Subchannel Allocation and Hybrid Precoding in Millimeter-Wave OFDMA Systems

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

***Abstract*— Constrained by the number of transmitted data streams, this paper proposes sub-carrier allocation (SA) and hybrid precoding (HP) designs for sum-rate maximization in mm-wave OFDMA systems. The optimization is first formulated as a computation sparsity-constrained HP design problem, which is non-convex and challenging to solve. Two two-stage solution approaches are proposed. In the first approach, a fully digital precoder (FDP) is optimized considering the computation sparsity constraint in the first stage. In the second approach, the sparsity constraint is only imposed in the second stage. To find the FDP, we employ the minimization of the weighted mean-squared error and the $\ell 1$-reweighted methods to tackle the non-convex objective function and sparsity constraints, respectively. In the second stage of each approach, we exploit an alternating weighted mean-squared error minimization algorithm to reconstruct HP's based on the FDP found in the first stage. Two novel analog precoding designs, namely semi-definite-relaxation-based and projected-gradient-descent-based, are then proposed to optimize the analog part of the obtained HP's. We also study the impacts of various system parameters on the system sum-rate and provide resource provisioning insights for HP systems. Numerical results show the superior performances of the proposed designs over joint SA and HP benchmark algorithms.**

***Index Terms*— Hybrid precoding, mm-wave, OFDMA, multi-user, resource allocation.**

I. Introduction

THE DEMAND for high capacity in the upcoming fifth-generation (5G) wireless cellular networks drives the search for new radio spectrum resources. In that process, unexploited spectrum in the millimeter-wave (mm-wave) band with GHz of bandwidth is becoming an attractive option. It is therefore expected that mm-wave communications will play a key role in 5G networks [1]–[3]. One major issue of mm-wave communications is the low link budget due to the small antenna apertures in mm-wave systems, compared to that in microwave band systems. Thus, mm-wave systems require large antenna arrays to reap the benefit of beamforming gain and thus to mitigate the high propagation loss [2]. Due to the band's short wavelength, a large number of antenna elements in a small space can be employed for mm-wave systems. In addition, multiple data streams for multiple users can be transmitted via spatial multiplexing which potentially results in a significant improvement in spectral efficiency [4], [5].

In multi-user multiple-input multiple-output systems, downlink multi-user precoding is enabled at the base station by assigning weight vectors to the transmitted signals intended for different users. Specifically, these precoding vectors adjust the magnitude and the phase of users' signals to enable spatial separation and multiplexing of multiple data streams [6]. Multi-user precoding is typically performed in the baseband by a digital signal processing unit. Precoded signals are then converted to analog signals through a radio frequency (RF) chain at each antenna [6]. However, this implementation demands prohibitively high cost and power consumption in systems with large antenna arrays. As a result, such transceiver architecture is not suitable for the current mm-wave mixed-signal hardware technologies [3]–[5].

Recently, hybrid precoding (HP) has been considered as a practical alternative to reduce the complexity of mm-wave systems [3], [5], [8]–[14]. This transceiver architecture requires a lower number of RF chains by accomplishing first a digital precoder at baseband and then an analog precoder at RF domain. RF analog precoding is often implemented with lower cost phase-shifters to constructively co-phase the transmitted RF signals and thus enable array gains with lower cost and complexity than with a fully digital precoder [3], [5], [7]. On the other hand, baseband digital precoding enables multiplexing gain by smartly precoding multiple data streams to the limit on the number of RF chains. In this paper, we focus on the designs on HP in a Multi-User Orthogonal Frequency-Division Multiple Access (MU-OFDMA) mm-wave system that serves multiple users over multiple frequency resources.

A. Related Works

While research on HP for the mm-wave system is plentiful, limited work has been studied for multi-user wideband HP system. The work in [15] has proposed a hybrid beamforming architecture that combines an analog beamforming with array antennas and a digital precoding with multiple RF chains

Manuscript received August 1, 2017; revised February 19, 2018 and May 25, 2018; accepted June 12, 2018. Date of publication July 6, 2018; date of current version September 10, 2018. This work was supported in part by the National Sciences and Engineering Research Council of Canada under Grant RGPIN-2016-06401 and in part by the Startup Fund from San Diego State University. The associate editor coordinating the review of this paper and approving it for publication was L. K. Rasmussen. *(Corresponding author: Vu Nguyen Ha.)*

V. N. Ha and J.-F. 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).

D. 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).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2018.2851230

for single-stream MIMO-OFDM. The objective is to maximize either the signal strength or the sum-rate over different sub-carriers. In [16], HP and beam-switching techniques for OFDM-based wireless personal area networks (WPAN) were investigated. The approach in [16] relied on a predefined beam codebook at the transmitter while trying to optimize the per-sub-carrier beamformers at the receiver. Likewise, the work in [17] proposed a low complexity codebook-based beamforming scheme that consists of multiple levels and level-adaptive antenna selection. The transmit and receive antennas, and the weight vectors in the codebook are then selected to maximize the effective signal-to-noise ratio. In [18], a closed-form solution for fully connected OFDM-based HB system was proposed for frequency-selective mm-wave systems based on which a novel long-term-channel-statistics hybrid sub-array technique is developed. In [19], a practical subspace construction algorithm based on partial channel state information is proposed for a massive MIMO-OFDM system in order to support multiple groups of users. On a different approach, Kwon *et al.* [40] proposes HP designs for MU-OFDMA mm-wave systems by utilizing the signal-to-leakage-plus-noise ratio (SLNR) instead of the signal-to-interference-plus-noise ratio (SINR).

A few existing works have studied user sub-carrier and spatial stream resource allocation problems for MU-OFDMA HB systems. Specifically, [20] considered time-slot allocation for time division multiple access mm-wave WPANs. The study in [21] and [22] investigated the user scheduling problem for downlink multi-user HP massive MIMO systems. The work in [23] proposed resource allocation algorithms to maximize the proportional fairness spectral efficiency under the per sub-carrier power and the beamforming rank constraints.

B. Research Contributions

Precoding design and baseband signal processing require a certain amount of computations at the base station. Meanwhile, mm-wave MU-OFDMA systems can use a large number of sub-carriers and RF chains to transmit to a large number of users over a large bandwidth, which can potentially lead to a number of data streams exceeding the computation capabilities of the base station. To the best of our knowledge, downlink MU-OFDMA HB design with considerations on the limited total number of data streams that can be transmitted has not been studied in existing literature. Extending the preliminary results presented in [39], this paper aims to fill this gap where the joint sub-carrier allocation (SA) and HP design for mm-wave MU-OFDMA system are studied to maximize the system sum-rate under a constraint on the number of data streams. This constraint is set over all sub-carriers for all users to limit the computation cost of the base station. Unfortunately, the constraint presents a major obstacle in solving the optimization since it is composed of integer variables. In this paper, we present multiple novel solution approaches to this difficult non-convex mixed-integer optimization problem. Those novel solutions can also find applications in solving other similar constrained problems. Specifically, the contributions of our work are as follows:

