IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 2, 2020, accepted July 28, 2020, date of publication July 31, 2020, date of current version August 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013305

Spectrum and Energy Efficiency Optimization for Hybrid Precoding-Based SWIPT-Enabled mmWave mMIMO-NOMA Systems

ANTHONY NGOZICHUKWUKA UWAECHIA AND NOR MUZLIFAH MAHYUDDIN, (Member, IEEE)
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Nibong Tebal 14300, Malaysia

Corresponding author: Nor Muzlifah Mahyuddin (eemnmuzlifah@usm.my)

This work was supported in part by the Postdoctoral Fellowship Program of the School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Malaysia, and in part by the Ministry of Education through the Fundamental Research Grant Scheme, Malaysia, under Grant FRGS 203/PELECT/6071373.

ABSTRACT Non-orthogonal multiple access (NOMA) in millimeter-wave (mmWave) multiple-input multiple-output (MIMO) (i.e., mmWave MIMO-NOMA) systems, is a promising technology to significantly enhance the spectrum efficiency of the fifth-generation (5G) mobile communication systems. Furthermore, enabling simultaneous wireless information and power transfer (SWIPT), where the energy-constrained user equipment (UE) equipped with power splitting receiver harvest both information and energy from the ambient radio-frequency (RF) signal, is crucial for energy-efficiency-maximization. In this paper, we initially design a user grouping algorithm, which preferentially groups UEs based on their channel correlation. Then, we design the analog RF precoder based on the selected user grouping for all beams, followed by a low-dimensional digital baseband precoder design, to further mitigate inter-beam interference and maximize the achievable sum-rate for the considered system. Subsequently, we equivalently transform the original optimization problem into a joint power allocation and power splitting maximization problem. Then, we propose to decouple the joint power allocation and power splitting nonconvex optimization problem into four separate optimization problems and then solved iteratively via an alternating optimization (AO) algorithm. Simulation results show that high spectrum and energy efficiency can be realized with the proposed algorithms than those of state-of-the-art designs and the conventional SWIPT-enabled mmWave MIMO-OMA system.

INDEX TERMS SWIPT, millimeter wave communication, MIMO, NOMA, hybrid analog-digital precoding, multiuser channels, optimal scheduling, power splitting, optimization methods.

I. INTRODUCTION

Millimeter-wave (mmWave) communications which operates in the $30-300$ gigahertz (GHz) band combined with massive multiple-input multiple-output (mMIMO) (i.e., mmWave mMIMO) system [1]–[4], can realize orders of magnitude increase in system throughput owing to broader bandwidth availability and higher spectral efficiency. Hence, considered a promising technology for the fifth-generation (5G) and beyond wireless communication systems [2], [5], [6]. In conventional multiantenna systems (i.e., for cellular frequency band $2-3$ GHz), multi-stream beamforming, i.e., precoding, is entirely realized in the digital domain at baseband to cancel interference among several data streams [5], [7]. For such conventional full-digital precoding, in which each antenna connects to a dedicated energy-intensive radio frequency (RF) chain [2], [5], [7]–[9], may not be a practical solution for mmWave mMIMO systems due to the excessive power consumption and prohibitive hardware complexity. Meanwhile, the large antenna array in mMIMO systems, which can be pack in small-form-factor at mmWave frequencies enables precoding multiple data streams and provides the beamforming gain needed to overcome path loss [2], [5]. Hence, mmWave mMIMO systems

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

preclude such full digital precoding transceiver architecture at present, and will almost certainly use hybrid precoding to realize beamforming with fewer RF chains to reduce energy consumption [5], [7], [10]. Specifically, the main idea of hybrid precoding is employing a linear network of variable phase shifters in the RF domain along with the baseband digital precoding [5], [7], [10]–[16], which makes analog processing in the RF domain more attractive. Based on the RF-antenna mapping strategy for mmWave mMIMO systems, which defines the number of required phase shifters [2], [5], [11], the hybrid precoding transceiver architectures can be classified as either partially-connected or fully-connected structures. The fully-connected structure experiences full beamforming gain for each RF chain (i.e., each RF chain connects to all antennas) [2], [8], while the partially-connected structure sacrifices some beamforming gain (i.e., each RF chain connects to a sub-antenna array) to improve energy efficiency and reduce hardware implementation complexity [7], [8], [17]. While the conventional mmWave mMIMO systems with orthogonal multiple access based hybrid precoding (i.e., mmWave mMIMO-OMA) improve energy efficiency and implementation cost, the channel time-frequency resource block in each beam cannot support multiple user equipments[1] (UEs) [8], [18]–[20]. Thus, leads to a limited number of served UEs [8], [18], since the number of served UEs cannot exceed the number of transmitter RF chains at the same time-frequency resource block.

Non-orthogonal multiple access (NOMA) is being considered a promising multiple access technique for 5G mmWave mobile communication systems [20]–[22], due to its superior spectrum efficiency. To further enhance the spectrum efficiency, additional analysis is required to combine NOMA with mmWave mMIMO systems [23] and, thus, the so-called mmWave mMIMO-NOMA systems. Unlike mmWave MIMO-OMA systems, mmWave MIMO-NOMA systems[2] serves multiple UEs per beam at the same time-frequency resource-block [8], [20], and employs intra-beam superposition coding at BS and successive interference cancellation (SIC) at the receivers of the UEs with better channel conditions. In other words, it effectively reduces the operational cost and energy consumption by requiring less number of RF chains and phase shifter (PS) network at BS and UE sides compared to the mmWave MIMO system with the same number of antennas and UEs.

The mmWave mMIMO-NOMA systems has received significant research attention from both academia and industry. Specifically, the application of NOMA to beamspace MIMO was initially investigated in the seminal work of [19], where they proposed an optimal power allocation strategy to maximize the achievable sum rate. In [20], a joint Tx-Rx beamforming and power allocation problem for the downlink mmWave MIMO-NOMA system for K-UE was solved, using boundary compressed particle swarm optimization algorithm. The work in [12] proposed a new hybrid precoding-based beamspace MIMO-NOMA transmission scheme, to serve more users than RF chains quasi simultaneously, using an iterative algorithm to find the optimal power allocation for maximizing the achievable sum rate. According to the channel correlations, the authors in [10] first proposed a K-means user grouping algorithm, where multiple UEs with distinct channel gains are grouped into mmWave mMIMO-NOMA clusters, such that one user per cluster is initially randomly selected as the representatives of the cluster. The authors in [10] then solve the joint hybrid analog-digital precoding and power allocation design problem, to maximize the achievable sum-rate per user constraints. However, since the traditional K-means has a sensitivity to the initial selection [24], then different starting initial users per cluster selected randomly may produce different clustering results. Hence, contrarily to randomly selecting initial UEs, an attempt to solve the cluster initialization problem for the traditional K-means user grouping algorithm, may minimize both inter and intra cluster interference and thus, improve the overall system performance.

Furthermore, besides improving the spectral efficiency of transmission, energy efficiency is becoming a significant key performance indicator (KPI) in evaluating 5G and beyond wireless communication networks [12], [13]. One key drawback of existing hybrid precoding solutions lies in the assumption of infinite- (or high-) resolution phase shifters [8], [25], [26]. Typically, the power consumption of a $30-300$ GHz $8-12$ bits, phase shifter is approximately 500 mW [25], while a $30-300$ GHz 4-bit, 22.5° phase resolution phase shifter requires $45-106$ mW [2], [28]. To solve this challenging problem, a promising solution is to substitute power-hungry high-resolution phase shifters (e.g., $8-12$ bits) with low power low-resolution phase shifters (e.g., $1-4$ bits) [25]–[27], to reduce the hardware cost and power consumption [2]. However, significant signal processing problems and complicated front-end designs (i.e., phase/frequency synchronization and multiuser detection) inevitably arise, owing to the strong nonlinearity distortion imposed by coarse quantization [25], [26]. To realize a better energy-rate trade-off, the authors in [26] work on the idea of mixed-analog-to-digital converters (ADCs)/digital-to-analog converters (DACs) architecture, by combining low- and high-resolution ADCs/DACs. However, the work in [26] was realized, with a fixed resolution for each ADC, without considering hybrid analog-digital precoding. In the meanwhile, the notion of energy harvesting (EH) has recently been proposed to provide valuable and continuous energy supplies. On that premise, examining the information-carrying RF signals as an additional energy source for the EH, has been regarded as a promising solution to realize this goal [8], [10]. The pioneering work

[1]We interchangeably use the acronym *user equipment (UE)* and the term *user* in the paper.

[2]We remark here that NOMA implementation for mmWave MIMO systems, in no case lead to additional delay resulting from channel estimation (CE) and feedback in comparison to OMA [8], [21]. While the SIC at the receiver utilizes profuse computation for signal demodulation and decoding at the NOMA user equipment (UE), the equivalent physical layer latency is insignificant in comparison to the network layer delay.

on simultaneous wireless information and power transfer (SWIPT), proposed in the seminal work of [29], enables wireless devices to harvest energy from receiving RF signals while guaranteeing information transmission. As manifested in [29] and [30] groundbreaking seminal works, which characterized from an information-theoretical perspective, the boundaries on rate and energy of point-to-point SWIPT systems, has resulted in a growing interest in the practical implementation of SWIPT transmission schemes and protocols. In [31], a joint transceiver design and power splitting for the downlink hybrid precoding-based multiuser MIMO system, with SWIPT, is investigated. The authors in [31] then proposed an iterative optimization scheme that solves the joint transmitter and power splitting factor optimization subproblem and a receiver side minimum mean-square error (MMSE) minimization subproblem. In [10], a SWIPT hybrid precoding-based mmWave mMIMO-NOMA system is introduced, where user grouping, hybrid precoding, power allocation, and power splitting factors are designed to enable spectrum- and energy-efficient systems. However, the system performance can be enhanced further with more sophisticated hybrid precoding design and user grouping algorithm.

This paper takes a different approach to solving the joint hybrid precoding, power allocation and power splitting optimization problem with user grouping for SWIPT-enabled mmWave mMIMO-NOMA Systems. The contributions of this paper are summarized as follows.

1) To realize the SWIPT-enabled mmWave mMIMO-NOMA systems with hybrid analog-digital Precoding, we investigate the joint hybrid analog/digital precoding and power splitting optimization. Specifically, we initially propose a modified K-means user grouping algorithm to solve the cluster initialization problem, which uses the normalized channel correlation among user channels as the parameter for grouping. Then, we formulate the achievable sum-rate maximization problem that jointly optimizes the hybrid analog-digital precoders, power allocation, and power slitting factors, under the total power and minimum rate constraints at each UE.
2) To address this problem, we now aim to design the hybrid mmWave MIMO-NOMA precoding matrix. First, the analog precoder is designed to maximize the equivalent channel gain based on the selected user grouping for all beams. Next, we design the digital precoding vector for each UE, which aims at cancelling the inter-user interference (at the expense of losing some beamforming gain) by selecting the UEs with the strongest equivalent channel gain per beam. With the hybrid analog/digital precoding matrices fixed, we then reformulate the achievable sum-rate maximization problem as a joint optimization of power allocation and power splitting factors, subject to the total power and minimum rate constraints at each UE.
3) To address this problem, in contrast to [8], we propose to decouple the joint power allocation and power splitting nonconvex optimization problem into four separate optimization problems and then solved via an alternating optimization (AO) algorithm that sequentially finds the solution to the problem by updating the transceiver coefficients, power slitting factor, power allocation and a real-valued slack variable, iteratively.
4) From the perspective of spectrum and energy efficiency, the proposed SWIPT-enabled mmWave mMIMO-NOMA system with hybrid analog-digital precoding is evaluated by simulations. Simulation results demonstrate that our proposed algorithms can significantly enhance the spectrum and energy efficiency performance of the considered system than those of state-of-the-art designs and the (conventional) SWIPT-enabled mmWave mMIMO-OMA system.

The remainder of the paper is organized as follows. In Section II, we introduce the system model of the SWIPT-enabled mmWave mMIMO-NOMA system with hybrid analog-digital precoding and the sum-rate problem formulation. In Section III, we design the user grouping algorithm, hybrid analog-digital precoders, and then formulate the problem. In Section III, we find a solution to the joint power allocation and power splitting nonconvex optimization problem via an AO algorithm under the total power and minimum rate constraints at each UE. Finally, Section IV concludes the paper.

