Energy Efficiency of Multi-User mmWave Rate-Splitting Multiple Access With Hybrid Precoding

Jared S. Everett and Brian L. Mark
Dept. of Electrical and Computer Engineering, George Mason University, Fairfax, Virginia, United States
Email: jared.everett@ieee.org, bmark@gmu.edu

2024 IEEE International Conference on Communications Workshops (ICC Workshops) | 979-8-3503-0405-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCWORKSHOPS59551.2024.10615508

***Abstract*—We investigate the energy efficiency (EE) problem in a downlink multi-user millimeter wave (mmWave) rate-splitting multiple access (RSMA) system and propose an energy-efficient one-layer RSMA hybrid precoder design for K users with quality of service constraints. This scheme is applicable to the design of sustainable sixth generation (6G) cellular networks. To make the problem tractable, the analog and the digital precoder designs are decoupled. First, the analog precoder is designed to maximize the desired signal power of each user while ignoring multi-user interference. Second, the digital precoder is designed to manage multi-user interference according to the EE optimization design criterion. We adopt a successive convex approximation-based algorithm for joint optimization of the digital precoders, power, and common rate allocation. Simulation results show that the proposed RSMA scheme always performs at least as well as a baseline spatial division multiple access (SDMA) hybrid precoding scheme and outperforms it under certain channel conditions. These results suggest that RSMA is suitable as a flexible physical layer design for future 6G mmWave networks.**

***Index Terms*—Rate-splitting multiple access (RSMA), millimeter wave (mmWave) communication, energy efficiency (EE), MIMO, hybrid precoding, 6G mobile communication**

I. Introduction

Millimeter-wave (mmWave) communication emerged as a new cellular technology in fifth generation (5G) cellular networks. As we move forward to sixth generation (6G) and beyond, efficient support for mmWave spectrum will be an important part of any new cellular physical layer (PHY) design. Ideally, it is desirable that the same multiple access scheme can be utilized across all available spectrum. As was the case in 5G, this may be done through the use of different system parameters (i.e., numerologies) that adapt to the needs of different frequency bands, but the overall scheme should be flexible enough to support this adaptation. At the same time, the international telecommunications union (ITU) recently identified sustainability as a foundational aspiration in their vision for 6G [1], which will increase the importance of energy-efficient system design across all frequency bands.

Rate-splitting multiple access (RSMA) has emerged in recent years as a powerful interference management approach and multiple access scheme for future wireless networks [2], [3]. Past research has shown that a key advantage of RSMA is its ability to generalize and softly bridge other multiple access schemes, e.g., orthogonal multiple access (OMA), space-division multiple access (SDMA), non-orthogonal multiple access (NOMA), and multicasting [4], [5]. Additional advantages of RSMA include performance advantages for both spectral efficiency (SE) [4] and energy efficiency (EE) [6], [7] in a variety of practical scenarios. It is no surprise that RSMA has been proposed as a candidate technology for 6G PHY design [2], [3], [8]. A key question for 6G is whether RSMA can provide a flexible framework that is adaptable across all existing spectrum (e.g., low-, mid-, and high-band including mmWave) and future spectrum (e.g., terahertz). With its many advantages, it may not be necessary for RSMA to outperform other schemes at mmWave in every metric, so long as RSMA performs at least as well as other schemes while offering suitability across all 6G bands.

While the literature on RSMA has advanced tremendously in recent years, the performance of RSMA specifically for mmWave spectrum [e.g., Third Generation Partnership Project (3GPP) frequency range 2 (FR2)] is not as extensively studied. Prior research shows that rate-splitting can be used in mmWave hybrid precoding systems to increase spectral efficiency [9]–[12], reduce channel training and feedback complexity [9], and improve robustness to imperfect channel state information at the transmitter (CSIT) caused by channel estimation error [11]. However, prior work on EE of RSMA for mmWave systems is limited and highly application-specific. A cellular-connected mmWave unmanned aerial vehicle (UAV) downlink is considered in [13], which primarily studies the EE of RSMA vs. NOMA as a function of UAV altitude. A Ka-band multibeam geostationary satellite communication downlink is considered in [12], which studies the EE of RSMA vs. SDMA under both perfect and imperfect CSIT assumptions. However, both [13] and [12] consider digital precoding only. Hybrid precoding is considered in [14], but the results are specific to an active intelligent reflecting surface (IRS)-aided mmWave RSMA downlink system. Therefore, none of these papers provide direct insights into the EE of mmWave RSMA systems with hybrid precoding in general.

This paper presents the first analysis of the EE of the mmWave RSMA downlink in a general setting. We propose a novel energy-efficient one-layer RSMA hybrid precoding scheme and compare EE performance against a baseline SDMA scheme assuming a geometric mmWave channel model with limited paths. The precoder is designed in two stages. First, the analog precoder is designed as in prior works [9]–[11]. Second, the digital precoders, power allocation, and common rate allocation are designed by adapting the state-of-the-art successive

This work was supported in part by the U.S. National Science Foundation under Grant 2034616.