- We first transform the integer constraint into an $\ell 0$-norm form to omit these complex integer variables. We then develop two two-stage solution approaches to solve the considered problem based on the premise that near-optimal HP can be designed by approximating an optimal fully digital design [4], [5]. The difference between the two proposed solution approaches arises from the method to address the $\ell 0$-norm constraint, i.e., the sparsity constraint. Specifically, this constraint is first considered in stage one of the first approach (Approach A) when optimizing the sparse FDP based on which the HP is re-constructed in stage two. In the second approach (Approach B), an FDP is first obtained without imposing the sparsity constraint. The sparse HP is then devised as closely to the FDP as possible by carefully addressing the sparsity constraint in stage two.
- To deal with the sparsity constraint, we develop a general $\ell 1$-norm re-weighted solution by regularizing the sparsity function into an approximated linear form and iteratively solving the approximated problem. In addition, an alternative method is employed for designing the HP in stage two of both approaches. In this method, the optimization of the analog precoding matrix is decoupled into multiple simpler problems to reduce the computation complexity. In addition, two novel analog precoding designs, named "Semi-Definite-Relaxation based" and "Projected-Gradient-Descent based," are proposed.
- For performance evaluation of the developed algorithms, we also present two joint SA and HP algorithms as referencing benchmarks. 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 system sum-rate. In addition, we also study the trade-off between the number of RF chains and that of data streams when the base station computation capacity is limited. This trade-off study can determine an optimal operating point for HP design.

The remaining of this paper is organized as follows. We describe the system model, mm-wave channel model, and formulations of the joint SA and HP design problem in Section II. In Section III, we outline two two-stage solution approaches to the problem. Two algorithmic solutions are then proposed in Sections IV and V. The complexity analysis and the reference joint SA and HP design algorithms are presented in Section VI. Numerical results are presented in Section VII followed by conclusions in Section VIII.

Notations: $(\mathbf{X})'$, $(\mathbf{X})^T$ and $(\mathbf{X})^H$ denote the conjugate, transpose, and conjugate transpose (Hermitian operator) of the matrix $\mathbf{X}$, respectively; $\|\mathbf{x}\|_0$, $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|$ denote the norm-0, norm-1, and Euclidean norm of a vector x, respectively whereas $\|\mathbf{X}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{X}$.

II. System Model

A. Multi-User OFDMA MIMO System Model

Consider the downlink mm-wave MU-OFDMA HB system as illustrated in Fig. 1 where a base station (BS) equipped

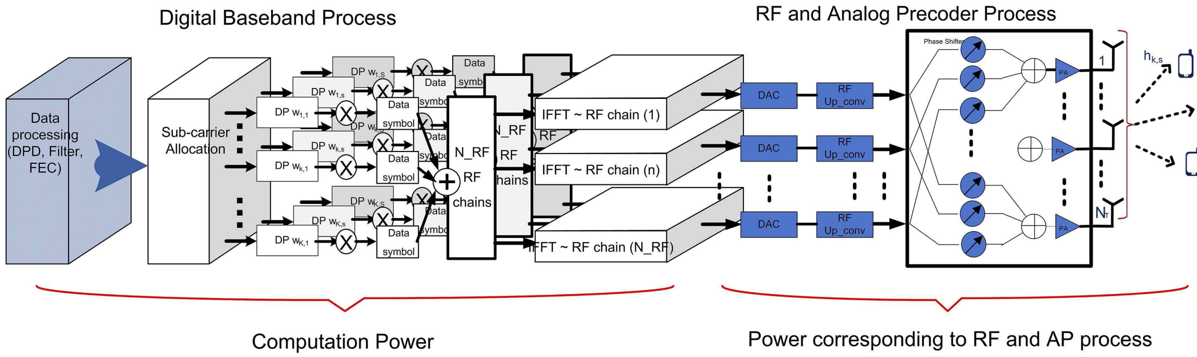

Fig. 1. Diagram of a mm-wave OFDMA multi-user system with sub-carrier allocation and hybrid precoding design.

with $N_{\rm T}$ antennas and $N_{\rm RF}$ RF chains serves K remote single-antenna mobile stations (users) over S sub-carriers. Let $\mathcal{K}$ and $\mathcal{S}$ be the sets of all users and sub-carriers, respectively. If sub-carrier s is assigned to user k, a digital precoding (DP) vector $\mathbf{w}_{k,s} \in \mathbb{C}^{N_{\rm RF}}$ is applied to the data symbol $x_{k,s} \in \mathbb{C}$, intended for user k at this very sub-carrier. Without loss of generality, we assume $\mathbb{E}_{x_{k,s}}\{|x_{k,s}|\} = 1$. Denote $a_{k,s}$ as a binary indicator where

$$a_{k,s} = \begin{cases} 1, \text{ if user } k \text{ is assigned sub-carrier } s, \\ 0, \text{ otherwise.} \end{cases} \tag{1}$$

Utilizing the HP, the BS first applies all DP vector $\mathbf{w}_{k,s}$'s to the corresponding symbol sequences $x_{k,s}$'s. Following the frequency to time domain transformation of the digitally precoded sequences and the RF processing steps, the BS then employs an $N_{\rm T} \times N_{\rm RF}$ analog precoding (AP) matrix $\mathbf{A}$ to map the RF signals from the $N_{\rm RF}$ RF chains to the $N_{\rm T}$ antennas. In this work, we consider a fully-connected RF chains to antennas structure for the AP matrix in which $\mathbf{A}$ is implemented using unit-modulus analog phase shifters, i.e, $\left|(\mathbf{A})_{i,j}\right| = 1 \; \forall (i,j)$. We denote $\mathcal{A}_{\rm R}$ as the set of matrices with all unit-modulus entries.

Assuming coherent detection at the users, the signal-to-interference-plus-noise ratio (SINR) at user k over sub-carrier s is given by [4] and [15]:

$$\text{SINR}_{k,s} = \frac{a_{k,s}\left|\mathbf{h}_{k,s}^H \mathbf{A}\mathbf{w}_{k,s}\right|^2}{\sum_{j\neq k} a_{j,s}\left|\mathbf{h}_{k,s}^H \mathbf{A}\mathbf{w}_{j,s}\right|^2 + \sigma^2}. \tag{2}$$

where σ^2 is the power of additive Gaussian noise at the users and the vector $h_{k,s} \in \mathbb{C}^{N_{\rm T}}$ is the frequency channel for sub-carrier s from the BS to user k, $\forall (k,s)$. Assuming Gaussian signaling between the BS and the users, the achievable data-rate for the transmission to user k over sub-carrier s is then given by $R_{k,s} = \log(1 + \text{SINR}_{k,s})$.

