System Energy-Efficient Hybrid Beamforming for mmWave Multi-User Systems

Vu Nguyen Ha, *Member, IEEE*, Duy H. N. Nguyen, *Senior Member, IEEE*,
and Jean-François Frigon, *Senior Member, IEEE*

***Abstract*—This paper develops energy-efficient hybrid beamforming designs for mmWave multi-user systems where analog precoding is realized by switches and phase shifters such that radio frequency (RF) chain to transmit antenna connections can be switched off for energy saving. By explicitly considering the effect of each connection on the required power for baseband and RF signal processing, we describe the total power consumption in a sparsity form of the analog precoding matrix. However, these sparsity terms and sparsity-modulus constraints of the analog precoding make the system energy-efficiency maximization problem non-convex and challenging to solve. To tackle this problem, we first transform it into a subtractive-form weighted sum rate and power problem. A compressed sensing-based reweighted quadratic-form relaxation method is employed to deal with the sparsity parts and the sparsity-modulus constraints. We then exploit alternating minimization of the mean-squared error to solve the equivalent problem where the digital precoding vectors and the analog precoding matrix are updated sequentially. The energy efficiency upper bound and a heuristic algorithm are also examined for comparison purposes. Numerical results confirm the superior performances of the proposed algorithm over benchmark energy-efficiency hybrid precoding algorithms and heuristic one.**

***Index Terms*—Hybrid precoding, mmWave, energy efficiency, MIMO, multi-user.**

I. Introduction

RECENTLY, mmWave has been considered as a promising technology for emerging wireless networks to deal with the increasing wireless traffic demands [1]–[4]. Operating in the frequency bands from 30-300 GHz, this technology can empower multi-Gbps transmission speed. Thanks to the band's short wavelength, a large number of antenna elements can be leveraged in a small space at the transceivers. Hence, multiple data streams for multiple users can be transmitted via spatial multiplexing which potentially results in a significant improvement in spectral efficiency [5], [6].

Employing one single RF chain for each antenna as in the conventional fully digital precoder design typically requires high implementation cost and complexity [7]. Thus, hybrid precoding (HP) has been proposed as a cost-efficient beamforming technique for the mmWave system [8]–[15]. This proposed transceiver architecture adds the analog precoder (AP) to the conventional digital precoder (DP); hence, the number of RF chains can be reduced to save CAPEX (capital expenditures, due to implementation complexity) and OPEX (operational expenditures, electrical energy costs) [6], [11]–[16]. Additional components typically consist of analog phase shifters connecting RF chains to antennas to achieve analog beamforming gain. Interestingly, HP enables near-optimal performance thanks to the low-rank characteristics of mmWave channels [12]. However, the design of mmWave transceivers still raises concerns on the system energy efficiency (SEE), which is an important aspect of mmWave systems [17]–[19]. Besides the conventional achievable rate and transmit power trade-off, the study in HP implementation for mmWave system should cover the impact of the design structure on the SEE.

There are two main HP structures, named fully-connected and partially-connected [19]. The first structure activates all the phase shifters between RF chains and antennas while only a subset of phase shifters is activated for analog signal processing in the second structure. Many studies on these two structures have been reported that turning on larger set of phase shifters can achieve the high capacity by enhancing more degree-of-freedom; however, employing the large number of phase shifters in AP component may result in high prohibitive power dissipation. In addition, turning off a number of phase shifters can also lessen the hardware complexity by lowering the number of RF paths since each RF chain is allowed to connect to a subset of all antennas, which hence can ease the implementation and drop down the consumed energy [20]. Therefore, the design of efficient HP taking care of optimizing the subset of phase shifters for SEE maximization in mmWave multiple-input and multiple-output (MIMO) systems is an interesting and challenging problem, which is the focus of this paper.

A. Related Works

While research on HP for mmWave systems is plentiful, limited work has studied maximizing the SEE. Several papers

Manuscript received January 25, 2019; revised October 14, 2019; accepted May 9, 2020. Date of publication May 18, 2020; date of current version November 20, 2020. This work was supported in part by the National Sciences and Engineering Research Council of Canada under Grant RGPIN-2016-06401; in part by the Québec's Merit Scholarship Program for Foreign Students from Ministère de l'Éducation, de l'Enseignement Supérieur et de la Recherche du Québec under Grant FQRNT-PBEEE-2018; and in part by the Startup Fund from San Diego State University. The associate editor coordinating the review of this article and approving it for publication was A. Tulino. *(Corresponding author: Vu Nguyen Ha.)*

Vu Nguyen Ha and Jean-François Frigon are with the Poly-Grames Research Center, École Polytechnique de Montréal, Montreal, QC H3T 1J4, Canada (e-mail: vu.ha-nguyen@polymtl.ca; j-f.frigon@polymtl.ca).

Duy H. N. Nguyen is with the Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA 92182 USA (e-mail: duy.nguyen@sdsu.edu).

Digital Object Identifier 10.1109/TGCN.2020.2995227

have considered the energy-efficient HP designs for mmWave systems or massive MIMO systems without optimizing the SEE, such as [21], [22]. In particular, an energy-efficient HP design for mmWave MIMO systems is proposed in [21] based on the successive interference cancellation method. A hybrid analog-digital architecture for HP in mmWave MIMO systems is proposed in [22], where multiple sub-arrays are employed at the transmit and receive antennas. Both works investigate SEE achieved by the proposed HP designs. HP design maximizing the SEE is directly studied in [23]–[25]. Specifically, [23] develops a novel HP design to maximize the SEE of massive MIMO system. In this work, the SEE is formulated as the ratio between the achievable rate and the total power consumption which is the sum of transmit power and constant power dissipation/consumption terms. The upper-bound fully digital precoding (FDP) is first optimized, based on which the HP is reconstructed by minimizing the Euclidean distance between two precoding designs. The numbers of RF chains and antennas are then optimized based on the statistical analysis values of SEE for very large array antenna systems. Optimizing the number of RF chains is also considered in [24] to maximize the SEE of a HP mmWave system. In this work, the power radiation is formulated as a linear function of the number of RF chains, then, an efficient codebook-based hybrid precoding design is proposed by jointly selecting the AP in a codebook set and optimizing the baseband ones. Using a different method, Gao *et al.* [25] studied the SEE HP for mmWave massive MIMO system by employing machine learning tools. In particular, a hardware-efficient analog network structure has been developed in this work where a new group-connected mapping strategy for HP is introduced. Among predetermined activated phase-shifter groups corresponding to different hardware implementations, the most efficient group together with the corresponding HP is then selected.

To the best of our knowledge, the SEE HP design for mmWave multi-user system which optimizes all the numbers of utilized RF chains, transmission antennas, and the connections among the RF chains and the antennas has not been studied in the literature. Filling this gap, this paper studies a novel HP structure for mmWave MIMO system where each of the connections between RF chains and antennas can be optimally activated or deactivated by utilizing ON-OFF switches to maximize the SEE.

B. Research Contributions

We consider a fully-connected HP structure where a switch is integrated with a phase shifter over every RF chain to antenna connection. The switches can be activated or deactivated to reduce the power dissipation of the phase shifters. In addition, an RF chain (or/and an antenna) can be deactivated for energy saving if all of its corresponding connections are turned off. This hardware structure is transferred into sparsity-modulus constraints for AP matrix design. In particular, the absolute value of an AP matrix element can be one or zero. The sparsity-modulus design also enables us to exploit the total power consumption as a sparsity function of AP matrix. The SEE is then calculated as the ratio between the achievable rate and total consumption power based on which the SEE maximization (SEEM) problem is formulated as a non-convex problem. In preliminary work reported in [42], we proposed an iterative algorithm to tackle this optimization problem without a detailed proof of convergence. Performance comparison between the proposed algorithm and heuristic algorithms was also not given. In solving the SEEM problem and exposing energy-efficient HP designs, the contributions of our paper are given as follows:

- To tackle the SEEM problem, we first transform it into a subtractive-form weighted sum rate and power (WSRP) problem based on the *"Dinkelbach's method"* [29]. Then, we exploit an alternating minimization of the mean-squared error (MMSE) algorithm to solve the WSRP problem where the DP vectors and AP matrix are updated alternatively. In each iteration, we employ compressed sensing-based relaxation method to deal with the sparsity-modulus constraints and sparse formulation of total power consumption. In particular, this method helps to transform the MMSE problem into a convex one, for which a locally optimal solution can be found efficiently. The analysis on the convergence of the proposed algorithm is also given.
- The SEE upper bound and an heuristic algorithm are also studied for comparison purpose. Extensive numerical studies are conducted where we examine the convergence and efficiency of the proposed algorithms as well as the impacts of different system parameters on the SEE.

The remaining of this paper is organized as follows. We describe the system model, and formulations of the SEEM HP design problem in Section II. In Section III, we transfer the SEEM problem into the WSRP problem and point out the general design algorithm. The compress-sensing-based method is then proposed in Section IV to solve the WSRP problem based on which we developed the novel HP design algorithm to maximize the SEE. The upper bound of SEE and an heuristic algorithm are presented in Section V. Numerical results are presented in Section VI followed by conclusions in Section VII.

Notations: $(\mathbf{X})^T$ and $(\mathbf{X})^H$ denote the transpose and conjugate transpose of the matrix $\mathbf{X}$, respectively; $\|\mathbf{x}\|_0$ and $\|\mathbf{x}\|$ denote the norm-0 and Euclidean norm of a vector $\mathbf{x}$, respectively.

II. System Model

A. Multi-User Hybrid Precoding System Model

Consider a downlink mmWave multi-user HP system where a base station (BS) equipped with N_{T} antennas and N_{RF} RF chains serves *K* remote single-antenna users. Utilizing the HP, the BS first applies the DP vectors to the corresponding symbol sequences for the users. Specifically, a DP vector $\mathbf{w}_k \in \mathbb{C}^{N_{\mathrm{RF}}\times 1}$ is applied to the data symbol $s_k \in \mathbb{C}$, intended for user *k*. Without loss of generality, we assume $|s_k| = 1$. Following the digitally precoded sequences, the BS then employs an AP matrix, $\mathbf{A} \in \mathbb{C}^{N_{\mathrm{T}}\times N_{\mathrm{RF}}}$, to map the RF signals from N_{RF} RF chains to N_{T} antennas. Let a_t^n be the

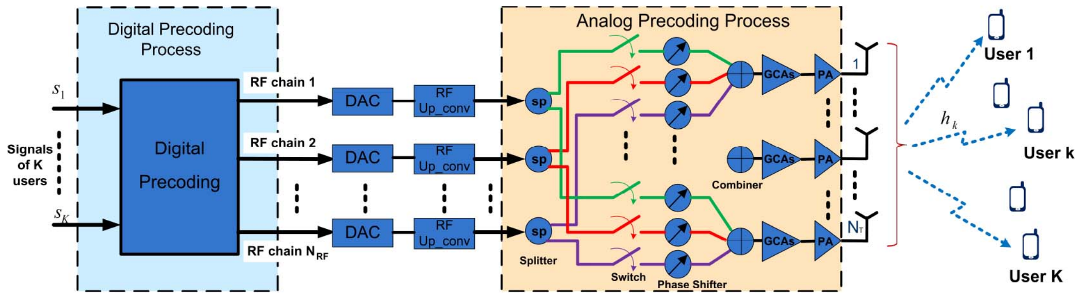

Fig. 1. Diagram of a mmWave multi-user system with hybrid analog/digital precoding design.

element allocated on the t^{th} row and the n^{th} column of $\mathbf{A}$ and x_n be the signal on RF chain n which can be determined as $x_n = \sum_{\forall k} w_k^n s_k$ where w_k^n is the n^{th} element of DP vector $\mathbf{w}_k$. Then, the signal of RF chain n at antenna t after going through the AP block can be written as $x_n a_t^n$. In this work, we consider dynamic fully-connected RF chains to antennas structure in which $\mathbf{A}$ is implemented by integrating switches and the phase shifters. The ON-OFF switch deployed on each RF chain to antenna connection can allow (or disallow) the corresponding RF signal be forwarded (or not be forwarded) to that antenna for transmission. When the connection between RF chain *n* and antenna *t* is deactivated (turned off), we can set $|a_t^n|^2 = 0$. Inversely, this connection is activated (turned on), the corresponding RF signal will be phase shifted, combined to others, and transmitted by antenna *t*. In this case, $|a_t^n|^2$ should be 1, and the phase-shifted version of x_n will be $x_n e^{j\theta_t^n}$ where $a_t^n = e^{j\theta_t^n}$. Hence, the following sparsity-modulus condition for the elements of $\mathbf{A}$ can help the AP matrix mathematically present well the implementation of switches and phase shifters in our system.

$$\left|a_t^n\right|^2 = 1 \text{ or } 0 \quad \forall (t,n). \tag{1}$$

By taking into account the HP design for multi-user system in [11], the signal received by user *k* can be given as

$$y_k = \mathbf{h}_k^H \sum_{\forall j} \mathbf{A}\mathbf{w}_j s_j + n_k^{\text{noise}}, \tag{2}$$

where n_k^{noise} is the additive Gaussian noise at user *k* and $\mathbf{h}_k \in \mathbb{C}^{N_{\text{T}}}$ is the multiple-input and single-output (MISO) channel from the BS to user *k*. Assuming coherent detection at the users, the signal-to-interference-plus-noise ratio (SINR) at user *k* can be given as

$$\text{SINR}_k = \frac{\left|\mathbf{h}_k^H \mathbf{A}\mathbf{w}_k\right|^2}{\sum_{j \neq k} \left|\mathbf{h}_k^H \mathbf{A}\mathbf{w}_j\right|^2 + \sigma^2}, \tag{3}$$

where σ^2 is the power of additive Gaussian noise. Assuming Gaussian signaling between the BS and the users, the total achievable data-rate of the system can be described as

$$R(\mathbf{W}, \mathbf{A}) = \sum_{\forall k} \log(1 + \text{SINR}_k), \tag{4}$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ is denoted as the matrix generated by all DP vectors.