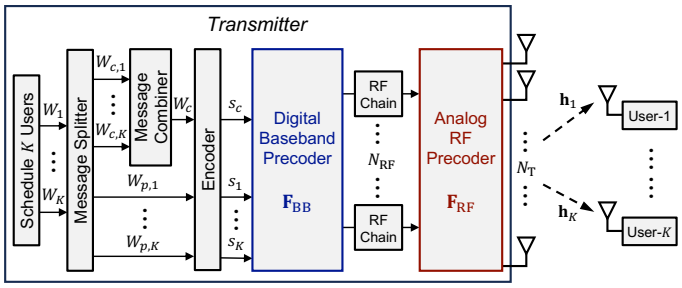

Fig. 1. Downlink multi-user mmWave RSMA system model.

convex approximation (SCA)-based algorithm from [6], [7] for hybrid precoding. In contrast to [9], we optimize the digital precoders jointly to account for interaction among users, and we do not restrict the power allocation among streams. This approach is similar to [10], except our one-layer RSMA scheme only requires one common stream, which avoids multiple layers of successive interference cancellation (SIC) processing at the receivers, making our scheme more practical in terms of user equipment complexity. Our simulation results show that the proposed RSMA scheme offers modest improvement in average EE over SDMA. However, for individual channel realizations, the RSMA scheme always performs at least as well as the SDMA scheme and significantly outperforms it for certain channel scenarios. These results suggest that the well-known EE benefits of RSMA at lower frequencies extend to the mmWave regime as well.

The remainder of this paper is organized as follows. Sec. II introduces the system model, channel model, and power consumption model. Sec. III describes the proposed energy-efficient RSMA hybrid precoder design. Sec. IV presents simulation results. Lastly, Sec. V gives concluding remarks.

Notation: Boldface uppercase and lowercase letters denote matrices and column vectors, respectively. Superscripts $(\cdot)^T$ and $(\cdot)^H$ denote transpose and conjugate-transpose (Hermitian) operators, respectively. $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. $\|\cdot\|$ denotes the Euclidean norm of a vector. $\mathbb{E}[\cdot]$ denotes the expected value of a random variable. $\mathbf{I}$ denotes the identity matrix. $\mathbb{C}$ denotes the set of complex numbers.

II. System Model

A. System Model

We consider a single-cell multi-user mmWave downlink system shown in Fig. 1 in which a base station (BS) communicates with K users in the user set $\mathcal{K} = \{1, ..., K\}$. The BS is equipped with N_{T} antennas and N_{RF} radio frequency (RF) chains satisfying $N_{\mathrm{RF}} < N_{\mathrm{T}}$. Each user is equipped with a single antenna. For simplicity, we make the common assumption that the number of users equals the number of RF chains, i.e., $K = N_{\mathrm{RF}}$ [9] [15].

In this paper, we study the one-layer RSMA hybrid precoding architecture first proposed in [9]. A hybrid precoding scheme is used at the BS, comprising an analog RF precoder $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{T}} \times N_{\mathrm{RF}}}$ and a digital baseband precoder $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{RF}} \times N_{\mathrm{S}}}$, where N_{S} is the number of independent streams to be transmitted. We consider a fully-connected architecture at the BS in which each RF chain is connected to all antennas through a group of phase shifters [15]. The phase shifting network imposes a constant modulus constraint on the entries of the analog precoder. We assume that the angles of the phase shifters are quantized with a finite set of possible values, as in prior works [9] [10]. Without loss of generality, we assume $[\mathbf{F}_{\mathrm{RF}}]_{m,n} = \frac{1}{\sqrt{N_{\mathrm{T}}}} e^{\varphi_{m,n}}$, where $\varphi_{m,n}$ is a quantized angle.

The BS communicates K messages, $W_1, ..., W_K$, to the K users. According to the rate-splitting strategy, the BS splits W_k, the message of user-k, into a common part and a private part, $W_{c,k}$ and $W_{p,k}$, respectively. The common parts of all K users' messages are combined into a single common message, $W_c = \{W_{c,1}, ..., W_{c,K}\}$, which is encoded into a common stream, s_c, using a shared codebook and decoded by all users. The private parts are independently encoded into private streams, $s_1, ..., s_K$, and decoded by their respective users. This results in $N_{\mathrm{S}} = K + 1$ streams at the input of the digital precoder. The streams are linearly precoded according to the hybrid precoding scheme. The resulting transmit signal is given by

$$\mathbf{x} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} = \mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},c}s_c + \sum_{k\in\mathcal{K}} \mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},k}s_k, \tag{1}$$

where $\mathbf{F}_{\mathrm{BB}} = [\mathbf{f}_{\mathrm{BB},c}, \mathbf{f}_{\mathrm{BB},1}, ..., \mathbf{f}_{\mathrm{BB},K}]$ are the digital precoding vectors for each stream, and $\mathbf{s} = [s_c, s_1, ..., s_K]^T$ is the stream vector for a given channel use. Assuming that $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}$, the average transmit power constraint is written as

$$P_c + \sum_{k\in\mathcal{K}} P_k \le P_{\max}, \tag{2}$$

where $P_c = \|\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},c}\|^2$ and $P_k = \|\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},k}\|^2, \forall k \in \mathcal{K}$.

