the number of RF chains N_{RF} are fixed to 64 and 4, respectively. It is observed that the SE performance of both the CN-based clustering and the random clustering techniques reduces as K increases. This is because both the inter-cluster and intra-cluster interferences increase when the number of users in the network increases. It is also evident that the relative performance of the proposed clustering over random clustering improves, as K increases. This is because when $(K - N_{RF})$ reduces, the number of users assigned in each cluster decreases, thereby reducing the dominance of the clustering algorithm in improving spectral efficiency.

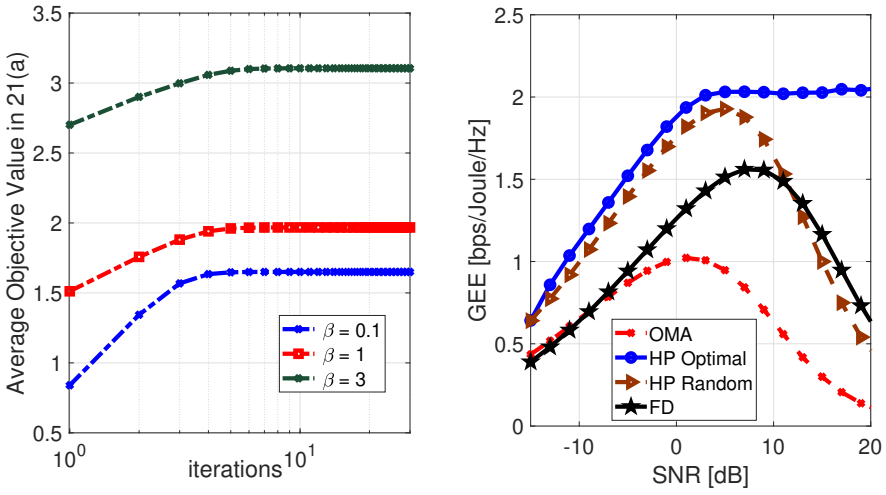
B. Performance of the joint algorithm

Considering only the LoS component of the mmWave channel in (1), i.e. $L_k = 0$, the SE performance of the proposed joint analog precoding, combining, and user clustering technique of Algorithm 3 is investigated now. Fig. 3a plots the SE versus SNR for the proposed joint clustering technique by varying the value of $(K - N_{RF})$. The number of BS antennas N_B is set to 64. For comparison, the SE performance of CN-based and random user clustering are also plotted. It is observed that when a small number of active users are present in the network, i.e., for a small value of $(K - N_{RF})$, the joint technique of Algorithm 3 yields improved performance compared to the clustering techniques provided in Algorithm 1 and 2. Conversely, for a large number of active users in the network, i.e, for a high value of $(K - N_{RF})$, the CN-based clustering technique of Algorithm 2 has an edge over the joint clustering technique. *The system designer can exploit this information to select an appropriate clustering algorithm based on the specific number of users in the network.*

C. SE-GEE trade-off performance

Fig 3b characterizes the SE-GEE trade-off performance of the proposed clustering Algorithm 2 by comparing it with i) random clustering; ii) baseline clustering [25]; iii) K-means [21]; and iv) other existing [42] clustering schemes. For this setup, the ZF-based HP precoding and RPA schemes are employed, with the SNR range set between -15 dB to 25 dB. The performance comparison with Algorithm 1 is omitted for this investigation as it always performs inferior to Algorithm 2. It is discovered that the SE-GEE trade-off of the proposed Algorithm 2 is always better than that of the other existing state-of-the-art clustering techniques. This shows the effectiveness of the proposed clustering scheme in enhancing both the SE as well as the GEE of the network.