Remark 1: The ON-OFF switches employed in the mmWave transmitter allow a selection of termination (inactive port) or pass-through (active port) for the RF chain signals [43]. Each switch is matched to an output of one splitter which is implemented to divide an RF chain signal to N_{T} antennas as shown in Fig. 1. Once the active port is selected, the signal is passed through, which is indicated as ON state. On another hand, when an inactive port is selected by the switch, the OFF state occurs. In the considered system, a good match is assumed, which means that a matched termination is implemented at the inactive port and the inactive signal is terminated by 50-Ohm load [43].

B. Power Consumption Model

In this section, the power consumption model is analyzed by counting the required power of each system component. In general, the total power consumption in the system is comprised of the power consumed by the digital signal processing (DSP) hardware, the RF signal processing hardware, and the RF signal radiation [28].

1) DSP Power Consumption: For the DSP component, the static power consumption corresponding to each user's signal is due to parts of the baseband signal process. In this paper, we assume that the power consumption for the DSP hardware is unchanged, which is given by

$$P_{\text{DP}} = K P_{\text{BB}}, \tag{5}$$

where P_{BB} represents the power consumption for the baseband signal processing of one user.

2) Power Consumption by RF Signal Processing Hardware: As illustrated in Fig. 1, the baseband signals are first converted to an analog signals and up-converted to RF band. Then, an $(N_{\text{T}} + 1)$-port splitter (so called divider) is implemented over each RF chain to divide the RF signal to N_{T} outputs corresponding to N_{T} antennas. Each of these N_{T} outputs will be matched to an ON-OFF switch. Here, the signal will be terminated or passed to the phase-shifter accordingly if $|a_t^n|^2 = 0$ or 1, respectively. All signals heading to one antenna are then combined by employing an $(N_{\text{RF}} + 1)$-port combiner. The combined signal is passed through the gain-compensation amplifier (GCA) and the power amplifier (PA) before being propagated by that very antenna. Denote P_{DAC}, P_{RFC}, P_{SW} and P_{PS} as the power consumption of the digital-to-analog converter (ADC), RF converter, ON-OFF switch, and phase shifter, respectively. In practice, the power dissipated by each of these components can be assumed to be unchanged [26]. However, the power consumed by the amplifiers varies due to the loss caused by the splitters as well as combiners [44], and due to the transmission power.

a) Power Consumption by GCAs: In this system, a number of GCAs are deployed to boost the RF signals, which lost part of their power after passing through splitters, switches, phase shifters and combiners; and to attain sufficient power for driving the PAs at the antennas. To estimate the GCA's power consumption, we revisit the power loss of each connection between RF chains and antennas which is caused by splitters and combiners. Assume that a multi-port splitter (or combiner) is implemented based on a cascade of simple three-port splitters (or combiners) as in [44]. Then, the losses of the outputs of $(N_{\mathsf{T}}+1)$-port splitter and $(N_{\mathsf{RF}}+1)$-port combiner can be estimated as $\lceil \log_2(N_{\mathsf{T}})\rceil L_{\mathsf{d}}$ and $\lceil \log_2(N_{\mathsf{RF}})\rceil L_{\mathsf{c}}$ [44] where $\lceil . \rceil$ stands for ceiling function and L_{d} and L_{c} (in dB) represent the power losses of a three-port splitter and a three-port combiner, respectively. In addition, denote L_{sw} and L_{ps} (in dB) as the losses when RF signal passes through the switch and phase shifter. Let G_{amp} (in dB) be the maximum amplification gain of one GCA. Then, the number of GCAs required for compensating signal before going to a PA can be given as

$$M_{\mathrm{GCA}} = \left\lceil \frac{\lceil \log_2(N_{\mathsf{T}})\rceil L_{\mathsf{d}} + \lceil \log_2(N_{\mathsf{RF}})\rceil L_{\mathsf{c}} + L_{\mathsf{sw}} + L_{\mathsf{ps}}}{G_{\mathsf{amp}}} \right\rceil, \tag{6}$$

Then, the power consumption of the GCAs corresponding to one activated antenna is given by

$$P_{\mathrm{GCA}} = M_{\mathrm{GCA}} P_{\mathrm{amp}}, \tag{7}$$

where P_{amp} represents the power consumption of one GCA.

b) Power Consumption by PAs: The power consumption of PAs can be modeled in a linear form of the radiation power [44], [45]. In particular, the power consumption by the PA implemented at antenna *t*, named $P_{\mathrm{PA},t}$, can be calculated as

$$P_{\mathrm{PA},t} = P_t/\rho_{\mathrm{pa}}, \tag{8}$$

where P_t represents the transmission power at antenna *t* and ρ_{pa} stands for the power amplifier efficiency [45].

c) Power Consumption of RF Signal Processing Hardware: An RF-chain-to-antenna connection is activated or deactivated by the ON-OFF switch allocated on the corresponding connecting link. An activated connection would consume certain amount of power for the RF chain and the antenna connected to it, as illustrated in Fig. 1. Furthermore, one RF chain can be turned off for power saving if there is no connection from that RF chain to any antenna. Likewise, an antenna may be inactive when all connections from RF chains to that antenna are turned off. It is recalled that $|a_t^n|^2 = 1$ implies an active connection between RF chain *n* and antenna *t* and $|a_t^n|^2 = 0$ implies otherwise. Hence, we can express the total power consumption for RF signal processing in a sparse form of $\mathbf{A}$ as follows:

$$\begin{aligned} P_{\mathrm{RF}} = {} & (P_{\mathrm{DAC}} + P_{\mathrm{RFC}})\|\mathbf{c}(\mathbf{A})\|_0 + N_{\mathrm{T}} N_{\mathrm{RF}} P_{\mathrm{SW}} \\ & + P_{\mathrm{PS}}\|\mathbf{A}\|_0 + P_{\mathrm{GCA}}\|\mathbf{r}(\mathbf{A})\|_0 + 1/\rho_{\mathrm{pa}} \sum_{\forall t} P_t, \end{aligned} \tag{9}$$

where $\mathbf{c}(\mathbf{A}) = [c_1, \ldots, c_{N_{\mathrm{RF}}}]^T \in \mathbb{C}^{N_{\mathrm{RF}}}$ and $c_n = \sum_{\forall t} |a_t^n|^2$, $\mathbf{r}(\mathbf{A}) = [r_1, \ldots, r_{N_{\mathrm{T}}}]^T \in \mathbb{C}^{N_{\mathrm{T}}}$ and $r_t = \sum_{\forall n} |a_t^n|^2$.

3) Transmission Power and Total Power Consumption: Next, the transmission power can be calculated based on $(\mathbf{W}, \mathbf{A})$ as follows:

$$P_{\mathrm{T}}(\mathbf{W}, \mathbf{A}) = \sum_{\forall k} \mathbf{w}_k^H \mathbf{A}^H \mathbf{A} \mathbf{w}_k. \tag{10}$$

Then, total power consumption can be calculated by taking the summation of P_{DP}, $P_{\mathrm{RF}}(\mathbf{A})$, and P_{T}. In addition, the last component in (9) represents the total transmission power over all antennas, which yields $\sum_{\forall t} P_t = \sum_{\forall k} \mathbf{w}_k^H \mathbf{A}^H \mathbf{A} \mathbf{w}_k$. Hence, the system power consumption can be described as

$$\begin{aligned} P_{\mathrm{tot}}(\mathbf{W}, \mathbf{A}) &= P_{\mathrm{DP}} + N_{\mathrm{T}} N_{\mathrm{RF}} P_{\mathrm{SW}} \\ &\quad + (P_{\mathrm{DAC}} + P_{\mathrm{RFC}})\|\mathbf{c}(\mathbf{A})\|_0 \\ &\quad + P_{\mathrm{PS}}\|\mathbf{A}\|_0 + P_{\mathrm{GCA}}\|\mathbf{r}(\mathbf{A})\|_0 \\ &\quad + \left(1 + \frac{1}{\rho_{\mathrm{pa}}}\right) \sum_{\forall k} \mathbf{w}_k^H \mathbf{A}^H \mathbf{A} \mathbf{w}_k \\ &= P_{\mathrm{cons}} + P_{\mathrm{spr}}(\mathbf{A}) + \left(1 + \frac{1}{\rho_{\mathrm{pa}}}\right) P_{\mathrm{T}}(\mathbf{W}, \mathbf{A}). \end{aligned} \tag{11}$$

where $P_{\mathrm{cons}} = P_{\mathrm{DP}} + N_{\mathrm{T}} N_{\mathrm{RF}} P_{\mathrm{SW}}$ and $P_{\mathrm{spr}}(\mathbf{A}) = (P_{\mathrm{DAC}} + P_{\mathrm{RFC}})\|\mathbf{c}(\mathbf{A})\|_0 + P_{\mathrm{PS}}\|\mathbf{A}\|_0 + P_{\mathrm{GCA}}\|\mathbf{r}(\mathbf{A})\|_0$.

C. Problem Formulation

We now ready to define the SEE (in bits/Hz/W) as the ratio of the achievable sum-rate to the total power consumption as follows.

$$\eta(\mathbf{W}, \mathbf{A}) = \frac{R(\mathbf{W}, \mathbf{A})}{P_{\mathrm{tot}}(\mathbf{W}, \mathbf{A})}. \tag{12}$$

In this paper, we are interested in jointly optimizing the DP vectors and the sparsity-modulus AP matrix to maximize the SEE under the constraint on the transmit power budget at each antenna. This SEEM problem can be stated as

$$\max_{\mathbf{W},\mathbf{A}} \ \eta(\mathbf{W}, \mathbf{A}) = \frac{\sum_{\forall k} \log(1 + \mathrm{SINR}_k)}{P_{\mathrm{cons}} + P_{\mathrm{spr}}(\mathbf{A}) + \left(1 + \frac{1}{\rho_{\mathrm{pa}}}\right) P_{\mathrm{T}}(\mathbf{W}, \mathbf{A})} \tag{13a}$$

$$\text{s.t.} \ |a_t^n| = 1 \ \text{or} \ 0, \forall (t, n), \tag{13b}$$

$$\sum_{\forall k} \mathbf{w}_k^H \mathbf{a}_t \mathbf{a}_t^H \mathbf{w}_k \le P_t^{\max}, 1 \le t \le N_{\mathrm{T}}, \tag{13c}$$

where $\mathbf{a}_t \in \mathbb{C}^{N_{\mathrm{RF}} \times 1}$ is denoted as the vector corresponding to the t^{th} column of $\mathbf{A}^H$, and $P_t^{\max}$ is transmit power limit at antenna *t* of the BS. The challenges for solving problem (13) come from the fractional form of the objective function and the sparsity terms associated with the power consumption model. To overcome these challenges, we first apply Dinkelbach's method [29] for solving problem (13) by transforming it into a sequence of parameterized subtractive-form problems. Then, compressed sensing based solution approaches [36] are employed to dual with the sparsity terms in the parameterized problems. In addition, MMSE-based transformation [37]

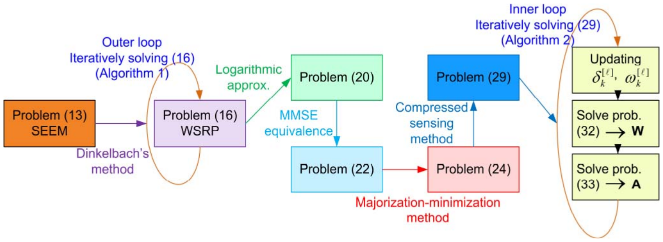

Fig. 2. Diagram of solution approach.

Algorithm 1 Overview of the Proposed Algorithm

```
Initialize $\eta^{(0)} = 0$, set $m = 0$, and choose predetermined tolerate $\tau^{\text{out}}$.
repeat
    Solve (16) with $\eta^{(m)}$ to achieve $(\mathbf{W}^{(m)}, \mathbf{A}^{(m)})$.
    Set $\eta^{(m+1)} = \frac{R(\mathbf{W}^{(m)},\mathbf{A}^{(m)})}{P_{\text{tot}}(\mathbf{W}^{(m)},\mathbf{A}^{(m)})}$.
    Update $m := m + 1$.
until $|\eta^{(m)} - \eta^{(m-1)}| \leq \tau^{\text{out}}$.
Return $(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)})$.
```

and majorization-minimization based method [46] are also enhanced for proposing a framework solving problem (13). The overall solution approach is summarized in Fig. 2, in which each of stages will be given in the subsequent sections.