B. Digital Baseband Processing and Problem Formulation

In this section, we discuss the digital baseband process at the base station. Then, we raise a new constraint on the number of served data streams based on which a new HP design problem is formulated. If $a_{k,s} = 1$, the BS will spend a certain amount of computation effort (CE) for the digital baseband process according to the data stream for user k over sub-carrier s. This process is modelled based on the required complexity in *Giga Operation Per Second* (GOPS) which is split into many sub-components, e.g., digital pre-distortion, up/down-sampling and filtering, IFFT and OFDM-specific processing, frequency-domain process scaling and corresponding to the digital precoder $\mathbf{w}_{k,s}$, forward error correction (FEC), and platform control processor (CPU) [41]. In general, the required CE can be quantified as a function of the number of RF chains, modulation bits, coding rate, number of data streams, and number of allocated resource blocks [31]–[33]. Let $C^{\rm eff}$ denote the CE (in GOPS) corresponding to one data stream which is assumed to be the same for all users and all subcarrier. Then, if the system supports D (i.e., $\sum_{(k,s)} a_{k,s} = D$) concurrent data streams, the required computation effort for each transmission is given as $DC^{\rm eff}$. Denote $\kappa_{\rm pe}$ as the power efficiency factor (GOPS/W) representing for the base station's computation ability. Then, the power (in Watt) consumed by the digital baseband process can be estimated as

$$P_{\rm comp} = \frac{DC^{\rm eff}}{\kappa_{\rm pe}}. \tag{3}$$

It is worth noting that D is as large as KS; hence, the power consumption by the digital baseband process for all users over all sub-carrier could be a huge amount. Moreover, numerical results in Section VII indicate that if the number of data streams is large enough, its increment will not result in a higher achievable rate. In addition, spending more computation power for more data streams may reduce the system energy efficiency. Hence, letting the system operate at a rightly tuned number of served data streams is a seasonal choice for the network operators. Considering the limitations on the available computation power in the BS, we are interested in this paper in jointly optimizing the SA and the HP matrix to maximize the system sum-rate under the constraint on the number of transmitted data streams.

Let $P_{\rm T}$ be the transmit power budget at the BS and $\bar{D}$ be the highest number of concurrent data streams which the system can support in one transmission slot. This optimization problem can be stated as

$$\max_{\{a_{k,s}\},\{\mathbf{w}_{k,s}\},\mathbf{A}} \sum_{\forall (k,s)} \log\left(1 + \text{SINR}_{k,s}\right) \tag{4a}$$

$$\text{s. t.} \quad \mathbf{A} \in \mathcal{A}_{\rm R}, \tag{4b}$$

$$\sum_{\forall(k,s)} |a_{k,s}|^2 \mathbf{w}_{k,s}^H \mathbf{A}^H \mathbf{A} \mathbf{w}_{k,s} \le P_{\mathrm{T}}, \tag{4c}$$

$$\sum_{\forall(k,s)} a_{k,s} \le \bar{D}, \tag{4d}$$

Remark 1: The problem (4) *with the sparsity constraint* (4d) *can be considered as a general form of sparse HP design, and its solutions, which are proposed in this paper, can be employed for some specific network scenarios without considering the limited computational capacity. A such scenario is user-scheduling design where the number of users is large, i.e., $K > N_{\mathrm{RF}}$, and the number of users allocated the same sub-carrier should be limited due to the lack of degree-of-freedom. In this scenario, the sparsity constraint over sub-carrier s can be formulated, i.e., $\sum_{\forall k} a_{k,s} \le N_{\mathrm{RF}}$.*

C. mm-wave Channel Model

The mm-wave channel is generally not rich in scattering because mm-wave signals do not reflect well in the surrounding environment. There are thus only few dominant paths in mm-wave transmission channel. In this paper, we adopt the extended Saleh-Valenzuela geometric channel model for the numerical evaluation of MU-OFDMA mm-wave system as in [8]. Specifically, the channel $\mathbf{h}_{k,s} \in \mathbb{C}^{N_{\mathrm{T}}}$ is modelled as

$$\mathbf{h}_{k,s} = \sqrt{\frac{N_{\mathrm{T}}}{P_{\mathrm{L}} CL}} \sum_{c=1}^{C} \sum_{\ell=1}^{L} \alpha_{c,\ell} a_{\mathrm{r}}(\phi^{\mathrm{r}}_{c,\ell}, \theta^{\mathrm{r}}_{c,\ell}) \mathbf{a}_{\mathrm{t}}(\phi^{\mathrm{t}}_{c,\ell}, \theta^{\mathrm{t}}_{c,\ell}) e^{\frac{-j2\pi cs}{S}}, \tag{5}$$

where P_{L} is the pathloss, and C and L are the number of clusters and number of propagation subpaths in each cluster, respectively. In addition, $\alpha_{c,\ell}$ is the complex gain of the ℓ-th path of cluster c, and $(\phi^{\mathrm{r}}_{c,\ell}, \theta^{\mathrm{r}}_{c,\ell})$ and $(\phi^{\mathrm{t}}_{c,\ell}, \theta^{\mathrm{t}}_{c,\ell})$ are its (azimuth, elevation) angles of arrival and departure corresponding, respectively. Then, $a_{\mathrm{r}}(\phi^{\mathrm{r}}_{c,\ell}, \theta^{\mathrm{r}}_{c,\ell})$ and $\mathbf{a}_{\mathrm{t}}(\phi^{\mathrm{t}}_{c,\ell}, \theta^{\mathrm{t}}_{c,\ell})$ represent the normalized receive response factor and transmit array response vectors at (azimuth, elevation) angles of $(\phi^{\mathrm{r}}_{c,\ell}, \theta^{\mathrm{r}}_{c,\ell})$ and $(\phi^{\mathrm{t}}_{c,\ell}, \theta^{\mathrm{t}}_{c,\ell})$, respectively. Finally, $\alpha_{c,\ell}$ is assumed to be i.i.d. Gaussian distributed and the normalization factor $\sqrt{N_{\mathrm{T}}/P_{\mathrm{L}}CL}$ is added to enforce $\mathbb{E}_{\mathbf{h}_{k,s}}\{\|\mathbf{h}_{k,s}\|_2^2\} = N_{\mathrm{T}}/P_{\mathrm{L}}$.

Remark 2: In this paper, we assume that perfect channel state information (CSI) between BS and all users is available at the BS. In practice, the CSI can be estimated by all users or by BS. Detailed realization of CSI channel estimation varies depending if the time division duplex (TDD) [43] or frequency division duplex (FDD) [44] strategy is employed. We do not explicitly consider the error of CSI estimation in the scope of this paper, which is left for our future works.