We adopt a narrowband block fading mmWave channel model as in [9]–[11], [15], [16]. The received signal at user-k is then written as

$$y_k = \underbrace{\mathbf{h}_k^H\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},c}s_c}_{\text{common stream}} + \underbrace{\mathbf{h}_k^H\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},k}s_k}_{\text{private stream}} + \underbrace{\sum_{j\in\mathcal{K},j\neq k} \mathbf{h}_k^H\mathbf{F}_{\mathrm{RF}}\mathbf{f}_{\mathrm{BB},j}s_j}_{\text{multi-user interference}} + \underbrace{n_k}_{\text{noise}}, \tag{3}$$

where $\mathbf{h}_k \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$ is the channel vector from the BS to user-k, and $n_k \sim \mathcal{CN}(0, \sigma_{n,k}^2)$ is additive white Gaussian noise (AWGN) at user-k. Without loss of generality, we assume equal noise variance at the users, i.e., $\sigma_{n,k}^2 = \sigma_n^2$.

B. Channel Model

A key feature of mmWave channel models, such as those used for 3GPP FR2, is the assumption of a limited scattering environment. Therefore, we adopt a geometric channel model with L scatterers, where each scatterer contributes a single propagation path between the BS and the user [9]–[11], [15], [16]. The channel between the BS and user-k is modeled as

$$\mathbf{h}_k = \sqrt{\frac{N_{\mathrm{T}}}{L_k}} \sum_{l=1}^{L_k} \alpha_{k,l}\mathbf{a}(\theta_{k,l}), \tag{4}$$

where L_k is the number of channel paths to user-k, $\alpha_{k,l}$ is the complex gain of path-l, $\theta_{k,l} \in [0, \pi]$ is the azimuth angle-of-departure (AoD) of path-l, and $\mathbf{a}(\theta_{k,l}) \in \mathbb{C}^{N_\mathrm{T} \times 1}$ is the array steering vector of path-l. The path gains are independent and identically distributed (i.i.d.) $\alpha_{k,l} \sim \mathcal{CN}(0,1)$ and vary independently across different time slots.

While the algorithms and results in this paper can be applied to arbitrary antenna arrays, we consider a uniform linear array (ULA) at the BS and azimuth beamforming for simplicity. The model can easily be extended to a 2D array. Under the plane wave and balanced narrowband array assumptions, the array steering vector is written as

$$\mathbf{a}(\theta_{k,l}) = \frac{1}{\sqrt{N_\mathrm{T}}}[1, e^{j\frac{2\pi}{\lambda}d\cos\theta_{k,l}}, ..., e^{j(N_\mathrm{T}-1)\frac{2\pi}{\lambda}d\cos\theta_{k,l}}]^T, \quad (5)$$

where λ is the wavelength and d is the antenna spacing [9]. We assume $d = \lambda/2$. Lastly, we assume perfect CSIT and perfect CSI at the receiver (CSIR) (i.e., $\mathbf{h}_k$ is known to the BS and user-k) to study upper bound performance.

C. Power Consumption Model

Following standard methods for EE analysis [6], [7], [13], [16], [17], the total power consumption at the BS consists of two parts: the flexible transmit power and the fixed circuit power. Power consumption at the user side is omitted since it is negligible compared to the power consumption at the BS [6]. The total power consumption is modeled as [6], [7], [17]

$$P_{\text{total}} = \psi\left(P_c + \sum_{k\in\mathcal{K}} P_k\right) + P_{\text{cir}}, \quad (6)$$

where $\psi \geq 1$ accounts for the inefficiency of the power amplifier and P_{cir} is the fixed circuit power consumption. We model the fixed circuit power of a mmWave BS as [16]

$$P_{\text{cir}} = P_{\text{BB}} + N_{\text{RF}}P_{\text{RF}} + N_{\text{RF}}N_{\text{T}}P_{\text{PS}} + N_{\text{T}}P_{\text{PA}}, \quad (7)$$

where P_{BB}, P_{RF}, P_{PS}, and P_{PA} denote the power consumption of the baseband, RF chain, phase shifter, and power amplifier, respectively.

III. Energy-Efficient Precoder Design

In this section, we propose an energy-efficient one-layer RSMA hybrid precoder design. To make the problem tractable, the analog precoder and the digital precoder are determined in a decoupled manner, as in [9], [10], [15]. First, the analog precoder is designed to maximize the desired signal power of each user, ignoring multi-user interference. Second, the digital precoder is designed to manage multi-user interference according to the EE optimization design criteria. For the digital precoder, we adopt the SCA-based algorithm originally proposed in [6] as an approximate solution to the EE optimization problem. This approach jointly optimizes the digital precoder direction, the power allocation among streams, and the common stream rate allocation among users for a given analog precoder.

A. Analog Precoder Design

We adopt a beamsteering codebook design for the analog precoder [9], [10], [15]. Let $\mathcal{F}$ represent the analog codebook, with cardinality $|\mathcal{F}| = Q = 2^{B_{\text{RF}}}$, where $B_{\text{RF}} = \lceil \log_2(N_\mathrm{T}) \rceil$ is the number of quantization bits. Since the codebook should have the same form as the array steering vector, we define the codebook as $\mathcal{F} = \{\mathbf{a}(\theta_q) | \theta_q = \frac{\pi q}{Q}, q \in [1, Q]\}$. For each channel realization, the BS sets the codeword of user-k as

$$\mathbf{f}_{\text{RF},k} = \arg\max_{\mathbf{f}_{\text{RF},k}\in\mathcal{F}} |\mathbf{h}_k^H \mathbf{f}_{\text{RF},k}|^2. \quad (8)$$

In practical systems, codeword selection can be achieved using an efficient beam search algorithm with feedback [9]. The overall analog precoder matrix is then given by $\mathbf{F}_{\text{RF}} = [\mathbf{f}_{\text{RF},1}, ..., \mathbf{f}_{\text{RF},K}]$.