III. General Design Algorithm Based on Dinkelbach's Method

A. Transformation of SEE Maximization Problem

The objective (13a) is in a fractional form, which is difficult to tackle directly. Dinkelbach in [29] has proposed an efficient method for solving optimization problem with this type of objective function, which is well-known as the Dinkelbach algorithm. Specifically, the method first transforms the fractional problem into a parameterized subtractive form. An iterative solution approach to update the parameter of subtractive-form problem is then invoked to obtain its optimal solution. The foundation of this method can be given in the following theorem which summarizes the theoretical results in [29].

Theorem 1: Consider two following problems

$$(\mathcal{P}_I) \ \max_{\mathbf{x}} \ R(\mathbf{x})/P(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S}, \tag{14}$$

$$(\mathcal{P}_{II}^{\eta}) \ \max_{\mathbf{x}} \ R(\mathbf{x}) - \eta P(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S}. \tag{15}$$

which can represent any arbitrary fractional and subtractive forms, respectively. Let $\chi(\eta)$ be the optimal objective value of (15) for given η, which can be considered as a function of η. Assume that $P(\mathbf{x}) > 0 \ \forall \mathbf{x} \in \mathcal{S}$ and η^* is the optimal objective value of $(\mathcal{P}_I)$. Then, there are four observations as follows.

i) $\chi(\eta)$ is a strictly monotonic decreasing function.
ii) $\chi(\eta) > 0$ if and only if $\eta < \eta^*$.
iii) $\chi(\eta) < 0$ if and only if $\eta > \eta^*$.
iv) $(\mathcal{P}_I)$ and $(\mathcal{P}_{II}^{\eta^*})$ have the same set of optimal solutions.

Proof: The proof is given in Appendix A. ■

Theorem 1 has provided the foundation, based on which the transformation of SEEM problem can be performed and an algorithmic solution approach can be devised to solve the SEEM problem. First, let us consider the weighted power (WSRP) maximization problem for given value of η which is stated as follows.

$$\begin{aligned} \max_{(\mathbf{W},\mathbf{A})} \quad & \bar{\chi}(\eta, \mathbf{W}, \mathbf{A}) = R(\mathbf{W}, \mathbf{A}) - \eta P_{\text{tot}}(\mathbf{W}, \mathbf{A}) \\ \text{s.t.} \quad & (\mathbf{W}, \mathbf{A}) \in \Theta, \end{aligned} \tag{16}$$

where Θ stand for the feasible set of $(\mathbf{W}, \mathbf{A})$ in SEEM problem. As can be seen, parameter η in the WSRP problem acts as a negative weight on the total energy consumption; hence, it can be considered viewed as the "price" of the system's energy consumption which can be adjusted to meet the efficient point of our design. We also denote $\eta^\star$ as the maximum SEE which can be expressed as

$$\eta^\star = \eta(\mathbf{W}^\star, \mathbf{A}^\star) = \frac{R(\mathbf{W}^\star, \mathbf{A}^\star)}{P_{\text{tot}}(\mathbf{W}^\star, \mathbf{A}^\star)} = \max_{(\mathbf{W},\mathbf{A}) \in \Theta} \frac{R(\mathbf{W}, \mathbf{A})}{P_{\text{tot}}(\mathbf{W}, \mathbf{A})}, \tag{17}$$

where $(\mathbf{W}^\star, \mathbf{A}^\star)$ represents the optimal solution. Then, the observation (iv) in Theorem 1 yields that SEEM problem and WSRP with $\eta^\star$ have the same set of optimal solutions. Hence, SEEM problem can be solved by iteratively solving WSRP problem for a certain value of η and adjusting η until an optimal $\eta^\star \geq 0$ satisfying $\chi(\eta^\star) = 0$ is found.

B. Overview of the Solution Approach

In this section, we present an overview of the proposed solution approach which is summarized in Algorithm 1. The algorithm relies on updating η and solving the corresponding WSRP problem iteratively based on the well-known Dinkelbach-type solution method [29], [30]. We start by setting $\eta^{(0)} = 0$. In iteration *m*, corresponding to a certain value $\eta^{(m)}$, the WSRP problem is solved to achieve the optimal value $(\mathbf{W}^{(m)}, \mathbf{A}^{(m)})$. Then, the value of η is updated for the next iteration as $\eta^{(m+1)} = \frac{R(\mathbf{W}^{(m)},\mathbf{A}^{(m)})}{P_{\text{tot}}(\mathbf{W}^{(m)},\mathbf{A}^{(m)})}$. According to the proof given in [29], [30], this process ensures the monotonic increase of $\eta^{(m)}$ if the WSRP problem is solved optimally in each iteration, which yields the convergence of Algorithm 1. However, obtaining such an optimal solution is challenging due to the non-convexity of WSRP problem. The following theorem establishes the convergence of Algorithm 1 when only local optimal solutions are found at each iteration.

Theorem 2: For a given $\eta^{(m)}$ in the *m*-th iteration of Algorithm 1, if a local optimal solution of WSRP problem, $(\mathbf{W}^{(m)}, \mathbf{A}^{(m)})$, is found such that

$$\bar{\chi}\left(\eta^{(m)}, \mathbf{W}^{(m)}, \mathbf{A}^{(m)}\right) \geq \bar{\chi}\left(\eta^{(m)}, \mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)}\right), \tag{18}$$

Algorithm 1 will converge after a finite number of iterations.

Proof: The proof is given in Appendix B. ■

The next task is to find an efficient method to solve the WSRP problem. This problem is NP-hard as a result of the non-convex sum rate, the ℓ_0-norm of matrix $\mathbf{A}$ in the objective function, and the sparsity-modulus constraint (13b). Hence, finding its globally optimal solution is prohibitively complex.

Theorem 2 encourages us to develop an efficient (probably sub-optimal) solution to indicate a local optimal solution of the WSRP problem in each iteration of Algorithm 1.

Remark 2: It is worth noting that the bisection searching approaching can be applied for updating $\eta^{(m)}$ in each iteration of Algorithm 1 instead of utilizing **Step 4** since $\chi(\eta)$ is monotonic decreasing function respect to η. However, bisection searching method, while being simple in the context, is quite complex the implementation, since there is no exact method of finding the initial upper and lower bounds. In addition, existing studies [31], [32] have demonstrated that updating $\eta^{(m)}$ as in **Step 4** based on Dinkelbach method can converge faster than bisection approaches in many circumstances.

IV. Energy-Efficiency Hybrid Precoding Design

This section presents an efficient solution approach to the WSRP problem via compressed-sensing (CS) based methods. In particular, the sparsity terms in P_{spr} are first transformed into the approximated continuous logarithmic forms. Then, the CS method is employed to relax the sparsity-modulus constraints of variables a_t^n's, elements of $\mathbf{A}$. To do so, the WSRP problem containing sparsity can be relaxed to non-sparsity problem which can be solved efficiently by iteratively dealing with number of quadratically constrained quadratic programming problems. The details are given as follows.

A. Sparsity Relaxation and MMSE-Based Transformation

It is worth to note that the ℓ_0-norms of matrix $\mathbf{A}$, $\mathbf{r}(\mathbf{A})$, and $\mathbf{c}(\mathbf{A})$ can be defined directly from $\|a_t^n\|_0$. Interestingly, the sparsity solutions can be obtained by employing the re-weighted ℓ_1-norm minimization methods, originally proposed to enhance the data acquisition in compressed sensing. In particular, for a given non-negative real vector $\mathbf{v} \in \mathbb{R}_+^M$, its ℓ_0-norm can be approximated as

$$\|\mathbf{v}\|_0 = \lim_{\varepsilon \to 0} \sum \frac{\log\left(1 + |v_i|\varepsilon^{-1}\right)}{\log(1 + \varepsilon^{-1})}, \tag{19}$$

where $\varepsilon \ll 1$. Then, the WSRP problem can be approximated to the following problem

$$\begin{aligned}\max_{\mathbf{A},\mathbf{W}} \ l(\mathbf{W},\mathbf{A}) = \ & \kappa \sum_{\forall k} \log(1 + \text{SINR}_k) \\ & - \eta\left(1 + \frac{1}{\rho_{\text{pa}}}\right)\mathbf{w}_k^H \mathbf{A}^H \mathbf{A} \mathbf{w}_k \\ & - \eta\left[\kappa P_{\text{cons}} + P_{\text{PS}} \sum_{\forall (t,n)} \log\left(|a_t^n|^2 + \varepsilon\right)\right. \\ & \quad + P_{\text{GCA}} \sum_{\forall n} \log(c_n + \varepsilon) \\ & \quad \left. + (P_{\text{DAC}} + P_{\text{RFC}}) \sum_{\forall t} \log(r_t + \varepsilon)\right] \\ \text{s.t.} \ \ & \text{constraints (13b) and (13c).} \end{aligned} \tag{20}$$

where $\kappa = \log(1 + \varepsilon^{-1})$. To simplify (20), we denote $P_{\text{Log}}(\mathbf{W},\mathbf{A}) = \sum_{\forall X} P_X \log(X + \varepsilon)$, where X stands for $\{|a_t^n|^2\}$'s, $\{c_n\}$'s, and $\{r_t\}$'s, and P_X represents to P_{PS}, P_{GCA}, and $(P_{\text{DAC}} + P_{\text{RFC}})$. Then, we can address the non-convex problem (20) by relating it to a weighted sum-mean square error (MSE) minimization problem as mentioned in the following theorem.

Theorem 3: Problem (20) is equivalent to the following weighted sum-MSE minimization problem, i.e., two problems have same optimal solutions,

$$\begin{aligned}\min_{\mathbf{A},\mathbf{W},\{\delta_k,\omega_k\}} \ g(\mathbf{W},\mathbf{A},\delta_k,\omega_k) = \ & \kappa \sum_{\forall k} (\omega_k e_k - \log \omega_k - 1) \\ & + \eta\left[\kappa\left(1 + \frac{1}{\rho_{\text{pa}}}\right) \sum_{\forall k} \mathbf{w}_k^H \mathbf{A}^H \mathbf{A} \mathbf{w}_k\right. \\ & \quad \left. + \kappa P_{\text{cons}} + P_{\text{Log}}(\mathbf{W},\mathbf{A})\right] \\ \text{s.t.} \ \ & \text{constraints (13b) and (13c).} \end{aligned} \tag{21}$$

where $e_k = \mathbb{E}[|s_k - \delta_k y_k|^2]$, ω_k and δ_k represent the MSE weight and the receive coefficient for user *k*, respectively.

Proof: The proof is given in Appendix C. ■

B. MMSE-Based Hybrid Precoding Design

1) Majorization-Minimization Based Method: As can be observed, $P_{\text{Log}}(\mathbf{W},\mathbf{A})$ term in objective function of problem (21) is the summation of concave functions which makes this minimization problem hard to be tackled. To overcome this challenge, one can employ the well-known majorization-minimization solution approach which implements an iteration algorithm to solve the optimization problem [46]–[50]. The principle of this algorithm consists of the following processes in each iteration: (i) approximating the objective function to tackle-able forms at a predetermined fixed point; (ii) solving the approximate problems; and (iii) updating the fixed point for the next iteration based on the recent achieved optimal solution. Employing this approach to solve problem (21), let $X^{(\ell)}$ represent a fixed value of X of the ℓ-th iteration. Then, the concavity of $\log(X + \varepsilon)$ function can be majored by the following result at $X^{(\ell)}$:

$$\log(X + \varepsilon) \le \log\left(X^{(\ell)} + \varepsilon\right) + \frac{1}{X^{(\ell)} + \varepsilon}\left(X - X^{(\ell)}\right), \tag{22}$$

This result implies that for given value of $X^{(\ell)}$, the right hand side (RHS) of (22) can be employed as a majoring function for $\log(X + \varepsilon)$. Then, in the ℓ-th iteration, the majorization-minimization method [46]–[49] focuses on minimizing the approximate problem of (21) which is formed by replacing $\log(X + \varepsilon)$ terms by the RHS of (22) [46]–[48]. Specifically, an iterative solution can be employed where problem (21) can be approximated to the following problem by replacing (22) into (21) at iteration $(\ell + 1)$.