In the next subsections, we evaluate the SE-GEE trade-off of the proposed HP-based MIMO NOMA network (HP MIMO NOMA) by comparing it to that of its MIMO-OMA counterpart (MIMO OMA) and fully digital MIMO counterpart (MIMO FD). Specifically, the MIMO-OMA system exploits the TDMA scheme within each cluster, wherein the BS transmits multiple user signals in each cluster on a time-sharing basis. The SE expression of the MIMO-OMA system is given by $\sum_{g=1}^{G} \frac{1}{|\mathcal{U}_g|} \sum_{m=1}^{|\mathcal{U}_g|} \log_2(1 + \text{SINR}_m^g)$, where $\frac{1}{|\mathcal{U}_g|}$ is the equal fraction of time allocated to all users within the gth cluster. On the other hand, the FD MIMO system exploits a



(a) **(b)**

Fig. 5: (a) Convergence behavior of the proposed optimization Algorithm 4 for different values of weight factor β; (b) GEE versus SNR performance of the proposed mmWave-based NOMA system compared with OMA and FD counterparts by setting $N_B = 64$.

fully digital architecture, where each RF chain is connected to each antenna element and only digital precoding is performed at the BS. In the FD MIMO scheme, each RF chain serves a single user only. The SE expression of the FD scheme is $\sum_{k=1}^{K} \log_2(1 + \text{SINR}_k)$. Here, SINR_m^g and SINR_k are the SINR expressions for the MIMO-OMA and FD counterparts, respectively. Furthermore, the power consumption models of the corresponding MIMO-OMA and fully digital architecture are given as $P_{c,\text{FD}} = P_{\text{FIX}} + 2N_B P_{\text{DAC}} + P_{\text{RF}} + N_B P_{\text{PA}} + K P_{\text{RX}}$ [3] and $P_{c,\text{OMA}} = P_{\text{FIX}} + P_{\text{TX}} + K P_{\text{RX}}$, respectively.

Fig. 3c compares the SE versus GEE trade-off of the proposed mmWave-based MIMO NOMA system to the different beamforming techniques. The SNR range is set as -15 dB to 25 dB for this simulation setup. It is observed that the proposed HP MIMO NOMA system has an improved SE-GEE trade-off compared to the MIMO OMA system. This is because the HP MIMO-NOMA system attains significantly higher values of GEE and SE compared to the MIMO-OMA system. It is also observed that the FD MIMO scheme has a poor SE-GEE trade-off for low SNR values compared to both the HP MIMO NOMA and MIMO OMA counterparts. This is because, in the low SNR regime, the slope of the SE-GEE trade-off curve is inversely proportional to the circuit power consumption [55] and the FD scheme consumes higher power than the HP MIMO NOMA and MIMO OMA systems. On the other hand, the MIMO FD scheme attains a superior SE-GEE trade-off at higher SNR values than the other two scenarios, which can be attributed to the fact that in the high SNR regime the slope of the SE-GEE plot is directly proportional to the multiplexing gain [55], coupled with the fact that the MIMO FD scheme has a higher multiplexing gain compared to the other systems under consideration.

Fig. 4a compares the SE versus GEE trade-off of the proposed ZF-based MIMO NOMA system (MIMO NOMA, ZF) to that of its MMSE-based digital counterpart (MIMO NOMA, MMSE) for different antenna configurations. The SNR range for this simulation setup is ranges from -15 dB to 25 dB. It is first noted that the MMSE-based digital precoding provides minimal improvement in the GEE-SE trade-off compared to the ZF-based digital precoding. This is because digital precoding is designed to cancel interference only among cluster heads. However, both the ZF- and the MMSE-based digital precoding perform ineffectively for the remaining users, resulting in the undesired phenomenon that significant inter-cluster interference persists irrespective of the specific choice of precoding. Furthermore, it is observed that as the number of antennas increases, the effectiveness of MMSE precoding over ZF-based digital precoding diminishes.

Fig. 4b characterizes the efficiency of the optimization framework proposed in Section V to balance the GEE-SE trade-off in the proposed system for different values of the weighting factor β. Similar to the previous study, the transmit SNR is varied from -15 dB to 25 dB. It is demonstrated that as β increases, the optimal value of the GEE-SE trade-off achieved by Algorithm 4 allocates a higher weight to the SE and for $\beta = 10$, the optimal value attained is close to that when only the SE is optimized i.e., SE^{opt}. Similarly, for $\beta = 0.1$, the optimal value attained approximately equals the one obtained when only the GEE is optimized i.e., GEE^{opt}. For $\beta = 1$, the problem $\mathbf{P3}$ assigns equal weights to both the components, and the performance gains attained by Algorithm 4 are 11% and 7% for the GEE and SE, respectively.