III. Two-Stage Based Solution Frameworks

The main challenge for solving problem (4) comes from the binary variables $\{a_{k,s}\}$'s associated with the computation complexity constraint. To overcome this challenge, let us first denote the new variable $\tilde{\mathbf{w}}_{k,s}$ as $a_{k,s}\mathbf{w}_{k,s}$, then we examine the relation between these new variables and the binary variables $\{a_{k,s}\}$'s. Denote $\tilde{p}_{k,s} = \tilde{\mathbf{w}}_{k,s}^H \tilde{\mathbf{w}}_{k,s} = |a_{k,s}|^2 \mathbf{w}_{k,s}^H \mathbf{w}_{k,s}$ as the transmission power for user k over sub-carrier s. Certainly, $\tilde{p}_{k,s} = 0$ implies that user k does not utilize sub-carrier s; hence, $a_{k,s} = 0$. In contrast, $\tilde{p}_{k,s} > 0$ means that $a_{k,s} = 1$. Therefore, the total number of transmitted data streams can be written mathematically as $\|\tilde{\mathbf{p}}\|_0$ where $\tilde{\mathbf{p}} = [\tilde{p}_{1,s}, \ldots, \tilde{p}_{k,s}, \ldots, \tilde{p}_{K,S}]$ is the vector representing the power of all DP vectors. Thus, the inequality (4d) can be transformed into a norm $\ell 0$ constraint as

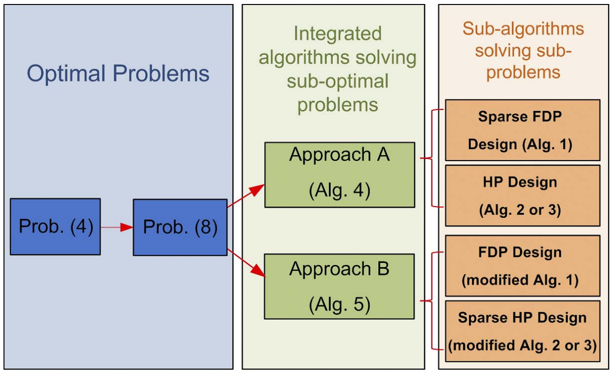
Optimal Problems
Prob. (4)
Prob. (8)
Integrated algorithms solving sub-optimal problems
Approach A (Alg. 4)
Approach B (Alg. 5)
Sub-algorithms solving sub-problems
Sparse FDP Design (Alg. 1)
HP Design (Alg. 2 or 3)
FDP Design (modified Alg. 1)
Sparse HP Design (modified Alg. 2 or 3)

Fig. 2. Diagram of two two-stage solution approaches.

$$\|\tilde{\mathbf{p}}\|_0 \le \bar{D}. \tag{6}$$

Furthermore, $a_{k,s}$ and $\mathbf{w}_{k,s}$ can be defined based on $\tilde{\mathbf{w}}_{k,s}$ as follows.

$$\begin{cases} a_{k,s} = 0 \text{ and } \mathbf{w}_{k,s} = \mathbf{0}, & \text{if } \tilde{p}_{k,s} = 0, \\ a_{k,s} = 1 \text{ and } \mathbf{w}_{k,s} = \tilde{\mathbf{w}}_{k,s}, & \text{if } \tilde{p}_{k,s} > 0. \end{cases} \tag{7}$$

Then, problem (4) can be rewritten as follows:

$$\max_{\{\tilde{\mathbf{w}}_{k,s}\},\mathbf{A}} \sum_{\forall(k,s)} \log\left(1 + \mathrm{SINR}_{k,s}\right) \quad \text{s. t. (4b), (4c), (6).} \tag{8}$$

Many studies [5], [8] have shown that HP can be obtained by employing a framework consisting of two main stages as follows: (i) obtaining the FDP in stage one, (ii) re-constructing a near-optimal performance HP from the FDP in stage two [5]. Based on this framework and the complex sparsity constraint in (6), we propose as shown in Fig. 2 two solution approaches for solving problem (8) which differ on whether the sparsity constraint is taken into account in the stage one FDP design (Approach A) or in stage two HP design (Approach B). In the following we present the optimization problems corresponding to each approach. Algorithms for solving both approaches are given in the subsequent sections.

A. Approach A: Sparse-FDP-Based Design

In this approach, we aim to optimize the sparse FDP with constraint (6) being imposed, then reconstruct HP without that constraint.

1) Stage One – Sparse FDP Optimization: Let $\mathbf{u}_{k,s} = \mathbf{A}\tilde{\mathbf{w}}_{k,s}$ be the actual FDP for user k over sub-carrier s where $\mathbf{u}_{k,s} \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$. Clearly, if $\tilde{p}_{k,s} = 0$ (or > 0), we also have $p^{\mathsf{F}}_{k,s} = \mathbf{u}_{k,s}^H \mathbf{u}_{k,s} = 0$ (or > 0). Hence, the sparse FDP optimization problem can be expressed as

$$\max_{\{\mathbf{u}_{k,s}\}} \sum_{\forall(k,s)} \log\left(1+\frac{\left|\mathbf{h}_{k,s}^H\mathbf{u}_{k,s}\right|^2}{\sum_{j\neq k}\left|\mathbf{h}_{k,s}^H\mathbf{u}_{j,s}\right|^2+\sigma^2}\right)$$
$$\text{s. t.} \sum_{\forall(k,s)} \mathbf{u}_{k,s}^H\mathbf{u}_{k,s} \le P, \quad \|\mathbf{p}^{\mathsf{F}}\|_0 \le \bar{D}. \tag{9}$$

where $\mathbf{p}^{\mathsf{F}} = [p_{1,s}^{\mathsf{F}},\ldots,p_{k,s}^{\mathsf{F}},\ldots,p_{K,S}^{\mathsf{F}}]$.

2) Stage Two – HP Design: Let $\mathbf{u}_{k,s}^{\mathrm{Spar}}$ be the optimal sparse FDP based on which the HP will be reconstructed. $\tilde{\mathbf{w}}_{k,s}$'s and $\mathbf{A}$ can then be approximated via a minimum mean square error (MMSE) approximation as follows.

$$\min_{\{\tilde{\mathbf{w}}_{k,s}\},\mathbf{A}} \sum_{\forall(k,s)} \left\|\mathbf{u}_{k,s}^{\mathrm{Spar}} - \mathbf{A}\tilde{\mathbf{w}}_{k,s}\right\|_2^2 \quad \text{s.t. (4b), (4c).} \tag{10}$$

B. Approach B: Sparse-HP-Based Design

In Approach B, the FDP in stage one is optimized without imposing the sparsity constraint (6). Instead, the constraint is taken into account when reconstructing the HP in the second stage.

1) Stage One – Traditional FDP Optimization: In stage one, we optimize the general FDP by solving the following traditional max-sum-rate problem:

$$\max_{\{\mathbf{u}_{k,s}\}} \sum_{\forall(k,s)} \log\left(1+\frac{\left|\mathbf{h}_{k,s}^H\mathbf{u}_{k,s}\right|^2}{\sum_{j\neq k}\left|\mathbf{h}_{k,s}^H\mathbf{u}_{j,s}\right|^2+\sigma^2}\right)$$
$$\text{s.t.} \sum_{\forall(k,s)} \mathbf{u}_{k,s}^H\mathbf{u}_{k,s} \le P_{\mathrm{T}}. \tag{11}$$

2) Stage Two – Sparse HP Design: Let $\mathbf{u}_{k,s}^{\mathrm{opt}}$ be the outcome of problem (11). Now, a sparse HP, including $\tilde{\mathbf{w}}_{k,s}$'s and $\mathbf{A}$, is then optimized by solving the MMSE problem

$$\min_{\{\tilde{\mathbf{w}}_{k,s}\},\mathbf{A}} \sum_{\forall(k,s)} \left\|\mathbf{u}_{k,s}^{\mathrm{opt}} - \mathbf{A}\tilde{\mathbf{w}}_{k,s}\right\|_2^2 \quad \text{s.t. (4b), (4c), (6).} \tag{12}$$

Remark 3: It is worth noting that the proposed solutions and algorithms can be employed for the joint SA and HP design without considering the sparsity constraint (6). *This can be done by solving problem* (11) *and problem* (10) *consecutively. The SA can then be obtained based on the outcome solution* $\tilde{\mathbf{w}}_{k,s}$*'s.*

IV. Approach A: Sparse-Fully-Digital-Precoder Based Design

In this section, we propose two algorithms solving problems (9) and (10) for Approach A in order to design the sparse FDP and HP, respectively.