B. Digital Precoder Design

We now design the digital precoder and common rate allocation to maximize the EE for a given $\mathbf{F}_{\text{RF}}$. Assuming SIC receivers, each user first decodes the common stream s_c while treating interference from all private streams as noise. This enables the ability to partially decode interference from other users. The signal-to-interference-plus-noise ratio (SINR) of the common stream at user-k is

$$\gamma_{c,k} = \frac{\left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},c}\right|^2}{\sum_{j\in\mathcal{K}} \left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j}\right|^2 + \sigma_n^2}. \quad (9)$$

Next, the common message estimate is re-encoded and subtracted from the received signal. The residual signal at user-k is then used to decode the private stream, s_k, while treating any residual multi-user interference from the other users' private streams as noise. Hence, RSMA enables partially treating interference from other users as noise. Assuming perfect SIC, the SINR of the private stream at user-k is

$$\gamma_{p,k} = \frac{\left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},k}\right|^2}{\sum_{j\in\mathcal{K}, j\neq k} \left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j}\right|^2 + \sigma_n^2}. \quad (10)$$

Assuming Gaussian signaling, the achievable rate (per Hz) of the common stream at user-k is

$$R_{c,k} = \log_2\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},c}\right|^2}{\sum_{j\in\mathcal{K}} \left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j}\right|^2 + \sigma_n^2}\right). \quad (11)$$

To ensure decodability of the common stream at all users, the common stream rate is limited to

$$R_c = \min\left(R_{c,1}, ..., R_{c,K}\right) = \sum_{k\in\mathcal{K}} C_k, \quad (12)$$

where C_k is the common stream rate allocation for user-k. Similarly, the achievable rate of the private stream at user-k is

$$R_{p,k} = \log_2\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},k}\right|^2}{\sum_{j\in\mathcal{K}, j\neq k} \left|\mathbf{h}_k^H \mathbf{F}_{\text{RF}} \mathbf{f}_{\text{BB},j}\right|^2 + \sigma_n^2}\right). \quad (13)$$

The total rate of user-k is then $R_k = C_k + R_{p,k}$. The spectral efficiency (SE) of the system, defined as the sum rate, is

$$R_{\text{sum}} = \sum_{k\in\mathcal{K}} R_k = R_c + \sum_{k\in\mathcal{K}} R_{p,k}. \quad (14)$$

As in prior work [6], [7], [12]–[14], [16], [17], we define the EE of the system as the sum rate divided by the total power consumption defined in (6). By this model, the EE is

$$\eta_{\text{EE}} = \frac{R_{\text{sum}}}{P_{\text{total}}} = \frac{\sum_{k\in\mathcal{K}}(C_k + R_{p,k})}{\psi(P_c + \sum_{k\in\mathcal{K}} P_k) + P_{\text{cir}}}. \tag{15}$$

The flexible transmit power can equivalently be written as

$$P_c + \sum_{k\in\mathcal{K}} P_k = \text{tr}\left(\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^H\mathbf{F}_{\text{RF}}^H\right). \tag{16}$$

For a given $\mathbf{F}_{\text{RF}}$, the resulting EE optimization problem is

$$\begin{aligned}
\max_{\mathbf{F}_{\text{BB}},\mathbf{c}} \quad & \eta_{\text{EE}} = \frac{\sum_{k\in\mathcal{K}}(C_k + R_{p,k})}{\psi\text{tr}\left(\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^H\mathbf{F}_{\text{RF}}^H\right) + P_{\text{cir}}} && \text{(17a)}\\
\text{s.t.} \quad & C_k + R_{p,k} \geq R_k^{\text{th}}, \quad \forall k \in \mathcal{K} && \text{(17b)}\\
& \textstyle\sum_{j\in\mathcal{K}} C_j \leq R_{c,k}, \quad \forall k \in \mathcal{K} && \text{(17c)}\\
& \text{tr}\left(\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^H\mathbf{F}_{\text{RF}}^H\right) \leq P_{\max} && \text{(17d)}\\
& C_k \geq 0, \quad \forall k \in \mathcal{K}, && \text{(17e)}
\end{aligned}$$

where (17b) is the quality of service (QoS) constraint, (17c) ensures the common stream can be decoded by all users, (17d) is the power constraint, and (17e) constrains the common stream rate allocations to be non-negative. For this analysis, we assume that all users have the same QoS threshold rate, i.e., $R_k^{\text{th}} = R^{\text{th}}, \forall k \in \mathcal{K}$.

Problem (17) is a non-convex fractional program, which is hard to solve directly. Therefore, we adopt the SCA approach to approximate the optimal solution [6], [7]. For a given $\mathbf{F}_{\text{RF}}$, we define the effective channel of the digital precoder from the BS to user-k as $\mathbf{h}_{\text{eff},k} = \mathbf{F}_{\text{RF}}^H\mathbf{h}_k$. Next, we apply the SCA-based beamforming algorithm with rate-splitting that was originally proposed in [6] to solve the 2-user EE optimization problem for digital precoding without QoS. The algorithm was extended in [7] to a K-user EE RSMA scenario in which a multicast stream and QoS rate constraints were introduced. Comparing with the EE optimization problem in [7, eq. (10)], the main differences of problem (17) in the hybrid precoding scheme lie in the removal of the multicast stream (as in [6]), the application of the digital precoder optimization to the effective channels $\mathbf{h}_{\text{eff},k}$, and the modification of the total power consumption (6) and the power constraint (17d) to reflect the total hybrid precoder $\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}$.

