Received 14 February 2023; accepted 1 March 2023. Date of publication 8 March 2023; date of current version 21 March 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3254157

Angle-Domain Hybrid Beamforming-Based mmWave Massive MIMO-NOMA Systems

ISRAA KHALED [1,2], AMMAR EL FALOU [2,3] (Senior Member, IEEE), CHARLOTTE LANGLAIS [1], MICHEL JEZEQUEL [1], AND BACHAR ELHASSAN [2] (Member, IEEE)

[1] IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238 Brest, France

[2] Faculty of Engineering, Lebanese University, Tripoli 1300, Lebanon

[3] CEMSE, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

CORRESPONDING AUTHOR: I. KHALED (e-mail: israa.khaled@imt-atlantique.fr)

This work was supported in part by Institut Mines-Télécom (IMT) Atlantique and in part by the Lebanese University Research Fund Program and the AZM Association.

ABSTRACT The millimeter-wave (mmWave) large-scale antenna arrays (LSAAs) systems play a vital role in increasing the beamforming (BF) gain and acquiring highly directional propagation. Recently, non-orthogonal multiple access (NOMA) has been integrated into these systems to manage massive connectivity and achieve spectral-efficient communications. This paper focuses on angle-domain (AD) hybrid beamforming (BF) for mmWave LSAAs and NOMA systems, thanks to the low complexity, power consumption, and channel estimation overhead. However, with limited radio-frequency chains, the hybrid BF-based single-beam (SB)-NOMA scheme generating a single beam to serve the NOMA users fails to exploit the multi-user diversity due to narrow beams with LSAAs. To tackle this limitation, we design schemes offering additional degrees of freedom. More importantly, they require only the knowledge of angular information and are suitable for either linear or rectangular antenna arrays, unlike those proposed in the literature. The first scheme exploits the time-domain resources to schedule groups having high spatial interference within distinct time slots. To minimize the need for fast and precise synchronization when applying time division multiple access (TDMA) with mmWave NOMA, we leverage the multi-beam (MB)-NOMA framework. And we propose a joint SB- and MB-NOMA scheme to benefit from NOMA multi-user diversity, whatever the cell load and the users' positions. Using the New York University channel simulator (NYUSIM), we further validate the performance of the proposed schemes compared to the solution proposed in the literature and others using fully digital BF. Specifically, the proposed TDMA-based scheme achieves a sum-rate gain of up to 83% over the TDMA-based one existing in the literature. Moreover, we verify the superiority of applying both SB- and MB-NOMA instead of only MB-NOMA.

INDEX TERMS Millimeter-wave communications, massive multiple-input multiple-output, hybrid beamforming, non-orthogonal multiple access, angle-domain information.

I. INTRODUCTION

A. MOTIVATIONS

MOVING to the millimeter-wave (mmWave) band is a promising solution offering a vast amount of unused spectrum and enabling high data rates for future network systems [1], [2]. Channel measurements reveal that mmWave signals suffer significant propagation losses and are best suited for small cell coverage [1]. Fortunately, the corresponding tiny wavelength facilitates the implementation of large-scale antenna arrays (LSAAs) in a small form factor. And the massive multiple-input multiple-output (MIMO) technology combats the severe path loss and blockage with a high beamforming (BF) gain. However, using fully digital BF (DBF), the number of required radio-frequency (RF) chains is equal to that of antennas, which is unrealistic with massive MIMO (mMIMO), especially in the context of mmWaves,

due to its high energy consumption and unaffordable hardware complexity and cost [3], [4]. Alternatively, hybrid BF (HBF), with both sub- and full-connected structures, is considered an effective solution for the possible implementation of mmWave mMIMO systems [4], [5]. This technique separates the signal processing into a low-dimensional digital precoder (addressing a small number of RF chains) and a high-dimensional analog precoder in the RF band to increase the array gain. On the other hand, mMIMO requires a massive channel state information (CSI) overhead, which generates a long training sequence and causes delays in signaling. Exploiting the main features of mmWave channels, i.e., high directionality and significant blockage, the user's angle-of-departure (AoD), w.r.t. the base station (BS), is considered a promising partial CSI for mmWave mMIMO systems [6]. Indeed, it only depends on the direction of the line of sight (LoS), so it varies slowly over time and is not proportional to the antennas number. This information has been the subject of much attention from the research community in various communication scenarios, leading to several proposed angle-domain channel estimation and tracking techniques like [7] for high-speed railways, [8] for indoor 60 GHz mMIMO systems, [9] for mmWave hybrid mMIMO systems, etc.

Massive connectivity is a critical requirement in future cellular networks. However, conventional orthogonal multiple access (OMA) techniques, such as space division multiple access (SDMA), time division multiple access (TDMA), frequency division multiple access (FDMA), etc., support a single user in the same space-time-frequency-code resource block (RB). Recently, non-OMA (NOMA) has emerged to improve network capacity and accommodate massive connectivity by exploiting additional non-orthogonal resources (e.g., power domain resources). Using superposition coding on the transmitter and successive interference cancellation (SIC) at the receiver, the BS can serve multiple users with different channel conditions at the same orthogonal RB.

Motivated by these observations, we focus on angle-domain (AD) mmWave HBF-based mMIMO-NOMA systems to take the benefits of three essential technologies, namely mmWave, mMIMO, and NOMA, with only the knowledge of angular information and low-complex BF technique. Next, we review the various HBF-based MIMO-NOMA systems presented in the literature.

B. RELATED WORKS

HBF-based mMIMO-NOMA systems are extensively studied in the literature to reduce energy consumption and hardware complexity [10], [11], [12], [13], [14], [15], [16], [17], [18]. In [10], the authors design a new HBF-based mmWave beamspace MIMO-NOMA scheme to support more users than the RF chains. And an iterative algorithm that solves the power allocation optimization problem is designed to maximize the system sum-rate. In addition, the authors in [11] apply NOMA with the fully-connected HBF and develop a new HBF technique by modifying the conventional block diagonalization scheme. This was done to improve the achievable spectral efficiency by reducing the co-channel interference. The authors in [12] propose a user clustering algorithm according to the users' channel correlation and then formulate a joint HBF and power allocation problem for maximizing the system sum-rate under a minimum rate constraint. Yet, this problem is non-convex. To this end, they first apply an arbitrary fixed HBF and find the power allocation solution. Then, they set the analog precoder technique and create a digital precoder that reduces inter-group interference by applying the approximate zero-forcing (ZF) method. Finally, they solve the analog precoder problem using the constant-modulus constraint with a proposed boundary-compressed particle swarm optimization algorithm. In [13], the authors consider the sub- and fully-connected structures in the RF stage and ZF in the digital baseband. And they propose an iterative low-complex power allocation algorithm to maximize energy efficiency. To improve spectrum- and energy-efficiencies, the authors in [14], [15] integrate HBF-based MIMO-NOMA with simultaneous wireless information and power transfer (SWIPT). The authors in [16] propose an optimal analog precoder with the aid of ZF in the baseband to maximize both the sum-rate and the energy efficiency. This was done for the two hybrid structures operating under LoS and non-LoS (NLoS) mmWave environments. In [17], the authors use the signal-to-leakage ratio (SLNR) for the first time as the performance index to tackle the issues of resource optimization in HBF-based mMIMO-NOMA. Specifically, they formulate a joint optimization problem concerning power allocation and HBF in order to maximize the minimum user SLNR, thus ensuring fairness between users.

However, previous works of HBF-based MIMO-NOMA adopt very complex digital precoders and require complete knowledge of CSI. To reduce the complexity of signal processing in baseband and the overhead of channel estimation with LSAAs, we consider AD MIMO-NOMA. Most existing works on AD MIMO-NOMA using analog BF (ABF), DBF, or HBF partition the users into groups according to their angle difference and form a single beam toward each group [6], [19], [20], [21], [22]. These schemes are referred to here as single-beam (SB)-NOMA. Unlike these works, we defined in [23], [24] a spatial interference metric, denoted as β, built based on the array factor definition. And we developed β-based 2-user and multi-user clustering algorithms to schedule the users with high spatial interference in the same SB-NOMA group. In these works, we considered the AD-DBF technique, while in this paper, we focus on implementing the AD-HBF technique with LSAAs. However, the beams are very narrow in such systems, so using SB-NOMA, only users with similar AoDs are served simultaneously in the same group. This means that in mmWave hybrid systems, due to the limited number of RF chains, SB-NOMA with LSAAs cannot provide connectivity to all users, especially in congested cells.

To exploit more degrees of freedom (DoFs) and accommodate more users using limited RF chains, Hu et al. in [19] suggest a joint SB-NOMA and TDMA scheme, denoted here as SB-NOMA-TDMA. Specifically, they cluster the users within single-user (SU) and SB-NOMA groups with angles belonging to a predefined set. Subsequently, a group clustering algorithm is used in the time domain to limit interference in each time slot while considering that increasing spatial direction distance can significantly reduce inter-group interference. To take advantage of the decreased complexity and cost from HBF-based SB-NOMA-TDMA, fast and accurate synchronization in time between users is essential since mmWave communication generally offers a high symbol rate. Moreover, there is more synchronization complexity on both sides of the BS and receiver.

Recently, the authors in [25] discuss the concept of the multi-beam (MB) NOMA framework in mmWave hybrid systems, and the authors in [26] offer its implementation details. Specifically, they propose a beam-splitting technique (BST) that divides the entire transmit array into various sub-arrays. Accordingly, within the same RF chain, the BS can generate multiple analog beams toward multiple NOMA users with arbitrary AoDs. The authors in [26] show that MB-NOMA efficiently exploits the multi-user diversity, unlike SB-NOMA, by performing NOMA transmission even when the users have separated AoDs, especially with LSAAs. In [27], they designed a suboptimal two-stage resource allocation that maximizes the system sum-rate based on a full CSI. Considering only ABF in the first stage, they suggest a joint user grouping and antenna allocation algorithm that maximizes the conditional system sum-rate by leveraging the coalition formation game theory. In the second stage, they adopt ZF in the digital baseband. Subsequently, they formulate a non-convex power allocation optimization problem to maximize the system sum-rate subject to the QoS constraints. A suboptimal solution is devised to solve this problem. However, the scheme in [26], [27] requires complete knowledge of CSI, and both [19] and [26], [27] are designed only for uniform linear array (ULA) architectures.

C. CONTRIBUTIONS

In this paper, we focus on AD HBF-based mMIMO-NOMA systems where only the users' angles are known at the BS. Due to the narrowness of beams with LSAAs and the limited number of RF chains, SB-NOMA fails to provide full connectivity to all users in overloaded scenarios. To address this issue, we propose two schemes offering additional DoFs for the SB-NOMA scheme. While the former leverages the time-domain resources using TDMA, the latter only adopts NOMA. More importantly, the proposed schemes only require the knowledge of angular information and use either ULA or uniform rectangular array (URA). The main contributions of our work are as follows

- We propose two-phase schemes, namely, joint β-based SB-NOMA and TDMA and joint AD SB- and MB-NOMA schemes addressing the limitation of SB-NOMA with hybrid LSAAs to provide more DoFs in overloaded scenarios.
- The joint β-based SB-NOMA and TDMA scheme proposed in this work was inspired by the one introduced in [19]. In contrast, we utilize the β-based spatial interference metric derived in [24], which is based on the array factor to accurately calculate inter-group interference for any uniform array architecture, such as ULA, URA, or uniform circular array (UCA), rather than using only the angular distance as done in [19].
- Unlike previous work on AD HBF-based mMIMO-NOMA, the proposed joint AD SB- and MB-NOMA scheme leverages the potential of SB-NOMA when users are close to each other and the ability of MB-NOMA to accommodate several users with distinct AoDs. This extends our prior work in [28], which focused solely on ULA.

D. ORGANIZATION

The rest of this paper is organized as follows. The system model of HBF-based MIMO-NOMA is presented in Section II. The performance analysis in terms of the sum-rate of SDMA, SB-NOMA, and MB-NOMA is studied in Section III. The proposed joint β-based SB-NOMA and TDMA, and joint AD-MB-SB-NOMA schemes are presented in Sections IV and III, respectively. The performance of the proposed schemes is evaluated in Section VI. Finally, a summary is given in Section VII.

E. NOTATIONS

Throughout this paper, $\mathbf{A}$, $\mathbf{a}$ and a denote matrix, vector and scalar, respectively. $(.)^T$, $(.)^H$ and $\mathrm{Tr}(\cdot)$ represent the transpose, the Hermitian transpose and the trace, respectively. $\mathcal{N}(\nu, \sigma^2)$ is a Gaussian random variable with mean ν and variance σ^2. $\mathbb{P}(.)$ is the probability of an event. And $\vec{\Theta} = (\theta, \phi)$ denotes a couple of azimuth and elevation angles.