A. Stage One: Sparse Fully Digital Precoder Design

In this section, we use a compressed-sensing-based method [24], [25] to deal with problem (9). The main idea behind this method is the iterative update of the weights of the $\ell 1$-norm elements to approximate the complex $\ell 0$-norm form. First, we approximate the $\ell 0$-norm of $\mathbf{p}^{\mathsf{F}}$ as $\|\mathbf{p}^{\mathsf{F}}\|_0 \approx \sum_{\forall(k,s)} f_{\mathsf{apx}}^{(k,s)}(p_{k,s}^{\mathsf{F}})$ where $f_{\mathsf{apx}}^{(k,s)}(p_{k,s}^{\mathsf{F}})$ is the concave function that approximates the step function of $p_{k,s}^{\mathsf{F}}$. Then, problem (9) can be rewritten as

$$\max_{\{\mathbf{u}_{k,s}\}} \sum_{\forall(k,s)} \log\left(1+\frac{\left|\mathbf{h}_{k,s}^H\mathbf{u}_{k,s}\right|^2}{\sum_{j\neq k}\left|\mathbf{h}_{k,s}^H\mathbf{u}_{j,s}\right|^2+\sigma^2}\right) \tag{13a}$$
$$\text{s. t.} \sum_{\forall(k,s)} \mathbf{u}_{k,s}^H\mathbf{u}_{k,s} \le P_{\mathrm{T}} \tag{13b}$$
$$\sum_{\forall(k,s)} f_{\mathsf{apx}}^{(k,s)}(p_{k,s}^{\mathsf{F}}) \le \bar{D}. \tag{13c}$$

This approximated problem is still non-convex because of the non-concave objective function and the non-convex constraint (13c). We first tackle the latter obstacle by transforming the constraint (13c) into a linear form using its duality function [26]. Specifically, let $f_{\mathsf{cnj}}^{(k,s)}(z)$ be the conjugate function of $f_{\mathsf{apx}}^{(k,s)}(w)$, we can describe $f_{\mathsf{apx}}^{(k,s)}(p_{k,s}^{\mathsf{F}})$ as

$$\begin{aligned} f_{\mathsf{apx}}^{(k,s)}(p_{k,s}^{\mathsf{F}}) &\triangleq \inf_{z_{k,s}}\left[z_{k,s}p_{k,s}^{\mathsf{F}} - f_{\mathsf{cnj}}^{(k,s)}(z_{k,s})\right] \\ &= \hat{z}_{k,s}p_{k,s}^{\mathsf{F}} - f_{\mathsf{cnj}}^{(k,s)}(\hat{z}_{k,s}), \end{aligned} \tag{14}$$

where $\hat{z}_{k,s}$ is the minimal value of the right-hand-side function, $z_{k,s}p_{k,s}^{\mathsf{F}} - f_{\mathsf{cnj}}^{(k,s)}(z_{k,s})$, which can be expressed as

$$\hat{z}_{k,s} = \nabla f_{\mathsf{apx}}^{(k,s)}(w)|_{w=p_{k,s}^{\mathsf{F}}}. \tag{15}$$

Hence, for given $\hat{z}_{k,s}$'s, problem (13) can be approximated to

$$\max_{\{\mathbf{u}_{k,s}\}} \sum_{\forall(k,s)} \log\left(1+\frac{\left|\mathbf{h}_{k,s}^H\mathbf{u}_{k,s}\right|^2}{\sum_{j\neq k}\left|\mathbf{h}_{k,s}^H\mathbf{u}_{j,s}\right|^2+\sigma^2}\right) \tag{16a}$$
$$\text{s. t. constraint (13b),}$$
$$\sum_{\forall(k,s)} \hat{z}_{k,s}\mathbf{u}_{k,s}^H\mathbf{u}_{k,s} \le Z, \tag{16b}$$

where $Z = \bar{D} + \sum_{\forall(k,s)} f_{\mathsf{cnj}}^{(k,s)}(\hat{z}_{k,s})$. Interestingly, problem (16) is now the well-known sum-rate maximization problem with multiple quadratic constraints. Some solution approaches have been proposed to solve this problem in the literature, such as Dirty Paper Coding (DPC) method [45], weighted minimum sum-mean square error (WMMSE) [27], Zero-Forcing Water-Filling (ZF-WF) [46]. Although DPC method is reported as a high-capacity-enhancing strategy for multiuser precoding; unfortunately, it is hard to be implemented due to its high complexity involving random nonlinear coding [47]. While ZF-WF strategy is simple to implement, it can only be employed in the scenario that $M \le N_{\mathrm{T}}$ and its performance is poor, especially at low signal-to-noise ratio region [47]. Thus, this paper presents a new algorithm based on the WMMSE method to solve the problem (16). The modified version of ZF-WF method for solving problem (16) will be presented in Appendix F as a benchmark solution for comparison purposes.

1) Proposed WMMSE Precoder: In this section, we address the non-convex sum-rate maximization problem (16) by relating it to a weighted sum-mean square error (MSE) minimization problem as mentioned in the following Proposition.

Proposition 1: The sum-rate maximization problem (16) is equivalent to the following weighted sum-MSE minimization problem

$$\min_{\{\mathbf{u}_{k,s},\delta_{k,s},\omega_{k,s}\}} \sum_{\forall(k,s)} \Big(\omega_{k,s}\mathbb{E}_{x_{k,s}}\Big[\big|x_{k,s}-\delta_{k,s}y_{k,s}\big|^2\Big]-\log\omega_{k,s}\Big)$$
$$\text{s.t.}\quad \textit{(13b) and (16b).} \tag{17}$$

where $\omega_{k,s}$ and $\delta_{k,s}$ denote the MSE weight and the receive coefficient for user k over sub-carrier s, respectively.

Proof: The proof for this proposition is similar to that in [27] for the case of a single sum-power constraint. We omit the details for brevity. ■

It is noted that the optimization in problem (17) can be taken over the FDP $\mathbf{u}_{k,s}$'s, the receive coefficients $\delta_{k,s}$'s as well as the weights $\omega_{k,s}$'s. While problem (17) is not *jointly* convex, it is convex over each set of variables $\mathbf{u}_{k,s}$'s, $\delta_{k,s}$'s, and $\omega_{k,s}$'s. Thus, it is possible to solve problem (17) by alternately optimizing over one set of variables while keeping the other two fixed.