Following the approach in [6], [7], we introduce scalar variables t, ω^2, and z to represent the EE metric, sum rate, and total power consumption, respectively. Problem (17) becomes

$$\begin{aligned}
\max_{\mathbf{F}_{\text{BB}},\mathbf{c},\omega,z,t} \quad & t && \text{(18a)}\\
\text{s.t.} \quad & \frac{\omega^2}{z} \geq t && \text{(18b)}\\
& \textstyle\sum_{k\in\mathcal{K}}(C_k + R_{p,k}) \geq \omega^2 && \text{(18c)}\\
& z \geq \psi\text{tr}\left(\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^H\mathbf{F}_{\text{RF}}^H\right) + P_{\text{cir}} && \text{(18d)}\\
& (17b) - (17d). && \text{(18e)}
\end{aligned}$$

Problems (17) and (18) are equivalent because constraints (18b)–(18d) hold with equality at the optimum. Next, we introduce auxiliary variables to transform the non-convex constraints. Introducing variables $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K]^T$ to represent the set of private rates, constraints (17b) and (18c) become

$$(17b), (18c) \Leftrightarrow \begin{cases} C_k + \alpha_k \geq R_k^{\text{th}}, & \forall k \in \mathcal{K} \quad \text{(19a)}\\ \textstyle\sum_{k\in\mathcal{K}}(C_k + \alpha_k) \geq \omega^2 & \text{(19b)}\\ R_{p,k} \geq \alpha_k, & \forall k \in \mathcal{K}. \quad \text{(19c)} \end{cases}$$

Introducing variables $\boldsymbol{\vartheta} = [\vartheta_1, ..., \vartheta_K]^T$ to represent one plus the SINR of each private stream, constraints (19c) become

$$(19c) \Leftrightarrow \begin{cases} \vartheta_k \geq 2^{\alpha_k}, & \forall k \in \mathcal{K} \quad \text{(20a)}\\ 1 + \gamma_{p,k} \geq \vartheta_k, & \forall k \in \mathcal{K}. \quad \text{(20b)} \end{cases}$$

Introducing variables $\boldsymbol{\beta} = [\beta_1, ..., \beta_K]^T$ to represent the interference plus noise of the intended private stream at each user, constraints (20b) become

$$(20b) \Leftrightarrow \begin{cases} \dfrac{|\mathbf{h}_{\text{eff},k}^H\mathbf{f}_{\text{BB},k}|^2}{\beta_k} \geq \vartheta_k - 1, & \forall k \in \mathcal{K} \quad \text{(21a)}\\ \beta_k \geq \sum_{j\neq k} |\mathbf{h}_{\text{eff},k}^H\mathbf{f}_{\text{BB},j}|^2 + \sigma_n^2, & \forall k \in \mathcal{K}. \quad \text{(21b)} \end{cases}$$

The same method is used to transform (17c). Introducing variables $\boldsymbol{\alpha_c} = [\alpha_{c,1}, ..., \alpha_{c,K}]^T$, $\boldsymbol{\vartheta}_c = [\vartheta_{c,1}, ..., \vartheta_{c,K}]^T$, and $\boldsymbol{\beta}_c = [\beta_{c,1}, ..., \beta_{c,K}]^T$, constraint (17c) becomes

$$(17c) \Leftrightarrow \begin{cases} \textstyle\sum_{i\in\mathcal{K}} C_i \leq \alpha_{c,k}, & \forall k \in \mathcal{K} \quad \text{(22a)}\\ \vartheta_{c,k} \geq 2^{\alpha_{c,k}}, & \forall k \in \mathcal{K} \quad \text{(22b)}\\ \dfrac{|\mathbf{h}_{\text{eff},k}^H\mathbf{f}_{\text{BB},c}|^2}{\beta_{c,k}} \geq \vartheta_{c,k} - 1, & \forall k \in \mathcal{K} \quad \text{(22c)}\\ \beta_{c,k} \geq \sum_{j\in\mathcal{K}} |\mathbf{h}_{\text{eff},k}^H\mathbf{f}_{\text{BB},j}|^2 + \sigma_n^2, & \forall k \in \mathcal{K}. \quad \text{(22d)} \end{cases}$$

Linear approximation methods from [6] are used to approximate the remaining non-convex constraints (18b), (21a), and (22c) in each iteration. Constraints (18b) and (21a) become

$$\frac{\omega^2}{z} \geq \frac{2\omega^{[n]}}{z^{[n]}}\omega - \left(\frac{\omega^{[n]}}{z^{[n]}}\right)^2 z \triangleq \Omega^{[n]}(\omega, z), \tag{23}$$

$$\begin{aligned}
\frac{|\mathbf{h}_{\text{eff},k}^H\mathbf{f}_{\text{BB},k}|^2}{\beta_k} \geq\ & 2\text{Re}\left(\left(\mathbf{f}_{\text{BB},k}^{[n]}\right)^H\mathbf{h}_{\text{eff},k}\mathbf{h}_{\text{eff},k}^H\mathbf{f}_{\text{BB},k}\right)/\beta_k^{[n]}\\
& - \left(\left|\mathbf{h}_{\text{eff},k}^H\mathbf{f}_{\text{BB},k}^{[n]}\right|/\beta_k^{[n]}\right)^2\beta_k \triangleq \Psi_k^{[n]}(\mathbf{f}_{\text{BB},k}, \beta_k), \quad \text{(24)}
\end{aligned}$$