II. HBF-BASED SB-NOMA SYSTEM MODEL

The downlink HBF-based mMIMO-NOMA system consists of a BS equipped with $M = M_x M_z \gg 1$ antennas and $N_{RF} \ll M$ RF chains to serve $K \ll M$ single-antenna UEs. The BS adopts a hybrid fully-connected structure. Denote by M_x and M_z the number of antennas along the $x-$ and $z-$axis, respectively. In this work, we consider both 1D and 2D antenna arrays at the BS by adopting a ULA array along the x-axis and a URA array in the xoz plane, respectively. In the classical AD MIMO-NOMA scheme, the users are regrouped within SU and multi-user SB-NOMA groups, according to their AoDs, see Fig. 1.

A. MMWAVE CHANNEL MODEL

The mmWave channel vector $\mathrm{h}_k \in \mathbb{C}^{M\times 1}$ between the BS and user k can be expressed as follows

$$\mathbf{h}_k = \sum_{n=1}^{N_k} \alpha_{n,k} e^{j\varphi_{n,k}} \mathbf{a}^H\left(\vec{\Theta}_{n,k}, M_x, M_z\right), \tag{1}$$

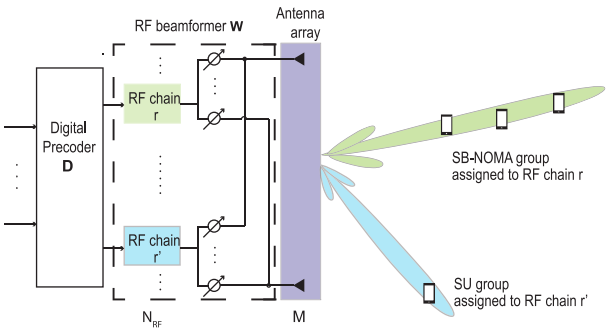


FIGURE 1. HBF-based SB-NOMA system model.

where N_k is the number of paths, $\alpha_{n,k}$ and $\varphi_{n,k}$ are the amplitude and the phase of the n-th path, and $\vec{\Theta}_{n,k} = (\theta_{n,k}, \phi_{n,k})$ is the couple of azimuth and elevation AoD of the n-th path. $\mathbf{a}(\vec{\Theta}, M_x, M_z) \in \mathbb{C}^{M\times 1}$ is the transmit array steering vector corresponding to the $\vec{\Theta}$ direction and is given by (2), shown at the bottom of the page for both ULA and URA. For a ULA array along the x-axis, $M_z = 1$ and $\phi = 0$, thus, $M = M_x$ and $\vec{\Theta} = (\theta, 0)$. For simplicity, we will use, in the rest of this paper, $\mathbf{a}_{n,k}$ instead of $\mathbf{a}(\vec{\Theta}_{n,k}, M_x, M_z)$. This work considers that the LoS path exists in each user's channel and has the highest power, labeled by $n = 1$. Thus, $\vec{\Theta}_{1,k} = (\theta_{1,k}, \phi_{1,k})$ represents the spatial direction of user k. Throughout this paper, we adopt the realistic and statistical channel simulator developed by New York University, called NYUSIM [29]. This simulator applies the time-cluster and spatial-lobe approach to generate different channel coefficients and is specified only for mmWave frequencies.

B. HYBRID BEAMFORMING DESIGN

Using HBF, the total number of groups G served simultaneously by the BS is restricted by the number of RF chains, i.e., $G \leq N_{RF}$. The overall downlink HBF precoding matrix $\mathbf{F} \in \mathbb{C}^{M\times G}$ is constructed in two stages as follows

$$\mathbf{F} = \mathbf{WD}, \tag{3}$$

where $\mathbf{D} = [\mathbf{d}_1 \cdots \mathbf{d}_G] \in \mathbb{C}^{G\times G}$ is the baseband digital component and $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_G] \in \mathbb{C}^{M\times G}$ is the RF analog component. Specifically, $\mathbf{d}_g$ and $\mathbf{w}_g$ denote respectively the digital and analog beamformers for the group g assigned to the g-th RF chain. A pure analog mmWave system is considered in this work to reduce the complexity of the signal processing in the baseband, i.e., $\mathbf{D} = \mathbf{I}_G$. Thus, the normalization BF factor η given by $\eta = \frac{1}{\mathrm{Tr}(\mathbf{F}^H\mathbf{F})}$ can be rewritten by $\eta = \frac{1}{\mathrm{Tr}(\mathbf{W}^H\mathbf{W})}$. The scenario with low-complex digital precoders is left for future extension.

C. HBF-BASED MIMO-NOMA SYSTEM MODEL

In the AD DBF-based MIMO-NOMA scheme as in [24], the K users are clustered into G_{SU} SU groups and G_{SB} multi-user SB-NOMA groups according to their AoDs. Denote $\mathcal{S}_g$, $g = 1, \ldots, G$, as the set of users scheduled on the group g such that $\bigcup_{g=1}^{G} \mathcal{S}_g = \mathcal{K}$ and $\mathcal{S}_g \bigcap \mathcal{S}_{g'} = \emptyset$, $\forall g \neq g'$. The user scheduling variable u_k^g is defined as follows

$$u_k^g = \begin{cases} 1, & \text{user } k \text{ belongs to group } g, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Note that each user will be served within one group, i.e., $\sum_{g=1}^{G} u_k^g = 1$, $\forall k \in \mathcal{K}$.

In contrast to the angle-based clustering algorithms proposed in the literature [19], [21], we design 2-user and multi-user β-UC algorithms that can be applied with any antenna array architectures [24]. This was done using the spatial interference $\beta_{k,u}$ metric which is defined as follows

$$\beta_{k,u} \overset{\text{def}}{=} \frac{1}{M}\left|\mathbf{a}_{1,k}^H \mathbf{a}_{1,u}\right|, \quad k, u \in \mathcal{K}, \tag{5}$$

As demonstrated in [24], $\beta_{k,u}$ for ULA can be rewritten by

$$\beta_{k,u} = \left| AF_{\left(\vec{\Theta}_{1,u}\right)}\left(\vec{\Theta}_{1,k}\right)\right| \tag{6}$$

where $AF_{(\vec{\Theta}_{1,u})}(\vec{\Theta})$ is the array factor of the beam directed at $\vec{\Theta}_{1,u}$. Similarly, for any uniformly excited array architecture, e.g., URA, UCA, (5) leads to (6).

For both SU and multi-user SB-NOMA groups, the BS generates a single beam in the spatial direction of each group. This work focuses on AD HBF-based MIMO-NOMA, where $N_{RF} \ll M$. Thereby, the analog beamformer of group g is given by the array steering vector corresponding to $\vec{\Theta}_g$

$$\mathbf{w}_g = \mathbf{a}\left(\vec{\Theta}_g, M_x, M_z\right), \tag{7}$$

where $\vec{\Theta}_g = (\theta_g, \phi_g)$ is the spatial direction of group g and is given by (8), shown at the bottom of the page. Following [24], for SB-NOMA groups, we take $\vec{\Theta}_g$ as the mean value between the minimum and maximum of all users' spatial direction in the group g. It's worth noting that according to [24], β and $\vec{\Theta}_g$ should be selected so that the main

$$\mathbf{a}\left(\vec{\Theta}, M_x, M_z\right) = \left[1, e^{-j2\pi\frac{d}{\lambda}\cos(\theta)\cos(\phi)}, \ldots, e^{-j\{(M_x-1)2\pi\frac{d}{\lambda}\cos(\theta)\cos(\phi)+(M_z-1)2\pi\frac{d}{\lambda}\sin(\phi)\}}\right]^T. \tag{2}$$

$$\vec{\Theta}_g = \begin{cases} (\theta_{1,k}, \phi_{1,k}), & \text{if SU with } \mathcal{S}_g = \{k\}, \\ \left(\theta_g = \frac{\min\limits_{k\in\mathcal{S}_g}\{\theta_{1,k}\}+\max\limits_{k\in\mathcal{S}_g}\{\theta_{1,k}\}}{2}, \phi_g = \frac{\min\limits_{k\in\mathcal{S}_g}\{\phi_{1,k}\}+\max\limits_{k\in\mathcal{S}_g}\{\phi_{1,k}\}}{2}\right), & \text{if SB-NOMA with } \mathbf{card}(\mathcal{S}_g) \geq 2. \end{cases} \tag{8}$$

beam covers all users to avoid severe beam misalignment. Thus, they should satisfy the following conditions

$$\begin{cases} \beta_{k,u} \geq \beta^{FSL}, \\ \min\limits_{k\in\mathcal{S}_g}\{\theta_{1,k}\} \leq \theta_g \leq \max\limits_{k\in\mathcal{S}_g}\{\theta_{1,k}\}, \\ \min\limits_{k\in\mathcal{S}_g}\{\phi_{1,k}\} \leq \phi_g \leq \max\limits_{k\in\mathcal{S}_g}\{\phi_{1,k}\}, \end{cases} \tag{9}$$

with β^{FSL} is the first side lobe level and is equal to 0.217 (or -13.26 dB) for both ULA and URA [30]. Note that the value of β^{FSL} does not depend on the beam's direction and the antennas number.

In short, we apply the multi-user β-UC algorithm to partition the K users within G_{SU} SU and G_{SB} multi-user SB-NOMA groups. After that, we classify the users in each SB-NOMA group according to the angular-based user ordering strategy [24].[1] This uses the angular-based channel quality $\breve{\zeta}$ metric. Denote $\breve{\zeta}_k^g$ as the angular-based channel quality of user k belonging to group g. Without loss of generality, we assume that the users are indexed in the descending order of their $\breve{\zeta}$, i.e., $\breve{\zeta}_k^g \geq \breve{\zeta}_{k'}^{g'}\ \forall\ k \leq k'$. Therefore, the SIC decoding realizes the signal separation at the users side in the increasing order of $\breve{\zeta}$. Assuming a successful SIC decoding, the user k from group g receives the following signal y_k^g

$$y_k^g = \underbrace{\sqrt{\eta\gamma_{k,g}p_g}\mathbf{h}_k\mathbf{w}_g s_k}_{\text{desired signal}} + \underbrace{\sum_{k'=1}^{k-1} u_{k'}^g\sqrt{\eta\gamma_{k',g}p_g}\mathbf{h}_k\mathbf{w}_g s_{k'}}_{\text{intra-group interference}} + \underbrace{\sum_{\substack{g'=1\\ g'\neq g}}^{G}\sum_{k'=1}^{K} u_{k'}^{g'}\sqrt{\eta\gamma_{k',g'}p_{g'}}\mathbf{h}_k\mathbf{w}_{g'}s_{k'}}_{\text{inter-group interference}} + z_k^g, \tag{10}$$

where s_k is the modulated signal relative to user k, p_g is the power allocated to group g such that $\sum_{g=1}^{G} p_g = P_e$ with P_e the total transmission power, $\gamma_{k,g}$ is the intra-group power allocation coefficient assigned to user k belonging to group g such that $\sum_{k=1}^{K} u_k^g\gamma_{k,g} = 1$ according to the NOMA principles, and $z_k^g \sim \mathcal{N}(0,\ \sigma_n^2)$ is the additive white Gaussian noise experienced at user k. Thereby, user k belonging to group g, has the SINR_k^g given in (11), shown at the bottom of the page, while decoding his own message.

1. In [24], we propose user ordering and power allocation techniques with only the knowledge of users' AoDs. We find that the proposed user ordering strategy outperforms other limited feedback strategies and that the AD power allocation technique provides an efficient SIC. For these reasons, and since we are interested in the feedback of angular information, we adopt them throughout this work.

In (11), the first term in the denominator is the residual intra-group interference after SIC, and the second one is the inter-group interference. Assuming a successful decoding and no propagation error, user k belonging to group g achieves the following data rate R_k^g

$$R_k^g = \log_2\left(1 + \text{SINR}_k^g\right). \tag{12}$$

And the total sum-rate can be expressed as follows

$$R_T = \sum_{g=1}^{G}\sum_{k=1}^{K} u_k^g R_k^g. \tag{13}$$

D. PROBLEM STATEMENT

In hybrid systems, up to N_{RF} groups can only be connected to the BS in the same orthogonal RB. However, when utilizing LSAAs with SB-NOMA, the narrowness of beams limits the capability of handling massive connectivity and adding more DoFs. As the number of antennas, M, increases, the beamwidth narrows, and the number of users served in the same NOMA groups gets smaller. Thus, in an overloaded scenario where $K > N_{RF}$, the probability that the cell is still overloaded using SB-NOMA, $P_{\text{os}} = \mathbb{P}(G_{SU} + G_{SB} > N_{RF})$, increases with increasing M. As M grows large but finite, $G_{SB} \overset{M\gg 1}{\to} 0$ and thus $G_{SU} \overset{M\gg 1}{\to} K$. This means that in the asymptotic limits when $K > N_{RF}$, the number of groups, G_{srv}, served by the BS is approximately equal to $G_{srv} \overset{M\gg 1}{\to} N_{RF}$, and $P_{\text{os}} \overset{M\gg 1}{\to} 1$. Consequently, $K - N_{RF}$ users will not be able to connect to the BS, making SB-NOMA insufficient for managing the connectivity of all users with LSAAs in an overloaded scenario.