For given FDP $\mathbf{u}_{k,s}$'s and $\omega_{k,s}$'s, the receive coefficient $\delta^{\star}_{k,s}$ to minimize the MSE for user k over sub-carrier s is the Wiener filter, i.e., MMSE receiver

$$\begin{aligned}\delta^{\star}_{k,s} &= \arg\min_{\delta_{k,s}} \mathbb{E}_{x_{k,s}}\big\{\,|x_{k,s}-\delta_{k,s}y_{k,s}|^2\,\big\}\\ &= \Big(\sum_{j\in\mathcal{K}}\big|\mathbf{h}^H_{k,s}\mathbf{u}_{j,s}\big|^2+\sigma^2_{k,s}\Big)^{-1}\mathbf{u}^H_{k,s}\mathbf{h}_{k,s}.\end{aligned} \tag{18}$$

Then, fixing $\mathbf{u}_{k,s}$'s and $\delta_{k,s}$'s, the MSE weights $\omega^{\star}_{k,s}$'s can be determined by the unconstrained optimization as follows:

$$\begin{aligned}\omega^{\star}_{k,s} &= \arg\min_{\omega_{k,s}>0}\ \omega_{k,s}e_{k,s}-\log\omega_{k,s}\\ &= e^{-1}_{k,s} = \frac{\sum_{j\in\mathcal{K}}\big|\mathbf{h}^H_{k,s}\mathbf{u}_{j,s}\big|^2+\sigma^2_{k,s}}{\sum_{j\in\mathcal{K}/k}\big|\mathbf{h}^H_{k,s}\mathbf{u}_{j,s}\big|^2+\sigma^2_{k,s}},\end{aligned} \tag{19}$$

where $e_{k,s} = \mathbb{E}_{x_{k,s}}\Big[\big|x_{k,s}-\delta_{k,s}y_{k,s}\big|^2\Big]$. Finally, for given receive coefficients $\delta_{k,s}$'s and MSE weights $\omega_{k,s}$'s, the optimal FDP $\mathbf{u}_{k,s}$'s can be obtained by solving the following Quadratically Constrained Quadratic Program (QCQP):

$$\min_{\{\mathbf{u}_{k,s}\}} \sum_{\forall(k,s)} \mathbf{u}^H_{k,s}\Bigg(\sum_{j\in\mathcal{K}}\omega_{j,s}|\delta_{j,s}|^2\mathbf{h}_{j,s}\mathbf{h}^H_{j,s}\Bigg)\mathbf{u}_{k,s}$$
$$\text{s.t. (13b) and (16b).} \tag{20}$$

This problem can be solved by employing standard optimization tools (e.g. CVX solver [29]). Alternately, its solution can be found by standard Lagrangian duality method [28].

By iteratively updating $\{\mathbf{u}_{k,s},\delta_{k,s},\omega_{k,s}\}$, we obtain the WMMSE FDP. Combined with the iterative approximation of the $\ell 0$-norm by updating $z_{k,s}$'s, the Sparse WMMSE-FDP design for sum-rate maximization is summarized in Algorithm 1. The properties of the converging solution in Algorithm 1 are stated in the following proposition.

Proposition 2: Algorithm 1 has the following properties:

1) *Algorithm 1 converges to a locally optimal solution.*

2) *That obtained solution satisfies all constraints in (13).*

Proof: See Appendix A. ■

Algorithm 1 Iterative Sparse WMMSE-FDP Design

```
Initialize by setting $\mathbf{u}^{(0)}_{k,s} = \theta\mathbf{1}_{N_{\mathrm{T}}\times 1}$ for all $(k,s)\in\mathcal{K}\times\mathcal{S}$, where $\theta$ $(>0)$ is small enough to satisfy the constraint (13c), and set $l_{\mathrm{o}} = l_{\mathrm{i}} = 0$.
repeat
  Calculate $\big\{z^{(l_{\mathrm{o}})}_{k,s}\big\}$ as in (15) and $B^{(l_{\mathrm{o}})}$.
  repeat
    Calculate $\big\{\delta^{(l_{\mathrm{i}})}_{k,s}\big\}$ and $\big\{\omega^{(l_{\mathrm{i}})}_{k,s}\big\}$ as in (18) and (19).
    Solve problem (20) to obtain $\big\{\mathbf{u}^{(l_{\mathrm{o}})}_{k,s}\big\}$ and increase $l_{\mathrm{i}} = l_{\mathrm{i}}+1$.
  until Convergence or a stopping criterion trigger.
  Update $l_{\mathrm{o}} = l_{\mathrm{o}}+1$.
until Convergence.
```

B. Stage Two: Hybrid Precoding Design

In this stage, we exploit an iterative WMMSE algorithm to reconstruct the HP, where one set of digital and analog precoders is optimized alternatively while keeping the others fixed.

1) Digital Precoders Design: For a given analog precoder matrix $\mathbf{A}$, problem (10) can be restated as

$$\min_{\{\tilde{\mathbf{w}}_{k,s}\}} \sum_{\forall(k,s)} \big\|\mathbf{u}^{\mathrm{Spar}}_{k,s}-\mathbf{A}\tilde{\mathbf{w}}_{k,s}\big\|^2_2 \quad \text{s. t. (4c).} \tag{21}$$

When the power constraint (4c) is temporarily removed, the well-known least squares solution can be achieved as

$$\hat{\mathbf{w}}_{k,s} = \mathbf{A}^{\dagger}\mathbf{u}^{\mathrm{Spar}}_{k,s} \tag{22}$$

Then, in order to satisfy the power constraints, the DP vector can be normalized as

$$\tilde{\mathbf{w}}_{k,s} = \frac{\sqrt{P}}{\Big(\sum_{\forall(k,s)}\big\|\mathbf{A}\hat{\mathbf{w}}_{k,s}\big\|^2_2\Big)^{-1/2}}\hat{\mathbf{w}}_{k,s}. \tag{23}$$

2) Analog Precoder Matrix Design: While fixing the DP vector $\{\tilde{\mathbf{w}}_{k,s}\}$'s, the analog precoder matrix $\mathbf{A}$ can be optimized by solving the following problem:

$$\min_{\mathbf{A}} \sum_{\forall(k,s)} \big\|\mathbf{u}^{\mathrm{Spar}}_{k,s}-\mathbf{A}\tilde{\mathbf{w}}_{k,s}\big\|^2_2 \quad \text{s. t.}\ \ |(\mathbf{A})_{i,j}| = 1\ \forall(i,j). \tag{24}$$

This problem is classified as a Unit-modulus Least Square (ULS) type, which is non-convex and NP-hard due to the unit modulus constraints. In [8], two methods, namely *"Manifold Optimization based Alternative Minimization Algorithm"* (MO-AltMin) and *"Alternative Minimization using Phase Extraction"* (PE-AltMin), are proposed to deal with this problem. However, the MO-AltMin algorithm has a relatively high complexity and the PE-AltMin algorithm is not suitable for problem (24) since $\mathbf{u}^{\mathrm{Spar}}_{k,s}$'s are not mutually orthogonal in a multi-user system. To this end, we propose two novel solution frameworks together with two algorithms to solve it.