respectively, where the points $(\omega^{[n]}, z^{[n]})$ and $(\mathbf{f}_{\text{BB},k}^{[n]}, \beta_k^{[n]})$ represent the values of the respective variables at the output of iteration-n. The approximation of constraint (22c) takes the same form as (24) by replacing $\mathbf{f}_{\text{BB},k}$, $\mathbf{f}_{\text{BB},k}^{[n]}$, β_k, $\beta_k^{[n]}$, and $\Psi_k^{[n]}$ with $\mathbf{f}_{\text{BB},c}$, $\mathbf{f}_{\text{BB},c}^{[n]}$, $\beta_{c,k}$, $\beta_{c,k}^{[n]}$, and $\Psi_{c,k}^{[n]}$, respectively. Therefore, problem (17) is approximated at iteration-n as

$$\begin{aligned}
\max_{\substack{\mathbf{F}_{\text{BB}},\mathbf{c},\omega,z,t,\\ \boldsymbol{\alpha},\boldsymbol{\alpha}_c,\boldsymbol{\vartheta},\boldsymbol{\vartheta}_c,\boldsymbol{\beta},\boldsymbol{\beta}_c}} \quad & t && \text{(25)}\\
\text{s.t.} \quad & \Omega^{[n]}(\omega, z) \geq t\\
& \Psi_k^{[n]}(\mathbf{f}_{\text{BB},k}, \beta_k) \geq \vartheta_k - 1, \quad \forall k \in \mathcal{K}\\
& \Psi_{c,k}^{[n]}(\mathbf{f}_{\text{BB},c}, \beta_{c,k}) \geq \vartheta_{c,k} - 1, \quad \forall k \in \mathcal{K}\\
& (17e), (17d), (18d), (19a), (19b),\\
& (20a), (21a), (22a), (22b), (22d).
\end{aligned}$$

Problem (25) is convex and can be solved efficiently using standard optimization toolboxes, such as YALMIP [18]. Initialization and convergence of the SCA-based algorithm are addressed in [6]. The algorithm is guaranteed to converge. Global optimality of the achieved solution cannot be guaranteed due to the approximation of the constraints. However, it was shown in [17] that the SCA-based algorithm in [6] often obtains a solution very close to the global optimum for a conventional digital precoding scenario. Details of the EE-optimized SCA-based hybrid precoding algorithm are specified in Algorithm 1.

Algorithm 1 Energy-Efficient SCA-Based Hybrid Precoder

1: **Analog Precoder:** BS finds $\mathbf{F}_{\text{RF}}$ from (8) as in single user case and calculates the effective channels $\mathbf{h}_{\text{eff},k}$, $\forall k \in \mathcal{K}$.
2: **Initialize:** $n \leftarrow 0$, $\mathbf{F}_{\text{BB}}^{[n]}$, $t^{[n]}$, $\omega^{[n]}$, $z^{[n]}$, $\boldsymbol{\beta}_c^{[n]}$, $\boldsymbol{\beta}^{[n]}$
3: **repeat**
4: $n \leftarrow n+1$
5: Solve (25) using $\mathbf{F}_{\text{BB}}^{[n-1]}$, $\omega^{[n-1]}$, $z^{[n-1]}$, $\boldsymbol{\beta}_c^{[n-1]}$, $\boldsymbol{\beta}^{[n-1]}$ and denote the optimal values as $\mathbf{F}_{\text{BB}}^*$, $\mathbf{c}^*$, t^*, ω^*, z^*, $\boldsymbol{\beta}_c^*$, $\boldsymbol{\beta}^*$
6: Update $\mathbf{F}_{\text{BB}}^{[n]} \leftarrow \mathbf{F}_{\text{BB}}^*$, $t^{[n]} \leftarrow t^*$, $\omega^{[n]} \leftarrow \omega^*$, $z^{[n]} \leftarrow z^*$, $\boldsymbol{\beta}_c^{[n]} \leftarrow \boldsymbol{\beta}_c^*$, $\boldsymbol{\beta}^{[n]} \leftarrow \boldsymbol{\beta}^*$
7: **until** $|t^{[n]} - t^{[n-1]}| < \epsilon$
8: **return** $\mathbf{F}_{\text{RF}}$, $\mathbf{F}_{\text{BB}}^*$, $\mathbf{c}^*$

C. Complexity Analysis

The complexity of the analog precoder design is equivalent to that of existing mmWave hybrid precoding schemes [15]. At the BS, the computational complexity of finding the RSMA digital precoder for a given $\mathbf{F}_{\text{RF}}$ is $\mathcal{O}([KN_{\text{RF}}]^{3.5}\log(\epsilon^{-1}))$, assuming each iteration is solved by the interior-point method [7]. At the user equipment, the computational complexity is equivalent to that of an SDMA receiver when no power is allocated to the common stream or a NOMA receiver with one layer of SIC when the common stream is used. No additional computational complexity is introduced in the receiver signal processing for scenarios in which RSMA performs equivalently to SDMA.