$$\begin{aligned}\min_{\mathbf{A},\mathbf{W},\{\delta_k,\omega_k\}} \ \tilde{g}(\mathbf{W},\mathbf{A},\delta_k,\omega_k) = \ & \kappa \sum_{\forall k} (\omega_k e_k - \log \omega_k - 1) \\ & + \eta\left[\kappa P_{\text{cons}} + \sum_{t=1}^{N_{\text{T}}} \mathbf{a}_t^H \mathbf{D}_t^{[\ell]} \mathbf{a}_t + \frac{\kappa\rho_{\text{pa}} + \kappa}{\rho_{\text{pa}}}\right. \end{aligned}$$

$$\times \sum_{t=1}^{N_{\mathrm{T}}} \mathbf{a}_t^H \left(\sum_{\forall k} \mathbf{w}_k \mathbf{w}_k^H\right) \mathbf{a}_t \Bigg]$$
$$\text{s.t. constraints (13b) and (13c)}. \tag{23}$$

In (23), $\mathbf{D}_t$ is calculated based on the outcomes of the previous iteration as

$$\mathbf{D}_t^{[\ell]} = P_{\mathrm{PS}} \Psi_t^{[\ell]} + P_{\mathrm{GCA}} \varphi_t^{[\ell]} \mathbf{I} + (P_{\mathrm{DAC}} + P_{\mathrm{RFC}}) \mathbf{F}_t \sum_{\forall n} \phi_n^{[\ell]}, \tag{24}$$

where $\Psi_t^{[\ell]} = diag(\psi_t^{1,[\ell]}, \ldots, \psi_t^{N_{\mathrm{RF}},[\ell]})$, $\mathbf{F}_t = diag\ (0_{1\times(t-1)}, 1\ , 0_{1\times(N_{\mathrm{RF}}-t)})$, and $\mathbf{I}$ is the identify matrix with the size of $N_{\mathrm{RF}} \times N_{\mathrm{RF}}$. In (24), $\psi_t^{[\ell]}$, $\phi_n^{[\ell]}$, and $\varphi_t^{[\ell]}$ are the weighting factors corresponding to $|a_t^{n,[\ell]}|^2$, $c_n^{[\ell]}$, and $r_t^{[\ell]}$ achieved from ℓ-th iteration, which are updated based on (22) as

$$\psi_t^{n,[\ell]} = \frac{1}{\left|a_t^{n,[\ell]}\right|^2 + \varepsilon}, \quad \phi_n^{[\ell]} = \frac{1}{c_n^{[\ell]} + \varepsilon}, \quad \varphi_t^{[\ell]} = \frac{1}{r_t^{[\ell]} + \varepsilon}. \tag{25}$$

By properly choosing and updating ψ_t's, ϕ_n's, and φ_t's iteratively, the non-convex discontinuous ℓ_0-norms can be effectively approximated to the quadratic forms [36]. This has confirmed that replacing the sparsity term $\|\mathbf{x}\|_0$ by $\sum_{\forall i}(1/|x_i|^2 + \varepsilon)^{1/2} x_i$ is useful to develop an algorithm solving sparsity problem which is also a very efficient Cs based benchmark approach in [34]–[36].

Problem (23) is however still very challenging due to the discontinuous constraint (13b). As can be seen, this challenging issue can be transferred to the following $|a_t^n| = \|a_t^n\|_0 \quad \forall(t,n)$ which is also in a sparsity form. Regarding CS technique, we can relax constraint (13b) in the iteration $(\ell+1)$ by approximating a_t^n as

$$a_t^n = \left(1/\left|a_t^{n,[\ell]}\right|^2 + \varepsilon\right)^{1/2} a_t^n = \left(\psi_t^{n,[\ell]}\right)^{1/2} a_t^n. \tag{26}$$

Then, problem (23) can be further approximated as

$$\min_{\mathbf{A},\mathbf{W},\{\delta_k,\omega_k\}} \tilde{g}(\mathbf{W},\mathbf{A},\delta_k,\omega_k) = \kappa \sum_{\forall k}(\omega_k e_k - \log \omega_k - 1) + \eta \Bigg[\kappa P_{\mathrm{cons}} + \sum_{t=1}^{N_{\mathrm{T}}} \mathbf{a}_t^H \mathbf{D}_t^{[\ell]} \Psi_t^{[\ell]} \mathbf{a}_t + \kappa\left(1 + \frac{1}{\rho_{\mathrm{pa}}}\right) \times \sum_{t=1}^{N_{\mathrm{T}}} \mathbf{a}_t^H \Psi_t^{[\ell]} \left(\sum_{\forall k} \mathbf{w}_k \mathbf{w}_k^H\right) \mathbf{a}_t \Bigg] \tag{27a}$$

$$\text{s.t.} \quad \left|a_t^n\right| \le 1 \quad \forall(t,n), \tag{27b}$$

$$\mathbf{w}_k^H \Psi_t^{[\ell]} \mathbf{a}_t \mathbf{a}_t^H \mathbf{w}_k \le P_t^{\max}, 1 \le t \le N_{\mathrm{T}}. \tag{27c}$$

It is noted that the objective function in problem (27) is not *jointly* convex, but it is convex over each set of variables $\mathbf{w}_k$'s, $\mathbf{a}_t$'s, δ_k's, and ω_k's. Hence, an efficient algorithm for solving this problem can be developed by alternately optimizing $\mathbf{w}_k$'s and $\mathbf{a}_t$'s, and the MSE weight update for δ_k's, and ω_k's.

2) Update MSE Weights and Receive Coefficients: For given $(\mathbf{W},\ \mathbf{A})$, δ_k's, and ω_k's can be determined according to the results in Appendix C. In particular, the MMSE receive filter at user k is given as

$$\delta_k^\star = \delta_k^{\mathrm{MMSE}} = \left(\sum_{\forall j} \left|\mathbf{h}_k^H \mathbf{A} \mathbf{w}_j\right|^2 + \sigma_k^2\right)^{-1} \mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k. \tag{28}$$

And, the optimum value of ω_k can be expressed as

$$\omega_k^\star = e_k^{-1} = 1 + \left(\sum_{\forall j} \left|\mathbf{h}_k^H \mathbf{A} \mathbf{w}_j\right|^2 + \sigma_k^2\right)^{-1} \left|\mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k\right|^2. \tag{29}$$

3) Digital Precoding Design: For given AP matrix $\mathbf{A}$, the optimal $\mathbf{w}_k$'s can be obtained by solving the following quadratically constrained quadratic program (QCQP) problem:

$$\min_{\{\mathbf{w}_k\}} \sum_{\forall(k)} \mathbf{w}_k^H \Pi_k \mathbf{w}_k - 2\omega_j \Re\left(\delta_k' \mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k\right) \quad \text{s.t. (27c)}. \tag{30}$$

where $\Pi_k = \sum_{\forall j} \omega_j |\delta_j|^2 \mathbf{A}^H \mathbf{h}_j \mathbf{h}_j^H \mathbf{A} + \eta(1 + \frac{1}{\rho_{\mathrm{pa}}}) \sum_{\forall t} \Psi_t^{[\ell]} \mathbf{a}_t \mathbf{a}_t^H$ and $\Re(.)$ stands for the real part. This QCQP problem can be solved by any standard convex optimization solvers or the Lagrangian duality method.

4) Sparsity-Modulus Analog Phase-Shifting Matrix Design: For given DP vectors $\mathbf{w}_k$'s, the AP matrix $\mathbf{A}$ can be achieved by solving the following problem:

$$\min_{\{\mathbf{a}_t\}'s} \sum_{t=1}^{N_{\mathrm{T}}} \left[\mathbf{a}_t^H \Phi_t \mathbf{a}_t - 2\Re\left(\mathbf{f}_t^H \mathbf{a}_t\right)\right] \quad \text{s.t. (27b) and (27c)}, \tag{31}$$

where $\Phi_t^{[\ell]} = \sum_{\forall k} \omega_k |\delta_k|^2 |h_{k,t}|^2 \Omega_{\mathbf{w}}^{[\ell]} + \eta \Omega_{\mathbf{w}}^{[\ell]} + \eta \mathbf{D}_t^{[\ell]} \Psi_t^{[\ell]}$, $\Omega_{\mathbf{w}}^{[\ell]} = \kappa(1 + \frac{1}{\rho_{\mathrm{pa}}}) \Psi_t^{[\ell]} \sum_{\forall k} \mathbf{w}_k \mathbf{w}_k^H$, and $\mathbf{f}_t = \sum_{\forall k} \omega_k \delta_k h_{k,t}' \mathbf{w}_k^H$. As can be observed, problem (31) can be decomposed into N_{T} simpler sub-problems corresponding to N_{T} vectors $\mathbf{a}_t$'s. Specifically, the sub-problem corresponding to $\mathbf{a}_t$ can be stated as follows:

$$(\mathcal{P}_t)\ \min_{\mathbf{a}_t}\ \mathbf{a}_t^H \Phi_t \mathbf{a}_t - 2\Re\left(\mathbf{f}_t^H \mathbf{a}_t\right) \quad \text{s.t. (27b) and (27c)} \tag{32}$$

Again, problem $(\mathcal{P}_t)$ is a QCQP-form problem which can be solved by employing any standard optimization solver.

5) MMSE-Based Hybrid Precoding Design: By iteratively updating $\{\mathbf{w}_k, \mathbf{a}_t, \delta_k, \omega_k\}$, we can obtain the MMSE hybrid precoding. Combined with the compressed sensing-based method and with updating the weight factors ψ_t's, φ_t's, and ϕ_n's, the DP vectors and sparsity-modulus AP matrix design for WSRP maximization is summarized in Algorithm 2. The convergence of this algorithm is analyzed in the following theorem.

Theorem 4: Algorithm 2 converges to a local optimum solution after a finite number of iterations.

Proof: The proof is given in Appendix D. ■

Algorithm 2 Iterative Weighted Rate and Power Maximization Hybrid Precoding

```
Initialize: Choose suitable $\mathbf{w}_k^{[0]}$'s and $\mathbf{A}^{[0]}$, set $\ell = 0$ and a predetermined tolerate $\tau^{\text{in}}$.
repeat
    Update $\psi_t^{[\ell]}$'s, $\phi_n^{[\ell]}$'s, $\varphi_t^{[\ell]}$'s as in (25).
    Calculate $\delta_k^{[\ell+1]}$'s as in (28).
    Calculate and $\omega_k^{[\ell+1]}$'s as in and (29).
    Determine $\mathbf{w}_k^{[\ell+1]}$'s by solving problem (30) corresponding to $\mathbf{A}^{[\ell]}$'s, $\delta_k^{[\ell]}$'s and $\omega_k^{[\ell]}$'s.
    for $t = 1$ to $N_{\text{T}}$ do
        Solve problem (32) corresponding to $\mathbf{W}^{[\ell+1]}$'s, $\delta_k^{[\ell]}$'s and $\omega_k^{[\ell]}$'s to determine $\mathbf{a}_t^{[\ell+1]}$.
    end for
    Update $\ell := \ell + 1$.
until $|\tilde{g}(\mathbf{W}^{[\ell]}, \mathbf{A}^{[\ell]}, \boldsymbol{\delta}^{[\ell]}, \boldsymbol{\omega}^{[\ell]}) - \tilde{g}(\mathbf{W}^{[\ell-1]}, \mathbf{A}^{[\ell-1]}, \boldsymbol{\delta}^{[\ell-1]}, \boldsymbol{\omega}^{[\ell-1]})| \leq \tau^{\text{in}}$.
Return ($\mathbf{W}$, $\mathbf{A}$) as $\mathbf{W}^{[\ell]}$ and $a_t^{n[\ell]} = a_t^{n[\ell]}\sqrt{\psi_t^{n[\ell]}}$, $\forall(t,n)$.
```

C. Energy-Efficiency Hybrid Precoding Implementation

Algorithm 2 can be employed in Step 3 of Algorithm 1 to solve the SEEM problem. To ease the notation, we denote the repeated steps of Algorithm 1 as the outer loop, and that due to Algorithm 2 as the inner loop. In this implementation, we choose the outcome of the $(m-1)$-th outer iteration, $(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)})$, as the initial point for running Algorithm 2 in the next outer iteration. This careful selection ensures the convergence condition of Algorithm 1 in Theorem 2, which is analyzed in the following proposition.

Proposition 1: Consider the implementation where Algorithm 2 is employed in the Step 3 of Algorithm 1. If in every outer iteration, such as the *m*-th one, initial point of the inner loop running Algorithm 2 is set as $(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)})$, then, the condition (18) is satisfied.

Proof: The proof is given in Appendix E. ∎

V. Upper Bound of SEE and Heuristic Solution

This section focuses on developing a framework which can exploit the existing HP designs to solve problem (13). Specifically, a two-stage heuristic algorithm is proposed based on works given in [6] and [23]. First, the upper bound of SEE is analyzed according to modifying an algorithm presented in [23] designing the FDP to maximize the ratio between the achievable rate and transmission energy. The FDP obtained in this stage is so-called upper-bound FDP. In the next stage, an heuristic iterative solution is developed by step-by-step updating the RF chain to antenna connection structure. Particularly, in each iteration, one link between RF chains and antennas is considered to be turned off if after deactivating it, the updated connection structure together with a near-optimal performance HP re-constructed from the upper-bound FDP based on the method given in [6] can improve the SEE.