This paper aims to overcome the limitation of SB-NOMA in an overloaded scenario, i.e., $G_{SU} + G_{SB} > N_{RF}$, to exploit the multi-user diversity with LSAAs. To do so, we design two schemes providing different types of additional DoFs to handle the connectivity of all users so that the G_f total groups served at the same orthogonal RB satisfy $G_f \leq N_{RF}$. Fig. 2 illustrates the flowchart of the proposed schemes. The first one, i.e., joint β-based SB-NOMA and TDMA, leverages the time-domain resources and is inspired by [19]. The other scheme, i.e., joint SB- and MB-NOMA, leverages the MB-NOMA framework [26], in which users with any AoDs can be served in the same group, i.e., by the same RF chain. Both schemes allow the connectivity of all users in an overloaded scenario and require only the knowledge of the user's spatial direction. Our analysis is restricted to cases with $2 \leq K \leq 2N_{RF}$ due to the assumption that each MB-NOMA group supports only two users. Going beyond this and allowing the BS to serve more than $2N_{RF}$ users through multi-user MB-NOMA groups is left as a future work.

$$\text{SINR}_k^g = \frac{\eta\gamma_{k,g}p_g|\mathbf{h}_k\mathbf{w}_g|^2}{\sum_{k'=1}^{k-1} u_{k'}^g\eta\gamma_{k',g}p_g|\mathbf{h}_k\mathbf{w}_g|^2 + \sum_{\substack{g'=1\\ g'\neq g}}^{G}\sum_{k'=1}^{K} u_{k'}^{g'}\eta\gamma_{k',g'}p_{g'}|\mathbf{h}_k\mathbf{w}_{g'}|^2 + \sigma_n^2}. \tag{11}$$

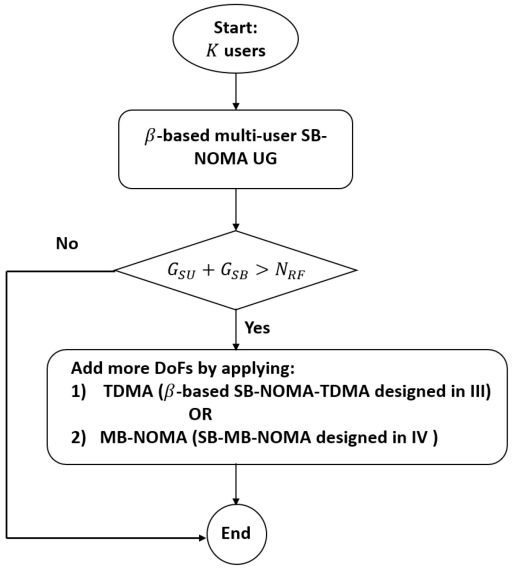


FIGURE 2. Flowchart of the proposed schemes.

III. PERFORMANCE ANALYSIS OF SPATIAL AND NON-ORTHOGONAL MULTIPLE ACCESS TECHNIQUES

In this section, we derive the sum-rate expressed in (12) using SDMA, SB-NOMA, and MB-NOMA for a two-user scenario and analyze their performance to see how we can exploit them in our proposed scheme. Before that, we first introduce the concept of MB-NOMA and how we extend it to address URA.

A. MULTIPLE ANALOG BEAMS USING BEAM SPLITTING

The authors in [26], [27] propose an MB-NOMA framework for mmWave hybrid systems to serve multiple users having arbitrary AoDs within the same RF chain. Specifically, they design a BST that divides the antenna array into various sub-arrays to generate multiple analog beams. Interestingly, MB-NOMA allows for more exploitation of multi-user diversity than SB-NOMA in mmWave hybrid systems. Indeed, with MB-NOMA, the number of users served simultaneously in the same NOMA cluster is not restricted by the users' AoD distribution as it is with SB-NOMA. In [27], their proposed BST can only be applied to the ULA architecture. However, since URA is more feasible with mMIMO in practice [31], we extend this technique to also deal with URA. In this work, we consider a 2-user MB-NOMA framework, where only two users belong to the same MB-NOMA group. So, we divide the antenna array connected by an RF chain into two sub-arrays to form two analog beams. And we assume that both sub-arrays have the same size, i.e., the same number of antennas M_x^{sa} and M_z^{sa} along the x-axis and z-axis, respectively, i.e., $M_x^{sa} M_z^{sa} = M/2$. Next, we separately present the BSTs with either ULA or URA. To facilitate the understanding of the principle of this technique, we first rewrite the corresponding array steering vector in (2). Then, we consider user k and user k' belonging to the MB-NOMA group g, and we reformulate the RF analog beamformer $\mathbf{w}_g$ assigned to the RF chain g performing beam splitting to generate two different beams toward each user.

1) BEAM SPLITTING WITH ULA

We start with a ULA array having $M = M_x$ antennas along the x-axis, for which the array steering vector $\mathbf{a}(\vec{\Theta}, M_x, 1)$ corresponding to the angle $\vec{\Theta} = (\theta, 0)$ is given by

$$\mathbf{a}\left(\vec{\Theta}, M_x, 1\right) = \left[1, \ldots, e^{j2\pi(M-1)\omega_x\left(\vec{\Theta}\right)}\right]^T, \tag{14a}$$

$$= \left[\mathbf{a}^T\left(\vec{\Theta}, M_x/2, 1\right), e^{j2\pi(M_x/2)\omega_x\left(\vec{\Theta}\right)}\mathbf{a}^T\left(\vec{\Theta}, M_x/2, 1\right)\right]^T. \tag{14b}$$

From (14), it is clear that the $\mathbf{a}(\vec{\Theta}, M_x, 1)$ vector can be constructed as a set of the steering vectors of two sub-arrays with $M_x/2$ antennas separated by a phase shift, i.e., $e^{j2\pi(M_x/2)\omega_x(\vec{\Theta})}$. This construction facilitates the understanding of the BST in [27] with ULA. Now assume that the BS adopts the BST to simultaneously serve user k and user k' belonging to group g. Recall that the ULA array is split into two sub-arrays with $M_x^{sa} = M_x/2$ and $M_z^{sa} = 1$. Therefore, the RF analog beamformer $\mathbf{w}_g$ assigned to the RF chain g performing beam splitting with two beams is given by [27]

$$\mathbf{w}_g = \left[\mathbf{a}^T\left(\vec{\Theta}_k, M_x^{sa}, M_z^{sa}\right), e^{j2\pi(M_x^{sa})\omega_x\left(\vec{\Theta}_k\right)}\mathbf{a}^T\left(\vec{\Theta}_{k'}, M_x^{sa}, M_z^{sa}\right)\right]^T. \tag{15}$$

From (15), each sub-array generates a beam in the spatial direction of one user.

2) BEAM SPLITTING WITH URA

Similarly, using URA with M_x and M_z antennas along the x- and the z-axis, respectively, the array steering vector $\mathbf{a}(\vec{\Theta}, M_x, M_z) \in \mathbb{C}^{M_x M_z \times 1}$ corresponding to the angle $\vec{\Theta} = (\theta, \phi)$ can be rewritten by

$$\mathbf{a}\left(\vec{\Theta}, M_x, M_z\right) = \begin{bmatrix} \mathbf{a}^T\left(\vec{\Theta}, M_x, M_z/2\right) \\ e^{j2\pi(M_z/2)\omega_z\left(\vec{\Theta}\right)}\mathbf{a}^T\left(\vec{\Theta}, M_x, M_z/2\right) \end{bmatrix}. \tag{16}$$

We now extend the BST to handle URA arrays as well. We vertically divide the antenna array into two sub-arrays, each with $M_x^{sa} = M_x$ and $M_z^{sa} = M_z/2$ antennas along the x- and the $z-$axis, respectively, as shown in Fig. 3. According to (16), the RF analog beamformer $\mathbf{w}_g$ performing the BST can be constructed as follows

$$\mathbf{w}_g = \begin{bmatrix} \mathbf{a}^T\left(\vec{\Theta}_k, M_x^{sa}, M_z^{sa}\right) \\ e^{j2\pi(M_z^{sa})\omega_z\left(\vec{\Theta}_k\right)}\mathbf{a}^T\left(\vec{\Theta}_{k'}, M_x^{sa}, M_z^{sa}\right) \end{bmatrix}. \tag{17}$$

From (17), it's clear that the first sub-array can successfully form a beam toward user k. Furthermore, the steering vector of the second sub-array is multiplied by a constant phase shift, i.e., $e^{j2\pi(M_z^{sa})\omega_z(\vec{\Theta}_k)}$. In other words, we add the same phase shift to all antennas in this sub-array, which have

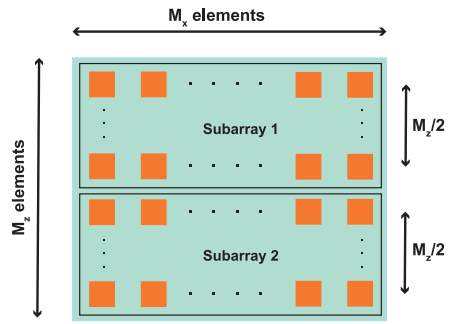


FIGURE 3. Vertically array splitting technique with URA.

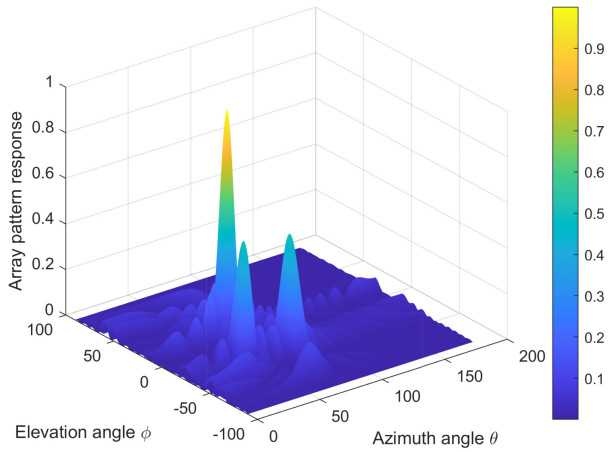


FIGURE 4. 32 × 6 URA with $\vec{\Theta}_k = (50°, -20°)$, $\vec{\Theta}_{k'} = (80°, -30°)$, and $\vec{\Theta}_{k''} = (60°, 10°)$.

different phase weights to direct a beam to $\vec{\Theta}_{k'}$. Thus, it can also successfully form a beam toward user k'.

Once the number of sub-arrays and the amount of horizontal and vertical antennas in each are defined, we can rewrite the expression of $\mathbf{a}(\vec{\Theta}_k, M_x, M_z)$ as we did in (17). Knowing that the construction of $\mathbf{w}_g$ performing the BST is mainly determined by $\mathbf{a}(\vec{\Theta}_k, M_x, M_z)$, this allows us to easily extend the BST to support multiple users per RF chain by considering the same or different sub-array configurations.

3) ILLUSTRATIVE REPRESENTATIONS

To illustrate the generation of two analog beams via the BST, Fig. 4 depicts the normalized array pattern responses for $\mathbf{w}_g$ in (17) at user k and user k' and of a single beam pointed at user k'' using URA. Compared to the single beam for user k'', the maximum magnitude of the array response for $\mathbf{w}_g$ is halved due to two sub-arrays of equal size with $M/2$ antennas each. Moreover, the width of the beams pointed at users k and k' increases. In the following, we will show that generating multiple analog beams via the BST manages the connectivity of all users in HBF-based mMIMO-NOMA systems. Indeed, MB-NOMA serves more users on each RF chain, while SB-NOMA only considers users with close directions.

B. SYSTEM SUM-RATE OF SDMA, SB-NOMA, AND MB-NOMA IN A 2-USER SCENARIO

Previous work on mmWave MB-NOMA [26], [27] requires full CSI to perform user clustering and does not exploit the potential benefits of SB-NOMA when users are close together. As discussed earlier, angular information is a promising partial CSI for mmWave channels. Therefore, throughout this subsection, we investigate the spatial behavior of SDMA, SB-NOMA, and MB-NOMA, and answer the following question: how can we benefit from both SB-NOMA and MB-NOMA in the angle domain? Here, we consider a special case scenario where the BS serves only two users in a mono-path environment. Thereby, $K = 2$ and only one SB-NOMA or MB-NOMA group exists, i.e., $G = 1$. Denote by user 1 the strong user, i.e., $\gamma_{1,g} \geq \gamma_{2,g}$ when NOMA is applied. For SDMA, we consider two RF chains at the BS to serve the two users with one beam each.