Fig 4c demonstrates the impact of the weighting factor β on the GEE and SE of the proposed system, when $P_{\max} = 15$ dB. It is observed that increasing β improves the SE, while leading to a reduction in the GEE of the network. This is because a low value of β assigns a higher consideration to the GEE, whereas a high value of β gives precedence to the SE. Thus, as $\beta \to 0$ and $\beta \to 100$, applying Algorithm 4 leads to GEE and SE maximization, respectively. Therefore, the GEE-SE trade-off adopted is more comprehensive in comparison to the conventional approach that optimizes only one of these objectives. Thus, by choosing an appropriate weighting factor β, the BS has the flexibility to trade-off the GEE against the SE, as per the requirements of the system.

Fig 5a investigates the convergence performance of the proposed technique by plotting the value of the objective function versus the number of iterations, for different values

of the weighting factor β. The transmit SNR $P_{\max}$ for this setting is fixed to be 10 dB. It is observed that a non-decreasing sequence of objective values is obtained as the number of iterations increases, with the algorithm converging to its optimal value within as few as 10 iterations. Interestingly, it is also observed that the convergence speed decreases upon increasing the weight β.

D. GEE performance

Fig. 5b characterizes the GEE performance of the scheme advocated, while setting $\beta = 0$ in the optimization problem $\mathbf{P3}$. The number of users and the RF chains are set to $K = 8$ and $N_{RF} = 4$, respectively. It is observed that the GEE obtained for the optimal power allocation (OPA) scheme is enhanced over the one obtained via the naive random power allocation (RPA) procedure for all SNR values. It is also observed that the GEE performance of the proposed mmWave-based NOMA scheme always surpasses that of its OMA counterpart, while being inferior in comparison to its FD counterpart in the high SNR regime. This is explained by the fact that both the circuit power consumption and spatial multiplexing gain of the FD technique are higher than that of the HP scheme. However, in the low SNR regime, the SE gap between the FD and HP techniques is smaller, whereby the larger circuit power consumption of the former results in lower GEE in contrast to the latter.

VII. Summary and Conclusions

This work investigated the SE-GEE trade-off in a mmWave NOMA network by designing hybrid precoding/combining, user clustering, and power optimization techniques. A joint hybrid precoding/combining was designed next, followed by the novel clustering techniques developed for minimizing inter-cluster interference. Thereafter, a MOO problem was formulated by considering the SE and GEE objectives, with the aim of achieving the optimal SE-GEE trade-off in the system. The above MOO problem was transformed into a non-convex SOO problem using a weight-sum approach, and further converted to concave optimization via QT. Our simulation findings demonstrated the impact of the proposed user clustering schemes on the i) SE; ii) GEE; and iii) SE-GEE trade-off of the proposed mmWave-NOMA network by comparing them with random and other existing state-of-the-art clustering techniques [21], [25], [42]. To summarize the performance of the proposed user clustering algorithms, simulation findings validated that the CN-based user clustering algorithm outperformed both random and similarity-based clustering with a moderate number of antennas and a large number of users. However, as the number of antennas increased, the SE performance of the random and similarity-based clustering started to converge with that of the CN-based clustering. Similarly, when the number of users became small, the dominance of the proposed clustering algorithm to enhance SE was reduced, limiting the advantage of CN-based clustering over the random and similarity-based methods. Moreover, the effectiveness of the proposed optimization framework was validated through simulations, demonstrating its ability to balance GEE and SE by fine-tuning the weights.

References

[1] C.-X. Wang, X. You, X. Gao *et al.*, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 905–974, 2023.

[2] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar *et al.*, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2018.

[3] L. N. Ribeiro, S. Schwarz, M. Rupp *et al.*, "Energy efficiency of mmwave massive MIMO precoding with low-resolution DACs," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 2, pp. 298–312, 2018.

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

[5] J. Zhu, Q. Li, Z. Liu *et al.*, "Enhanced User Grouping and Power Allocation for Hybrid mmWave MIMO-NOMA Systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 2034–2050, 2022.