Notations: We use a to denote a scalar, whereas lower- ($\mathbf{a}$) and upper-case boldface ($\mathbf{A}$) letters denote vectors and matrices, respectively. By $A \succeq 0$, we means that matrix A is positive semidefinite. $\emptyset$ denotes the empty set. $\mathbb{C}$ denotes the complex field. We use $\mathbf{h}^H$, $\mathbf{h}^T$ and $\mathbf{h}^{-1}$ to denote the Hermitian transpose, transpose and inverse of $\mathbf{h}$, respectively. We use $\|\cdot\|_2$ and $\mathbb{E}[\cdot]$ to denote the ℓ_2 vector norm and statistical expectation, respectively. We use $|\cdot|$ interchangeably to denote the absolute value and as $|\Omega|$ to denote the number of elements in set Ω. We use $[\mathbf{f}^{\mathrm{RF}}]_i$, $[\mathbf{F}^{\mathrm{RF}}]_{i,:}$, $[\mathbf{F}^{\mathrm{RF}}]_{:,j}$ and $[\mathbf{F}^{\mathrm{RF}}]_{i,j}$ to denote the ith entry of $\mathbf{f}^{\mathrm{RF}}$, the ith row, the jth column, and the entry in the ith row and the jth column of $\mathbf{F}^{\mathrm{RF}}$, respectively. While $\mathbf{0}^M$ and $\mathbf{0}^{M\times N}$ denote the zero vector with dimension M and the $M \times N$ zero matrix, respectively.

II. SYSTEM MODEL AND SUM-RATE FORMULATION

A. SYSTEM MODEL

Consider the multiuser downlink mmWave MIMO-NOMA system depicted in Fig. 1 in which the base station (BS) is equipped with N transmit antennas but only N^{RF} RF chains to support K single-antenna user equipment (UEs) through spatial multiplexing such that $N^{\mathrm{RF}} \leq K < N$. Here, each UE is equipped with a power splitter[3] for RF energy harvesting to prolong their lifetime. By exploiting hybrid analog-digital precoding, the number of beams $\mathcal{D}$ that can be simultaneously transmitted cannot surpass the number of RF chains, N^{RF}.

[3]In practice, the power-splitting protocol is more efficient with a higher transmission rate than the time-switching protocol since it does not waste the time resource for transmission [32].

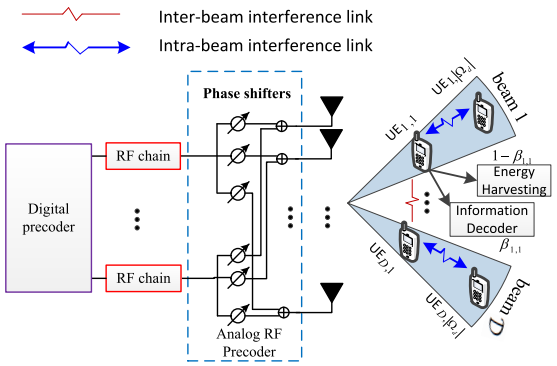

FIGURE 1. The system model for SWIPT-enabled mmWave MIMO-NOMA with fully-connected hybrid precoding architecture. Each receiver has two circuits to perform information decoding (ID) and energy harvesting (EH) separately, shown in $\mathrm{UE}_{1,1}$ as an example with $\beta_{1,1}$ denoting the power slitting factor.

Therefore, to realize efficient spatial multiplexing in spite of the limited RF chains, we assume $\mathcal{D} = N^{\mathrm{RF}}$. Thus, using mmWave mMIMO-NOMA, the K UEs need to be scheduled into D groups/clusters corresponding to the number of beams. This hardware structure permits the transmitter to apply an $N^{\mathrm{RF}} \times 1$ baseband precoder $\mathbf{f}_d^{\mathrm{BB}}$, $d \in [1, \mathcal{D}]$ for the dth beam using its N^{RF} transmit chains, and an $N \times N^{\mathrm{RF}}$ RF precoder $\mathbf{F}^{\mathrm{RF}}$ using analog circuitry. As the RF precoder $\mathbf{F}^{\mathrm{RF}}$ is realized using analog phase shifters, its elements are constrained to satisfy $([\mathbf{F}^{\mathrm{RF}}]_{:,j}[\mathbf{F}^{\mathrm{RF}}]_{:,j}^{H})_{\ell,\ell} = N^{-1}$ with the constant modulus constraint of $|[\mathbf{F}^{\mathrm{RF}}]_{i,j}| = 1/\sqrt{N}$, $1 \le i \le N$, $1 \le j \le N^{\mathrm{RF}}$ and quantized phases: $[\mathbf{F}^{\mathrm{RF}}]_{i,j} = \frac{1}{\sqrt{N}} \exp(j[\Theta]_{i,j})$ in which the phase $[\Theta]_{i,j}$ is quantized as $[\Theta]_{i,j} \in \mathcal{B} := \{\frac{2\pi n}{2^B} : n = 0, \ldots, 2^B - 1\}$. We have used $(\cdot)_{\ell,\ell}$ to denote the ℓth diagonal element of a matrix. Furthermore, the total power constraint is enforced by normalizing the baseband precoder $\mathbf{f}_d^{\mathrm{BB}}$, $d = 1, \ldots, \mathcal{D}$ for the dth beam to satisfy $\|\mathbf{F}^{\mathrm{RF}}\mathbf{f}_d^{\mathrm{BB}}\|_2 = 1$ for $d = 1, \ldots, \mathcal{D}$; no additional mmWave MIMO-NOMA hardware-related constraints are imposed on $\mathbf{f}_d^{\mathrm{BB}}$, $d = 1, \ldots, \mathcal{D}$ for the dth beam. Without loss of generality, let us define by $K = \sum_{d=1}^{\mathcal{D}} |\Omega_d|$, for $\Omega_i \cap \Omega_j = \emptyset$, $i \neq j$ and $|\Omega_d| \ge 1$ the total number of UEs served by the BS in the considered system. We have used Ω_d, $d = 1, \ldots, \mathcal{D}$ to denote the set of users served by the dth beam and $|\cdot|$ to denote the number of elements in set Ω_d for $d = 1, \ldots, \mathcal{D}$.

1) SUPERPOSITION CODING

From the NOMA transmission protocol, let us denote by $s_{d,m}$, where $\mathbb{E}[|s_{d,m}|^2] = 1$, the signal transmitted to the mth UE in the dth beam. During transmission, the BS precodes the signals $s_{d,m}$, $d = 1, \ldots, \mathcal{D}$, $m = 1, \ldots, |\Omega_d|$, according to the superposition coding technique to enable the transmission of multiple UEs' information at the same time. On that premise, all K UEs must share the total transmit power P_T. However, each UE must employ SIC to detect its own signal.

2) SUCCESSIVE INTERFERENCE CANCELLATION

After transmission, the received baseband signal $y_{d,m}$ for $d = 1, \ldots, \mathcal{D}$, $m = 1, \ldots, |\Omega_d|$ at the mth UE in the dth beam is written as

$$
\begin{aligned}
y_{d,m} &= \mathbf{h}_{d,m}^{H}\mathbf{F}^{\mathrm{RF}}\sum_{i=1}^{\mathcal{D}}\sum_{j=1}^{|\Omega_i|}\mathbf{f}_i^{\mathrm{BB}}\sqrt{\rho_{i,j}}s_{i,j} + n_{d,m} \\
&= \underbrace{\mathbf{h}_{d,m}^{H}\mathbf{F}^{\mathrm{RF}}\mathbf{f}_d^{\mathrm{BB}}\sqrt{\rho_{d,m}}s_{d,m}}_{\textbf{Desired signal}} \\
&\quad + \underbrace{\mathbf{h}_{d,m}^{H}\mathbf{F}^{\mathrm{RF}}\mathbf{f}_d^{\mathrm{BB}}\left(\sum_{j=1}^{m-1}\sqrt{\rho_{d,j}}s_{d,j} + \sum_{j=m+1}^{|\Omega_d|}\sqrt{\rho_{d,j}}s_{d,j}\right)}_{\textbf{Intra-beam interference}} \\
&\quad + \underbrace{\mathbf{h}_{d,m}^{H}\mathbf{F}^{\mathrm{RF}}\sum_{i\neq d}^{}\sum_{j=1}^{|\Omega_i|}\mathbf{f}_i^{\mathrm{BB}}\sqrt{\rho_{i,j}}s_{i,j}}_{\textbf{Inter-beam interference}} + \underbrace{n_{d,m}}_{\textbf{Noise signal}} \\
&= \mathbf{h}_{d,m}^{H}\mathbf{F}^{\mathrm{RF}}\mathbf{F}^{\mathrm{BB}}\mathbf{P}\mathbf{s} + n_{d,m} \qquad (1)
\end{aligned}
$$

where $\mathbf{h}_{d,m}^{H} \in \mathbb{C}^{1\times N}$ is channel response vector, $\mathbf{F}^{\mathrm{RF}} \in \mathbb{C}^{N\times N^{\mathrm{RF}}}$ is the analog precoder matrix, $\mathbf{F}^{\mathrm{BB}} = [\mathbf{f}_1^{\mathrm{BB}}, \ldots, \mathbf{f}_{\mathcal{D}}^{\mathrm{BB}}] \in \mathbb{C}^{N^{\mathrm{RF}}\times\mathcal{D}}$ is the digital precoder matrix, $\mathbf{P} = \mathrm{diag}\{\mathbf{p}_1, \ldots, \mathbf{p}_{\mathcal{D}}\}$ is the power allocation matrix with $\mathbf{p}_d = [\sqrt{\rho_{d,1}}, \sqrt{\rho_{d,2}}, \ldots, \sqrt{\rho_{d,|\Omega_d|}}]$ for $d = 1, \ldots, \mathcal{D}$, $\mathbf{s} = [s_{1,1}, \ldots, s_{1,|\Omega_d|}, \ldots, s_{\mathcal{D},1}, \ldots, s_{\mathcal{D},|\Omega_d|}]^T \in \mathbb{C}^{K\times 1}$ is the transmission vector signal, $\rho_{d,m}$ represents the average received power of the mth UE in the dth beam, and $n_{d,m} \sim \mathcal{CN}(0, \sigma_n^2)$ is the white complex Gaussian noise (AWGN) with zero mean and variance σ_n^2 added at the mth UE in the dth beam. Moreover, on the right-hand side of the second equation of (1), if we consider detecting $s_{d,m}$, the first term is the desired signal while the third term is the interference caused by adjacent beams, i.e., inter-beam interference. Since the mth UE cannot effectively eliminate the detected jth UE's signals $\forall 1 \le j \le |\Omega_d| - 1$, intra-beam interference results denoted by the second term. Therefore, SIC will be performed at UEs. Consequently, in the dth beam, the mth UE will detect the jth UE's signals for all $1 \le j < m \le |\Omega_d|$ and then effectively cancel the detected signals from its received signals, in a successive manner.

3) CHANNEL MODEL

Due to the sparse scattering nature of mmWave channels, we adopt a narrowband mmWave channel model, premised on the extended Saleh-Valenzuela model[4] [33], where the channel $\mathbf{h}_{d,m}$ of the mth UE in the dth beam is assumed to be a sum of the contributions of $\mathcal{L}_{d,m}$ scatters. Accordingly, the normalized single-cell mmWave mMIMO-NOMA channel between the mth UE in the dth beam denoted as $\mathbf{h}_{d,m}$,

[4]Specifically [15], the existing Saleh-Valenzuela model, proposed in [34], has typically been extended to the spatial domain in [35], [36], and generally termed the extended Saleh-Valenzuela [33] since it exploits angular domain information (i.e., AoA) in the channel model of [34].