SYSTEM SUM-RATE

1) SPACE DIVISION MULTIPLE ACCESS (SDMA)

We start by considering that the BS applies the SDMA technique. Thus, a single beam is pointed to serve each user. The $\mathrm{SINR}_1^{\mathrm{SD}}$ at user 1 is then given by ($\mathrm{SINR}_2^{\mathrm{SD}}$ can be obtained by symmetry)

$$\mathrm{SINR}_1^{\mathrm{SD}} = \frac{\eta P_e |\mathbf{h}_1 \mathbf{a}\left(\vec{\Theta}_{1,1}, M_x, M_z\right)|^2}{\eta P_e |\mathbf{h}_1 \mathbf{a}\left(\vec{\Theta}_{1,2}, M_x, M_z\right)|^2 + \sigma_n^2}, \tag{18a}$$

$$\overset{(a)}{=} \frac{\eta P_e |\alpha_{1,1}|^2 |\mathbf{a}^H\left(\vec{\Theta}_{1,1}, M_x, M_z\right)\mathbf{a}\left(\vec{\Theta}_{1,1}, M_x, M_z\right)|^2}{\eta P_e |\alpha_{1,1}|^2 |\mathbf{a}^H\left(\vec{\Theta}_{1,1}, M_x, M_z\right)\mathbf{a}\left(\vec{\Theta}_{1,2}, M_x, M_z\right)|^2 + \sigma_n^2}, \tag{18b}$$

$$\overset{(b)}{=} \frac{\rho\eta |\alpha_{1,1}|^2 M^2}{\rho\eta |\alpha_{1,1}|^2 \left|\mathbf{a}^H\left(\vec{\Theta}_{1,1}, M_x, M_z\right)\mathbf{a}\left(\vec{\Theta}_{1,2}, M_x, M_z\right)\right|^2 + 1}. \tag{18c}$$

(a) follows from the mono-path channel assumption, i.e., $\mathbf{h}_k = \alpha_{1,k}\mathbf{a}^H(\vec{\Theta}_{1,1}, M_x, M_z)$. And (b) by setting $\rho = P_e/\sigma_n^2$ and since $|\mathbf{a}^H(\vec{\Theta}, M_x, M_z)\mathbf{a}(\vec{\Theta}, M_x, M_z)| = M$.

We set $\beta_{1,2}(M_x, M_z)$ as the normalized spatial interference between the two users

$$\beta_{1,2}(M_x, M_z) = \frac{\left|\mathbf{a}^H\left(\vec{\Theta}_{1,1}, M_x, M_z\right)\mathbf{a}\left(\vec{\Theta}_{1,2}, M_x, M_z\right)\right|}{M_x M_z}. \tag{19}$$

And since $\eta = \frac{1}{KM} = \frac{1}{2M}$ with SDMA, the sum-rate $\mathrm{R}_T^{\mathrm{SD}}$ can be expressed as

$$\mathrm{R}_T^{\mathrm{SD}} = \sum_{k=1}^{K} \log_2\left(1 + \frac{\rho|\alpha_{1,k}|^2 M}{\rho|\alpha_{1,k}|^2 \beta_{1,2}^2(M_x, M_z)M + K}\right). \tag{20}$$

2) SINGLE-BEAM NOMA (SB-NOMA)

The BS now applies the SB-NOMA technique. So, it forms a single beam between the two users, with an angle calculated

as in (8). Thus, $\mathrm{SINR}_1^{\mathrm{SB}}$ and $\mathrm{SINR}_2^{\mathrm{SB}}$ of users belonging to group g $(=1)$ are respectively given by

$$\mathrm{SINR}_1^{\mathrm{SB}} = \frac{\eta\gamma_{1,g}p_g|\mathbf{h}_1\mathbf{w}_g|^2}{\sigma_n^2}, \tag{21a}$$

$$\overset{(a)}{=} \rho\eta\gamma_{1,g}|\alpha_{1,1}|^2|\mathbf{a}^H\left(\vec{\Theta}_{1,1}, M\right)\mathbf{a}\left(\vec{\Theta}_g, M\right)|^2, \tag{21b}$$

where (a) is obtained since $G=1$, i.e., $p_g = P_e$.

$$\mathrm{SINR}_2^{\mathrm{SB}} = \frac{\eta\left(1-\gamma_{1,g}\right)p_g|\mathbf{h}_2\mathbf{w}_g|^2}{\gamma_{1,g}p_g|\mathbf{h}_2\mathbf{w}_g|^2+\sigma_n^2}, \tag{22a}$$

$$= \frac{\rho\eta\left(1-\gamma_{1,g}\right)|\alpha_{1,2}|^2\left|\mathbf{a}^H\left(\vec{\Theta}_{1,2},M_x,M_z\right)\mathbf{a}\left(\vec{\Theta}_g,M_x,M_z\right)\right|^2}{\rho\eta\gamma_{1,g}|\alpha_{1,2}|^2\left|\mathbf{a}^H\left(\vec{\Theta}_{1,2},M_x,M_z\right)\mathbf{a}\left(\vec{\Theta}_g,M_x,M_z\right)\right|^2+1}. \tag{22b}$$

We set $\beta_{k,g}(M_x,M_z) = \frac{|\mathbf{a}^H(\vec{\Theta}_{1,k},M_x,M_z)\mathbf{a}(\vec{\Theta}_g,M_x,M_z)|}{MxMz}$, and since $\eta = \frac{1}{MG} = \frac{1}{M}$ with SB-NOMA, the sum-rate $\mathrm{R}_T^{\mathrm{SB}}$ can be expressed as follows

$$\mathrm{R}_T^{\mathrm{SB}} = \log_2\left(1+\rho\gamma_{1,g}|\alpha_{1,1}|^2\beta_{1,g}^2(M_x,M_z)M\right) + \log_2\left(1+\frac{\rho\left(1-\gamma_{1,g}\right)|\alpha_{1,2}|^2\beta_{2,g}^2(M_x,M_z)M}{\rho\gamma_{1,g}|\alpha_{1,2}|^2\beta_{2,g}^2(M_x,M_z)M+1}\right). \tag{23}$$

3) MULTIPLE-BEAM NOMA (MB-NOMA)

The BS now applies the MB-NOMA technique via the BST. So, it generates two beams at the same RF chain; each is directed in the AoD of each user. The analog beamformer $\mathbf{w}_g$ $(g=1)$ is given by (15) and (17) with ULA and URA, respectively. Denote by $\psi(\vec{\Theta}_k, M_x^{sa}, M_z^{sa})$ the additional phase shift experienced at the second sub-array using either ULA or URA and given by

$$\psi\left(\vec{\Theta}_k, M_x^{sa}, M_z^{sa}\right) = \begin{cases} j2\pi\left(M_x^{sa}\right)\omega_x\left(\vec{\Theta}_k\right) & \text{if ULA,} \\ j2\pi\left(M_z^{sa}\right)\omega_z\left(\vec{\Theta}_k\right) & \text{if URA.} \end{cases} \tag{24}$$

Thus, using either ULA or URA, $\mathrm{SINR}_1^{\mathrm{MB}}$ and $\mathrm{SINR}_2^{\mathrm{MB}}$ can be expressed respectively as follows

$$\mathrm{SINR}_1^{\mathrm{MB}} = \frac{\eta\gamma_{1,g}p_g|\mathbf{h}_1\mathbf{w}_g|^2}{\sigma_n^2}. \tag{25}$$

$$\mathrm{SINR}_2^{\mathrm{MB}} = \frac{\eta\left(1-\gamma_{1,g}\right)p_g|\mathbf{h}_2\mathbf{w}_g|^2}{\gamma_{1,g}p_g|\mathbf{h}_2\mathbf{w}_g|^2+\sigma_n^2}. \tag{26}$$

Eqs. (25) and (26) can be rewritten by (27) and (28), shown at the bottom of the page, where $\Delta\psi = \psi(\vec{\Theta}_{1,1}, M_x^{sa}, M_z^{sa}) - \psi(\vec{\Theta}_{1,2}, M_x^{sa}, M_z^{sa})$.

We set $\beta'_{1,2}(M_x^{sa}, M_z^{sa})$ as follows

$$\beta'_{1,2}(M_x^{sa},M_z^{sa}) = \frac{\mathbf{a}^H\left(\vec{\Theta}_{1,1},M_x^{sa},M_z^{sa}\right)\mathbf{a}\left(\vec{\Theta}_{1,2},M_x^{sa},M_z^{sa}\right)}{M_x^{sa}M_z^{sa}}. \tag{29}$$

And since $\eta = \frac{1}{MG} = \frac{1}{M}$ with MB-NOMA, the sum-rate $\mathrm{R}_T^{\mathrm{MB}}$ can be given by (30), shown at the bottom of the page.

PERFORMANCE ANALYSIS

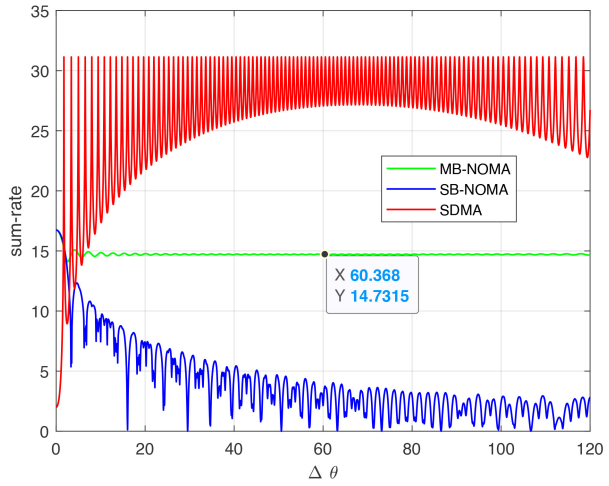
In this subsection, we provide illustrative results to analyze the spatial behavior of SDMA, SB-NOMA, and MB-NOMA using ULA. It is worth noting that similar results can be obtained with URA configurations. For a 128×1 ULA, Fig. 5(a) and Fig. 5(b) plot the sum-rate of the three schemes as a function of the angular distance $\Delta\theta$ and the spatial inter-user interference $\beta_{1,2}$ calculated in (19), respectively. The AoD $\theta_{1,1}$ of the first user is set to $10°$, and we uniformly change the AoD $\theta_{1,2}$ of the second user, such that $0° \leq \Delta\theta = \theta_{1,2} - \theta_{1,1} \leq 120°$. Note that $\gamma_{1,g} = 0.2$, $\gamma_{2,g} = 0.8$, $P_e = 30$ dBm and $\sigma_n^2 = -101$ dBm. $|\alpha_{1,1}| = 8.3 \times 10^{-6}$ and $|\alpha_{1,2}| = 7.4 \times 10^{-6}$.