[6] L. Dai, B. Wang, Z. Ding *et al.*, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 2018.

[7] X. Hu, C. Zhong, X. Chen *et al.*, "Cluster grouping and power control for angle-domain mmwave MIMO NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 5, pp. 1167–1180, 2019.

[8] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, 2016.

[9] Y. Cao, S. Wang, M. Jin *et al.*, "Joint user grouping and power optimization for secure mmwave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3307–3320, 2022.

[10] S. Norouzi, B. Champagne, and Y. Cai, "Joint optimization framework for user clustering, downlink beamforming, and power allocation in MIMO NOMA systems," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 214–228, 2023.

[11] L. Pang, W. Wu, Y. Zhang *et al.*, "Joint power allocation and hybrid beamforming for downlink mmwave-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10 173–10 184, 2021.

[12] L. Zhu, J. Zhang, Z. Xiao *et al.*, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, 2019.

[13] J. Zhu, Q. Li, Z. Liu *et al.*, "Enhanced user grouping and power allocation for hybrid mmwave MIMO-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 2034–2050, 2022.

[14] J. Ren, Z. Wang, M. Xu *et al.*, "An EM-based user clustering method in non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8422–8434, 2019.

[15] K. Wang, J. Cui, Z. Ding *et al.*, "Stackelberg game for user clustering and power allocation in millimeter wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2842–2857, 2019.

[16] M. R. G. Aghdam, R. Abdolee, F. A. Azhiri *et al.*, "Random user pairing in massive-MIMO-NOMA transmission systems based on mmwave," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–6.

[17] X. Qi, G. Xie, and Y. Liu, "Energy-efficient power allocation in multi-user mmwave-NOMA systems with finite resolution analog precoding," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 3750–3759, 2022.

[18] M. Zhang, Y. Guo, L. Salan *et al.*, "Proportional fair scheduling for downlink mmwave multi-user MISO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6308–6321, 2022.

[19] S. Solaiman, L. Nassef, and E. Fadel, "User clustering and optimized power allocation for D2D communications at mmwave underlaying MIMO-NOMA cellular networks," *IEEE Access*, vol. 9, pp. 57 726–57 742, 2021.

[20] A. Celik, M.-C. Tsai, R. M. Radaydeh *et al.*, "Distributed user clustering and resource allocation for imperfect NOMA in heterogeneous networks," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7211–7227, 2019.

[21] M. Liu, J. Zhang, K. Xiong *et al.*, "Effective user clustering and power control for multiantenna uplink NOMA transmission," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8995–9009, 2022.

[22] M. Katwe, K. Singh, P. K. Sharma *et al.*, "Dynamic user clustering and optimal power allocation in UAV-assisted full-duplex hybrid NOMA system," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2573–2590, 2022.

[23] M. K. Naeem, R. Abozariba, M. Asaduzzaman *et al.*, "Mobility support for MIMO-NOMA user clustering in next-generation wireless networks," *IEEE Transactions on Mobile Computing*, vol. 22, no. 10, pp. 6011–6026, 2023.

[24] I. Khaled, C. Langlais, A. E. Falou *et al.*, "Multi-user angle-domain MIMO-NOMA system for mmwave communications," *IEEE Access*, vol. 9, pp. 129 443–129 459, 2021.

[25] L. Dai, B. Wang, M. Peng *et al.*, "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.

[26] D. Marasinghe, N. Jayaweera, N. Rajatheva *et al.*, "Hierarchical user clustering for mmwave-NOMA systems," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[27] B. Lim, W. J. Yun, J. Kim *et al.*, "Joint user clustering, beamforming, and power allocation for mmwave-NOMA with imperfect SIC," *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2025–2038, 2024.

[28] J. Tang, D. K. C. So, E. Alsusa *et al.*, "Resource efficiency: A new paradigm on energy efficiency and spectral efficiency tradeoff," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4656–4669, 2014.

[29] J. Ghosh, "A trade-off between energy efficiency and spectral efficiency in macro-femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 10 914–10 924, 2020.