$d \in [1, \ldots, \mathcal{D}]$, $m \in [1, \ldots, K]$ can be expressed as

$$\mathbf{h}_{d,m} = \gamma \sum_{l=1}^{\mathcal{L}_{d,m}} \alpha_{d,m}^{(l)} \Lambda\big(\theta_{d,m}^{(l)}, \Theta_{d,m}^{(l)}\big)\mathbf{a}\big(\theta_{d,m}^{(l)}, \Theta_{d,m}^{(l)}\big), \quad (2)$$

where γ denotes the normalization factor such that $\gamma = \sqrt{N/\mathcal{L}_{d,m}}$ in $\mathbf{h}_{d,m}$ ensures standard channel power normalization in MIMO system, $\alpha_{d,m}^{(l)}$ is the complex gain[5] of the lth path for the mth UE in the dth beam, whereas $\theta_{d,m}^{(l)}(\Theta_{d,m}^{(l)})$ are its azimuth (elevation) angles of departure (AoD) of the lth path, and $\mathbf{a}(\theta_{d,m}^{(l)}, \Theta_{d,m}^{(l)})$ denotes the normalized transmit array response vector at an azimuth (elevation) angle of $\theta_{d,m}^{(l)}(\Theta_{d,m}^{(l)})$. Assuming ideal sectored elements of the transmitting antenna, then $\Lambda\big(\theta_{d,m}^{(l)}, \Theta_{d,m}^{(l)}\big)$ would be modeled as

$$\Lambda\big(\theta_{d,m}^{(l)}, \Theta_{d,m}^{(l)}\big) = \begin{cases} 1, & \theta_{d,m}^{(l)} \in [\theta_{\min}^{(l)}, \theta_{\max}^{(l)}], \Theta_{d,m}^{(l)} \in [\Theta_{\min}^{(l)}, \Theta_{\max}^{(l)}], \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Without loss of generality, we have assumed that the sector described by $\theta_{d,m}^{(l)} \in [\theta_{\min}^{(l)}, \theta_{\max}^{(l)}]$ and $\Theta_{d,m}^{(l)} \in [\Theta_{\min}^{(l)}, \Theta_{\max}^{(l)}]$ have unit gain. Assuming the typical N-element uniform linear array (ULA), the array response vector $\mathbf{a}(\theta)$ is given by

$$\mathbf{a}(\theta) = \frac{1}{\sqrt{N}}\left(1, e^{jk\breve{d}\sin(\theta)}, \ldots, e^{j(N-1)k\breve{d}\sin(\theta)}\right)^T, \quad (4)$$

with the inter-element distance $\breve{d} = \lambda/2$, $k = 2\pi/\lambda$. We have written $(\cdot)^T$ to denote the transposition of a vector. Notice that (4) omits Θ since the array response in the elevation domain is invariant. In practice, the channels of the mth UE in the dth beam i.e., $\mathbf{h}_{d,m}$, $d = 1, \ldots, \mathcal{D}$, are not constant due to rapidly-varying channel state information (CSI). In the meanwhile, in formulating (1), we implicitly assume that carrier frequency recovery and timing are ideal. Furthermore, to enable SWIPT in mmWave MIMO-NOMA systems with hybrid precoding, we assume that the channel,[6] $\mathbf{h}_{d,m}$, is perfectly and instantaneously known to both the BS and UEs via channel estimation (CE).

As the power splitter receiver enables SWIPT, this splits the received signal $y_{d,m}$ into two parts for information decoding (ID) and energy harvesting (EH) in the proportion of $\beta_{d,m} : (1 - \beta_{d,m})$, where $\beta_{d,m} \in (0, 1)$ is termed as the power splitting factor for the mth UE in the dth beam. Using (1) the first part $\sqrt{\beta_{d,m}}y_{d,m}$ forms an input for ID at the mth UE in the gth beam as

$$y_{d,m}^{\text{ID}} = \sqrt{\beta_{d,m}}\big(\mathbf{h}_{d,m}^H \mathbf{F}^{\text{RF}}\mathbf{F}^{\text{BB}}\mathbf{P}\mathbf{s} + n_{d,m}\big) + n'_{d,m}. \quad (5)$$

[5]For $\alpha_{d,m}^{(l)}$, we assume a complex Gaussian model for illustration sake of the main results. Nonetheless, the entire results straightforwardly extend beyond the original area of application to more general models.

[6]Since we focus on designing the UE grouping, hybrid digital and analog precoders, power allocation, and power splitting factors for the SWIPT-enabled mmWave mMIMO-NOMA system, on probing the downlink time-varying channel is beyond the scope of this paper. We assume that information regarding the channel state is known to both the BS and UEs. Several mmWave channel estimation approaches have been proposed; for details, the interested reader is referred to [9], [37]–[39].

where $n'_{d,m} \sim \mathcal{CN}(0, \sigma^2_{n'_{d,m}})$ is the addition of noise induced by the ID (i.e., power splitter). In the meanwhile, the second part $\sqrt{1 - \beta_{d,m}}y_{d,m}$ forms an input for EH at the mth UE in the gth beam as

$$y_{d,m}^{\text{EH}} = \sqrt{1 - \beta_{d,m}}\big(\mathbf{h}_{d,m}^H \mathbf{F}^{\text{RF}}\mathbf{F}^{\text{BB}}\mathbf{P}\mathbf{s} + n_{d,m}\big). \quad (6)$$

The harvested energy at the EH receiver at the mth UE of the dth beam extracted from (6) can be expressed as

$$\mathcal{P}_{d,m}^{\text{EH}} = \eta(1 - \beta_{d,m})\left(\sum_{i=1}^{\mathcal{D}}\sum_{j=1}^{|\mathcal{S}_i|} \|\mathbf{h}_{d,m}^H \mathbf{F}^{\text{RF}}\mathbf{f}_d^{\text{BB}}\|_2^2 p_{i,j} + \sigma_n^2\right), \quad (7)$$

We have written $\|\cdot\|_2$ for the ℓ_2 vector norm. Here, $0 \le \eta \le 1$ represents the energy conversion efficiency at each UE.

To deliver higher spectral efficiency, multiple UEs are supported in each beam via NOMA mechanism with the aid of intra-beam superposition coding at the transmitter and SIC at each of the receivers. To design the digital baseband precoder for mmWave-NOMA systems with SWIPT, we use the strong effective channel-based method, which design for only LoS channels. The channel fading conditions are critical to the realization of NOMA. We assume that the effective mmWave mMIMO-NOMA channel gains for the set of UEs in the dth beam are ordered as follows:

$$\|\mathbf{h}_{d,1}^H \mathbf{F}^{\text{RF}}\mathbf{f}_d^{\text{BB}}\|_2 \ge \|\mathbf{h}_{d,2}^H \mathbf{F}^{\text{RF}}\mathbf{f}_d^{\text{BB}}\|_2 \ge \cdots \ge \|\mathbf{h}_{d,|\Omega_d|}^H \mathbf{F}^{\text{RF}}\mathbf{f}_d^{\text{BB}}\|_2 \quad (8)$$

for $d = 1, \ldots, \mathcal{D}$. Without loss of generality and for notational simplicity, one can write the effective channel gains of the first UE in the dth beam as $\breve{\mathbf{h}}_{d,m}^H = \mathbf{h}_{d,m}^H \mathbf{F}^{\text{RF}}$. In this paper, we also optimize the power allocation coefficients and power slitting factors according to instantaneous channel conditions to further improve the achievable rate of the SWIPT-enabled mmWave mMIMO-NOMA systems. Hence, the optimal order for NOMA signal detection is the decreasing order of the effective channel gains normalized by noise. In practice, as in traditional single-input single-output (SISO) NOMA systems namely, SISO-NOMA systems [40], the strong UE can eliminate intercarrier interference (ICI) via SIC. However, weak UEs do not effectively cancel ICI. Giving by (1) namely, in term of zero-forcing beamforming, we can cancel inter-beam interference because $\breve{\mathbf{h}}_{d,m}^H = 0$ for $i \ne d$. Specifically, the mth UE in the dth beam can decode process $s_{d,j}$, $m + 1 \le j \le |\Omega_d|$ and then subtracts them from the received signal successively to yield the desired signal, with the remaining signal processed as interference. From (8) the signal for ID of the mth UE in the dth beam after performing SIC can be formulated as

$$\hat{y}_{d,m}^{\text{ID}} = \sqrt{\beta_{d,m}}\left(\breve{\mathbf{h}}_{d,m}^H \mathbf{f}_d^{\text{BB}}\sqrt{\rho_{d,m}}s_{d,m} + \breve{\mathbf{h}}_{d,m}^H \mathbf{f}_d^{\text{BB}}\sum_{j=1}^{m-1}\sqrt{\rho_{d,j}}s_{d,j} + \breve{\mathbf{h}}_{d,m}^H \sum_{i\ne d}\sum_{j-1}^{\Omega_i} \mathbf{f}_i^{\text{BB}}\sqrt{\rho_{i,j}}s_{i,j} + n_{d,m}\right) + n'_{d,m} \quad (9)$$

Let $f_{d,m}$ denote the receive filter for ID at the mth UE in the dth beam; the detected signal is express as[7]

$$\begin{aligned}\hat{\tilde{s}}_{d,m} &= \frac{1}{\sqrt{\rho_{d,m}}} f_{d,m} \hat{y}^{\mathrm{ID}}_{d,m} \\ &= f_{d,m} \check{\mathbf{h}}^{H}_{d,m} \mathbf{f}^{\mathrm{BB}}_{d} \sqrt{\rho_{d,m}} s_{d,m} + f_{d,m} \check{\mathbf{h}}^{H}_{d,m} \mathbf{f}^{\mathrm{BB}}_{d} \sum_{j=1}^{m-1} \sqrt{\rho_{d,j}} s_{d,j} \\ &\quad + f_{d,m} \check{\mathbf{h}}^{H}_{d,m} \sum_{i \neq d} \sum_{j-1}^{\Omega_i} \mathbf{f}^{\mathrm{BB}}_{i} \sqrt{\rho_{i,j}} s_{i,j} + f_{d,m} n_{d,m} \\ &\quad + \frac{1}{\sqrt{\rho_{d,m}}} f_{d,m} n'_{d,m} \end{aligned} \tag{10}$$

B. SUM-RATE FORMULATION

Accordingly, the signal-to-interference plus-noise ratio (SINR) for the mth ordered UE in the dth beam can be express as

$$SINR_{d,m} = \frac{\|\check{\mathbf{h}}^{H}_{d,m} \mathbf{f}^{\mathrm{BB}}_{d}\|_2^2 \rho_{d,m}}{\Upsilon_{d,m}} \tag{11}$$

where

$$\Upsilon_{d,m} = \|\check{\mathbf{h}}^{H}_{d,m} \mathbf{f}^{\mathrm{BB}}_{d}\|_2^2 \sum_{j=1}^{m-1} \rho_{d,j} + \sum_{i \neq d} \|\check{\mathbf{h}}^{H}_{d,m} \mathbf{f}^{\mathrm{BB}}_{d}\|_2^2 \sum_{j=1}^{|\Omega_i|} \rho_{i,j} \sigma_n^2 + \frac{\sigma_{n'}}{\beta_{d,m}} \tag{12}$$

Hence, the instantaneous achievable rate of the mth UE in the dth beam is known to be $R_{d,m} = \log_2(1 + SINR_{d,m})$. We thus have the achievable sum-rate as

$$R_{\mathrm{sum}} = \sum_{d=1}^{\mathcal{D}} \sum_{m=1}^{|\Omega_d|} R_{d,m}, \tag{13}$$

which can be improved by designing the analog RF precoder matrix $\mathbf{F}^{\mathrm{RF}}$, digital baseband precoders for the d-th UE $\mathbf{f}^{\mathrm{BB}}_{d}$ for $d = 1, \ldots, \mathcal{D}$ having that $\mathcal{D} = N^{\mathrm{RF}}$, power allocation $\rho_{d,m}$ for $d = 1, \ldots, \mathcal{D}, m = 1, \ldots, |\Omega_d|$, and power splitting factors $\beta_{d,m}$ for $d = 1, \ldots, \mathcal{D}, m = 1, \ldots, |\Omega_d|$.

The following proposition is useful in the analysis of this work.

Proposition 2.1 (Proposition 1 in [8]): Let $a \in \mathbb{R}^{1\times 1}$ be a positive real number and $f(a) = -(ab/\ln 2) + \log_2 a + (1/\ln 2)$, we have

$$-\log_2 b = \max_{a \in \mathbb{R}^{1\times 1},\ a>0} f(a) \tag{14}$$

where $a = 1/b$ is the optimal solution on the right-hand side (RHS) of (14).