From (20), we find that $\mathrm{R}_T^{\mathrm{SD}}$ is a decreasing function of $\beta_{1,2}$. This also can be seen from Fig. 5(b). In other words, when users are very close to each other, i.e., $\Delta\vec{\Theta} \to (0,0)$ and $\beta_{1,2} \to 1$, the system suffers from high inter-user interference and the SDMA performance degrades. For instance, for a large but finite number of antennas or in a high SNR regime, we can obtain this approximation that $\mathrm{R}_T^{\mathrm{SD}}(\beta_{1,2} \to 1) \overset{M\gg 1}{\to} K = 2$. From Fig. 5, we can see this equality and that SDMA performs very well when the users are well separated in space. With a large but finite value of M, $|\mathbf{a}^H(\vec{\Theta}_{1,1}, M_x, M_z)\mathbf{a}(\vec{\Theta}_{1,2}, M_x, M_z)| \to 0$ when

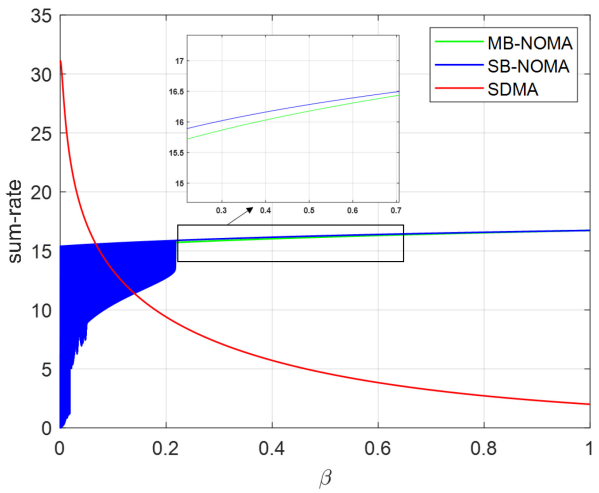
$$\mathrm{SINR}_1^{\mathrm{MB}} = \rho\eta\gamma_{1,g}|\alpha_{1,1}|^2\left|M/2 + \mathbf{a}^H\left(\vec{\Theta}_{1,1}, M_x^{sa}, M_z^{sa}\right)\mathbf{a}\left(\vec{\Theta}_{1,2}, M_x^{sa}, M_z^{sa}\right)\right|^2. \tag{27}$$

$$\mathrm{SINR}_2^{\mathrm{MB}} = \frac{\rho\eta\left(1-\gamma_{1,g}\right)|\alpha_{1,2}|^2\left|\mathbf{a}^H\left(\vec{\Theta}_{1,2}, M_x^{sa}, M_z^{sa}\right)\mathbf{a}\left(\vec{\Theta}_{1,1}, M_x^{sa}, M_z^{sa}\right) + M/2e^{\Delta\psi}\right|^2}{\rho\eta\gamma_{1,g}|\alpha_{1,2}|^2\left|\mathbf{a}^H\left(\vec{\Theta}_{1,2}, M_x^{sa}, M_z^{sa}\right)\mathbf{a}\left(\vec{\Theta}_{1,1}, M_x^{sa}, M_z^{sa}\right) + M/2e^{\Delta\psi}\right|^2 + 1}. \tag{28}$$

$$\mathrm{R}_T^{\mathrm{MB}} = \log_2\left(1+\rho\gamma_{1,g}|\alpha_{1,1}|^2\left|\frac{1}{2}\left(1+\beta'_{1,2}(M_x,M_z/2)\right)\right|^2 M\right) + \log_2\left(1+\frac{\rho\left(1-\gamma_{1,g}\right)|\alpha_{1,2}|^2\left|\frac{1}{2}\left(\beta'_{1,2}(M_x^{sa},M_z^{sa})+e^{\Delta\psi}\right)\right|^2 M}{\rho\gamma_{1,g}|\alpha_{1,2}|^2\left|\frac{1}{2}\left(\beta'_{1,2}(M_x^{sa},M_z^{sa})+e^{\Delta\psi}\right)\right|^2 M+1}\right). \tag{30}$$



(a) Sum-rate as a function of $\Delta\theta$ (in degree).



(b) Sum-rate as a function of $\beta_{1,2}$.

FIGURE 5. SB-NOMA versus MB-NOMA versus SDMA: Sum-rate as a function of $\Delta\theta$ and $\beta_{1,2}$ with a 128 × 1 ULA.

$\Delta\vec{\Theta} \neq (0,0)$ [32]. $\mathrm{R}_T^{\mathrm{SD}}$ can be then approximated by

$$\mathrm{R}_T^{\mathrm{SD}} \overset{M\gg 1}{\to} \sum_{k=1}^{2} \log_2\left(1+\frac{\rho|\alpha_{1,k}|^2 M}{2}\right). \tag{31}$$

From (23), $\mathrm{R}_T^{\mathrm{SB}}$ is restricted by the $\beta_{k,g}$ term. From [24, Lemma 2], if user 2 is in the main lobe of the beam pointed to user 1, i.e., $\Delta\vec{\Theta} \leq \vec{\Omega}_1^{3\mathrm{dB}} \overset{\mathrm{def}}{=} (\Omega_{x_1}^{3\mathrm{dB}}, \Omega_{z_1}^{3\mathrm{dB}})$[2] and $\beta_{1,2} \geq \beta^{\mathrm{FSL}} = 0.217$, then both users will be located in the main lobe of the beam directed at $\vec{\Theta}_g$. Therefore, SB-NOMA performs well and the system sum-rate is maximized. More the users are separated, more SB-NOMA degrades due to the beam misalignment. This is also observed in Fig. 5(b). With a large but finite value of M, $\left|\mathbf{a}^H(\vec{\Theta}_g, M_x, M_z)\mathbf{a}(\vec{\Theta}_{1,k}, M_x, M_z)\right| \to 0$ for $\Delta\vec{\Theta} \neq (0,0)$, thereby, $\mathrm{R}_T^{\mathrm{SB}} \overset{M\gg 1}{\to} 0$.

From (30), $\mathrm{R}_T^{\mathrm{MB}}$ is restricted by the $\beta'_{1,2}(M_x^{sa}, M_z^{sa})$ term. When the two users have similar AoD, i.e., $\Delta\vec{\Theta} \to (0,0)$ and $\beta_{1,2} \to 1$, the analog beamformer $\mathbf{w}_g$ degenerates to the single-beam case, thus, $\mathrm{R}_T^{\mathrm{MB}} \to \mathrm{R}_T^{\mathrm{SB}}$. At this point, $\mathrm{R}_T^{\mathrm{MB}}$ and $\mathrm{R}_T^{\mathrm{SB}}$ represents the highest sum-rate obtained with NOMA as seen in Fig. 5. We can see also that when $0 < \Delta\vec{\Theta} \leq \vec{\Omega}_1^{3\mathrm{dB}}$, SB-NOMA outperforms MB-NOMA since the two beams generated by MB-NOMA scheme do not overlap well. Otherwise, MB-NOMA outperforms SB-NOMA and has a constant sum-rate, as given in Lemma 1.

Lemma 1: Using LSAAs with a large but finite values of M_x and M_z, MB-NOMA has the following constant sum-rate R_∞^{MB} when $\Delta\vec{\Theta} > \vec{\Omega}_1^{3\mathrm{dB}}$ for the 2-user scenario

$$\mathrm{R}_T^{\mathrm{MB}} \overset{M_x\gg 1, M_z\gg 1}{\longrightarrow} \log_2\left(1+\frac{\gamma_{1,g}\rho|\alpha_{1,1}|^2 M}{4}\right) + \log_2\left(1+\frac{1-\gamma_{1,g}}{\gamma_{1,g}}\right) = R_\infty^{\mathrm{MB}}. \tag{32}$$

Proof: Using LSAAs with a large values of M_x and M_z, $\mathbf{a}^H(\vec{\Theta}_{1,1}, M_x^{sa}, M_z^{sa})\mathbf{a}(\vec{\Theta}_{1,2}, M_x^{sa}, M_z^{sa}) \overset{M_x\gg 1, M_z\gg 1}{\longrightarrow} 0$ for $\vec{\Theta}_{1,1} \neq \vec{\Theta}_{1,2}$. Thereby, by setting $\beta'_{1,2}(M_x^{sa}, M_z^{sa}) \overset{M_x\gg 1, M_z\gg 1}{\longrightarrow} 0$, (30) can be rewritten by

$$\mathrm{R}_T^{\mathrm{MB}} = \log_2\left(1+\frac{\gamma_{1,g}\rho|\alpha_{1,1}|^2 M}{4}\right) + \log_2\left(1+\frac{\rho\left(1-\gamma_{1,g}\right)|\alpha_{1,2}|^2 M}{\rho\gamma_{1,g}|\alpha_{1,2}|^2 M+4}\right). \tag{33}$$

The second term in (33) can be approximated to $\log_2(1+\frac{1-\gamma_{1,g}}{\gamma_{1,g}})$ with a large but finite number of antennas. Thus, Lemma 1 is verified. ■

The corresponding value of R_∞^{MB} in (32) is approximately equal to 14.7 when using 128×1 ULA. From Fig. 5(a), we find that the simulation results verify Lemma 1.

Remark 1: Based on the above analysis, we conclude that SB-NOMA is only suitable when users are very close. Otherwise, MB-NOMA has a significant constant sum-rate independently of $\Delta\vec{\Theta}$, compared to SB-NOMA. These results allow us to extend to the more general multi-user scenario by first clustering users with high spatial interference within SB-NOMA groups and then clustering the remaining users in MB-NOMA groups.

Remark 2: SDMA exhibits a significant sum-rate performance with LSAAs when the users are well separated in the angle domain. However, for HBF-based mMIMO-NOMA systems, SDMA fails to satisfy all users in overloaded scenarios, i.e., when $K > N_{RF}$. Indeed, SDMA can accommodate one user at each RF chain. In contrast, MB-NOMA can serve multiple users with arbitrary AoDs within one RF chain and has a considerable system sum-rate.

2. $\vec{\Omega}^{3\mathrm{dB}} = (\Omega_x^{3\mathrm{dB}}, \Omega_z^{3\mathrm{dB}})$ is the 3dB-width vector in azimuth and elevation axes. With 128×1 ULA, the 3dB-width $\Omega_1^{3\mathrm{dB}}$ of the beam generated at $\theta_1 = 10°$ is equal to $\Omega_1^{3\mathrm{dB}} = 2.25°$.

Considering these observations, we propose a joint SB- and MB-NOMA scheme in Section V that requires only user angles. Additionally, we demonstrate in Section VI-C that combining the benefits of both SB- and MB-NOMA is more advantageous than using one of these schemes alone.

IV. JOINT β-BASED SB-NOMA AND TDMA SCHEME

One possible solution to leverage more DoFs is the implementation of TDMA which may support more than N_{RF} groups within different time slots. Inspired by [19], the joint β-based SB-NOMA and TDMA is a two-phase scheme adopting TDMA only in an overloaded scenario. In the first phase, the multi-user β-UC algorithm proposed in [24] partitions the K users into G_{SU} SU and G_{SB} multi-user SB-NOMA groups. Only in an overloaded scenario, i.e., $G_{SU}+G_{SB} > N_{RF}$, a second phase that exploits the time-domain resources is applied. In fact, during each time slot, at most N_{RF} groups can be served concurrently. For this, it is necessary to design a group clustering algorithm that schedules the $G_{SU} + G_{SB} > N_{RF}$ groups in the time domain. In [19], the authors design a two-stage group clustering algorithm to reduce interference between groups served simultaneously. Specifically, they use the spatial angular distance as a metric to measure the corresponding inter-group interference. They found that inter-group interference can be significantly reduced by increasing the spatial angular distance. Note that this algorithm is applied only with ULA, and the cluster angles belong to a predefined set of azimuth angles with a fixed search step size $J = M$. In this section, we have updated this group clustering to reduce the inter-group interference measured by β in each time slot. Contrary to this work [19], β determines the level of spatial interference and includes both angular distance and beamwidth information [24]. Thereby, it is more accurate than the angular distance for calculating inter-group interference. Moreover, it is built based on the array factor, and thus our β-based group clustering algorithm in time-domain can be applied to any array architecture. Since $\beta_{g,g'} = \beta_{g',g}$, we first define the triangular matrix $\mathbf{B} \in \mathbb{C}^{G\times G}$ describing the spatial inter-group interference, with the (g, g')-element given by

$$\mathbf{B}(g,g') = \begin{cases} \beta_{g,g'}, & g < g', \\ -\infty, & \text{otherwise}, \end{cases} \tag{34}$$

where $\beta_{g,g'}$ is the spatial interference between group g and group g', and is given by $\beta_{g,g'} = \frac{|\mathbf{a}^H(\vec{\Theta}_g,M_x,M_z)\mathbf{a}(\vec{\Theta}_{g'},M_x,M_z)|}{M_xM_z}$, as defined in (5).

Before detailing the proposed scheme, some notations and definitions are presented as follows

- N_{TS} is the total number of time slots and is given by $\lfloor \frac{G}{N_{RF}} \rfloor$ with $G = G_{SU} + G_{SB}$ and $\lfloor . \rfloor$ the integer part.
- $\mathcal{G}_m^{\text{TS}}$, $m = 1, \ldots, N_{\text{TS}}$, is the group set including all groups scheduled in the time slot m and is initialized by an empty set, i.e., $\mathbf{card}(\mathcal{G}_m^{\text{TS}}) = 0$, $\forall m$.
- $\mathcal{G}^{NA}$ is the group set including all groups not yet assigned to any time slot and is initialized by a set containing all the G groups, i.e., $\mathcal{G}^{NA} = \{\mathcal{S}_1, \ldots, \mathcal{S}_G\}$.
- $\mathcal{G}^{AV}$ is the group set including all the available groups sets and is initialized by $\mathcal{G}^{AV} = \{\mathcal{G}_1^{\text{TS}}, \ldots, \mathcal{G}_{N_{\text{TS}}}^{\text{TS}}\}$.