A. Upper-Bound of System Energy Efficiency

This section introduces the upper bound of SEE and a framework obtaining the upper-bound FDP. Let $\mathbf{B} \in \mathbb{R}^{N_{\text{T}} \times N_{\text{RF}}}$ be the mapping matrix, where the elements of $\mathbf{B}$ represent whether the RF chain to antenna connections are activated or not. Specifically, the mapping matrix corresponding to given AP matrix $\mathbf{A}$ can be defined as

$$b_t^n = |a_t^n|^2, \forall(t,n), \tag{33}$$

where b_t^n be the element allocated on the t^{th} row and the n^{th} column of $\mathbf{B}$. According to constraint (13b), we have that b_t^n can be 1 or 0. In addition, the total power consumption according to the RF process can be expressed based on $\mathbf{B}$ as

$$P_{\text{spr}}(\mathbf{A}) = \bar{P}_{\text{spr}}(\mathbf{B}) = (P_{\text{DAC}} + P_{\text{RFC}})\|\mathbf{c}(\mathbf{B})\|_0 + P_{\text{PS}}\|\mathbf{B}\|_0 + P_{\text{GCA}}\|\mathbf{r}(\mathbf{B})\|_0. \tag{34}$$

Hence, for given matrix $\mathbf{B}$, if (33) holds, the SEE can be rewritten as

$$\eta(\mathbf{W}, \mathbf{A}) = \frac{\sum_{\forall k} \log(1 + \text{SINR}_k)}{P_{\text{cons}} + \bar{P}_{\text{spr}}(\mathbf{B}) + (1 + 1/\rho_{\text{pa}}) \sum_{\forall k} \mathbf{w}_k^H \mathbf{A}^H \mathbf{A} \mathbf{w}_k}. \tag{35}$$

Then, the upper bound of $\eta(\mathbf{W}, \mathbf{A})$ can be determined based on the following proposition.

Proposition 2: When $\mathbf{A}$ satisfies (33), $\eta(\mathbf{W}, \mathbf{A})$ can be upper bounded by

$$\bar{\eta}\left(\mathbf{U}^{\text{Up}}, \mathbf{B}\right) = \frac{\sum_{\forall k} R_k^{\text{FDP}}\left(\mathbf{U}^{\text{Up}}\right)}{P_{\text{cons}} + \bar{P}_{\text{spr}}(\mathbf{B}) + (1 + 1/\rho_{\text{pa}}) \sum_{\forall k} \mathbf{u}_k^{\text{Up}\,H} \mathbf{u}_k^{\text{Up}}}, \tag{36}$$

where $\mathbf{u}_k^{\text{Up}}$'s be the upper-bound FDPs which can be obtained by solving the following problem,

$$\max_{\{\mathbf{u}_k\}} \ \tilde{\eta}(\mathbf{U}) = \frac{R_k^{\text{FDP}}(\mathbf{U})}{\sum_{\forall k} \mathbf{u}_k^H \mathbf{u}_k} \quad \text{s.t.} \ \sum_{\forall k} \mathbf{u}_k^H \mathbf{E}_t \mathbf{u}_k \leq P_t^{\max} \ \forall(t). \tag{37}$$

in which $\mathbf{U}$ is denoted as the matrix generated by all FDP vectors $\mathbf{u}_k$'s $\in \mathbb{C}^{N_{\text{T}} \times 1}$, $\mathbf{E}_t = diag(0_{1 \times (t-1)}, 1\ , 0_{1 \times (N_{\text{T}} - t)})$, and

$$R_k^{\text{FDP}}(\mathbf{U}) = \log\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{u}_k\right|^2}{\sum_{j \neq k} \left|\mathbf{h}_k^H \mathbf{u}_j\right|^2 + \sigma^2}\right). \tag{38}$$

Proof: The proof is given in Appendix F. ∎

1) Proposed Framework Solving Problem (37): We are now ready to determine the upper-bound of $\eta(\mathbf{W}, \mathbf{A})$ by proposing a framework solving problem (37) based on EEHP-A algorithm given in [23]. Denoting the upper-bound of SEE $\tilde{\eta}(\mathbf{U})$ and the rate in (38) as the functions of $\mathbf{u}_k$, i.e., $\tilde{\eta}(U)$ and $R_k^{\text{FDP}}(\mathbf{U})$. The gradient of $\tilde{\eta}(\mathbf{U})$ with respect to $\mathbf{u}_k$ is derived by

$$\frac{\partial \tilde{\eta}(\mathbf{U})}{\partial \mathbf{u}_k} = \frac{2}{\tilde{P}_{\text{T}}^2}(\Upsilon_k - \Phi_k)\mathbf{w}_k, \tag{39a}$$

$$\text{with} \quad \Upsilon_k = \frac{P_{\text{T}} \mathbf{h}_k \mathbf{h}_k^H}{\sum_{\forall j} \mathbf{h}_k^H \mathbf{u}_j \mathbf{u}_j^H \mathbf{h}_k + \sigma^2}, \tag{39b}$$

Algorithm 3 Iterative Weighted Rate and Power Maximization Hybrid Precoding

```
1: Initialize by choosing any $\mathbf{U}^{(0)}$ satisfying the power constraint, selecting a tolerance for converging $\tau^{\mathrm{up}}$, and setting $l = 0$.
2: repeat
3:     Update $\Upsilon_k$'s and $\Phi_k$'s as in (39) based on $\mathbf{U}^{(l)}$.
4:     Determine the step size $\mu_k^{(l+1)}$'s by solving the following problem.
           $\mu_k^{(l+1)} = \arg\max_{\mu_k \in [0,1]} \bar{\eta}\Big\{\big[\mathbf{I} + \mu_k(\Upsilon_k^{(l)}/\Phi_k^{(l)} - \mathbf{I})\big]\mathbf{u}_k^l\Big\}$   (40a)
           $\text{s.t.} \sum_{\forall k} \big\|\big[\mathbf{I} + \mu_k(\Upsilon_k^{(l)}/\Phi_k^{(l)} - \mathbf{I})\big]\mathbf{E}_t\mathbf{u}_k^l\big\|^2 \leq P_t^{\max}, \forall t.$   (40b)
5:     Update $\mathbf{u}_k^{(l+1)} = \big[\mathbf{I} + \mu_k^{(l+1)}(\Upsilon_k^{(l)}/\Phi_k^{(l)} - \mathbf{I})\big]\mathbf{u}_k^l, \forall l.$
6: until Convergence, i.e., $|\tilde{\eta}(\mathbf{U}^{(l)}) - \tilde{\eta}(\mathbf{U}^{(l-1)})| \leq \tau^{\mathrm{up}}$.
```

$$\Phi_k = \frac{\sum_{\forall j} R_j^{\mathrm{FDP}}}{\ln 2}\mathbf{I} + P_{\mathrm{T}} \sum_{j \neq k} \frac{p_k \mathbf{h}_k \mathbf{h}_k^H}{m_k(m_k - p_k)}, \tag{39c}$$

$$P_{\mathrm{T}} = \sum_{\forall k} \mathbf{u}_k^H \mathbf{u}_k, \tag{39d}$$

where $p_k = \mathbf{h}_k^H \mathbf{u}_k \mathbf{u}_k^H \mathbf{h}_k$ and $m_k = \sum_{\forall j} \mathbf{h}_k^H \mathbf{u}_j \mathbf{u}_j^H \mathbf{h}_k + \sigma^2$. Then, the local optimization solution for $\mathbf{u}_k$, can be defined by applying the zero-gradient condition $\partial\tilde{\eta}(\mathbf{U})/\partial\mathbf{u}_k = 0$ as $\mathbf{u}_k = \Phi_k^{-1}\Upsilon_k\mathbf{u}_k$. To obtain the optimal FDP vectors $\mathbf{u}_k$'s, an iterative algorithm is developed as in Algorithm 3 which is similar to the EEHP-A algorithm in [23]. Let $\mathbf{U}^{\mathrm{Up}}$ be the optimal solution of problem (37). Then, for any feasible solution $(\mathbf{W}', \mathbf{A}')$, $\eta(\mathbf{W}', \mathbf{A}')$ is upper-bounded by $\bar{\eta}(\mathbf{U}^{\mathrm{Up}}, \mathbf{B}')$, where $\mathbf{B}'$ is defined based on $\mathbf{A}'$.

B. Heuristic Energy Efficiency Maximization Method

In this section, we first re-construct a near-optimal performance HP based on $\mathbf{U}^{\mathrm{Up}}$ [6] for a given $\mathbf{B}$. Then, we develop an heuristic algorithm for solving problem (13).

1) Hybrid Precoding Design for Given $\boldsymbol{B}$*:* We aim to reconstruct the HP where $\mathbf{w}_k$'s and $\mathbf{A}$ can be defined via MMSE approximation as follows.

$$\min_{\mathbf{W},\mathbf{A}} \sum_{\forall k} \|\mathbf{u}_k^{\mathrm{Up}} - \mathbf{A}\mathbf{w}_k\|_2^2 \ \ \text{s.t.}\ (13\text{c})\ \text{and}\ (33). \tag{41}$$

For given $\mathbf{A}$, $\mathbf{w}_k$'s can be determined based on the well-known least squares method as

$$\mathbf{w}_k = \beta_k \mathbf{A}^{\dagger}\mathbf{u}_k^{\mathrm{Up}}, \tag{42}$$

where β_k's are the factors satisfying the power constraint (13c). While fixing $\mathbf{w}_{k,s}$'s, $\mathbf{A}$ can be obtained by solving the following problem:

$$\min_{\mathbf{A}} \sum_{\forall k} \left\|\mathbf{u}_k^{\mathrm{Up}} - \mathbf{A}\mathbf{w}_k\right\|_2^2 \ \ \text{s.t.}\ (33). \tag{43}$$

This problem is classified as a unit-modulus least square type, which is non-convex and NP-hard. This problem can be efficiently solved by employing the "Projected Gradient Descent Method" proposed in [38], [39].

2) A Heuristic Algorithmic Approach: Here, we develop an heuristic algorithm to solve the SEEM problem. First, we define the upper-bound FDP as in Algorithm 3. Then, we start with $\mathbf{B} = \mathbf{1}_{N_{\mathrm{T}} \times N_{\mathrm{RF}}}$. In each iteration, we define $(\mathbf{W}, \mathbf{A})$ based on $\mathbf{B}$ and $\mathbf{U}^{\mathrm{Up}}$ before deactivating the active connection without which one can increase the SEE most. The process stops if no activated connection can be turned off to increase the SEE.

VI. Simulation Results

A. mmWave Channel Model

The mmWave channel is generally not rich in scattering because mmWave signals do not reflect well in the surrounding environment [27]. Hence, there are only few dominant paths in mmWave transmission channel. Similar to [8], the Saleh-Valenzuela geometric channel model is adopted for the numerical evaluation in this paper as

$$\mathbf{h}_k = \sqrt{\frac{N_{\mathrm{T}}}{P_{\mathrm{L}} N_{\mathrm{C}} N_{\mathrm{L}}}} \sum_{c=1}^{N_{\mathrm{C}}} \sum_{\ell=1}^{N_{\mathrm{L}}} \alpha_{c,\ell}\, a_{\mathrm{r}}\left(\phi_{c,\ell}^{\mathrm{r}}, \theta_{c,\ell}^{\mathrm{r}}\right) \mathbf{a}_{\mathrm{t}}\left(\phi_{c,\ell}^{\mathrm{t}}, \theta_{c,\ell}^{\mathrm{t}}\right), \tag{44}$$

where P_{L} is the path-loss, and N_{C} and N_{L} are the number of clusters and number of propagation sub-paths in each cluster, respectively. In addition, $\alpha_{c,\ell}$ is the complex gain of the ℓ-th path of cluster *c*, and $(\phi_{c,\ell}^{\mathrm{r}}, \theta_{c,\ell}^{\mathrm{r}})$ and $(\phi_{c,\ell}^{\mathrm{t}}, \theta_{c,\ell}^{\mathrm{t}})$ are the (azimuth, elevation) angles of arrival and departure corresponding, respectively. Herein, $a_{\mathrm{r}}(\phi_{c,\ell}^{\mathrm{r}}, \theta_{c,\ell}^{\mathrm{r}})$ and $\mathbf{a}_{\mathrm{t}}(\phi_{c,\ell}^{\mathrm{t}}, \theta_{c,\ell}^{\mathrm{t}})$ represent the normalized receive response factor and transmit array response vectors at (azimuth, elevation) angles of $(\phi_{c,\ell}^{\mathrm{r}}, \theta_{c,\ell}^{\mathrm{r}})$ and $(\phi_{c,\ell}^{\mathrm{t}}, \theta_{c,\ell}^{\mathrm{t}})$, respectively [8], [27]. Finally, $\alpha_{c,\ell}$ is assumed to be i.i.d. Gaussian distributed and the normalization factor $\sqrt{N_{\mathrm{T}}/(P_{\mathrm{L}} N_{\mathrm{C}} N_{\mathrm{L}})}$ is added to get $\mathbb{E}_{\mathbf{h}_k}\{\|\mathbf{h}_k\|_2^2\} = N_{\mathrm{T}}/P_{\mathrm{L}}$.