Proof: By letting $\partial f(a)/\partial a = 0$, the optimal solution a^o on the RHS of (14) is obtained, since $f(a)$ is concave. ■

[7]We can obtain the MSE between the transceiver signals as $\mathrm{MSE}_{d,m} = \mathbb{E}[\|\hat{s}_{d,m} - s_{d,m}\|_2^2]$.

III. PERFORMANCE OPTIMIZATION

In the propose system, the number of served UEs is larger than N^{RF}, that is $K > N^{\mathrm{RF}}$, with $\mathcal{D} = N^{\mathrm{RF}}$ beams. Thus, we first allocate the UEs into $\mathcal{D}$ groups. Then we formulate a problem to optimize the hybrid analog-digital precoders, power allocation and power splitting factors, jointly.

A. USER GROUPING AND PROBLEM FORMULATION

1) USER GROUPING

Since at most $\mathcal{D}$ beams can be simultaneously served at each transmission period, we propose a modified K-Means user grouping algorithm, for SWIPT-based mmWave mMIMO-NOMA system, which uses the normalized channel correlation among user channels as the parameter for grouping. Specifically, the normalized channel correlation between UE i and UE j is denoted as $\mathrm{Corr}_{i,j} = |\mathbf{h}_i^H \mathbf{h}_j|/(\|\mathbf{h}_i\|_2 \|\mathbf{h}_j\|_2)$. We have written $|\cdot|$ and $\|\cdot\|_2$ for the absolute value and ℓ_2 vector norm operations, respectively. The modified K-Means user grouping algorithm starts by selecting one user representative for each beam by minimizing the normalized channel correlation among the beam selected representatives. Subsequently, the UEs whose channels are highly correlated are allocated to the same beam to minimize intra-beam interference, while the UEs with low channel correlation are allocated to different beams to minimize inter-beam interference.

The pseudocode for the proposed algorithm is shown as Algorithm 1. Contrarily to the traditional K-means user grouping algorithm proposed in [10] which randomly selects user representatives (or cluster heads), we start by selecting (in **step 11**) one optimal user representative for each beam by minimizing the normalized channel correlation among the beam chosen representatives. Hence, after completing **steps 9-15**, the cluster heads i.e., $\Omega = \{\Omega_1, \ldots, \Omega_{\mathcal{D}}\}$, is selected. Now, let $\Omega_d = k_d \in \mathcal{K} = \{1, \ldots, K\}$. Then, using **steps 18 to 26**, the remaining UEs are scheduled to cluster groups as determined by the normalized channel correlation. For example, the kth UE is scheduled to the d^* beam if $d^* = \arg \max_{1\le d\le \mathcal{D}} \mathrm{Corr}_{k,\hat{\Omega}}$. Next, the cluster heads is updated such that they possess low correlation with other cluster heads. Let the correlation between a UE to the other beams be defined as $\breve{\mathrm{Corr}}_k = \sum_{1\le j\le K}^{j\notin \mathcal{G}^k} \mathrm{Corr}_{k,j}$. We have written $\mathcal{G}^k$ to denote the group/cluster where the kth UE resides. Thud, the cluster head is updated as

$$\Omega_d = \arg \min_{1\le m\le |\mathcal{G}_d|} \breve{\mathrm{Corr}}_m. \tag{15}$$

Finally, we denote the set of selected UE groups as $\{\mathcal{G}_1, \ldots, \mathcal{G}_{\mathcal{D}}\}$, where $\sum_{d=1}^{\mathcal{D}} |\mathcal{G}_d| = K$ UEs.

In addition, when operating Algorithm 1, the complexity of measuring the channel correlation and the channel vector norm is $\mathcal{O}(K^2 N)$ and $O(KN)$, respectively. Specifically, in each iteration, the maximal complexity is $2(K-1)$ from **step 10∼14**, while the complexities of updating the cluster heads and the user grouping are $O(K^2)$ and $O(K\mathcal{D})$, respectively. Since $N \gg \mathcal{D}$ then, Algorithm 1 maximal complexity results to $O(K^2 N)$.

Algorithm 1 Modified K-Means User Grouping Algorithm

Input: Number of UEs $K(> N_t^{\mathrm{RF}})$, number of beams $\mathcal{D}$, channel vectors $\mathbf{h}_k$ for $k \in [1, K]$, set predefined beam correlation threshold $0 \leq \xi \leq 1$.
Initialization: $\Omega = \mathbf{0}^{\mathcal{D}}$

Output: Optimized user grouping $\{\mathcal{G}_1, \ldots, \mathcal{G}_{\mathcal{D}}\}$

$\mathcal{K} = \{1, \ldots, K\}$
$\acute{\mathbf{h}} = [\|\mathbf{h}_1\|_2, \ldots, \|\mathbf{h}_K\|_2]$ {channel gains for each user}
$\check{\mathbf{h}} = [\frac{\mathbf{h}_1}{\|\mathbf{h}_1\|_2}, \ldots, \frac{\mathbf{h}_K}{\|\mathbf{h}_K\|_2}]$ {user's channel normalization }
$[\sim, \mathcal{S}] = \mathrm{sort}(\acute{\mathbf{h}}, \ 'descend')$;
$\Omega(1) = \mathcal{S}(1)$,
$\mathcal{S}(1) = [\ ]$; {remove the first element of $\mathcal{S}$}
$\mho = \mathcal{S}$ {remaining user set}
$d = 2$
while $d \leq \mathcal{D}$ **do**
  $\mathbf{w} = |\ \check{\mathbf{h}}_i^H \check{\mathbf{h}}_j|, i \in \mho, \ \forall j \in \Omega$
  $[\sim, \check{\mathcal{S}}] = \min(\mathrm{find}(\mathbf{w} < \xi))$
  $\Omega(d) = \ \mho\big(\check{\mathcal{S}}(1)\big)$ {Merge new cluster head to Ω}
  $\mathcal{S}\big(\mathrm{find}(\mathcal{S} == \Omega(d))\big) = [\ ]$ {remove element $\Omega(d)$ from $\mathcal{S}$}
  $d = d + 1$
end while
$\Omega = \{\Omega_1, \ldots, \Omega_{\mathcal{D}}\}$, with $\Omega_d = k_d \in \mathcal{K}$,{ cluster heads selected in **step 12**}
$t = 1$;
while $\Omega_d^t \neq \Omega_d^{t-1}$ **do**
  Set $\{\mathcal{G}_d\}_{d=1}^{\mathcal{D}} = \{\Omega_d^t\}_{d=1}^{\mathcal{D}}$
  for $k \in \mathcal{K}/\Omega_d^t$ **do**
    $d^* = \arg \max_{1 \leq d \leq \mathcal{D}} \mathrm{Corr}_{k, \Omega_d^t}$
    $\mathcal{G}_{d^*} = \mathcal{G}_{d^*} \bigcup k$
  end for
  $t = t + 1$
  Update $\Omega_d^t, d = 1, \ldots, \mathcal{D}$ in accordance with (15)
end while

2) PROBLEM FORMULATION

After performing user grouping $\{\mathcal{G}_1, \ldots, \mathcal{G}_{\mathcal{D}}\}$, we seek to maximize the achievable sum rate by solving the joint power allocation, power splitting, digital baseband precoder and analog RF precoder optimization problem which can be modeled as follows:

$$\max_{p_{d,m},\ \beta_{d,m},\ \mathbf{f}_d^{\mathrm{BB}},\ \mathbf{F}^{\mathrm{RF}}} \sum_{d=1}^{\mathcal{D}} \sum_{m=1}^{|\mathcal{G}_d|} R_{d,m} \tag{16a}$$

$$\text{s.t.} \quad R_{d,m} \geq R_{d,m}^{\min}, \quad \forall d, m, \tag{16b}$$

$$p_{d.m} \geq 0, \quad \forall d, m, \tag{16c}$$

$$\sum_{d=1}^{\mathcal{D}} \sum_{m=1}^{|\mathcal{G}_d|} p_{d,m} \leq P_T, \tag{16d}$$

$$P_{d,m}^{EH} \geq P_{d,m}^{\min}, \quad \forall d, m, \tag{16e}$$

$$|[\mathbf{F}^{\mathrm{RF}}]_{i,j}| = \frac{1}{\sqrt{N}}, 1 \leq i \leq N, 1 \leq j \leq N^{\mathrm{RF}} \tag{16f}$$

$$\|\mathbf{F}^{\mathrm{RF}} \mathbf{f}_d^{\mathrm{BB}}\|_2 = 1 \quad \forall d, \tag{16g}$$

where constraint (16b) guarantees that the mth UE in the dth beam achieves the predefined minimum data rate. The constraint (16c) signifies that the transmit power for each UE at the BS should be positive. The constraint (16d) denotes the BS total power constraint such that the BS total transmitted power is no more than P_T. The constraint (16e) is the EH constraint for the mth UE in the dth beam with $P_{d,m}^{min}$ being the minimum harvested energy level. Here, (16f) is the constant-modulus constraint for the analog RF precoding matrix, while (16g) is the unit power constraint for the hybrid analog/digital precoding matrix. Observe in (16g), that we need to optimize the digital baseband precoding matrix for different beams. Apart from the challenge of the joint optimization over all variables in (16), the optimization problem (16) is nonconvex due to the transmit power constraint, EH constraint, and the constant modulus constraints of the analog RF precoder due to the phase shifters. Thus, (16) is difficult to solve. The solution is, therefore, non-trivial and challenging to directly obtain globally optimum.

B. HYBRID ANALOG-DIGITAL PRECODER SOLUTION

To maximize (11) for each UE, we require to subdue the inter-beam interference and enhance the effective channel gain. In the traditional multiuser MIMO (MU-MIMO) system, zero forcing (ZF) can be employed [8], [11], [18]. To reduce the hardware constraints while achieving full potentials of mmWave mMIMO-NOMA systems, we propose to implement phase-only array response adjustment to link the N^{RF} RF chain outputs with N BS antennas, employing low-cost phase shifters. Nonetheless, as a result of the elementwise constant-magnitude constraint on the analog precoder i.e., $|[F^{\mathrm{RF}}]_{i,j}| = 1/\sqrt{N}$, $\forall i, j$, they cannot be straightforwardly used to the hybrid analog-digital precoding system [8], [18]. Since the constant-magnitude constraint makes the subsets of feasible regions non-convex. Therefore, we consider developing the analog RF precoder and digital baseband precoder separately. Motivated by [8] and [11], in dividing the hybrid precoding design into two step, we present an efficient analog RF precoding algorithm to design $\mathbf{F}^{\mathrm{RF}}$ and a low-dimensional digital baseband precoding algorithm to design $\mathbf{F}^{\mathrm{BB}}$, for downlink multi-user mmWave mMIMO-NOMA systems. We first focus on the design of the analog RF precoder matrix.

1) ANALOG RF PRECODER DESIGN

To design the analog RF precoder, we attempt to align the phases of $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$, such that the large array gain rendered by the mMIMO large antennas can be harvested. We summarize the analog RF precoder design in Algorithm 2. For simplicity, we address the essential part of the proposed algorithm without giving the obvious redundant illustration. We first initialize the analog precoder as an all-zero matrix. Step 4 extracts the phases of the conjugate transpose of the aggregate downlink mmWave MIMO-NOMA channel from the BS to multiple users. Step 10 aligns the phases of channel elements to design the analog RF precoder such that substantial array gain can be harvested. Subsequently, having

Algorithm 2 Analog RF Precoder Design

Input: Number of UEs $K(> N_t^{\mathrm{RF}})$, number of RF chains N^{RF}, channel matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$, optimized user grouping $\{\mathcal{G}_1, \ldots, \mathcal{G}_{\mathcal{D}}\}$, number of BS antennas N.
Initialization: $\mathbf{F}^{\mathrm{RF}} = \mathbf{0}^{N \times N^{\mathrm{RF}}}$, number of quantization bits B

Output: $\mathbf{F}^{\mathrm{RF}}$ {designed analog RF precoder}

1: $\Lambda = \{\frac{2\pi n}{2^B}, n = 0, 1, \ldots, 2^{B-1}\}$; {phase set}
2: **for** $d = 1$ to N^{RF} **do**
3: $\breve{\mathbf{H}} = [\mathbf{H}]_{:,\mathcal{G}_d}$ {recall, $\mathcal{G}_d, d = 1, \ldots, \mathcal{D}$ are user sets}
4: $\Theta = \angle\breve{\mathbf{H}}$ {extract phases of the aggregate downlink channel matrix $\breve{\mathbf{H}}$}
5: $\theta = \mathbf{0}^N$
6: **for** $n = 1$ to N **do**
7: $[\sim, i] = \min |[\Theta]_n - \Lambda|$
8: $\theta(n) = [\Lambda]_i$ {select the ith entry of Λ}
9: **end for**
10: $\mathbf{F}^{\mathrm{RF}}(:, d) = \exp(j * \theta)$ {j denotes the imaginary unit}
11: **end for**

the effective baseband channel linked with the acquired optimal analog RF precoder, the digital baseband precoder design then follows to mitigate the interference further and maximise the achievable sum-rate.