Algorithm 1 β-Based Group Clustering in the Time Domain Algorithm

Input: $\vec{\Theta}_g = (\theta_g, \phi_g)$, $\beta_{g,g'}$.
Output: N_{TS}, $\mathcal{G}_{\text{TS}}$
Initialization: $t = 0$, $\mathbf{card}(\mathcal{G}_m^{\text{TS}}) = 0$ $\forall m \in \{1, \cdots, N_{\text{TS}}\}$, $\mathcal{G}^{NA} = \{\mathcal{S}_1, \cdots, \mathcal{S}_G\}$, $\mathcal{G}^{AV} = \{\mathcal{G}_1^{\text{TS}}, \cdots, \mathcal{G}_{\text{TS}}^{\text{TS}}\}$.
Compute $\mathbf{B}$ as in (34).
Stage 1: Clustering of groups with high inter-group spatial interference in different time slots
repeat
  Locate in $\mathbf{B}$ the 2 groups (e.g., groups o and r) having the largest spatial interference
  if $o \in \mathcal{G}^{NA}$ **then**
    $t = t + 1$, add group o to $\mathcal{G}_t^{\text{TS}}$, and remove group o from $\mathcal{G}^{NA}$.
  end if
  if $r \in \mathcal{G}^{NA}$ and $t < N_{\text{TS}}$ **then**
    $t = t + 1$, add group r to $\mathcal{G}_t^{\text{TS}}$, and remove group r from $\mathcal{G}^{NA}$.
  end if
  Set $B(o, r) = -\infty$
until $t = N_{\text{TS}}$
Stage 2: Clustering of groups with low inter-group spatial interference in the same time slots
repeat
  Select one arbitrary group g' from $\mathcal{G}^{NA}$
  Locate, at each time-slot n, the group $\hat{i}_n$ having the highest spatial interference with the group g', i.e., $\beta_{\hat{i}_n,g'} = \max_{i_n \in \mathcal{G}_n^{\text{TS}}} \beta_{i_n,g'}$, $\forall n \in \{1, \ldots, \mathbf{card}(\mathcal{G}_n^{\text{TS}})\}$
  Find the time-slot $\check{n}$, in which the maximum interference with group g' obtained in Step 14 is minimized, i.e., $\beta_{\hat{i}_{\check{n}},g'} = \min_{\{n=1,\ldots,\mathbf{card}(\mathcal{G}_n^{\text{TS}})\}} \beta_{\hat{i}_n,g'}$
  Add group g' to $\mathcal{G}_{\check{n}}^{\text{TS}}$, and remove group g' from $\mathcal{G}^{NA}$.
  if $\mathbf{card}(\mathcal{G}_{\check{n}}^{\text{TS}}) \geq N_{RF}$ **then**
    remove $\mathcal{G}_{\check{n}}^{\text{TS}}$ from $\mathcal{G}^{AV}$.
  end if
until $\mathcal{G}^{NA} = \emptyset$.

In hybrid systems, it is possible to simultaneously schedule at most N_{RF} groups, i.e., $\mathbf{card}(\mathcal{G}_m^{\text{TS}}) \leq N_{RF}$. Therefore, if $\mathbf{card}(\mathcal{G}_m^{\text{TS}}) < N_{RF}$, then $\mathcal{G}_m^{\text{TS}} \in \mathcal{G}^{AV}$. Thus, $\mathcal{G}_m^{\text{TS}}$ is considered an available group set if we could add at least one more group in the m-th time slot.

Initially, the first N_{TS} groups with high spatial interference β are scheduled in different time slots. Then, in the second stage, the remaining groups are clustered so that the inter-group interference is reduced in each time slot. Further details of the β-based group clustering algorithm in the time domain are given in Algorithm 1.

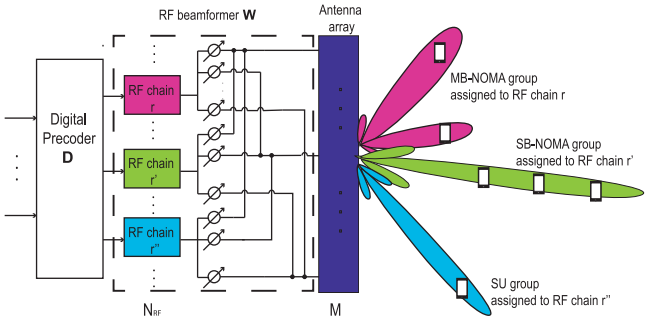


FIGURE 6. The HBF-based joint SB- and MB-NOMA scheme.

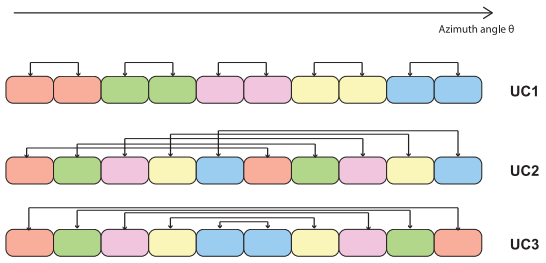


FIGURE 7. Three different 2-by-2 selection strategies. The users having the same color are selected to be in the same MB-NOMA group.

In Section VI, we show that the joint β-based SB-NOMA and TDMA scheme solves the problem of the SB-NOMA limitation in overloaded scenarios. It works well and offers a considerable sum-rate performance. However, synchronization is a major hurdle for implementing TDMA in mmWave hybrid systems, as quick and precise timing is essential for attaining high data rates. Since applying TDMA with NOMA introduces temporal synchronization complexity, a scheme that grants more DoFs by utilizing just NOMA without using other multiple access (MA) techniques is needed. To that end, we assess the potential of MB-NOMA in the following section, allowing users within the same group to be served, whatever their spatial distribution.

V. JOINT SB- AND MB-NOMA SCHEME

To provide more DoFs for the SB-NOMA in an overloaded scenario without relying on any extra MA techniques, we suggest a scheme based solely on NOMA. Recall from Remark 1 that SB-NOMA is suitable only for nearby users in space, whereas MB-NOMA offers a significant sum-rate regardless of the spatial distribution of users. Taking inspiration from our findings, we combine the benefits of SB- and MB-NOMA to tackle the limitation of SB-NOMA in overloaded scenarios. As a first step, as seen in joint SB-NOMA and TDMA, users with high spatial interference are clustered into the same SB-NOMA group according to the multi-user β-UC algorithm from [24]. If $G_{SU} + G_{SB} > N_{RF}$, then the remaining users are partitioned between two-by-two MB-NOMA groups and SU groups such that the total number of groups $G_f = G_{SB} + G'_{SU} + G_{MB}$ is equal to N_{RF}. Since G_f needs to be less than or equal to N_{RF} to serve all users, the highest possible value, N_{RF}, was selected as it generally yields a higher overall rate. This is because more users served by SDMA with its high BF gain leads to reduced interference and an enhanced system sum-rate.

In the proposed joint SB- and MB-NOMA scheme, the transmit RF beamformer $\mathbf{w}_g$ related to group g depends on its type and is given by (35) at the bottom of the page. In Fig. 6, we present the system model of the proposed scheme.

To the best of our knowledge, this is the first work that jointly leverages the potentiality of SB- and MB-NOMA. And this scheme is applicable for both linear and rectangular antenna arrays with only the knowledge of users' AoDs. We develop a low-complex MB-NOMA UC technique to reduce their complexity using only angular information. From Section III-B, once users with high spatial interference are all grouped in SB-NOMA groups, the manner in selecting two users into MB-NOMA groups based on their AoD is not critical regarding the sum-rate. To this end, we first sort the G_{SU} users in an array in the ascending order of their azimuth AoD θ with either ULA or URA. Then, we adopt three different selection strategies as shown in Fig. 7, denoted as UC1, UC2, and UC3, to select 2-by-2 the users from the sorted array and cluster them in MB-NOMA groups until all the G_{MB} MB-NOMA groups are defined.

We now carry out a comparison in terms of the sum-rate to corroborate the best selection strategy. In Fig. 8, we plot the sum-rate of the different MB-NOMA UC strategies and that of the random MB-NOMA UC, where the users are selected randomly to belong within the MB-NOMA groups. We assume a ULA array with $N_{BS} = 128$ antennas. As expected for $K \leq N_{RF}$, the K users are spatially separated thanks to the high BF gain with LSAAs. So, there is no need to exploit MB-NOMA, and the impact of the selection strategy does not appear here. Otherwise, we adopt MB-NOMA to offer more DoFs. We find that UC1 outperforms other strategies. This superiority is due to lower inter-group interference achieved when using UC1 compared to UC2 and UC3. In Section VI, UC1 is considered as the reference.

Both SB- and MB-NOMA employ the same power allocation and user ordering techniques at the BS and the

$$\mathbf{w}_g = \begin{cases} \mathbf{a}\left(\vec{\Theta}_g, M_x, M_z\right) \text{ with } \vec{\Theta}_g \text{ calculated as in (8)} & \text{if SB-NOMA with } \mathbf{card}\left(\mathcal{S}_g\right) \geq 2, \\ \mathbf{w}_g \text{ is calculated as in (15) and (17)} & \text{if MB-NOMA with } \mathcal{S}_g = \left\{k, k'\right\}, \\ \mathbf{a}\left(\vec{\Theta}_{1,k}, M_x, M_z\right) & \text{if SU with } \mathcal{S}_g = \{k\}. \end{cases} \tag{35}$$

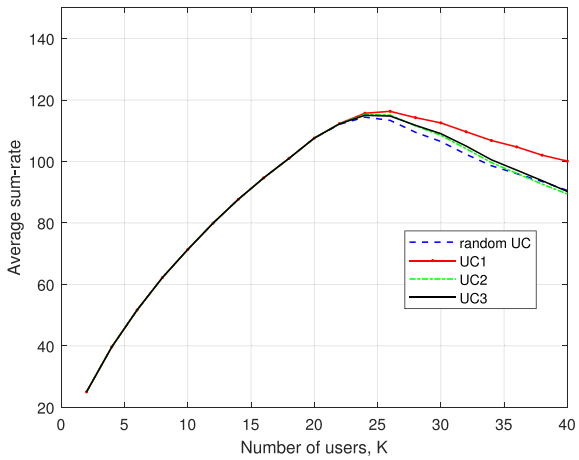


FIGURE 8. The sum-rate performance of different MB-NOMA UC algorithms versus the number of users for the joint SB- and MB-NOMA scheme. The sum-rate is averaged over 5,000 trials.

TABLE 1. Simulation parameters.

Parameters	Value
Number of transmit antennas, M	$\in \{128, 256\}$
Number of RF chains, N_{RF}	20
Carrier frequency	28 GHz
Channel bandwidth	20 MHz
Cell radius	100 m
Total transmission power, P_e	30 dBm
Noise power, σ_n^2	-101 dBm
Minimum SIC power difference, $P_{\min}$	1 mW
Number of paths per time cluster in a rural environment	$\in \{1, 2\}$

same SIC decoding strategy at the receivers, where the primary distinction lies in their analog beamforming design. Whereas SB-NOMA generates a single beam for NOMA users, MB-NOMA creates multiple beams for each user. Adding MB-NOMA to SB-NOMA allows the BS to serve more users using NOMA while introducing additional complexity such as SIC decoding at the receiver. Nonetheless, this system enables users to use the same time-frequency resources without the trouble and need for fast and accurate synchronization.

VI. ILLUSTRATIVE RESULTS AND DISCUSSIONS

In this section, we numerically evaluate the performance of the proposed schemes for mmWave hybrid systems. Specifically, we consider a rural environment using the statistical mmWave channel model and simulator NYUSIM. In Table 1, we summarize the specific values of the adopted simulation parameters.

To illustrate the effectiveness of our proposed schemes, we adopt three baseline schemes; two apply fully DBF (thus requiring $N_{RF}^{D} = M$ RF chains) and the other one uses HBF with $N_{RF}^{H} \leq M$. The following acronyms will be used to refer to the different schemes.

- DBF: denotes the scheme considered in [33], where the BS generates K directive beams, each one is steered in the AoD of the intended user.
- DBF-SB-NOMA: denotes the scheme proposed in [24], where the β-UC algorithm clusters the users within SU and multi-user SB-NOMA groups.
- HBF-SB-NOMA-TDMA: denotes the scheme proposed in [19].[3]
- β-HBF-SB-NOMA-TDMA: denotes the scheme proposed in Section IV.
- HBF-SMB-NOMA: denotes the scheme proposed in Section V.
- HBF-SB-NOMA: denotes the scheme applying HBF that solely employs SB-NOMA. For $G_{SU} + G_{SB} \leq N_{RF}$, HBF-SB-NOMA, HBF-SMB-NOMA, and β-HBF-SB-NOMA-TDMA are all equivalent. Otherwise, HBF-SB-NOMA is unable to manage connectivity for all users. For that, we select N_{RF}^{H} groups and prioritize the SB-NOMA groups to fulfill the requirements of maximum user handling.