a) Problem Decoupling: Denote $\mathbf{v}_r(\mathbf{A}) = \mathbf{row}_r(\mathbf{A})^H$, ($\mathbf{v}_r(\mathbf{A}) \in \mathbb{C}^{N_{\mathrm{RF}}}$) where $\mathbf{row}_r(\mathbf{A})$ is the r-th row of $\mathbf{A}$. Then, the cost function of problem (24) can be rewritten as

$$\begin{aligned}\Phi(\mathbf{A}) &= \sum_{r=1}^{N_{\mathrm{T}}} \sum_{\forall(k,s)} \left\| u_{k,s,r}^{\mathrm{Spar}} - \mathbf{w}_{k,s}^H \mathbf{v}_r(\mathbf{A}) \right\|_F^2 \\ &= \sum_{r=1}^{N_{\mathrm{T}}} \left(\mu_r - 2\Re(\boldsymbol{\pi}_r^H \mathbf{v}_r(\mathbf{A})) + \mathbf{v}_r(\mathbf{A})^H \mathbf{\Pi}_r \mathbf{v}_r(\mathbf{A}) \right),\end{aligned} \tag{25}$$

where $u_{k,s,r}^{\mathrm{Spar}}$ denotes the r-th element of vector $\mathbf{u}_{k,s}^{\mathrm{Spar}}$, $\mu_r = \sum_{(k,s)} u_{k,s,r}^{\mathrm{Spar}} u_{k,s,r}^{\mathrm{Spar}\prime}$, $\boldsymbol{\pi}_r = \sum_{(k,s)} u_{k,s,r}^{\mathrm{Spar}\prime} \tilde{\mathbf{w}}_{k,s}$, and $\mathbf{\Pi}_r = \sum_{(k,s)} \tilde{\mathbf{w}}_{k,s} \tilde{\mathbf{w}}_{k,s}^H$. Problem (24) can then be decomposed into N_{T} $(\mathcal{P}_r)$ sub-problems as

$$(\mathcal{P}_r) \quad \min_{\mathbf{v}} \mathbf{v}^H \mathbf{\Pi}_r \mathbf{v} - 2\Re(\boldsymbol{\pi}_r^H \mathbf{v}) \quad \text{s. t. } |(\mathbf{v})_i| = 1, \ \forall i. \tag{26}$$

This problem is a special type of non-convex QCQP, which can be solved by applying the Semi-Definite (rank) Relaxation (SDR) or the projected gradient descent method as following.

b) Semi-Definite Relaxation Method: First, we transfer problem $(\mathcal{P}_r)$ into the Unit-modulus Quadratic Programming (UQP) form based on the following proposition.

Proposition 3: Problem $(\mathcal{P}_r)$ can be recast as the following UQP problem and vice-versa

$$(\mathcal{Q}_r) \quad \min_{\tilde{\mathbf{v}} \in \mathbb{C}^{N_{\mathrm{RF}}+1}} \tilde{\mathbf{v}}^H \mathbf{\Omega}_r \tilde{\mathbf{v}} \quad \textit{s. t. } |(\tilde{\mathbf{v}})_i| = 1 \ \forall i, \tag{27}$$

where $\mathbf{\Omega}_r = [\mathbf{\Pi}_r \ \ - \boldsymbol{\pi}_r; -\boldsymbol{\pi}_r^H \ \ 0]$. Let $\mathbf{v}^{\mathrm{opt}}$ and $\tilde{\mathbf{v}}^{\mathrm{opt}}$ be the optimal solutions of $(\mathcal{P}_r)$ and $(\mathcal{Q}_r)$, respectively. Then, we have $\tilde{\mathbf{v}}^{\mathrm{opt}} = [e^{j\theta^\star} \mathbf{v}^{\mathrm{opt}}; e^{j\theta^\star}]$ where $(\tilde{\mathbf{v}}^{\mathrm{opt}})_{N_{\mathrm{RF}}+1} = e^{j\theta^\star}$.

Proof: See Appendix B. ■

Problem $(\mathcal{P}_r)$ is now ready to be solved by applying Semi-Definite Programming (SDP) tool through the following proposition.

Proposition 4: Problem $(\mathcal{Q}_r)$ is equivalently rewritten as

$$\min_{\mathbf{V}} \operatorname{Tr}(\bar{\mathbf{\Omega}}_r \mathbf{V}) \quad \textit{s. t. } (\mathbf{V})_{i,i} = 1 \ \forall i, \mathbf{V} \succeq 0, \ \operatorname{Rank}(\mathbf{V}) = 1, \tag{28}$$

where $\bar{\mathbf{\Omega}}_r = \mathbf{\Omega}_r - \lambda_r^{\min} \mathbf{I}$ and $\lambda_r^{\min}$ is the minimum eigenvalue of $\mathbf{\Omega}_r$.

Proof: See Appendix C. ■

As given in Proposition 4, $\bar{\mathbf{\Omega}}_r$ is positive semi-definite, which ensures the convexity of the objective function in (28). However, problem (28) is still non-convex because of its rank-one constraint. Therefore, we firstly remove the rank-one constraint, relax problem (28) into a convex SDP, and solve it using standard convex solvers such as the CVX solver [29]. It is noted that the optimal solution to problem (28), named $\mathbf{V}_r^{\mathrm{opt}}$, may not be rank one. Hence, we then estimate a good feasible solution of $(\mathcal{Q}_r)$ from $\mathbf{V}_r^{\mathrm{opt}}$ by performing additional Algorithm 2 **Step 6–10** similar to [34] to extract a rank-one solution. Herein, $\angle(.)$ denotes the element-wise angle of the argument, and $\mathbf{V}_r^{\mathrm{opt}} = \Lambda \Sigma \Lambda^H$ is the eigen-decomposition of $\mathbf{V}_r^{\mathrm{opt}}$. A detailed discussion of the reasoning behind the rank-one modification in Algorithm 2 can be found in [34].

Algorithm 2 SDR-Based Analog Precoder Design

```
1: Initialize: given $\mathbf{u}_{k,s}^{\mathrm{Spar}}$'s and $\mathbf{w}_{k,s}$.
2: for $r = 1$ to $N_{\mathrm{T}}$ do
3:   Solve the SDP problem (28) to achieve solution $\mathbf{V}_r^{\mathrm{opt}}$.
4:   if $\operatorname{Rank}(\mathbf{V}_r^{\mathrm{opt}}) = 1$ then
5:     Find $\tilde{\mathbf{v}}_r$ so that $\tilde{\mathbf{v}}_r \tilde{\mathbf{v}}_r^H = \mathbf{V}_r^{\mathrm{opt}}$.
6:   else if $\operatorname{Rank}(\mathbf{V}_r^{\mathrm{opt}}) > 1$ then
7:     Generate $M$ candidate vectors $\varsigma_m$'s as $\varsigma_m = \Lambda \Sigma^{1/2} \mathbf{t}_m, \ m = 1, \ldots, M$ where each $\mathbf{t}_m$ is independently chosen from a circularly symmetric zero-mean complex Gaussian distribution of unit variance.
8:     Project these $M$ vectors onto the feasible set via $\tilde{\mathbf{v}}_m = e^{j\angle(\varsigma_m)}, m = 1, \ldots, M$.
9:     Set $\tilde{\mathbf{v}}_r$ as the one yielding the smallest cost in $(\mathcal{Q}_r)$.
10:  end if
11:  Return $\mathbf{v}_r(\mathbf{A})$ from $\tilde{\mathbf{v}}_r$ based on Proposition 3.
12: end for
13: Infer $\mathbf{A}$ from all vectors $\mathbf{v}_r(\mathbf{A})$'s where $\mathbf{v}_r^H(\mathbf{A})$ is the $r$-th row of $\mathbf{A}$.
```