2) DIGITAL BASEBAND PRECODER DESIGN

To cancel inter-user interference, we design the digital baseband precoding matrix such that, UEs in each beam with strong channels are selected. For the digital precoder design $\mathbf{F}^{\mathrm{BB}} = [\mathbf{f}_1^{\mathrm{BB}}, \ldots, \mathbf{f}_{N^{\mathrm{RF}}}^{\mathrm{BB}}]$, while regarding $\mathbf{F}^{\mathrm{RF}}$ as fixed (i.e., using the already designed $\mathbf{F}^{\mathrm{RF}}$), we introduce an algorithmic solution inspired by the concept of [8], [11] which after designing the analog RF precoder $\mathbf{F}^{\mathrm{RF}}$, implements a low-dimensional baseband digital zero-forcing (ZF) precoding relying on the effective channels seen from baseband. The pseudocode for the digital baseband precoder solution is given in Algorithm 3. In summary, the precoding algorithm starts by employing in **step 3**, zero-forcing digital precoding, to cancel the inter-user interference. Thus, the digital baseband precoder is calculated as $\hat{\mathbf{F}}^{\mathrm{BB}} = \breve{\mathbf{H}}^H(\breve{\mathbf{H}}\breve{\mathbf{H}}^H)^{-1}$. Then, the digital baseband precoder is normalized in **step 5** as $\hat{\mathbf{F}}^{\mathrm{BB}} = [\hat{\mathbf{f}}_1^{\mathrm{BB}}./\mathbf{f}_1^{\mathrm{BB}^*}, \ldots, \hat{\mathbf{f}}_{N^{\mathrm{RF}}}^{\mathrm{BB}}./\mathbf{f}_{N^{\mathrm{RF}}}^{\mathrm{BB}^*}]$, with $\mathbf{F}^{\mathrm{BB}^*} = [\mathbf{f}_1^{\mathrm{BB}^*}, \ldots, \mathbf{f}_{N^{\mathrm{RF}}}^{\mathrm{BB}^*}]$ where $\mathbf{f}_n^{\mathrm{BB}^*} = \texttt{repmat}(\|\mathbf{F}^{\mathrm{RF}}\hat{\mathbf{f}}_n^{\mathrm{BB}}\|_2, N^{\mathrm{RF}}, 1), n = 1, \ldots, N^{\mathrm{RF}}$ and ./ denote element-wise dot operator, to satisfy the total power constraint. $\|\mathbf{F}^{\mathrm{RF}}\hat{\mathbf{f}}_1^{\mathrm{BB}}\|_2$ is realized as $\sqrt{\sum |\mathbf{F}^{\mathrm{RF}}\hat{\mathbf{f}}_1^{\mathrm{BB}}|^2}$. Here, $|\cdot|$ denotes the absolute value. Moreover, $\mathbf{f}_n^{\mathrm{BB}^*} = \texttt{repmat}(\|\mathbf{F}^{\mathrm{RF}}\hat{\mathbf{f}}_n^{\mathrm{BB}}\|_2, N^{\mathrm{RF}}, 1)$ creates a vector $\mathbf{f}_n^{\mathrm{BB}^*}$ consisting of an N^{RF}-by-1 tiling of copies of $\|\mathbf{F}^{\mathrm{RF}}\hat{\mathbf{f}}_1^{\mathrm{BB}}\|_2 \in \mathbb{R}^{1\times 1}$. Notice that in **step 6** $[\mathbf{F}^{\mathrm{BB}}]_{:,[\mathcal{G}]_{:,1}} = \hat{\mathbf{F}}^{\mathrm{BB}}$. Finally, using **steps** $7 \sim 13$, after N^{RF} iterations the digital baseband precoder $F^{\mathrm{BB}} = (f_1^{\mathrm{BB}}, \ldots, f_{\mathcal{D}}^{\mathrm{BB}})$, is designed. Notice that there are $\mathcal{D} = N^{\mathrm{RF}}$ beams.

Algorithm 3 Digital Baseband Precoder Design

Input: Number of UEs $K(> N_t^{\mathrm{RF}})$, number of RF chains N^{RF}, channel matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{C}^{N\times K}$, optimized analog RF precoder $\mathbf{F}^{\mathrm{RF}} \in \mathbb{C}^{N\times N^{\mathrm{RF}}}$, optimized user grouping $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_{\mathcal{D}}\}$, number of BS antennas N, we assume $\mathcal{D} = N^{\mathrm{RF}}$.
Initialization: number of quantization bits B

Output: $\mathbf{F}^{\mathrm{BB}} \in \mathbb{C}^{N^{\mathrm{RF}}\times K} = [\mathbf{f}_1^{\mathrm{BB}}, \ldots, \mathbf{f}_K^{\mathrm{BB}}]$ {designed digital baseband precoder}

1: $\breve{\mathbf{H}} \in \mathbb{C}^{K\times N^{\mathrm{RF}}} = \mathbf{H}^H\mathbf{F}^{\mathrm{RF}}$, {equivalent channel matrix}
2: $\tilde{\mathbf{H}} = [\breve{\mathbf{H}}]_{[\mathcal{G}]_{:,1},:}$ {where, $\tilde{\mathbf{H}} \in \mathbb{C}^{N^{\mathrm{RF}}\times N^{\mathrm{RF}}}$}
3: $\hat{\mathbf{F}}^{\mathrm{BB}} = \tilde{\mathbf{H}}^H(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H)^{-1}$ {where, $\hat{\mathbf{F}}^{\mathrm{BB}} \in \mathbb{C}^{N^{\mathrm{RF}}\times N^{\mathrm{RF}}}$ and $\hat{\mathbf{F}}^{\mathrm{BB}} = \hat{\mathbf{f}}_1^{\mathrm{BB}}, \ldots, \hat{\mathbf{f}}_{N^{\mathrm{RF}}}^{\mathrm{BB}}$, where, $[\hat{\mathbf{f}}^{\mathrm{BB}}]_n \in \mathbb{C}^{N^{\mathrm{RF}}\times 1}$, $n = 1, \ldots, N^{\mathrm{RF}}$}
4: $\hat{\mathbf{F}}^{\mathrm{BB}} = [\hat{\mathbf{f}}_1^{\mathrm{BB}}./\mathbf{f}_1^{\mathrm{BB}^*}, \ldots, \hat{\mathbf{f}}_{N^{\mathrm{RF}}}^{\mathrm{BB}}./\mathbf{f}_{N^{\mathrm{RF}}}^{\mathrm{BB}^*}]$ {./ denote element-wise dot operator and $\mathbf{F}^{\mathrm{BB}^*} = [\mathbf{f}_1^{\mathrm{BB}^*}, \ldots, \mathbf{f}_{N^{\mathrm{RF}}}^{\mathrm{BB}^*}]$ where $\mathbf{f}_n^{\mathrm{BB}^*} = \texttt{repmat}(\|\mathbf{F}^{\mathrm{RF}}\hat{\mathbf{f}}_n^{\mathrm{BB}}\|_2, N^{\mathrm{RF}}, 1)$, $n = 1, \ldots, N^{\mathrm{RF}}$ }
5: $\mathbf{F}^{\mathrm{BB}} = \mathbf{0}^{N^{\mathrm{RF}}\times K}$
6: $[\mathbf{F}^{\mathrm{BB}}]_{:,[\mathcal{G}]_{:,1}} \in \mathbb{C}^{N^{\mathrm{RF}}\times K} = \hat{\mathbf{F}}^{\mathrm{BB}}$
7: **for** $d = 1$ to N^{RF} **do**
8: $\Lambda = [\mathcal{G}]_{d,:}$
9: $\Lambda = \text{nonzeros}(\Lambda)^T$ {remove the zero elements}
10: **for** $n = 2$ to $|\Lambda|$ **do**
11: $[\mathbf{F}^{\mathrm{BB}}]_{:,\Lambda_n} = [\mathbf{F}^{\mathrm{BB}}]_{:,[\mathcal{G}]_{d,1}}$
12: **end for**
13: **end for**

In addition, based on Algorithm 3, we need to compute the effective channel matrices in step 3 and digital precoders in step 12. Therefore, the complexity of Algorithm 3 is $O(\mathcal{D}NK^2)$.

Now, given an arbitrarily fixed analog RF precoder $\mathbf{F}^{\mathrm{RF}} \in \mathbb{C}^{N\times N^{\mathrm{RF}}}$ and an arbitrarily fixed digital baseband precoder $\mathbf{F}^{\mathrm{BB}} = [\mathbf{f}_1^{\mathrm{BB}}, \ldots, \mathbf{f}_{\mathcal{D}}^{\mathrm{BB}}] \in \mathbb{C}^{N^{\mathrm{RF}}\times\mathcal{D}}$ using Algorithm 2 and Algorithm 3, respectively, the order of effective channel gains no longer varies. Since the designed precoders (i.e., $\mathbf{F}^{\mathrm{RF}}$ and $\mathbf{F}^{\mathrm{BB}}$) are now arbitrarily fixed. Without loss of generality, we assume that the UEs in each cluster are sorted such that $|\breve{\mathbf{h}}_{d,1}^H\mathbf{f}_d^{\mathrm{BB}}|^2 \geq |\breve{\mathbf{h}}_{d,2}^H\mathbf{f}_d^{\mathrm{BB}}|^2 \geq \cdots \geq |\breve{\mathbf{h}}_{d,|\mathcal{G}_d|}^H\mathbf{f}_d^{\mathrm{BB}}|^2$, for all $1 \leq d \leq \mathcal{D}$.

C. POWER ALLOCATION AND POWER SPLITTING SOLUTION

Now, seek to maximize the achievable sum rate by solving the joint power allocation and power splitting optimization problem. Mathematically, (16) is equivalently transformed into the following joint power allocation and power splitting maximization problem:

$$\max_{p_{d,m},\ \beta_{d,m}} \sum_{d=1}^{\mathcal{D}} \sum_{m=1}^{|\mathcal{G}_d|} R_{d,m}$$
$$\text{s.t. } (16b), (16c), (16d), (16e) \tag{17}$$

Consequently, in (17), an effective power allocation scheme to maximize the achievable sum-rate is required to solve the problem. It is obvious that (17) is still a non-convex optimization problem on account of the non-concave objective function and the constraint in (16b),(16c),(16d) and (16e) are still non-convex, making it challenging to find the globally optimal solution. Let us rewrite (9) as

$$\acute{y}_{d,m} = \check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\sqrt{\rho_{d,m}}s_{d,m} + \check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\sum_{j=1}^{m-1}\sqrt{\rho_{d,m}}s_{d,m} + \check{\mathbf{h}}_{d,m}^{H}\sum_{i\neq d}\sum_{j-1}^{\Omega_i}\mathbf{f}_{i}^{\mathrm{BB}}\sqrt{\rho_{d,m}}s_{d,m} + n_{d,m} + \frac{n'_{d,m}}{\sqrt{\beta_{d,m}}} \tag{18}$$

and define the mean square error (MSE) to estimate $s_{d,m}$, $d = 1,\ldots,\mathcal{D}$, $m = 1,\ldots,|\Omega_d|$ (given that the minimum MSE-receive filter is applied) as:

$$e_{d,m} = \mathbb{E}\left\{|s_{d,m} - c_{d,m}\acute{y}_{d,m}|^2\right\} \tag{19}$$

where $c_{d,m}$ denotes the equivalent channel equalization coefficient. If we substitute (18) into (19), we have

$$e_{d,m} = 1 - 2\mathbb{R}\left(c_{d,m}\sqrt{\rho_{d,m}}\check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\right) + |c_{d,m}|^2\left(\rho_{d,m}\|\check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\|_2^2 + \Upsilon_{d,m}\right). \tag{20}$$