B. Simulation Results

This section presents the performance of the proposed SEE HP algorithm which is Algorithm 1 integrated with Algorithm 2 (denoted "Proposed Alg." in the figures). For comparison purpose, the performances of other precoding designs are also presented, including: i.) the upper bound corresponding to the FDP achieved by Algorithm 3 presented in Section V-A, denoted as "FDP Upper Bound;" ii.) the heuristic SEE HP achieved by Algorithm 4, denoted as "Heuristic Alg.;" and iii.) two algorithms algorithm given in [23] and [24], denoted as "Zi's Alg." and "He's Alg.," respectively. An uniform polar array with antenna spacing equal to a half-wavelength is adopted at the base station. The channel to each user contains three clusters of 10 paths, i.e., $C = 3$, $L = 10$. All the channel path gains $\alpha_{c,\ell}$'s are assumed to be i.i.d. Gaussian random variables with variance σ_{α}^2. The azimuth angles are assumed to be uniformly distributed in $[0; 2\pi]$, and the AoA/AoD elevation angles are uniformly distributed in $[-\frac{\pi}{2}; \frac{\pi}{2}]$. The noise variance σ^2 is set at 1.2×10^{-13} W due to the noise factor of 4.79 dB, the power spectral density of -174 dBm/Hz, and 1 MHz bandwidth [33]. In this

Algorithm 4 Heuristic Energy-Efficiency Maximization Hybrid Precoding Algorithm

```
Initialize: Define $\mathbf{U}^{\mathrm{Up}}$ by employing Algorithm 3. Set $\mathbf{B} = \mathbf{1}_{N_{\mathrm{T}} \times N_{\mathrm{RF}}}$ and $\mathcal{L} = \{(t, n) | b_t^n = 1\}$.
repeat
    Solve $\mathbf{W}$ and $\mathbf{A}$ based on $\mathbf{U}^{\mathrm{Up}}$ and $\mathbf{B}$ as discussed in Section V-B1.
    For every $(t, n) \in \mathcal{L}$, define $\mathbf{A}^{(t,n)} = \mathbf{A}$ then set $(\mathbf{A}^{(t,n)})_{t,n} = 0$.
    Define $(t', n') = \arg\max_{(t,n) \in \mathcal{L}} \eta(\mathbf{W}, \mathbf{A}^{(t,n)})$.
    if $\eta(\mathbf{W}, \mathbf{A}^{(t',n')}) > \eta(\mathbf{W}, \mathbf{A})$ then
        Set $\mathcal{L} = \mathcal{L}/(t', n')$ and update $(\mathbf{B})_{t',n'} = 0$.
    end if
until $\eta(\mathbf{W}, \mathbf{A}^{(t',n')}) \leq \eta(\mathbf{W}, \mathbf{A})$.
```

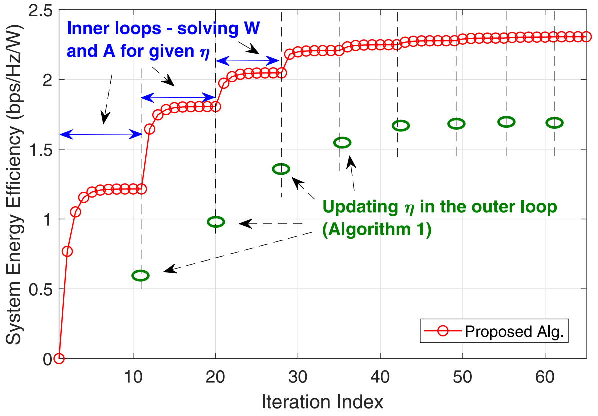

Fig. 3. Convergence of the proposed algorithm.

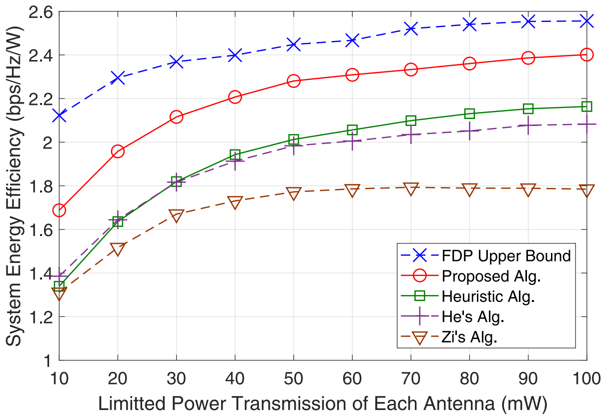

Fig. 4. The system energy efficiency *versus* the power budget of each antenna.

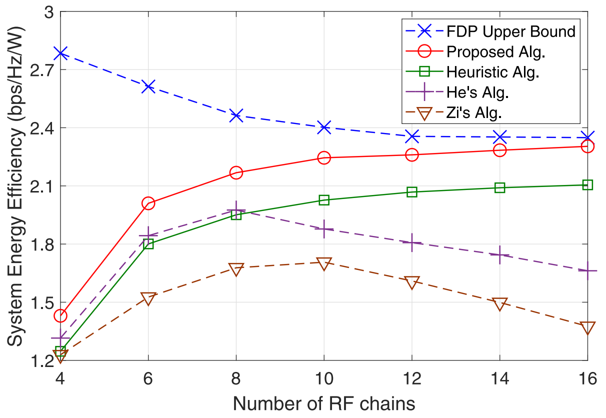

Fig. 5. The system energy efficiency *versus* the number of RF chains N_{RF}.

simulation, we employ the *path loss ABG model* for macro-cellular scenario with $f = 60$ GHz in [40, Table I], where distance is set as 100 m for all users. Set $P_{\mathrm{BB}} = 300$ mW, $P_{\mathrm{DAC}} = 200$ mW, $P_{\mathrm{RFC}} = 43$ mW, $P_{\mathrm{PS}} = 40$ mW, $P_{\mathrm{SW}} = 2$ mW, and $P_{\mathrm{amp}} = 40$ mW, $G_{\mathrm{amp}} = 20$ dB, $\rho_{\mathrm{pa}} = 0.3$, $L_{\mathrm{sw}} = L_{\mathrm{ps}} = 2$ dB [26], [41], [44]. For implementation, we choose $\varepsilon = 10^{-8}$, $\tau^{\mathrm{out}} = \tau^{\mathrm{in}} = \tau^{\mathrm{up}} = 10^{-4}$. Unless indicated otherwise, $P_t^{\max} = 50$ mW, both K and N_{RF} are set equal to 8, and $N_{\mathrm{T}} = 64$.

We illustrate the convergence of our proposed algorithm in Fig. 3 where the variations of SEE calculated based on the outcomes $(\mathbf{W}, \mathbf{A})$ of Algorithm 1 integrating with Algorithm 2 over the iterations are shown. As can be seen, the SEE increases monotonically after each iteration before reaching its the maximum value which is the outcome of Algorithm 1. Additionally, there are many fragments of the iteration sequence within of which the SEE firstly boosts speedily, then grows slower before saturating. This is because of that the iterations in each such fragment are due to the inner loop employing Algorithm 2 to solve WSRP for given value of η. At the end of each fragment, η is updated according to Algorithm 1 before the next inner loop is implemented in the next fragment. This simulation result hence confirms the convergence of the proposed algorithm proved in Theorem 2 and Proposition 1.

Fig. 4 presents the SEEs achieved by the proposed algorithm and other methods versus the transmit power at each antenna, $P_t^{\max}$. As can be seen, the SEEs achieved by all algorithms increase as $P_t^{\max}$ increases before saturating at the high regime of $P_t^{\max}$,while the SEE upper bound increases then slightly decreases. This is due to the larger feasible set that can attain better solutions. However, the FDP is designed to maximize the data rate to the transmit power ratio, not the data rate to the total power consumption ratio. Hence, when the increase of $P_t^{\max}$ may result in more active phase shifters which then reduces the SEE. In addition, the proposed algorithm achieves much higher achievable SEEs at all values of $P_t^{\max}$ than the heuristic algorithm and the two methods given in [23], [24] do since our design takes care of all the total consumption power while the works in [23], [24] do not. Interestingly, the method given in [23] obtains the smallest SEE in comparison to others while the heuristic algorithm outperforms He's algorithm in [24] when value of $P_t^{\max}$ increases.

Fig. 5 illustrates the SEE achieved by all schemes versus the number of RF chains, N_{RF}. As can be observed, the SEE upper bound (in Section V-A) decreases as N_{RF} increases since the achievable rate is unchanged while the system consumes more power. When N_{RF} becomes larger, the proposed algorithms can achieve better SEE. In contrast, the SEE obtained by Zi's and He's methods first increases with N_{RF} then decreases. Interestingly, the proposed algorithm significantly outperforms the other HP designs and

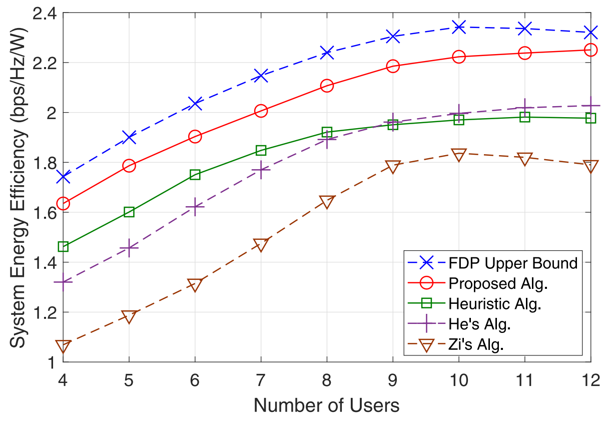

Fig. 6. The system energy efficiency *versus* the number of users *K*.

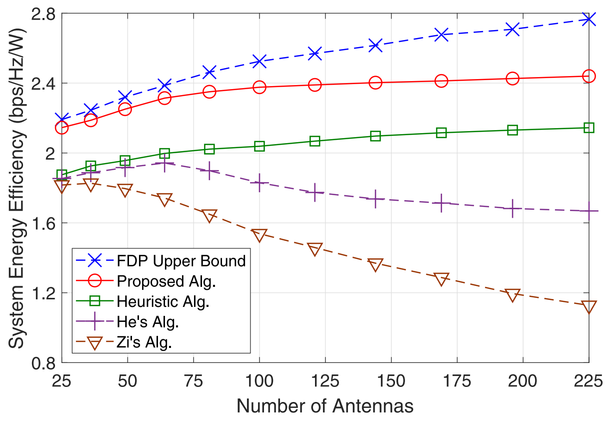

Fig. 7. The system energy efficiency *versus* the number of antennas N_{T}.

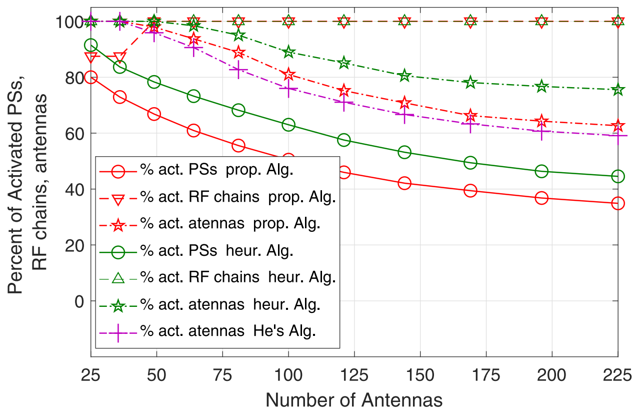

Fig. 8. The percentage of activated phase shifters, RF chains, and antennas *versus* the number of antennas N_{T}.

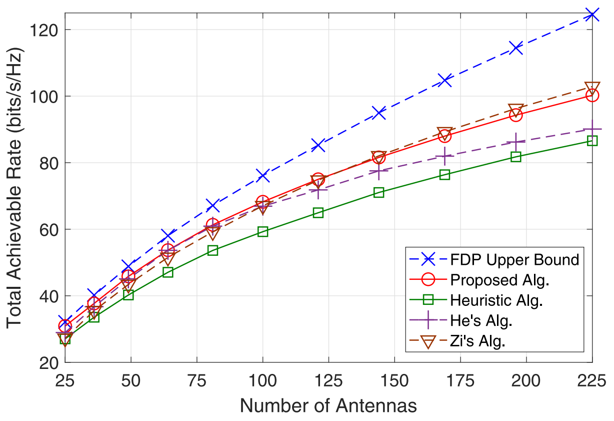

Fig. 9. Total achievable rate *versus* the number of antennas N_{T}.

achieves a SEE near to the upper bound at the high regime of N_{RF}, which again confirms the superior performance of the proposed design.

In Fig. 6, the SEE achieved by all schemes are displayed versus the number of users *K*. The SEE achieved by all schemes increases and then decreases with the number of users. This observation is arguably due to the increased power consumption for baseband signal processing to support more users. Furthermore, the increase rate in the required power is much faster than the rate in the sum-rate improvements. Again, our proposed algorithm can attain better SEE than other HP design methods.