IV. Simulation Results

In this section, we evaluate the performance of the proposed RSMA hybrid precoding scheme and compare it to SDMA. We define a baseline SDMA hybrid precoding scheme using Algorithm 1 but set $P_c = 0$ and $C_k = 0, \forall k \in \mathcal{K}$. By turning off the common stream, the EE optimization problem in (17) reduces to the problem of multi-user linear precoding (MU-LP) design for SDMA [2]. The definition of the normalized maximum transmit power constraint, P_{max}, follows the methods of [6], [7]. Without loss of generality, unit noise variance is considered.

The system parameters are shown in Table I. We use the power consumption model parameters from [16] for a mmWave BS. We generate 500 random i.i.d. channel realizations that are re-used for each simulation. The AoDs are randomly generated and uniformly distributed over $[0, \pi]$. As in [9], we assume full rank $\mathbf{F}_{\text{RF}}$, i.e., the BS uses a different analog beamforming direction for each user. It was observed in [10] that when a larger number of mmWave channel paths is assumed, i.e., $L = 15$, the non-orthogonality among signal spaces of the users makes neither nullifying interference nor treating interference as noise optimal, and the benefits of RSMA are more pronounced. Therefore, we consider $L = 15$ channel paths per user, as studied previously in [9], [10].

TABLE I
System Parameters

Parameter	Value
Number of BS Transmit Antennas (N_{T})	32
Number of BS Transmit RF Chains (N_{RF})	2
Number of Users (K)	2
Number of Channel Paths Per User (L)	15
PA Inefficiency Factor (ψ)	$1/0.38$
Baseband Power Consumption (P_{BB})	200 mW
RF Chain Power Consumption (P_{RF})	160 mW
PA Power Consumption (P_{PA})	40 mW
Phase Shifter Power Consumption (P_{PS})	20 mW

Fig. 2a shows the average EE versus P_{max} for RSMA and SDMA for fixed $R^{\text{th}} = 0.5$ bps/Hz. The label "EE-optimized" denotes the proposed energy-efficient RSMA scheme and corresponding baseline SDMA scheme. The label "SE-optimized" denotes the EE when the SE of the system is maximized by using Algorithm 1 and setting $\psi = 0$ and $P_{\text{cir}} = 1$ so the objective function η_{EE} reduces to R_{sum}. It is clear that the proposed EE-optimized RSMA scheme outperforms SDMA across the entire range of P_{max}. We also see a common trend for EE optimization curves. When transmit power is low, it is energy-efficient to use all available power, and an increase in P_{max} yields an increase in EE. Once an optimal transmit power is reached, i.e., $P_{\text{max}} = 30$ dBm in this case, further increase in transmit power diminishes EE. Beyond this point, the BS transmits below the maximum power constraint for the EE-optimized schemes, i.e., $(P_c + \sum_{k=1}^{K} P_k) < P_{\text{max}}$, and EE plateaus. In contrast, the SE-optimized schemes continue to utilize additional power and EE decreases.

Fig. 2b shows the average SE (i.e., sum rate) versus P_{max} for the same scenario as Fig. 2a. When $P_{\text{max}} > 30$ dBm, the EE-optimized curves and SE-optimized curves diverge, with the EE-optimized curves staying approximately constant and the SE-optimized curves increasing, as expected. Although it is optimized for EE, the proposed RSMA scheme outperforms the SDMA scheme in terms of average R_{sum} across the entire range of P_{max} by a small margin. Therefore, RSMA can outperform SDMA in both EE and SE at the same time.

Fig. 2c shows the average EE versus the QoS threshold rate constraint, R^{th}, for fixed $P_{\text{max}} = 40$ dBm. The proposed EE-optimized RSMA scheme outperforms SDMA across the entire range of R^{th}. We also note that the SDMA scheme was unable to meet the QoS rate constraint for some random channel realizations when $R^{\text{th}} > 1$ bps/Hz, whereas the RSMA scheme met the QoS constraint in these cases.

Another key outcome is that the proposed EE-optimized RSMA scheme always performed at least as well as SDMA in terms of EE for *individual channel realizations*. We de-