Clearly [8], [41], the optimal $c_{d,m}$ that minimizes $e_{d,m}$ is

$$c_{d,m}^{o} = \arg\min_{c_{d,m}} e_{d,m} = (\sqrt{\rho_{d,m}}\check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}})^{H}\left(\rho_{d,m}\|\check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\|_2^2 + \Upsilon_{d,m}\right)^{-1} \tag{21}$$

The minimum MSE obtainable plugging (21) into (20) is

$$e_{d,m}^{o} = 1 - \rho_{d,m}\|\check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\|_2^2\left(\rho_{d,m}\|\check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\|_2^2 + \Upsilon_{d,m}\right)^{-1}. \tag{22}$$

Using the extension of the so-called Sherman-Morrison-Woodbury (SMW) formula of matrix inversion [42], i.e., $(A + BCD)^{-1} = A^{-1}A^{-1}B(I + CDA^{-1}B)^{-1}CDA^{-1}$, we have

$$(1 + SINR_{d,m})^{-1} = e_{d,m}^{o} \tag{23}$$

Now, utilizing Proposition 2.1 and (23), the joint power allocation and power splitting problem of (17) can be restated equivalently as

$$\max_{\substack{\{p_{d,m}\},\\ \{\beta_{d,m}\}}} \sum_{d=1}^{\mathcal{D}}\sum_{m=1}^{|\mathcal{G}_d|} \max_{\{a_{d,m}>0\}} \max_{\{c_{d,m}\}} -\frac{a_{d,m}e_{d,m}}{\ln 2} + \log_2 a_{d,m} + \frac{1}{\ln 2}$$
$$\text{s.t. } (16b), (16c), (16d), (16e). \tag{24}$$

Notice that $a_{d,m} > 0$ has been introduced in problem (24) as a slack variable and that, when $c_{d,m}$ and $a_{d,m}$ for $d = 1,\ldots,\mathcal{D}$, $m = 1,\ldots,|\Omega_d|$ is optimal, the objective function of problem (24) is $\sum_{d=1}^{\mathcal{D}}\sum_{m=1}^{|\mathcal{G}_d|} R_{d,m}$. In contrast to [8], we propose to decouple problem (24) into four separate optimization problems and then optimize the transceiver coefficient $\{c_{d,m}\}$, slack variable $\{a_{d,m}\}$, power splitting factors $\{\beta_{d,m}\}$ and power allocation $\{p_{d,m}\}$ iteratively, via an alternating optimization (AO) algorithm.

Intuitively, given the power splitting factor $\{\beta_{d,m}\}$ and power allocation $\{p_{d,m}\}$ of the mth UE in the dth beam which is optimal in the $(t-1)$th iteration, we solve the following optimization problem:

$$\min_{c_{d,m}} e_{d,m} \tag{25}$$

to iteratively realize the optimal $c_{d,m}$ in the tth iteration. Notice that problem (25) is convex since the objective function $e_{d,m}$ is convex with reference to $c_{d,m}$.

Subsequently, after obtaining the equivalent channel equalization coefficient $\{c_{d,m}\}$ according to (25), which is optimal in the tth iteration, we turn to solve the following optimization problem:

$$\max_{\{a_{d,m}>0\}} -\frac{a_{d,m}e_{d,m}}{\ln 2} + \log_2 a_{d,m} + \frac{1}{\ln 2} \tag{26}$$

to obtain the optimal solution to the introduced slack variable $a_{d,m}$ in the tth iteration. Thus, by using Proposition 2.1, clearly suffices to obtain the closed-form solution to problem (26). That is, the optimal slack variable $a_{d,m}$ in the tth iteration results to $1/e_{d,m}^{o}$, where $e_{d,m}^{o}$ is the optimal minimum MSE obtainable in (22).

Next, after obtaining the optimal slack variable $\{a_{d,m}\}$ and the equivalent channel equalization coefficient $\{c_{d,m}\}$, which are optimal in the tth iteration, we turn to solve the following optimization problem:

$$\max_{\{\beta_{d,m}\}} \sum_{d=1}^{\mathcal{D}}\sum_{m=1}^{|\mathcal{G}_d|} -\frac{a_{d,m}e_{d,m}}{\ln 2} + \log_2 a_{d,m} + \frac{1}{\ln 2}$$
$$\text{s.t. } (16e). \tag{27}$$

to obtain the optimal power slitting factor $\{\beta_{d,m}\}$ in the tth iteration. Introducing another slack variable $\varphi_{d,m} \geq 1/\beta_{d,m}$, $d = 1,\ldots,\mathcal{D}$, $m = 1,\ldots,|\mathcal{G}_d|$, the optimization problem of (27) is equivalently reformulated as

$$\max_{\{\beta_{d,m}\}} \sum_{d=1}^{\mathcal{D}}\sum_{m=1}^{|\mathcal{G}_d|} \varphi_{d,m}$$
$$\text{s.t. } -\frac{a_{d,m}e_{d,m}}{\ln 2} + \log_2 a_{d,m} + \frac{1}{\ln 2} \geq \varphi_{d,m}, (16e). \tag{28}$$

To enable problem (28) solvable, we introduced another slack variable $\gamma_{d,m} = P^{\min}/\big(\eta(1-\beta_{d,m})\big)$ and we rewrite constraint (16e) in problem (28) as

$$\sum_{d=1}^{\mathcal{D}}\sum_{m=1}^{|\mathcal{G}_d|} \|\check{\mathbf{h}}_{d,m}^{H}\mathbf{f}_{d}^{\mathrm{BB}}\|_2^2 p_{d,m} + \sigma^2 \geq \gamma_{d,m} \tag{29}$$

Hence, transforming $\gamma_{d,m}$ and $\varphi_{d,m}$ into matrix form according to the Schur complement lemma, problem (28) is rewritten as

$$\max_{\{\beta_{d,m}\}} \sum_{d=1}^{\mathcal{D}} \sum_{m=1}^{|\mathcal{G}_d|} \varphi_{d,m} \tag{30a}$$

$$\text{s.t.} \; -\frac{a_{d,m} e_{d,m}}{\ln 2} + \log_2 a_{d,m} + \frac{1}{\ln 2} \geq \varphi_{d,m}, \tag{30b}$$

$$\begin{bmatrix} \varphi_{d,m} & 1 \\ 1 & \beta_{d,m} \end{bmatrix} \succeq 0, \quad \forall d, m \tag{30c}$$

$$\begin{bmatrix} \gamma_{d,m} & \sqrt{P_{d,m}^{\min}/\eta} \\ \sqrt{P_{d,m}^{\min}/\eta} & 1-\beta_{d,m} \end{bmatrix} \succeq 0, \quad \forall d, m, \tag{30d}$$

where, $A \succeq 0$ for example means that matrix A is positive semidefinite.

Finally, after obtaining the optimal slack variable $\{a_{d,m}\}$, equivalent channel equalization coefficient $\{c_{d,m}\}$ and power slitting factor $\beta_{d,m}$, which are optimal in the tth iteration, we solve the following optimization problem:

$$\max_{\{p_{d,m}\}} \sum_{d=1}^{\mathcal{D}} \sum_{m=1}^{|\mathcal{G}_d|} \varphi_{d,m}$$
$$\text{s.t. } (16c), (30b), (30c), (30d). \tag{31}$$

to obtain the optimal $p_{d,m}$ in the tth iteration. The optimization problems in (30) and (31), are convex, and can be solved via the interior-point technique [43].

It is worth noting that since problems (25), (30) and (31) are convex and that the obtained slack variable $a_{d,m}$ is the closed-form optimal solution per iteration, then iteratively updating $\{c_{d,m}\}$, $\{a_{d,m}\}$, $\{\beta_{d,m}\}$ and $\{p_{d,m}\}$ via our introduced alternating optimization algorithm will, however, result in a monotonically increasing sequence (upper bounded by maximum transmission power), but will preserve and possibly improve the solution to the optimization objective of problem (24).

IV. SIMULATION RESULTS

In this section, we provide simulation results to verify the performance of the proposed SWIPT-enabled mmWave mMIMO-NOMA system in terms of both spectrum efficiency and energy efficiency for the fully-connected hybrid precoding architecture. Throughout the simulation, we assume $N = 64$, $N^{\text{RF}} = 4$, $K \geq N^{\text{RF}}$ UEs, where all K UEs are grouped in $\mathcal{D} = N^{\text{RF}} = 4$ beams, $B = 4$-bit resolution PSs, while $P_{d,m}^{\min} = 0.1$ mW and $R_{d,m}^{\min} = R_{zf}/10$ are the minimum harvested energy and minimum data-rate for the mth UE in the dth beam, respectively. Here, R_{zf} is calculated as the minimum achievable sum-rate via full-digital ZF precoding among all K UEs. The BS and UEs are equipped with uniform linear arrays (ULAs) antenna configuration with half-wavelength antenna spacing, and the channel of the mth UE in the dth beam is generated as in (2) with $\mathcal{L}_{d,m} = 3$ for all $d = 1, \ldots, \mathcal{D}$, $m = 1, \ldots, |\Omega_d|$, where $|\Omega_d|$ denotes the number ot UEs in each beam. Note that for the $\mathcal{L}_{d,m}$ number of propagation paths $l = 1$ represents the line-of-sight (LOS) component with $\alpha_{d,m}^{(1)} \sim \mathcal{CN}(0,1)$, and $2 \leq l \leq \mathcal{L}_{d,m}$ is the $\mathcal{L}_{d,m} - 1$ non-line-of-sight (NLOS) components with $\alpha_{d,m}^{(l)} \sim \mathcal{CN}(0, \frac{1}{10})$. The azimuth (elevation) AoD, i.e., $\theta_{d,m}^{(l)}(\Theta_{d,m}^{(l)})$, $1 \leq l \leq \mathcal{L}_{d,m}$ of the lth path are assumed to be are uniformly distributed over $[-\pi, \pi]$. For fairness, the same maximum transmitted power $P_T = 30$ mW constraint is enforced on all precoding solutions and signal-to-noise ratio is defined as $\text{SNR} = P_T/\sigma_n^2$.

Following the energy consumption model in [13], we define the energy efficiency (EE) as

$$\text{EE} = \frac{\text{Achievable sum-rate}}{\text{Total power consumption}} = \frac{R_{\text{sum}}}{P_{\text{total}}} \; \text{(bps/Hz/W)}, \tag{32}$$

where $P_{\text{total}} \triangleq P_{tran} + N^{\text{RF}} P^{\text{RF}} + N^{\text{PS}} P^{\text{PS}} + P^{\text{BB}}$, $P_{tran} = \sum_{d=1}^{\mathcal{D}} \sum_{m=1}^{|\Omega_d|} \rho_{d,m}$ denotes the total transmitted power by all UEs, P^{RF} denotes the RF chain power consumption, P^{PS} denotes the power consumption by each PS, and P^{BB} denotes the baseband signal processing power consumption, where $N^{\text{PS}} = NN^{\text{RF}}$ denotes the number of fully-connected PSs for the hybrid precoding architecture and $N^{\text{PS}} = 0$ for the full-digital ZF Precoding structure. Here, the spectrum efficiency which provides the achievable sum-rate R_{sum} is defined as in (13). In the simulations,[8] we select the typical parameter settings of $P^{\text{RF}} = 300$ mW, $P^{\text{PS}} = 40$ mW, $P^{\text{BB}} = 200$ mW, and the maximum transmitted power $P_T = 30$ mW [14]. Further, all reported results are averaged over 100 random channel realizations.