Next, the impact of the number of antennas N_{T} is presented in Figs. 7, 8, and 9. Fig. 7 illustrates the SEEs achieved by different schemes while varying N_{T}. While the upper bound keeps increasing with N_{T}, the proposed algorithm increases then saturates at large N_{T}. On the other hand, the SEE achieved by the two methods given in [23] and [24] increases then deceases quickly. In this simulation, our proposed method again outperforms all other designs except the upper bound by the FDP. In Fig. 8, we present the percentage of phase shifters, RF chains, and antennas which are activated with the proposed algorithm and the heuristic algorithm in scenarios with different number of antennas. The percentage of utilized antennas with the method given in [24] is also illustrated for comparison. Interestingly, the percentage of activated phase shifters and antennas decreases as N_{T} increases while the percentage of activated RF chains increases. As can be observed, the proposed algorithm turns off more phase shifters, RF chains, and antennas than the heuristic one. In addition, He's algorithm utilizes fewer antennas than others since it focuses on reducing the number of antennas to maximize the SEE. Fig. 9 presents the achievable rate achieved by all schemes versus the number of antennas N_{T}. As expected, the system achievable rate achieved by all schemes increases with N_{T}. The FDP enhances well the "degree-of-freedom" to achieve the highest data rate. Zi's algorithm places itself in the second place since this scheme utilizes all the hardware components to serve all users. He's algorithm outperforms our proposed algorithm in the low regime of N_{T} while ours is superior to this method in the high regime. In addition, the proposed algorithm is better than the heuristic one in all scenarios with different number of antennas.

VII. Conclusion

This paper has considered the dynamic fully-connected AP structure for the HP mmWave multi-user systems where each connection between one RF chain and one antenna can be activated or deactivated for power saving. This dynamic fully-connected AP structure has been represented as a sparsity-modulus constraint for the AP matrix design. In addition, the total power consumption of this system is formulated generally as the sparsity form of AP matrix. Then, a new compressed sensing-based energy-efficiency HP algorithm has been proposed to maximize the non-convex SEE. Numerical results have confirmed the superior performances of the proposed energy-efficiency HP design over heuristic algorithms and benchmark algorithms in literature.

Appendix A
Proof of Theorem 1

Theorem 1 can be proved by employing an approach similar to [29] as follows.

i) Consider any η_1 and η_2 where $0 < \eta_1 < \eta_2$. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be the optimal solution of problem (15) corresponding to η_1 and η_2, respectively. Then, we have:

$$\begin{aligned}\chi(\eta_2) &= R(\mathbf{x}_2) - \eta_2 P(\mathbf{x}_2) < R(\mathbf{x}_2) - \eta_1 P(\mathbf{x}_2)\\ &\leq R(\mathbf{x}_1) - \eta_1 P(\mathbf{x}_1) = \chi(\eta_1). \qquad (45)\end{aligned}$$

Hence, $\chi(\eta)$ is a strictly monotonic decreasing function.

ii) Assume that $\chi(\eta) > 0$. So, there exists $\hat{\mathbf{x}}$ such that $R(\hat{\mathbf{x}}) - \eta P(\hat{\mathbf{x}}) > 0$ which means $\eta < \frac{R(\hat{\mathbf{x}}}{P(\hat{\mathbf{x}})}$. This implies that

$$\eta < \max_{\mathbf{x}\in\mathcal{S}} \frac{R(\mathbf{x})}{P(\mathbf{x})} = \eta^*. \qquad (46)$$

Conversely, if $\eta < \eta^*$, then, $\eta < \max_{\mathbf{x}\in\mathcal{S}} \frac{R(\mathbf{x})}{P(\mathbf{x})}$. Hence, there exists $(\hat{\mathbf{x}}')$ such that $\eta < \frac{R(\hat{\mathbf{x}}')}{P(\hat{\mathbf{x}}')}$, which implies that

$$\chi(\eta) = \max_{\mathbf{x}\in\mathcal{S}} R(\mathbf{x}) - \eta P(\mathbf{x}) > 0. \qquad (47)$$

iii) Assume that $\chi(\eta) < 0$. Then, we have $R(\mathbf{x}) - \eta P(\mathbf{x}) < 0$ $\forall \mathbf{x} \in \mathcal{S}$. Hence, $\eta > \frac{R(\mathbf{x})}{P(\mathbf{x})}$ $\forall \mathbf{x} \in \mathcal{S}$, which yields

$$\eta > \max_{\mathbf{x}\in\mathcal{S}} \frac{R(\mathbf{x})}{P(\mathbf{x})} = \eta^\star. \qquad (48)$$

Conversely, if $\eta > \eta^\star$, then, we have $\eta > \max_{\mathbf{x}\in\mathcal{S}} \frac{R(\mathbf{x})}{P(\mathbf{x})}$. Hence, $\eta > \frac{R(\mathbf{x})}{P(\mathbf{x})}$ $\forall \mathbf{x} \in \mathcal{S}$, which implies

$$\chi(\eta) = \max_{\mathbf{x}\in\mathcal{S}} R(\mathbf{x}) - \eta P(\mathbf{x}) < 0. \qquad (49)$$

iv) Statements *(i)*, *(ii)*, and *(iii)* yield $\chi(\eta^\star) = 0$. Then, if $\mathbf{x}^*$ is an optimal solution of $(\mathcal{P}_I)$, we have $R(\mathbf{x}^*) - \eta^* P(\mathbf{x}^*) = 0$ which implies that $\mathbf{x}^*$ is also an optimal solution of $(\mathcal{P}_{II}^{\eta^\star})$. Conversely, if $\mathbf{x}'$ is an optimal solution of $(\mathcal{P}_{II}^{\eta^*})$, then $\chi(\eta^*) = R(\mathbf{x}') - \eta^* P(\mathbf{x}')$ must equal to 0 due to statements *(ii)* and *(iii)*. Hence, we have $\frac{R(\mathbf{x}')}{P(\mathbf{x}')} = \eta^*$, which yields that $\mathbf{x}'$ is an optimal solution of problem $(\mathcal{P}_I)$. Thus, $(\mathcal{P}_I)$ and $(\mathcal{P}_{II}^{\eta^\star})$ have the same set of optimal solutions.

Appendix B
Proof of Theorem 2

We will prove that $\eta^{(m)}$ monotonically increases after each iteration. It is easy to see that if $\bar{\chi}(\eta^{(m)}, \mathbf{W}^{(m)}, \mathbf{A}^{(m)}) \geq \bar{\chi}(\eta^{(m)}, \mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)})$, one has

$$\begin{aligned}&R\Big(\mathbf{W}^{(m)}, \mathbf{A}^{(m)}\Big) - \eta^{(m)} P_{\text{tot}}\Big(\mathbf{W}^{(m)}, \mathbf{A}^{(m)}\Big)\\ &\geq R\Big(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)}\Big) - \eta^{(m)} P_{\text{tot}}\Big(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)}\Big) = 0, \qquad (50)\end{aligned}$$

since $\eta^{(m)} = R(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)})/P_{\text{tot}}(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)})$. Hence, the process updating $\eta^{(m+1)}$ in Algorithm 1 implies that

$$\eta^{(m+1)} = R\Big(\mathbf{W}^{(m)}, \mathbf{A}^{(m)}\Big)/P_{\text{tot}}\Big(\mathbf{W}^{(m)}, \mathbf{A}^{(m)}\Big) \geq \eta^{(m)}. \qquad (51)$$

The result implies the monotonic increase of η after each iteration. Furthermore, the value of energy efficiency η cannot be infinity. Hence, Algorithm 1 can converge after a finite number of iterations.

Appendix C
Proof of Theorem 3

According to the receive coefficient δ_k, the estimate of s_k at the user *k* can be given by $\hat{s}_k = \delta_k y_k$. Hence, the MMSE receive coefficient at user *k* is given by

$$\begin{aligned}\delta_k^{\text{MMSE}} &= \arg\min_{\delta_k} \mathbb{E}\Big\{|s_k - \delta_k y_k|^2\Big\}\\ &= \left(\sum_{\forall j} |\mathbf{h}_k^H \mathbf{A}\mathbf{w}_j|^2 + \sigma_k^2\right)^{-1} \mathbf{w}_k^H \mathbf{A}^H \mathbf{h}_k. \qquad (52)\end{aligned}$$

Then, the MSE for user *k* corresponding to the MMSE-receive filter can be written as $e_k = \mathbb{E}\{|s_k - \delta_k^{\text{MMSE}} y_k|^2\} = (1 + \text{SINR}_k)^{-1}$. Hence, $R_k(\mathbf{A}, \mathbf{W})$ can be expressed as a function of e_k as $R_k(\mathbf{A}, \mathbf{W}) = \log(e_k^{-1})$. Furthermore, thanks to the first-order Taylor approximation for the log-function, one can yield

$$R_k(\mathbf{A}, \mathbf{W}) = \log\Big(e_k^{-1}\Big) \geq -\Big[\log\Big(\omega_k^{-1}\Big) + \omega_k^{-1}(e_k - \omega_k)\Big]. \qquad (53)$$

In addition, the optimum value of δ_k and ω_k is expressed as

$$\delta_k^\star = \delta_k^{\text{MMSE}} \text{ and } \omega_k^\star = e_k^{-1}. \qquad (54)$$

At that point one has $g(\mathbf{A}, \mathbf{W}, \delta_k, \omega_k) = l(\mathbf{A}, \mathbf{W})$. The proof thus follows.

Appendix D
Proof of Theorem 4

Denote the $a_t^{n(t)}$ as the solution of a_t^n at the ℓth iteration in Algorithm 2. The majorization of the log function given in (22) in conjunction with updating $\delta_k^{[\ell]}$'s in Step 4 (as in (28)), determining $\omega_k^{[\ell]}$'s in Step 5 (as in (29)), optimizing $\mathbf{w}_k^{[\ell]}$'s in Step 6 (by solving problem (30)), and solving problem (32)'s to obtain $\mathbf{a}_t^{[\ell]}$ in Step 7–9 ensure the decrement of the objective function in problem (23) at the ℓth iteration. Then, we have

$$\begin{aligned}g^{[\ell+1]} &= \Sigma^{[\ell+1]} + \sum_{\forall X^{[\ell+1]}} \eta_X \log\Big(X^{[\ell+1]} + \varepsilon\Big)\\ &\leq \Sigma^{[\ell+1]} + \sum_{\forall X} \eta_X \left[\log\Big(X^{[\ell]} + \varepsilon\Big) + \frac{X^{[\ell+1]} - X^{[\ell]}}{X^{[\ell]} + \varepsilon}\right]\end{aligned}$$

$$\begin{aligned} &\leq \Sigma^{[\ell]} + \sum_{\forall X} \eta_X \left[\log\Big(X^{[\ell]} + \varepsilon \Big) + \frac{X^{[\ell]} - X^{[\ell]}}{X^{[\ell]} + \varepsilon} \right] \\ &\leq \Sigma^{[\ell]} + \sum_{\forall X} \eta_X \log\Big(X^{[\ell]} + \varepsilon \Big) = g^{[\ell]}, \end{aligned} \tag{55}$$

where $g^{[\ell]}$ stands for the value of the objective function in problem (21) at the ℓth iteration. Therefore, the convergence of Algorithm 2 is attained thanks to the monotonic decrease of the objective function in problem (21) after each iteration.

Appendix E
Proof of Proposition 1

In iteration m of the outer loop, Algorithm 2 is called for solving the WRSP problem for a given $\eta^{(m)}$. The starting point of the inner loop is chosen as $(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)})$, which is the outcome of the $(m-1)$-th iteration of the outer loop. Thanks to Theorem 4, the objective function of problem (21) monotonically decreases from the initial point to a convergence point. This convergence point is the set as the outcome of the m-th iteration of the outer loop. Hence, we have

$$\begin{aligned} &g\Big(\mathbf{W}^{(m)}, \mathbf{A}^{(m)}, \delta_k^{(m)}, \omega_k^{(m)}\Big) \\ &\quad \leq g\Big(\mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)}, \delta_k^{(m-1)}, \omega_k^{(m-1)}\Big). \end{aligned} \tag{56}$$

This result combined with Theorem 3 yields

$$\bar{\chi}\Big(\eta^{(m)}, \mathbf{W}^{(m)}, \mathbf{A}^{(m)}\Big) \geq \bar{\chi}\Big(\eta^{(m)}, \mathbf{W}^{(m-1)}, \mathbf{A}^{(m-1)}\Big).$$

Proposition 1 thus follows.