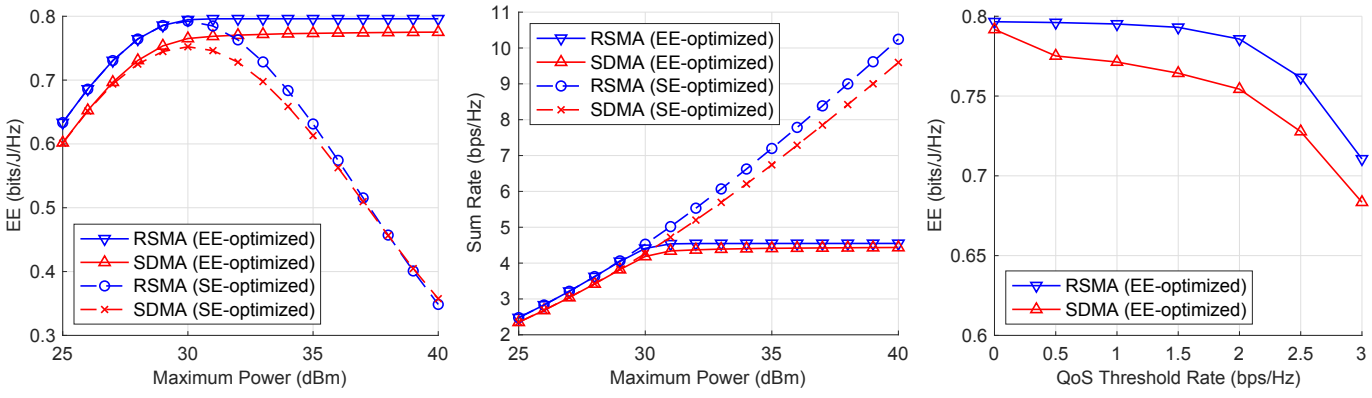

Fig. 2. Simulation results averaged over 500 random channel realizations. (a) EE (η_{EE}) versus normalized maximum power ($P_{\max}$) for $R^{\text{th}} = 0.5$ bps/Hz. (b) SE (R_{sum}) versus normalized maximum power ($P_{\max}$) for $R^{\text{th}} = 0.5$ bps/Hz. (c) EE (η_{EE}) versus QoS threshold rate (R^{th}) for $P_{\max} = 40$ dBm.

fine $\Delta\text{EE} = \eta_{\text{EE}}(\text{RSMA}) - \eta_{\text{EE}}(\text{SDMA})$ to denote the EE difference for a single channel realization. For all cases in which Algorithm 1 generated a feasible result for both RSMA and SDMA, we observe that ΔEE was always non-negative to within a small computational margin (i.e., $|\Delta\text{EE}| < 3\text{e}{-4}$ bits/J/Hz). Furthermore, considering the EE-optimized results from Fig. 2a, a maximum $\Delta\text{EE} = 0.66$ bits/J/Hz (135% increase) was observed for $R^{\text{th}} = 0.5$ bps/Hz and $P_{\max} = 29$ dBm. Thus, RSMA has the potential to produce large gains in EE over SDMA for some channel geometries. However, this result is tempered by the fact that the EE of RSMA and SDMA were nearly equal, e.g., $\Delta\text{EE} < 0.001$ bits/J/Hz, for the majority of the randomly generated channels.

V. Conclusion

We studied, for the first time in a general setting, the EE performance of RSMA in a downlink multi-user mmWave system and proposed a new energy-efficient one-layer RSMA hybrid precoder scheme with QoS by first fixing the analog precoder directions and then applying an SCA-based algorithm adapted from [6], [7]. Simulation results showed the proposed EE-optimized RSMA scheme was always at least as energy-efficient as an SDMA baseline scheme. For the majority of the randomly-generated channels, the EE of RSMA and SDMA were nearly equal. However, for some channel geometries, RSMA was significantly more energy-efficient than SDMA. Our results provide insights into the EE of mmWave RSMA systems and suggest that the well-known EE benefits of RSMA at lower frequencies extend to the mmWave regime as well. In future work, we will consider the impact of imperfect CSIT.

References

[1] *Framework and overall objectives of the future development of IMT for 2030 and beyond*, ITU Rec. ITU-R M.2160-0, Nov. 2023.

[2] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surv. and Tut.*, vol. 24, no. 4, pp. 2073–2126, 4th Quart. 2022.

[3] B. Clerckx, Y. Mao, E. Jorswieck, J. Yuan, D. Love, E. Erkip, and D. Niyato, "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1265–1308, May 2023.

[4] Y. Mao, B. Clerckx, and V. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP J. Wireless Commun. and Netw.*, vol. 2018, no. 133, May 2018.

[5] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 349–353, Mar. 2020.

[6] Y. Mao, B. Clerckx, and V. Li, "Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA," in *15th Int. Symp. on Wireless Commun. Syst.*, 2018.

[7] ——, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8754–8770, Dec. 2019.

[8] A. Mishra, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for 6G - Part I: Principles, applications and future works," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2232–2236, Oct. 2022.

[9] M. Dai and B. Clerckx, "Multiuser millimeter wave beamforming strategies with quantized and statistical CSIT," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7025–7038, Nov. 2017.

[10] Z. Li, S. Yang, and T. Clessienne, "A general rate splitting scheme for hybrid precoding in mmWave systems," in *IEEE ICC*, 2019.

[11] O. Kolawole, A. Papazafeiropoulos, and T. Ratnarajah, "A rate-splitting strategy for multi-user millimeter-wave systems with imperfect CSIT," in *IEEE SPAWC*, 2018.

[12] J. Liu, Y. Guan, Y. Ge, L. Yin, and B. Clerckx, "Energy efficiency of rate-splitting multiple access for multibeam satellite communications," in *IEEE VTC2023-Spring*, 2023.

[13] A. Rahmati, Y. Yapici, N. Rupasinghe, I. Guvenc, H. Dai, and A. Bhuyan, "Energy efficiency of RSMA and NOMA in cellular-connected mmWave UAV networks," in *IEEE ICC Workshops*, 2019.

[14] P. Liu, Y. Li, W. Cheng, X. Dong, and L. Dong, "Active intelligent reflecting surface aided RSMA for millimeter-wave hybrid antenna array," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5287–5302, Sep. 2023.

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

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

[17] B. Matthiesen, Y. Mao, A. Dekorsy, P. Popvski, and B. Clerckx, "Globally optimal spectrum- and energy-efficient beamforming for rate splitting multiple access," *IEEE Trans. Signal Process.*, vol. 70, pp. 5025–5040, Oct. 2022.

[18] J. Löfberg, "YALMIP: A Toolbox for Modeling and Optimization in MATLAB," in *Proc. CACSD*, Taipei, Taiwan, 2004.