In the simulation, we compare the spectrum and energy efficiency performance of our method referred to as "proposed SWIPT-Hybrid Precoding NOMA" against the following four typical mmWave communication schemes, considered as the benchmark methods: 1) "SWIPT-Full-digital ZF Precoding", where each antenna needs to be supported by a dedicated RF chain (i.e., $N^{\text{PS}} = 0$); 2) Hybrid precoding-based mmWave mMIMO-NOMA with SWIPT scheme [8] denoted as "SWIPT-Hybrid Precoding NOMA proposed in [8]"; 3) mmWave NOMA with user grouping, power allocation and hybrid beamforming [10] denoted as "SWIPT-Hybrid Precoding NOMA proposed in [10]" and 4) "SWIPT-Hybrid Precoding OMA", where OMA is implemented for UEs in each beam; Notably, we realize OMA with frequency division multiple access (FDMA), where equal bandwidth allocation is assigned to UEs in the same beam. Then, the achievable rate of the SWIFT-enabled mmWave-OMA scheme for the kth UE is

$$R_k^{\text{OMA}} = \frac{1}{|\Omega^k|} \log_2 \left(\frac{\|\check{\mathbf{h}}_k^H \mathbf{f}_k^{\text{BB}}\|_2^2 \rho_k}{\sum_{j \notin \Omega^k} \|\check{\mathbf{h}}_k^H \mathbf{f}_j^{\text{BB}}\|_2^2 \rho_j + \frac{\sigma^2}{|\Omega^k|}} \right) \tag{33}$$

where Ω^k denotes the group that the kth UE resides. It is worth mentioning that the last three schemes for

[8] While the proposed algorithms in Section III is for any B-bit resolution PSs, we only focus on the low-resolution case ($B = 4$) in the simulation studies.

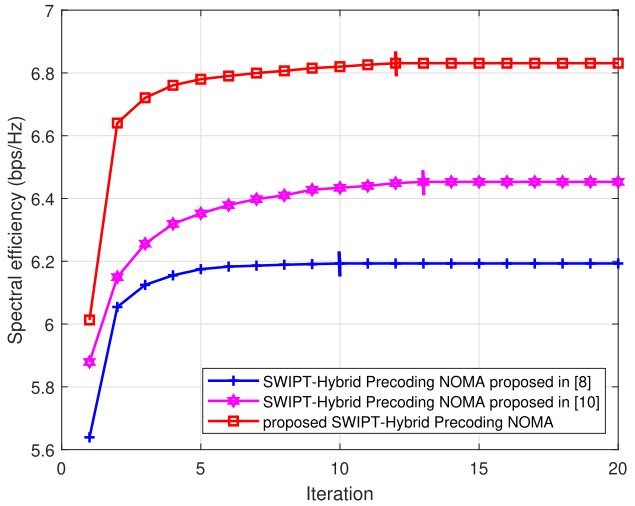

FIGURE 2. Spectrum efficiency versus iterations for the power allocation and power splitting optimization, where $N = 64$, $\mathcal{D} = 4$ beams, $K = 6$ UEs, $P_T = 30$ mW and SNR = 0 dB. The propagation medium is a $\mathcal{L}_{d,m} = 3$ paths comprising of $l = 1$ LOS and $2 \leq l \leq \mathcal{L}_{d,m}$ NLOS. $N^{\mathrm{RF}} = 4$ RF chains are assumed to be available for hybrid precoding with $B = 4$-bit resolution PSs.

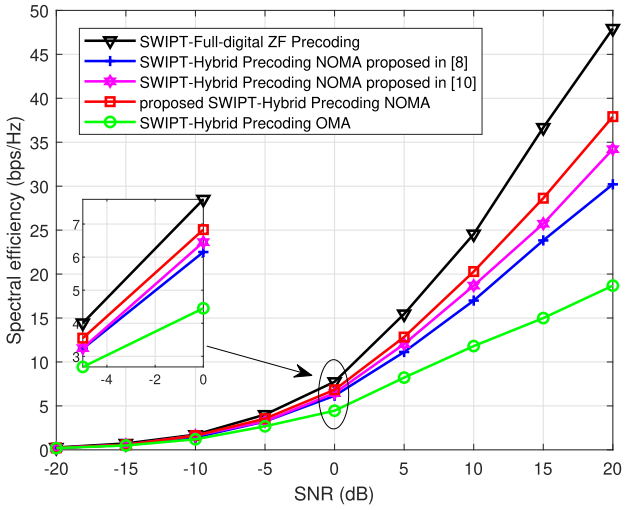

FIGURE 3. Spectral Efficiency achieved by various precoding solutions, where $N = 64$, $\mathcal{D} = 4$ beams, $K = 6$ UEs, $P_T = 30$ mW. The propagation medium is a $\mathcal{L}_{d,m} = 3$ paths comprising of $l = 1$ LOS and $2 \leq l \leq \mathcal{L}_{d,m}$ NLOS. $N^{\mathrm{RF}} = 4$ RF chains are assumed to be available for hybrid precoding with $B = 4$-bit resolution PSs.

comparison employ the fully-connected hybrid precoding architecture. For fair comparisons, the simulations[9] are performed, employing precisely the same computer hardware and software.

A. PERFORMANCE OF CONVERGENCE OF THE SYSTEM

We first provide insight on the convergence property of the proposed alternating optimization algorithm for the joint power allocation and power splitting optimization formulated to maximize the achievable sum-rate in Section III. Fig. 2, presents the spectrum efficiency versus the number of iterations of the considered system, where the number of UEs is set as $K = 6$, and SNR $= 0$ dB. We see that spectrum efficiency performance of all curves increases steadily and converges to the maximum value within some certain number of iterations. Specifically, from the curves depicted in Fig. 2, we find that SWIPT-Hybrid Precoding NOMA proposed in [8] converges with about 10 iterations, while our proposed SWIPT-Hybrid Precoding NOMA requires approximately 12 iterations for the joint power allocation and power splitting optimization to converge. However, our proposed scheme converges to a higher spectrum efficiency than SWIPT-Hybrid Precoding NOMA proposed in [8]. The SWIPT-Hybrid Precoding NOMA proposed in [10] requires approximately 13 iterations for the joint power allocation and power splitting optimization to converge. Therefore, ensuring fairness among different schemes, the number of iterations times for the power allocation and power slitting optimization is set to 13 to ensure all schemes are stable.

B. SPECTRUM EFFICIENCY OF THE SYSTEM

In this section, we evaluate the spectrum efficiency of the system. Fig. 3 shows the spectrum efficiency comparison against SNR for all five schemes under consideration, where the number of UEs is set as $K = 6$. From Fig. 3, the spectrum efficiency for all schemes under consideration increases with increasing SNR. It is worth noting that although the SWIPT-Full-digital ZF Precoding achieves the best spectral efficiency performance than all schemes, it has higher computational efficiency than the hybrid precoding alternatives [7]. The degradation in performance of the SWIPT-enable hybrid precoding schemes (with $N^{\mathrm{RF}} = 4$ RF chains) compared to the SWIPT-Full-digital ZF Precoding scheme (with $N^{\mathrm{RF}} = N$ dedicated RF chains) is that hybrid precoding utilizes analog phase shifters which impose constant modulus constraint on the entries of the analog RF precoding matrix. This results in a less degree of signal freedom and low precoding performance. However, we find that the spectral efficiency performance of the proposed SWIPT-Enabled mmWave mMIMO-NOMA Systems with Hybrid Precoding is closer to the performance of the SWIPT-Full-digital ZF Precoding than those of state-of-the-art designs and the conventional SWIPT-enabled mmWave MIMO-OMA system. For example, at SNR $= 10$ dB the proposed SWIPT-Hybrid Precoding NOMA, SWIPT-Hybrid Precoding NOMA proposed in [8], SWIPT-Hybrid Precoding NOMA proposed in [10] and SWIPT-Hybrid Precoding OMA can realize approximately (20.29 bps/Hz)/(24.55 bps/Hz) $\times$ 100 $=$ 82.65%, (16.99 bps/Hz)/(24.55 bps/Hz) $\times$ 100 $=$ 69.21%, (18.69 bps/Hz)/(24.55 bps/Hz) $\times$ 100 $=$ 76.13%, (11.81 bps/Hz)/(24.55 bps/Hz) $\times$ 100 $=$ 48.11%, respectively, of the spectral efficiency achieved by the optimal SWIPT-Full-digital ZF Precoding.

Fig. 4 shows the spectrum efficiency comparison against the number of UEs for all five schemes under consideration, where the SNR is set as 10 dB. From Fig. 4, we find that spectrum efficiency increases for all curves as the number of UE increases. This is because multiple UEs can

[9]The simulations are performed using MATLAB v9.7 (R2019b) on a laptop equipped with an Intel Core i5-8265U at 1.6 GHz (4 cores) and 8 GB of memory.

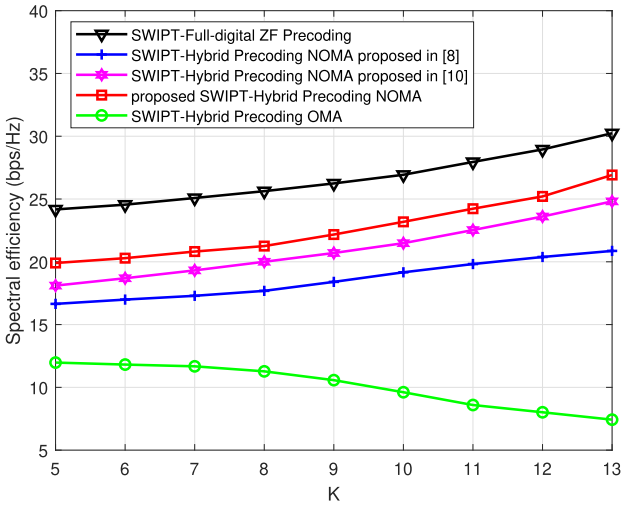

FIGURE 4. Spectral Efficiency achieved by various precoding solutions, where $N = 64$, $\mathcal{D} = 4$ beams, SNR $= 10$ dB, $P_T = 30$ mW. The propagation medium is a $\mathcal{L}_{d,m} = 3$ paths comprising of $l = 1$ LOS and $2 \leq l \leq \mathcal{L}_{d,m}$ NLOS. $N^{RF} = 4$ RF chains are assumed to be available for hybrid precoding with $B = 4$-bit resolution PSs.

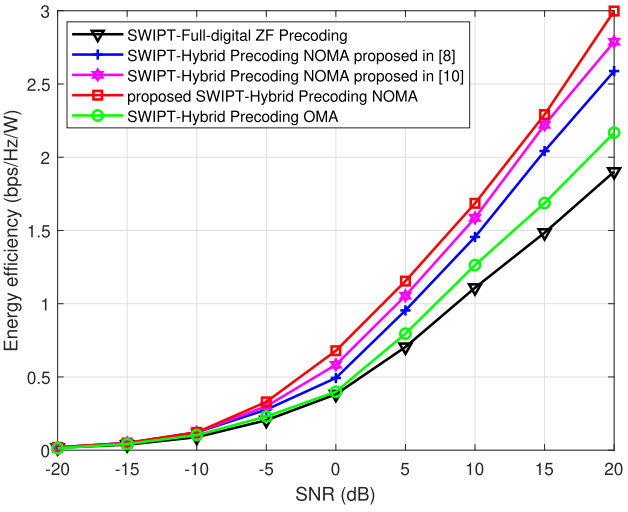

FIGURE 5. Energy Efficiency achieved by various precoding solutions, where $N = 64$, $\mathcal{D} = 4$ beams, $K = 6$ UEs, $P_T = 30$ mW. The propagation medium is a $\mathcal{L}_{d,m} = 3$ paths comprising of $l = 1$ LOS and $2 \leq l \leq \mathcal{L}_{d,m}$ NLOS. $N^{RF} = 4$ RF chains are assumed to be available for hybrid precoding with $B = 4$-bit resolution PSs.

use the same time-frequency resource-block by employing intra-beam superposition coding at BS and SIC at the receiver. However, we observe that our proposed SWIPT-Enabled mmWave mMIMO-NOMA Systems with Hybrid Precoding outperforms the other schemes and attains performance close to the SWIPT-Full-digital ZF Precoding. Therefore, this suggests that utilizing the proposed user grouping, analog RF precoder, digital baseband precoder design schemes is beneficial for inter-beam interference cancellation while optimizing (24).