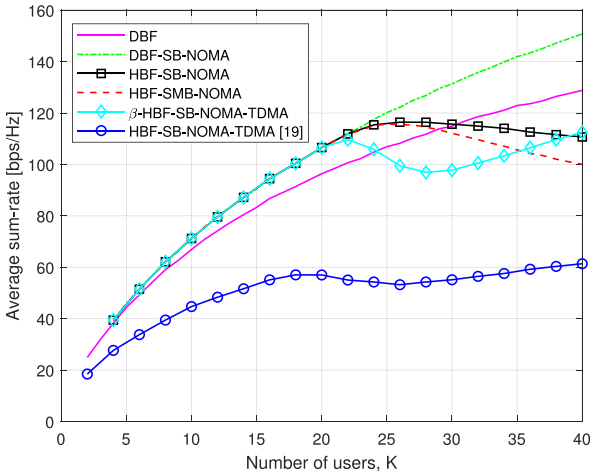
For a fair comparison, the different schemes apply our angle-domain user ordering and power allocation strategies designed in [24]. Therefore, all of them only require the users' AoDs.

A. 2D HBF-BASED MIMO-NOMA PERFORMANCE

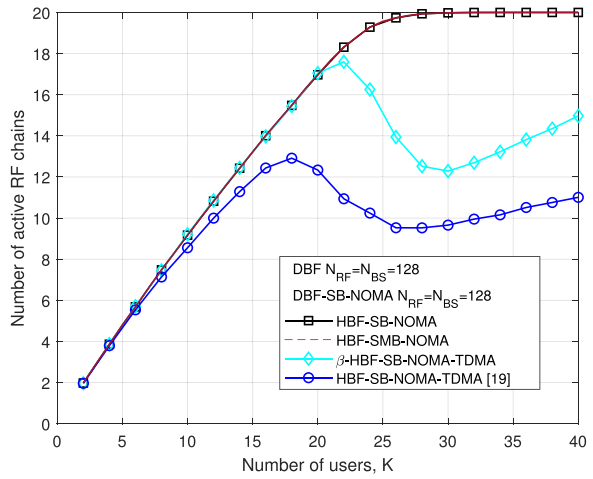
The BS adopts a ULA array with $M = 128$ transmit antennas to serve the $2 \leq K \leq 2N_{RF}^{H}$ users. While the fully DBF involves $N_{RF}^{D} = M = 128$ RF chains, we assume $N_{RF}^{H} = 20$ RF chains for HBF. Fig. 9 plots the sum-rate and the number of active RF chains, N_{RF}^{a}, per time slot versus the number of users for the different aforementioned schemes. And Fig. 10 plots the probability $P_{\text{os}} = \mathbb{P}(G_{SU} + G_{SB} > N_{RF}^{H})$ that the cell is overloaded using SB-NOMA in hybrid LSAAs systems.

Obviously, for a sparse cell with $K \leq N_{RF}^{H}$, the hybrid schemes are the optimal MIMO-NOMA solution, as they provide the same performance as DBF-SB-NOMA with much lower complexity, cost, and power consumption. This is evident from Fig. 9a, where the HBF-SB-NOMA, HBF-SMB-NOMA, β-HBF-SB-NOMA-TDMA curves merge with the DBF-SB-NOMA curve, implying that the cell is not yet overloaded, as also illustrated in Fig. 10, where $P_{\text{os}} = 0$ for $K \leq 20$. Otherwise, when the number of users exceeds the RF chains available, the cell begins to experience an overloaded scenario due to very narrow beams with LSAAs, as seen in Fig. 10. However, in this scenario, HBF-SB-NOMA fails to provide full connectivity, even if it outperforms other HBF schemes in terms of sum-rate performance. To fulfill the requirements of maximum

3. In [19], the authors propose a simple AD UC strategy, in which UEs with the same estimated angle belong to the same group. To offer new DoFs, the authors developed a group clustering algorithm in the time domain by scheduling the groups with small angular distances in distinct time slots. Note that the estimated angles belong to a predefined azimuth angle set with a fixed search step size $J = M$ [19].



(a) Sum-rate per time slot.



(b) Number of active RF chains per time slot.

FIGURE 9. (a) The sum-rate and (b) the number of active RF chains per time slot versus the number of users with a 128 × 1 ULA array and 20 RF chains at the BS.

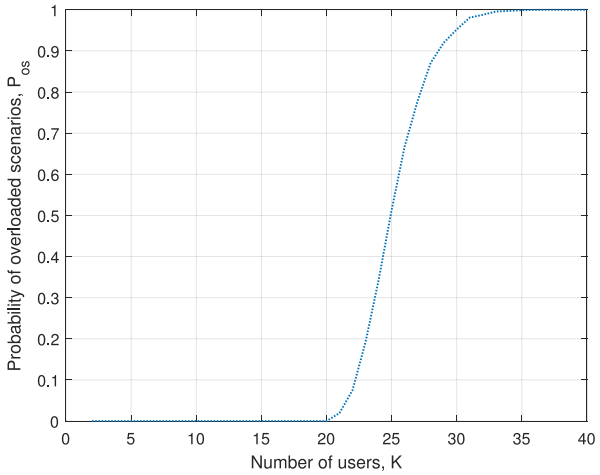


FIGURE 10. The impact of the number of users on the probability P_{os} that the cell is overloaded using SB-NOMA in hybrid LSAAs systems.

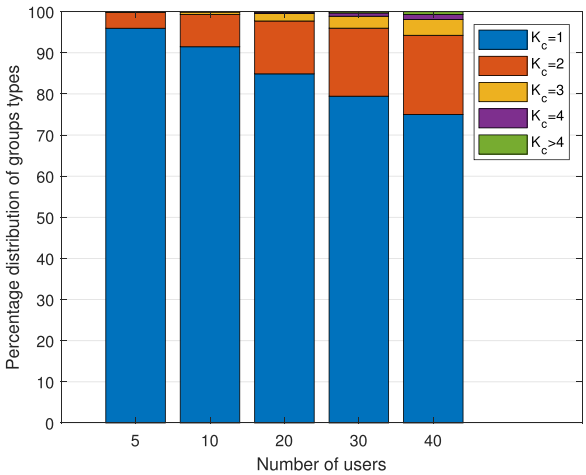


FIGURE 11. Group's percentage distribution per type versus the number of users, when the BS adopts HBF-SB-NOMA to serve $K \in \{5, 10, 20, 30, 40\}$ users.

user handling, HBF-SB-NOMA selects N_{RF}^H groups and prioritize the SB-NOMA groups. From Fig. 9(a), we can see that its system sum-rate slightly degrades for $K \geq N_{RF}^H$. Indeed, as more users are present in the cell, they are more likely to be clustered into the same NOMA groups. This is also illustrated in Fig. 11, which shows the percentage distribution of groups when the BS adopts HBF-SB-NOMA for different values of K. We can see that with increasing K, the number of NOMA groups increases, as well as the number of users per NOMA group. However, up to N_{RF}^H groups can only be connected simultaneously in hybrid schemes. Therefore, as K grows, the number of groups remains at N_{RF}^H, but the number of served users in each group increases, leading to increased interference. This can be observed in the slight degradation of the system sum-rate when comparing HBF-SB-NOMA for $K = 25$ and $K = 40$.

To ensure full connectivity for all users and maintain fairness among them, we apply other approaches, such as TDMA and MB-NOMA, which compensate for the limited RF chains. As seen in Fig. 9, they offer a good sum-rate performance compared to DBF with $6.4 \leq M/N_{RF}^a \leq 64$ times fewer active RF chains. In the following, we will separately analyze the evolution of their sum-rate curves for $N_{RF}^H \leq K \leq 2N_{RF}^H$.

1) PERFORMANCE OF β-HBF-SB-NOMA-TDMA

As seen from Fig. 9(a), for $N_{RF}^H < K \leq 2N_{RF}^H$, the sum-rate of β-HBF-SB-NOMA-TDMA first decreases, then increases with the number of users, K. Indeed, for cells where $N_{RF}^H < K \leq 2N_{RF}^H$, TDMA is used to serve the SU and SB-NOMA groups within different time slots and get the full connectivity. Scheduling the users in separate time slots reduces the system sum-rate per slot and the need for more active RF chains, as seen in Fig. 9(b). Then, in the second step by continuing to connect more users to the BS, the sum-rate increases again as well as the number of active RF

chains. For that reason, the evolution of the sum-rate matches that of the number of active RF chains in Fig. 9(b).

Furthermore, Fig. 9(a) shows that β-HBF-SB-NOMA-TDMA scheme outperforms the one proposed in [19]. For example, when $K = 2N_{RF}^{H} = 40$ users, β-HBF-SB-NOMA-TDMA achieves a sum-rate gain up to 83% over HBF-SB-NOMA-TDMA [19]. This significant gain reveals the performance of our proposed user and group clustering algorithms in SB-NOMA groups and time slots, respectively, against those proposed in [19]. Moreover, while the latter is proposed only for ULA, our proposed β-HBF-SB-NOMA-TDMA scheme uses β that can be applied to any array architecture. The performance evaluation with URA will be illustrated in Section VI-B.

2) PERFORMANCE OF HBF-SMB-NOMA

HBF-SMB-NOMA, is proposed to provide more DoFs using only NOMA. As seen in Fig. 9, it has superior performance compared to DBF for $K < 28$ with $6.4 \leq M/N_{RF}^{a} \leq 64$ times fewer active RF chains. Otherwise, DBF (with or w/o NOMA) has a greater sum-rate at the expense of higher cost and complex processing. Indeed, N_{RF}^{a} is smaller than or equal to $N_{RF}^{H} = 20$ for HBF-SMB-NOMA however, it always equals to $M = 128$ for DBF (with or w/o NOMA).

The sum-rate of HBF-SMB-NOMA increases up until $K = 26$, and then it decreases as shown in Fig. 9(a). This evolution can be explained by that of the number of active RF chains, N_{RF}^{a}, in Fig. 9(b). Once the cell reaches $K = 26$ users, N_{RF}^{a} starts to stabilize at N_{RF}^{H}, as seen in Fig. 9. To meet the requirement of $G_f = N_{RF}^{H}$ in the congested cell, the number of MB-NOMA groups must be increased and that of SU groups must be decreased. However, this comes with a decrease in performance, as seen in Fig. 9 for $26 < K \leq 40$, as the BF gain at each user within the MB-NOMA group is halved compared to that of the single beam in an SU group, as depicted in Fig. 4. Even with a basic antenna allocation, whereby the array is split into two sub-arrays of equal size, and a straightforward MB-NOMA user clustering strategy, HBF-SMB-NOMA achieves a promising sum-rate performance with few active RF chains when compared to digital methods. Future works should focus on designing angle-domain user clustering and antenna allocation methods that maximize the BF gain of each user about the other users and minimize inter-group interference to optimize the system's sum-rate performance.

For a better comparison between HBF-SB-NOMA and HBF-SMB-NOMA, we plot in Fig. 12 the cumulative distribution function (CDF) of the probability $\mathbb{P}(\text{rate} \leq \delta)$ that the user's rate is lower than a threshold, δ, for $K = 25$ and $K = 40$. We can observe that HBF-SB-NOMA fails to manage all of the users, leaving 4% and 25% unserved for $K = 25$ and $K = 40$, respectively, while HBF-SMB-NOMA offers full connectivity with $\mathbb{P}(\text{rate} = 0) = 0$. However, HBF-SB-NOMA provides higher rates to its served users than HBF-SMB-NOMA as the latter serves a greater number of users, leading to increased inter-user interference.

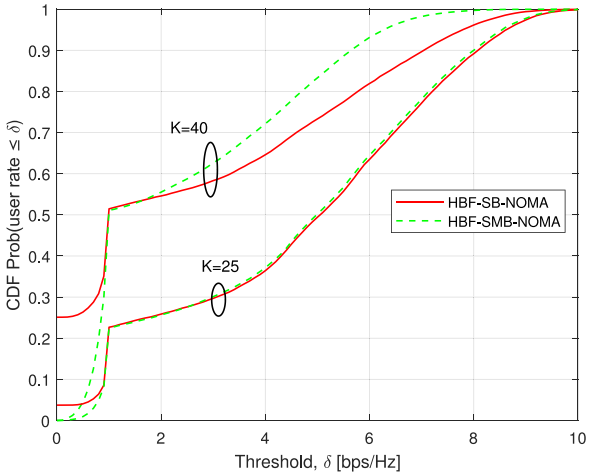


FIGURE 12. The CDF of the probability that the user's rate is lower than a threshold, δ, when the BS adopts either HBF-SB-NOMA or HBF-SMB-NOMA to serve $K \in \{25, 40\}$ users.

Furthermore, the BF gain when using MB-NOMA is halved against that of SB-NOMA. For example, if $K = 40$, the probability of HBF-SMB-NOMA providing a rate greater than 6 [bps/Hz] is 7%, while it is 18% for HBF-SB-NOMA. Consequently, even though HBF-SB-NOMA offers a higher system sum-rate, as also seen in Fig. 9(a), HBF-SMB-NOMA manages the connectivity of all users and ensures better fairness among them.