[30] Z. Xiao, L. Zhu, J. Choi *et al.*, "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, 2018.

[31] W. Shao, S. Zhang, H. Li *et al.*, "Angle-domain NOMA over multicell millimeter wave massive MIMO networks," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2277–2292, 2020.

[32] F. Zhao, W. Hao, L. Shen *et al.*, "Secure energy efficiency transmission for mmwave-NOMA system," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2226–2229, 2021.

[33] A. J. Muhammed, H. Chen, A. M. Seid *et al.*, "Energy-efficient resource allocation for NOMA Hetnets in millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3790–3804, 2023.

[34] Z. Wei, D. W. K. Ng, and J. Yuan, "Noma for hybrid mmwave communication systems with beamwidth control," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 3, pp. 567–583, 2019.

[35] L. You, J. Xiong, A. Zappone *et al.*, "Spectral efficiency and energy efficiency tradeoff in massive MIMO downlink transmission with statistical CSIT," *IEEE Trans. Signal Process.*, vol. 68, pp. 2645–2659, 2020.

[36] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4816–4828, 2022.

[37] J. Wang, C. Jiang, H. Zhang *et al.*, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 2020.

[38] Y. Yang, S. Dang, M. Wen *et al.*, "Millimeter wave MIMO-OFDM with index modulation: A pareto paradigm on spectral- energy efficiency trade-off," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6371–6386, 2021.

[39] H. M. Al-Obiedollah, K. Cumanan, J. Thiyagalingam *et al.*, "Spectral-energy efficiency trade-off-based beamforming design for MISO non-orthogonal multiple access systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6593–6606, 2020.

[40] W. U. Khan, F. Jameel, T. Ristaniemi *et al.*, "Joint spectral and energy efficiency optimization for downlink NOMA networks," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 645–656, 2020.

[41] F. Xu and H. Zhang, "Energy efficiency and spectral efficiency tradeoff in IRS-assisted downlink mmwave NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1433–1437, 2022.

[42] M. Shili, M. Hajjaj, and M. L. Ammari, "User clustering and power allocation for massive MIMO with NOMA-inspired cognitive radio," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7656–7664, 2022.

[43] Y. Liu, H. Xing, C. Pan *et al.*, "Multiple-antenna-assisted non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 17–23, 2018.

[44] B. Wang, L. Dai, Z. Wang *et al.*, "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, 2017.

[45] S. Hur, T. Kim, D. J. Love *et al.*, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, 2013.

[46] S. R. C. Magalhes, S. Bayhan, and G. Heijenk, "Impact of power consumption models on the energy efficiency of downlink NOMA systems," *IEEE Trans. Green Commun. Netw*, vol. 7, no. 4, pp. 1739–1753, 2023.

[47] M. Di Renzo, A. Zappone, T. T. Lam *et al.*, "System-level modeling and optimization of the energy efficiency in cellular networksA stochastic geometry framework," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2539–2556, 2018.

[48] ——, "Spectral-energy efficiency pareto front in cellular networks: A stochastic geometry framework," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 424–427, 2019.

[49] R. Wang, Z. Zhou, H. Ishibuchi *et al.*, "Localized weighted sum method for many-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 22, no. 1, pp. 3–18, 2018.

[50] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: new insights," *Structural and Multidisciplinary Optimization*, vol. 41, no. 6, pp. 853–862, Jun 2010. [Online]. Available: https://doi.org/10.1007/s00158-009-0460-7

[51] K. Shen and W. Yu, "Fractional Programming for Communication SystemsPart I: Power Control and Beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, 2018.

[52] M. Grant and S. Boyd, "CVX: matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[53] X. Zhu, Z. Wang, L. Dai *et al.*, "Adaptive hybrid precoding for multiuser massive MIMO," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 776–779, 2016.

[54] K. Hassan, M. Masarra, M. Zwingelstein *et al.*, "Channel estimation techniques for millimeter-wave communication systems: Achievements and challenges," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 1336–1363, 2020.

[55] Z. Xu, Z. Pan, and I. Chih-Lin, "Fundamental properties of the EE-SE relationship," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, 2014, pp. 1115–1120.