C. ENERGY EFFICIENCY OF THE SYSTEM

In this section, we evaluate the energy efficiency of the system. Fig. 5 shows the energy efficiency comparison against SNR for all five schemes under consideration,

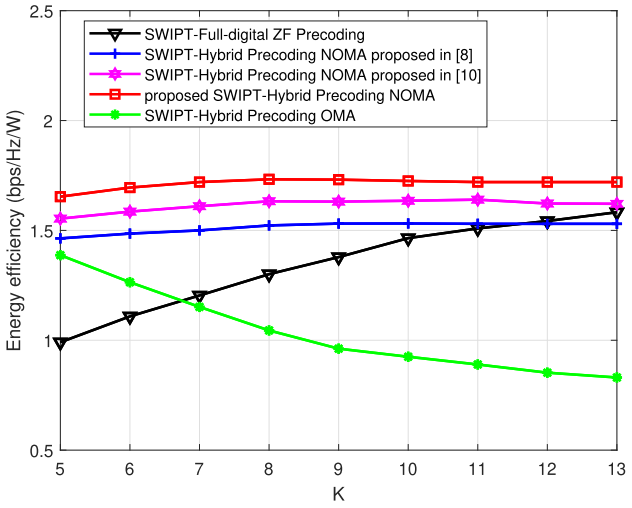

FIGURE 6. Energy Efficiency achieved by various precoding solutions, where $N = 64$, $\mathcal{D} = 4$ beams, SNR $= 10$ dB, $P_T = 30$ mW. The propagation medium is a $\mathcal{L}_{d,m} = 3$ paths comprising of $l = 1$ LOS and $2 \leq l \leq \mathcal{L}_{d,m}$ NLOS. $N^{RF} = 4$ RF c hains are assumed to be available for hybrid precoding with $B = 4$-bit resolution PSs.

where the number of UEs is set as $K = 6$. Fig. 5 shows that the SWIPT-Full-digital ZF Precoding scheme is as energy-efficient as the other hybrid precoding schemes at low SNR. In this setting, the SWIPT-Full-digital ZF Precoding scheme compensates for the notable power consumption by high data throughput, consequently increasing energy efficiency. However, we can observe that the proposed SWIPT-enabled mmWave mMIMO-NOMA system with Hybrid Precoding outperforms existing schemes under consideration in terms of energy efficiency in the moderate to high SNR regime, which profits from the implementation of NOMA to serve multiple UEs in each beam. For example, at SNR $= 10$ dB energy efficiency for the proposed SWIPT-Hybrid Precoding NOMA, SWIPT-Hybrid Precoding NOMA proposed in [8], SWIPT-Hybrid Precoding NOMA proposed in [10], SWIPT-Hybrid Precoding OMA and SWIPT-Full-digital ZF Precoding are 1.69 bps/Hz/W, 1.46 bps/Hz/W, 1.59 bps/Hz/W, 1.26 bps/Hz/W and 1.02 bps/Hz/W, respectively.

Fig. 6 shows the energy efficiency comparison against the number of UEs for all five schemes under consideration, where the SNR is set as 10 dB. From the curves depicted in Fig. 6, we find that the performance in terms of energy efficiency for the SWIPT-Hybrid Precoding OMA system deteriorates with an increasing number of UEs. We also observe that the energy efficiency of the SWIPT-Full-digital ZF Precoding scheme improves with an increasing number of UEs. We also observe that our proposed SWIPT-enabled mmWave mMIMO-NOMA system renders better energy efficiency in the low and medium number of UEs regimes, whereas its energy efficiency saturates as we increase the number of UEs.

V. CONCLUSION

This paper studied hybrid precoding for SWIPT-enabled mmWave mMIMO-NOMA systems to maximize the

achievable sum-rate and total energy efficiency. We first presented the optimization of user grouping and then designed the hybrid analog-digital precoders. We then showed that the resulting non-convex joint power allocation and power splitting optimization problem that maximizes the achievable sum-rate can be decoupled entirely into four separate optimization problems, where we then employ an alternating optimization algorithm to solve the problem. As shown in simulation results, our proposed SWIPT-enabled mmWave mMIMO-NOMA system with hybrid precoding converged reasonably fast within several iterations. In comparison with other state-of-the-art schemes, the proposed SWIPT-enabled mmWave mMIMO-NOMA system with hybrid precoding significantly improved the spectrum efficiency and energy efficiency performance of the considered system than those of state-of-the-art designs, which demonstrated its effectiveness. Moreover, the superiority of mmWave MIMO-NOMA over mmWave MIMO-OMA still holds.

REFERENCES

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

[2] A. N. Uwaechia, N. M. Mahyuddin, M. F. Ain, N. M. A. Latiff, and N. F. Za'bah, "On the spectral-efficiency of low-complexity and resolution hybrid precoding and combining transceivers for mmWave MIMO systems," *IEEE Access*, vol. 7, pp. 109259–109277, Aug. 2019.

[3] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, C.-L. I, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, Sep. 2017.

[4] J. Choi, G. Lee, and B. L. Evans, "User scheduling for millimeter wave hybrid beamforming systems with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2401–2414, Apr. 2019.

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

[6] A. N. Uwaechia, N. M. Mahyuddin, M. F. Ain, N. M. A. Latiff, and N. F. Za'bah, "Compressed channel estimation for massive MIMO-OFDM systems over doubly selective channels," *Phys. Commun.*, vol. 36, Oct. 2019, Art. no. 100771.

[7] T. Ding, Y. Zhao, L. Li, D. Hu, and L. Zhang, "Hybrid precoding for beamspace MIMO systems with sub-connected switches: A machine learning approach," *IEEE Access*, vol. 7, pp. 143273–143281, Sep. 2019.

[8] 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.

[9] F. Talaei and X. Dong, "Hybrid mmWave MIMO-OFDM channel estimation based on the multi-band sparse structure of channel," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1018–1030, Feb. 2019.

[10] 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.

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

[12] 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.

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

[14] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[15] A. N. Uwaechia and N. M. Mahyuddin, "A comprehensive survey on millimeter wave communications for fifth-generation wireless networks: Feasibility and challenges," *IEEE Access*, vol. 8, pp. 62367–62414, Mar. 2020.

[16] Q. Shi and M. Hong, "Spectral efficiency optimization for millimeter wave multiuser MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 455–468, Jun. 2018.

[17] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[18] R. Pal, K. V. Srinivas, and A. K. Chaitanya, "A beam selection algorithm for millimeter-wave multi-user MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 852–855, Apr. 2018.

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

[20] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Joint Tx-Rx beamforming and power allocation for 5G millimeter-wave non-orthogonal multiple access networks," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5114–5125, Jul. 2019.

[21] Y. Sun, Z. Ding, X. Dai, and G. K. Karagiannidis, "A feasibility study on network NOMA," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4303–4317, Sep. 2018.

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

[23] Q. Zhang, Q. Li, and J. Qin, "Robust beamforming for nonorthogonal multiple-access systems in MISO channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10231–10236, Dec. 2016.

[24] C. Xiong, Z. Hua, K. Lv, and X. Li, "An improved K-means text clustering algorithm by optimizing initial cluster centers," in *Proc. 7th Int. Conf. Cloud Comput. Big Data (CCBD)*, Nov. 2016, pp. 265–268.

[25] J. Zhang, L. Dai, X. Li, Y. Liu, and L. Hanzo, "On low-resolution ADCs in practical 5G millimeter-wave massive MIMO systems," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 205–211, Jul. 2018.

[26] J. Zhang, L. Dai, Z. He, S. Jin, and X. Li, "Performance analysis of mixed-ADC massive MIMO systems over Rician fading channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1327–1338, Jun. 2017.

[27] P. Raviteja, Y. Hong, and E. Viterbo, "Analog beamforming with low resolution phase shifters," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 502–505, Aug. 2017.

[28] Z. Wang, M. Li, Q. Liu, and A. L. Swindlehurst, "Hybrid precoder and combiner design with low-resolution phase shifters in mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 256–269, May 2018.

[29] L. R. Varshney, "Transporting information and energy simultaneously," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1612–1616.

[30] P. Grover and A. Sahai, "Shannon meets Tesla: Wireless information and power transfer," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 2363–2367.

[31] H. Zhang, A. Dong, S. Jin, and D. Yuan, "Joint transceiver and power splitting optimization for multiuser MIMO SWIPT under MSE QoS constraints," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7123–7135, Aug. 2017.

[32] T. N. Nguyen, M. Tran, P. T. Tran, P. T. Tin, T.-L. Nguyen, D.-H. Ha, and M. Voznak, "On the performance of power splitting energy harvested wireless full-duplex relaying network with imperfect CSI over dissimilar channels," *Secur. Commun. Netw.*, vol. 2018, pp. 1–11, Dec. 2018.

[33] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyösti, D. Laurenson, G. Matz, A. F. Molisch, C. Oestges, and H. Özcelik, "Survey of channel and radio propagation models for wireless MIMO systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2007, no. 1, Feb. 2007, Art. no. 019070.

[34] A. A. M. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 2, pp. 128–137, Feb. 1987.

[35] C.-C. Chong, C.-M. Tan, D. I. Laurenson, S. McLaughlin, M. A. Beach, and A. R. Nix, "A new statistical wideband spatio-temporal channel model for 5-GHz band WLAN systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 139–150, Feb. 2003.

[36] J. W. Wallace and M. A. Jensen, "Modeling the indoor MIMO wireless channel," *IEEE Trans. Antennas Propag.*, vol. 50, no. 5, pp. 591–599, May 2002.
[37] Q. Qin, L. Gui, P. Cheng, and B. Gong, "Time-varying channel estimation for millimeter wave multiuser MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9435–9448, Oct. 2018.
[38] L. Cheng, G. Yue, D. Yu, Y. Liang, and S. Li, "Millimeter wave time-varying channel estimation via exploiting block-sparse and low-rank structures," *IEEE Access*, vol. 7, pp. 123355–123366, Aug. 2019.
[39] M. Kokshoorn, H. Chen, P. Wang, Y. Li, and B. Vucetic, "Millimeter wave MIMO channel estimation using overlapped beam patterns and rate adaptation," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 601–616, Feb. 2017.
[40] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.
[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[42] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*. New York, NY, USA: Wiley, 1988.
[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

ANTHONY NGOZICHUKWUKA UWAECHIA received the B.Eng. and M.Sc. degrees in electrical and electronic engineering from Ahmadu Bello University, Zaria, Nigeria, in 2006 and 2013, respectively, and the Ph.D. degree in wireless and mobile systems from the Universiti Sains Malaysia (USM), Malaysia, in 2018.

He has been a Postdoctoral Research Fellow with the School of Electrical and Electronic Engineering, USM, since 2019. He has authored several technical articles and book chapters in major international journals and conferences. His research interests include signal processing for communications and particularly focusing in digital signal processing, sparse channel estimation, channel measurement, sparse representation, OFDM systems, robust precoding in massive MIMO systems, and mm-wave communications.

Dr. Uwaechia was a recipient of the Nigeria-Sao Tome and Principe Joint Development Postgraduate Scholarship Award during the M.Sc. degree, the Graduate on Time Award for the Ph.D. degree, and the Institute of Postgraduate Studies, Graduate Assistant Scheme Award from USM during the Ph.D. studies.

NOR MUZLIFAH MAHYUDDIN (Member, IEEE) received the B.Eng. degree from the Universiti Teknologi Malaysia, Malaysia, in 2005, the M.Sc. degree from the Universiti Sains Malaysia, Malaysia, in 2006, and the Ph.D. degree from Newcastle University, Newcastle upon Tyne, U.K., in 2011.

She has more than seven years of teaching experience in digital circuit and system encompassing various fields from wireless communication to integrated circuit. She is currently an Associate Professor with the School of Electrical and Electronic Engineering, Universiti Sains Malaysia. She has also involved with the projects revolving around sparse channel estimation, OFDM systems, massive MIMO systems, and mm-wave communications. Through Malaysia national grants and works in 5G technology, she would like to expand this further into the area of reconfigurable system, which encapsulates the importance of miniaturization and low-power design for current technology. She has about 29 publications in book, book chapters, and conference proceedings and journals. She has also supervised four Ph.D. students and more than 20 master's students. Her research interests include art of miniaturization, low-power in RF and microwave engineering, and digital integrated circuit.

Dr. Mahyuddin is a member of the IEEE Communications Society (ComSoc) and IET. She is also registered with the Board of Engineers Malaysia (BEM).