Appendix F
Proof of Proposition 2

Denote $(\mathbf{A}', \mathbf{W}')$ as an arbitrary feasible solution of problem (13). Let $\mathbf{u}'_k = \mathbf{A}'\mathbf{w}'_k$. Then, one has

$$\sum_{\forall k} \mathbf{u}'^H_k \mathbf{E}_t \mathbf{u}'_k = \mathbf{w}'^H_k \mathbf{a}'_t \mathbf{a}'^H_t \mathbf{w}'_k \leq P_t^{\max} \ \ \forall(t), \tag{57}$$

which yields $\mathbf{u}'_k$'s is a feasible solution of problem (37). It means

$$\begin{aligned} \tilde{\eta}\big(\mathbf{U}'\big) &= \frac{\sum_{\forall k} \log\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{A}'\mathbf{w}'_k\right|^2}{\sum_{j \neq k} \left|\mathbf{h}_k^H \mathbf{A}'\mathbf{w}'_j\right|^2 + \sigma^2}\right)}{\sum_{\forall k} \mathbf{w}'^H_k \mathbf{A}'^H \mathbf{A}'\mathbf{w}'_k} \\ &\leq \frac{\sum_{\forall k} \log\left(1 + \frac{\left|\mathbf{h}_k^H \mathbf{u}_k^{\mathrm{Up}}\right|^2}{\sum_{j \neq k} \left|\mathbf{h}_k^H \mathbf{u}_j^{\mathrm{Up}}\right|^2 + \sigma^2}\right)}{\sum_{\forall k} \mathbf{u}_k^{\mathrm{Up}\,H} \mathbf{u}_k^{\mathrm{Up}}} = \tilde{\eta}\Big(\mathbf{U}^{\mathrm{Up}}\Big), \end{aligned} \tag{58}$$

for all $(\mathbf{W}', \mathbf{A}') \in \Theta$, where $\mathbf{U}'$ which is the matrix generated by all vectors $\mathbf{u}'_k$'s. Thanks to (58), one can conclude that

$$\begin{aligned} \eta(\mathbf{W}, \mathbf{A}) &= \frac{\sum_{\forall k} \log(1 + \mathrm{SINR}_k)}{P_{\mathrm{cons}} + \bar{P}_{\mathrm{spr}}(\mathbf{B}) + (1 + 1/\rho_{\mathrm{pa}}) \sum_{\forall k} \mathbf{w}_k^H \mathbf{A}^H \mathbf{A} \mathbf{w}_k} \\ &\leq \frac{\sum_{\forall k} R_k^{\mathrm{FDP}}\big(\mathbf{U}^{\mathrm{Up}}\big)}{P_{\mathrm{cons}} + \bar{P}_{\mathrm{spr}}(\mathbf{B}) + (1 + 1/\rho_{\mathrm{pa}}) \sum_{\forall k} \mathbf{u}_k^{\mathrm{Up}\,H} \mathbf{u}_k^{\mathrm{Up}}} \\ &= \bar{\eta}\Big(\mathbf{U}^{\mathrm{Up}}, \mathbf{B}\Big), \quad \forall(\mathbf{W}, \mathbf{A}) \in \Theta, \end{aligned} \tag{59}$$

which finised the proof of Proposition 2.

References

[1] J. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] Q. V. Pham *et al.*, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," Jun. 2019. [Online]. Available: arXiv:1906.08452.

[3] T. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[4] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

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

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

[7] T. Werthmann, H. Grob-Lipski, and P. Proebster, "Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks," in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)* London, U.K., Sep. 2013, pp. 3328–3333.

[8] 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.

[9] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[10] A. Alkhateeb and R. W. Heath, "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1801–1818, May 2016.

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

[12] T. E. Bogale, L. B. Le, A. Haghighat, and L. Vandendorpe, "On the number of RF chains and phase shifters, and scheduling design with hybrid analog–digital beamforming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3311–3326, May 2016.

[13] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, "Hybrid MMSE precoding for mmWave multiuser MIMO systems," in *Proc. IEEE Int. Conf. Commun.*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[14] D. H. N. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, "Hybrid MMSE precoding and combining designs for mmWave Multiuser systems," *IEEE Access*, vol. 5, pp. 19167–19181, 2017.

[15] R. Mai, T. Le-Ngoc, and D. H. N. Nguyen, "Joint hybrid Tx–Rx design for wireless backhaul with delay-outage constraint in massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6736–6750, Oct. 2017.

[16] R. Mai, T. Le-Ngoc, and D. H. N. Nguyen, "Two-timescale hybrid RF-baseband precoding with MMSE-VP for multi-user massive MIMO broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4462–4476, Jul. 2018.

[17] K. N. R. S. V. Prasad, E. Hossain, and V. K. Bhargava, "Energy efficiency in massive MIMO-based 5G networks: Opportunities and challenges," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 86–94, Jun. 2017.

[18] T. M. Nguyen, V. N. Ha, and L. B. Le, "Resource allocation optimization in multi-user multi-cell massive MIMO networks considering pilot contamination," *IEEE Access*, vol. 3, pp. 1272–1287, 2015.

[19] J. Du, W. Xu, H. Shen, X. Dong, and C. Zhao, "Hybrid precoding architecture for massive multiuser MIMO with dissipation: Sub-connected or fully-connected structures?" *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5465–5479, Aug. 2018.

[20] I. Ahmed *et al.*, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart. 2018.

[21] X. Gao, L. Dai, S. Han, C.-L. I, and R. W. Heath, "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.

[22] D. Zhang, Y. Wang, X. Li, and W. Xiang, "Hybridly connected structure for hybrid beamforming in mmWave massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 662–674, Feb. 2018.

[23] R. Zi, X. Ge, J. Thompson, C.-X. Wang, H. Wang, and T. Han, "Energy efficiency optimization of 5G radio frequency chain systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 758–771, Apr. 2016.

[24] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, “Codebook-based hybrid precoding for millimeter wave multiuser systems,” *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5289–5304, Oct. 2017.

[25] 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)*, Paris, France, May 2017, pp. 1–6.

[26] R. Méndez-Rial, C. Rusu, A. Alkhateeb, N. Gonzalez-Prélcic, and R. W. Heath, “Channel estimation and hybrid combining for mmWave: Phase shifters or switches?” in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2015, pp. 90–97.

[27] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Upper Saddle River, NJ, USA: Pearson Education, 2014.

[28] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Boston, MA, USA: Cambridge Univ. Press, 2016.

[29] W. Dinkelbach, “On nonlinear fractional programming,” *Bull. Aust. Math. Soc.*, vol. 13, pp. 492–498, Mar. 1967.

[30] J. P. Crouzeix, J. A. Ferland, and S. Schaible, “An algorithm for generalized fractional programs,” *J. Optim. Theory Appl.*, vol. 47, no. 1, pp. 35–49, Sep. 1985.

[31] Z. Dai, Y. Wu, F. Zhang, and H. Wang, “A novel fast method for L_{inf} problems in multiview geometry,” in *Proc. 12th Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 116–129.

[32] C. Wu, Q. Shi, C. He, and Y. Chen, “Energy utilization efficient frame structure for energy harvesting cognitive radio networks,” *IEEE Wireless Commun. Lett.*, vol. 5, no. 5, pp. 488–491, Oct. 2016.

[33] P. Poshala, K. K. Rushil, and R. Gupta, “Signal chain noise figure analysis,” Texas Instrum., Dallas, TX, USA, Rep. SLAA652, Oct. 2014.

[34] J. Zhao, T. Q. S. Quek, and Z. Lei, “Coordinated multipoint transmission with limited backhaul data transfer,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.

[35] B. Dai and W. Yu, “Sparse beamforming and user-centric clustering for downlink cloud radio access network,” *IEEE Access*, vol. 2, pp. 1326–1339, 2014.

[36] E. Candes, M. Wakin, and S. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *J. Fourier Anal. Appl.* vol. 14, no. 5, pp. 877–905, Dec. 2008.

[37] S. S. Christensen, R. Argawal, E. de Carvalho, and J. M. Cioffi, “Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

[38] V. N. Ha, D. H. N. Nguyen, and J.-F. Frigon, “Subchannel allocation and hybrid precoding in millimeter-Wave OFDMA systems,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5900–5914, Sep. 2018.

[39] V. N. Ha, D. H. N. Nguyen, and J.-F. Frigon, “Joint subchannel allocation and hybrid precoding design for mmwave multi-user OFDMA systems,” in *Proc. IEEE 28th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun.*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.

[40] S. Sun *et al.*, “Propagation path loss models for 5G urban micro- and macro-cellular scenarios,” in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC-Spring)*, Nanjing, China, May 2016, pp. 1–6.

[41] W. Li, Y. Chiang, J. Tsai, H. Yang, J. Cheng, and T. Huang, “60-GHz 5-bit phase shifter with integrated VGA phase-error compensation,” *IEEE Trans. Microw. Theory Techn.*, vol. 61, no. 3, pp. 1224–1235, Mar. 2013.

[42] V. N. Ha, D. H. N. Nguyen, and J.-F. Frigon, “Energy-efficient hybrid precoding for mmWave multi-user systems,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[43] R. Cory and D. Fryklund, “Solid state RF/microwave switch technology: Part 2,” in *Microw. Products Dig.*, Jun. 2009, pp. 34–66.

[44] V. Jamali, A. M. Tulino, G. Fischer, R. Mller, and R. Schober, “Scalable and energy-efficient millimeter massive MIMO architectures: Reflect-array and transmit-array antennas,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, Feb. 2019. pp. 1–7.

[45] H. Yan, S. Ramesh, T. Gallagher, C. Ling, and D. Cabric, “Performance, power, and area design trade-offs in millimeter-wave transmitter beamforming architectures,” *IEEE Circuits Syst. Mag.*, vol. 19, no. 2, pp. 33–58, 2nd Quart., 2019.

[46] D. Hunter and K. Lange, “A tutorial on MM algorithms,” *Amer. Stat.* vol. 58, no. 1, pp. 30–37, 2004.

[47] Y. Sun, P. Babu, and D. P. Palomar, “Majorization-minimization algorithms in signal processing, communications, and machine learning,” *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[48] V. N. Ha, L. B. Le, and N.-D. Dáo, “Coordinated multipoint transmission design for cloud-RANs with limited fronthaul capacity constraints,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7432–7447, Sep. 2016.

[49] V. N. Ha, L. B. Le, and N.-D. Dáo, “Energy-efficient coordinated transmission for cloud-RANs: Algorithm design and trade-off,” in *Proc. IEEE 48th Annu. Conf. Inf. Sci. Syst. (CISS)*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.

[50] M. Figueiredo, J. Bioucas-Dias, and R. Nowak, “Majorization–minimization algorithms for wavelet-based image restoration,” *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, Dec. 2007.

Vu Nguyen Ha (Member, IEEE) received the B.Eng. degree from the French Training Program for Excellent Engineers in Vietnam, Ho Chi Minh City University of Technology, Vietnam, an Addendum degree from the École Nationale Supérieure des Télécommunications de Bretagne–Groupe des École des Télécommunications, Bretagne, France, in 2007, and the Ph.D. degree from the Institut National de la Recherche Scientifique–Énergie, Matériaux et Télécommunications, Université du Québec, Montréal, QC, Canada, in 2017. From 2008 to 2011, he was a Research Assistant with the School of Electrical Engineering, University of Ulsan, Ulsan, South Korea. He was a Postdoctoral Fellow with the École Polytechnique de Montréal, Montréal, from 2016 to 2019. His research interests include radio resource management and emerging enabling technologies for 5G wireless systems with special emphasis on HetNets, C-RANs, massive MIMO communications, MEC, 5G-NR, WiFi 802.11ax. He was a recipient of the FRQNT Postdoctoral Fellowship for International Researcher (PBEEE).

Duy H. N. Nguyen (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from the Swinburne University of Technology, Hawthorn, VIC, Australia, in 2005, the M.Sc. degree in electrical engineering from the University of Saskatchewan, Saskatoon, SK, Canada, in 2009, and the Ph.D. degree in electrical engineering from McGill University, Montréal, QC, Canada, in 2013. From 2013 to 2015, he held a joint appointment as a Research Associate with McGill University and a Postdoctoral Research Fellow with the Institut National de la Recherche Scientifique, Université du Québec, Montréal. He was a Research Assistant with the University of Houston in 2015 and a Postdoctoral Research Fellow with the University of Texas at Austin in 2016. Since 2016, he has been an Assistant Professor with the Department of Electrical and Computer Engineering, San Diego State University, San Diego, CA, USA. His current research interests include resource allocation in wireless networks, signal processing for communications, convex optimization, and game theory. He was a recipient of the Australian Development Scholarship, the FRQNT Doctoral Fellowship and Postdoctoral Fellowship, and the NSERC Postdoctoral Fellowship.

Jean-François Frigon (Senior Member, IEEE) received the B.Eng. degree from the École Polytechnique de Montréal, Montréal, QC, Canada, in 1996, the M.A.Sc. degree from the University of British Columbia, Vancouver, BC, Canada, in 1998, and the Ph.D. degree from the University of California at Los Angeles, in 2004. From 2001 to 2003, he worked as the Director of wireless communications systems with Innovics Wireless. He joined the Electrical Engineering Department, École Polytechnique de Montréal, Montréal, in 2004, where he is currently a Full Professor. His research interests include wireless networks, MAC and link layer protocols, MIMO communication systems, reconfigurable antennas, cognitive radios, and cross-layer design.