Moreover, it can be seen that HBF-SMB-NOMA brings a notable sum-rate gain when compared to HBF-SB-NOMA-TDMA [19]. For instance, with $K = 25$ and $K = 40$, the gain of HBF-SMB-NOMA is nearly 99% and 66.6% respectively. Additionally, for $N_{RF}^{H} < K < 35$, HBF-SMB-NOMA outperforms our proposed β-HBF-SB-NOMA-TDMA scheme. However, for $35 < K < 40$, β-HBF-SB-NOMA-TDMA gains slightly better performance but by requiring fast and accurate synchronization at the BS and the receiver sides.

B. 3D HBF-BASED MIMO-NOMA PERFORMANCE

Due to the size constraints of mMIMO antennas, URA is more practical for mMIMO than ULA [31]. In this section, we consider a 64×8 URA at the BS, meaning that there are $N_{RF}^{D} = 512$ RF chains for DBF. And we consider only $N_{RF}^{H} = 20$ RF chains for HBF. Fig. 13 plots the sum-rate for the different schemes versus the number of users. The TDMA-based scheme in [19] is specifically tailored for ULA and therefore is not included in this section. Instead, we consider our proposed TDMA-based scheme for comparison.

It is clear that the sum-rate curves for URA and ULA in Figs. 13 and 9 have a similar overall shape. However, the number of users K_i at which DBF curves separate from those of HBF is different. Specifically, $K_i = 20$ (resp. $K_i = 22$) for 128×1 ULA (resp. 64×8 URA). This comparison shows that, with URA, more users can be clustered in SB-NOMA despite the total number of antennas being four times greater than what ULA offers. Indeed, the 3D beam-width in both azimuth and elevation is determined by the number of

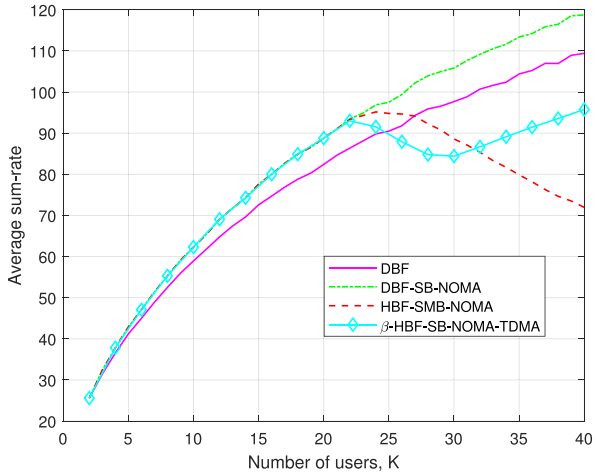


FIGURE 13. The sum-rate per time slot versus the number of users with 64 × 8 URA and 20 RF chains at the BS.

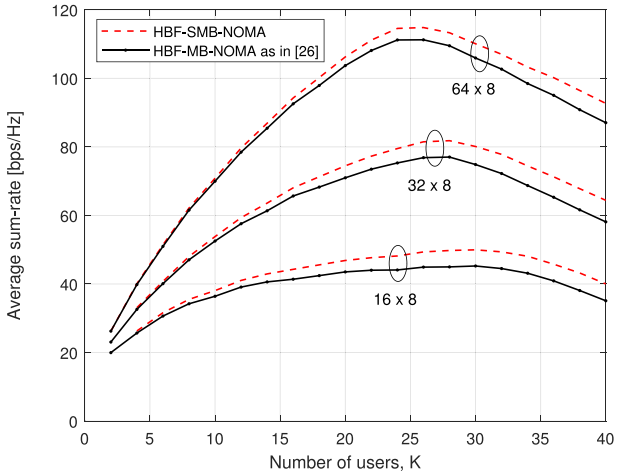


FIGURE 14. Average sum-rate of HBF-SMB-NOMA and HBF-MB-NOMA versus the number of users, K. The BS adopts a 16 × 8, 32 × 8 or 64 × 8 URA array with 20 RF chains.

horizontal and vertical antennas, respectively, that are smaller than the number of antennas in ULA.

C. JOINT SB- AND MB-NOMA VERSUS ONLY MB-NOMA

Now, we will verify the potentiality of using SB- and MB-NOMA frameworks w.r.t. those using only MB-NOMA. Fig. 5 shows that when both users are in each other's main lobe, SB-NOMA slightly performs better than MB-NOMA. Otherwise, MB-NOMA outperforms and has a constant sum-rate. Note that for a 2-user scenario, there is no inter-beam interference impact on the system sum-rate performance. However, SB-NOMA forms a beam using the whole array, while with the BST, the steered beam of MB-NOMA is the superposition of two wider beams. Hence, the beam generated with MB-NOMA when the users are in each other's main lobe is larger than that with SB-NOMA. Therefore, even when the users are in each other's main lobe, SB-NOMA outperforms MB-NOMA thanks to its narrow beams, which increase the BF gain and decrease inter-group interference.

To illustrate this, we compare HBF-SMB-NOMA with HBF-MB-NOMA, which applies only MB-NOMA without SB-NOMA, as in [26]. We assume that the BS adopts a 16×8, 32×8 or 64×8 URA array with 20 RF chains. For HBF-MB-NOMA, since we only consider 2-user MB-NOMA groups, we apply the 2-user β-UC algorithm [23] in the first phase. Fig. 14 plots the average sum-rate for the two schemes versus the number of users, K. We find that HBF-SMB-NOMA is superior to HBF-MB-NOMA in terms of average sum-rate, with the gap between the two increasing with K. The gap reaches its maximum point at K_a, depending on the antenna array configuration. For instance, $K_a = 32, 27$, and 25 for 16×8, 32×8, and 64×8 URA, respectively. Indeed, as the number of users in a cell increases, it is more likely that they are located near each other, causing the gain gap to rise with K until it reaches K_a, the point at which the NOMA groups in the first phase have saturated. The results verify the potentiality of using both SB- and MB-NOMA instead of only MB-NOMA to benefit from the multi-user NOMA diversity.

VII. SUMMARY

This paper considers HBF-based mMIMO-NOMA systems at mmWave frequencies. In particular, we address the limitation of SB-NOMA to exploit the multi-user diversity in mmWave hybrid systems with LSAAs. We have proposed two schemes offering additional DoFs. Contrary to the work done in the literature, we leverage the directionality of mmWave channels, and we use only angular information. We consider 2D and 3D systems using ULA and URA architectures, respectively. The first scheme adopts TDMA, and the other one leverages the potential of MB-NOMA. Simulation results have shown that they yield significant performance gains in terms of sum-rate, compared to the solution proposed in [19] and other schemes based on fully DBF. For instance, the proposed TDMA-based scheme achieves a sum-rate gain of up to 83% over the existing one. Furthermore, the results demonstrate the effectiveness of the proposed joint SB- and MB-NOMA scheme in providing more DoFs without applying additional multiple access techniques. Moreover, they verify the superiority of this scheme over those using only MB-NOMA. However, their spectral and energy efficiencies are not optimized. An angle-domain resource allocation method, including antenna selection, user clustering, and power allocation to achieve a significant gain, is left as a future work. Besides, it would be interesting to include more practical HBF structures, such as sub-connected HBF, which are known to have reduced power consumption compared to the fully-connected HBF and allow for more energy-efficient designs. Given the potential of the reconfigurable intelligent surfaces (RIS) technology to increase coverage in mmWave communications, which has attracted a great deal of attention, it is of particular interest to examine our proposed schemes in the presence of RIS, as done in [34].

REFERENCES

[1] W. Xiang, K. Zheng, and X. S. Shen, *5G Mobile Communications*. Cham, Switzerland: Springer, 2016.

[2] A. Osseiran, J. F. Monserrat, and P. Marsch, *5G Mobile and Wireless Communications Technology*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[3] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[4] A. F. Molisch et al., "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.

[5] W. Roh et al., "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

[6] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

[7] K. Xu, Z. Shen, Y. Wang, and X. Xia, "Location-aided mMIMO channel tracking and hybrid beamforming for high-speed railway communications: An angle-domain approach," *IEEE Syst. J.*, vol. 14, no. 1, pp. 93–104, Mar. 2020.

[8] D. Fan, F. Gao, G. Wang, Z. Zhong, and A. Nallanathan, "Angle domain signal processing-aided channel estimation for indoor 60-GHz TDD/FDD massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1948–1961, Sep. 2017.

[9] D. Fan et al., "Angle domain channel estimation in hybrid millimeter wave massive MIMO systems," *IEEE Trans. Commun.*, vol. 17, no. 12, pp. 8165–8179, Dec. 2018.

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

[11] W. Yuan, V. Kalokidou, S. M. D. Armour, A. Doufexi, and M. A. Beach, "Application of non-orthogonal multiplexing to mmWave multi-user systems," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, 2017, pp. 1–6.

[12] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Millimeter-wave NOMA with user grouping, power allocation and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5065–5079, Nov. 2019.

[13] X. Yu, F. Xu, K. Yu, and X. Dang, "Power allocation for energy efficiency optimization in multi-user mmWave-NOMA system with hybrid precoding," *IEEE Access*, vol. 7, pp. 109083–109093, 2019.

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

[15] A. N. Uwaechia and N. M. Mahyuddin, "Spectrum and energy efficiency optimization for hybrid precoding-based SWIPT-enabled mmWave mMIMO-NOMA systems," *IEEE Access*, vol. 8, pp. 139994–140007, 2020.

[16] A. A. Badrudeen, C. Y. Leow, and S. Won, "Performance analysis of hybrid beamforming precoders for multiuser millimeter wave NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8739–8752, Aug. 2020.

[17] L. Pang et al., "Joint power allocation and hybrid beamforming for downlink mmwave-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10173–10184, Oct. 2021.

[18] A. A. Badrudeen, C. Y. Leow, and S. Won, "Sub-connected structure hybrid precoding for millimeter-wave NOMA communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 6, pp. 1334–1338, Jun. 2021.

[19] X. Hu, C. Zhong, X. Chen, W. Xu, and Z. Zhang, "Cluster grouping and power control for angle-domain mmWave MIMO NOMA systems," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1167–1180, Sep. 2019.

[20] J. Wang, Y. Li, C. Ji, Q. Sun, S. Jin, and T. Q. S. Quek, "Location-based MIMO-NOMA: Multiple access regions and low-complexity user pairing," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2293–2307, Apr. 2020.

[21] E. M. Mohamed, "Joint users selection and beamforming in downlink millimetre-wave NOMA based on users positioning," *IET Commun.*, vol. 14, no. 8, pp. 1234–1240, 2020.

[22] X. Lu, Y. Zhou, and V. W. S. Wong, "A joint angle and distance based user pairing strategy for millimeter wave NOMA networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2020, pp. 1–6.

[23] I. Khaled, C. Langlais, A. El Falou, M. Jezequel, and B. ElHasssan, "Joint SDMA and power-domain NOMA system for multi-user mm-wave communications," in *Proc. IEEE Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2020, pp. 1112–1117.

[24] I. Khaled, C. Langlais, A. El Falou, B. A. ElHassan, and M. Jezequel, "Multi-user angle-domain MIMO-NOMA system for mmWave communications," *IEEE Access*, vol. 9, pp. 129443–129459, 2021.

[25] Z. Xiao, L. Dai, Z. Ding, J. Choi, and P. Xia, "Millimeter-wave communication with non-orthogonal multiple access for 5G," 2017, *arXiv:1709.07980*.

[26] Z. Wei, L. Zhao, J. Guo, D. W. K. Ng, and J. Yuan, "A multi-beam NOMA framework for hybrid mmWave systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–7.

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

[28] I. Khaled, A. El Falou, C. Langlais, B. El Hassan, and M. Jezequel, "Joint single-and multi-beam angle-domain NOMA for hybrid mmWave MIMO systems," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, 2021, pp. 1–6.

[29] T. S. Rappaport, S. Sun, and M. Shafi, "Investigation and comparison of 3GPP and NYUSIM channel models for 5G wireless communications," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, 2017, pp. 1–5.

[30] H. J. Visser, *Array and Phased Array Antenna Basics*. Hoboken, NJ, USA: Wiley, 2005.

[31] X. Su et al., "Limited feedback precoding for massive MIMO," *Int. J. Antennas Propag.*, vol. 2013, pp. 1–9, Aug. 2013.

[32] G. Lee, Y. Sung, and J. Seo, "Randomly-directional beamforming in millimeter-wave multiuser MISO downlink," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1086–1100, Feb. 2016.

[33] I. Khaled, A. El Falou, C. Langlais, B. El Hassan, and M. Jezequel, "Multi-user digital beamforming based on path angle information for mm-wave MIMO systems," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, 2020, pp. 1–6.

[34] Y. Xiu et al., "Reconfigurable intelligent surfaces aided mmWave NOMA: Joint power allocation, phase shifts, and hybrid beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8393–8409, Dec. 2021.