Algorithm 3 Projected-Gradient-Descent-Based Analog Precoder Design

```
1: Initialize: Choose step sizes $\alpha_r$'s, and set $\mathbf{v}_r^{(0)} = e^{j\angle(\mathbf{\Pi}_r^\dagger \boldsymbol{\pi}_r)} \ \forall r$.
2: for $r = 1$ to $N_{\mathrm{T}}$ do
3:   Set $n = 0$.
4:   repeat
5:     Calculate $\boldsymbol{\xi}_r^{(n+1)} = \mathbf{v}_r^{(n)} + 2\alpha_r \left( \boldsymbol{\pi}_r - \mathbf{\Pi}_r \mathbf{v}_r^{(n)} \right)$.
6:     Project $\mathbf{v}_r^{(n+1)} = e^{j\angle(\boldsymbol{\xi}_r^{(n+1)})}$.
7:     Update $n = n + 1$.
8:   until Convergence, return $\mathbf{v}_r(\mathbf{A}) = \mathbf{v}_r^{(n)}$.
9: end for
10: Infer $\mathbf{A}$ from all vectors $\mathbf{v}_r(\mathbf{A})$'s where $\mathbf{v}_r^H(\mathbf{A})$ is the $r$-th row of $\mathbf{A}$.
```

c) Projected Gradient Descent Method: Taking a different approach, we propose the projected gradient descent-based method in this section, where the updates of the variables are performed with the unit modulus constraint being intact. Specifically, this solution algorithm has shown superior performances as confirmed by the numerical results in the Section VII. Moreover, the complexity analysis is Section VI shows lower complexity by the latter approach under certain conditions. The details of this method are summarized in Algorithm 3 where the derivative of the objective function in $(\mathcal{P}_r)$ with respected to $\mathbf{v}_r$ can be calculated as $2\left(\boldsymbol{\pi}_r - \mathbf{\Pi}_r \mathbf{v}_r\right)$. In Algorithm 3, α_r's are the step sizes along the opposite directions of the gradient values. In addition, the convergence issue due to the projection onto a unit modulus constraint in **Step 6** can be relieved based on the following proposition.

Proposition 5: Let $\lambda_r^{\max}(\mathbf{\Pi}_r)$ be the maximum eigenvalue of matrix $\mathbf{\Pi}_r$. For any step size satisfying $\alpha_r \leq \dfrac{1}{4\lambda_r^{\max}(\mathbf{\Pi}_r)}$,

Algorithm 4 Approach A: Sparse-FDP-Based Design

```
Initialize: Run Algorithm 1 to obtain $\mathbf{u}^{\mathrm{Spar}}_{k,s}$'s.
Select any $\mathbf{A}^{(0)} \in \mathcal{A}$, set $l = 0$
repeat
    Fix $\mathbf{A}^{(l)}$, update $\tilde{\mathbf{w}}^{(l)}_{k,s} = \mathbf{A}^{(l)\dagger}\mathbf{u}^{\mathrm{Spar}}_{k,s}$ for all $(k,s)$.
    Fix $\tilde{\mathbf{w}}^{(l)}_{k,s}$'s, update $\mathbf{A}^{(l+1)}$ using Algorithm 2 (or Algorithm 3).
    Update $l = l + 1$.
until Convergence or a stopping criterion trigger.
Return $\mathbf{A}^{\star}$ and $\tilde{\mathbf{w}}^{\star}_{k,s}$'s.
Normalize $\tilde{\mathbf{w}}^{\star}_{k,s}$'s as in (23).
Define $a^{\star}_{k,s}$'s and $\mathbf{w}^{\star}_{k,s}$'s based on $\tilde{\mathbf{w}}^{\star}_{k,s}$'s as in (7).
```

the process in ***Step 4–8*** *of Algorithm 3 converges to the KKT point of the non-convex and NP-hard problem* $(\mathcal{P}_r)$.

Proof: See Appendix D. ■

3) Hybrid Precoder Design: Having proposed Algorithms 2 and 3 to obtain the analog precoder, we can integrate them with Algorithm 1 into an algorithm to obtain the HP solving problem (8). In Algorithm 4, problems (21) and (24) are solved iteratively to derive the DP vector and the AP matrix. For ease of reference, if Algorithm 4 is implemented using Algorithm 2, it will be referred to as Algorithm A-2. On the other hand, if Algorithm 3 is used, the overall algorithm is referred to as Algorithm A-3.

Remark 4: The convergence of Algorithm 3 has been analysed in Proposition 5 and its proof in Appendix D also ensures that the MSE between FDP and HP decreases after each iteration of this algorithm (named inner loop). Therefore, the convergence of Algorithm A-3 can be guaranteed due to the fact that the objective function (MSE between FDP and HP) monotone decreases in each iteration.

V. Approach B: Sparse-Hybrid-Precoder-Based Design

In this section, we propose algorithms solving problem (8) for Approach B where we first determine the FDP and then optimize the sparse HP by carefully addressing constraint (6).

A. Stage One: Fully Digital Precoder Design

Here, we tackle the non-convex sum-rate maximization problem (13), without the sparsity constraint (13c) by relating it to the weighted sum-MSE minimization problem as follows:

$$\min_{\{\mathbf{u}_{k,s},\delta_{k,s},\omega_{k,s}\}} \sum_{\forall(k,s)} (\omega_{k,s}e_{k,s} - \log\omega_{k,s}) \quad \text{s.t. (13b).} \tag{29}$$

Similarly to the process given in Section IV-A, this problem can be solved by iteratively updating each set of variables $\mathbf{u}_{k,s}$'s, $\delta_{k,s}$'s and $\omega_{k,s}$'s. Specifically, $\delta_{k,s}$'s, and $\omega_{k,s}$'s can be updated as (18) and (19), respectively, while $\mathbf{u}_{k,s}$'s can be obtained by solving the following QCQP:

$$\min_{\{\mathbf{u}_{k,s}\}} \sum_{\forall(k,s)} \mathbf{u}^{H}_{k,s}\left(\sum_{j\in\mathcal{K}}\omega_{j,s}|\delta_{j,s}|^2\mathbf{h}_{j,s}\mathbf{h}^{H}_{j,s}\right)\mathbf{u}_{k,s} \quad \text{s.t. (13b).} \tag{30}$$

Then, problem (29) can be solved by employing the same Algorithm 1 except for changing problem (20) by (30) in **Step 6** and omitting **Step 3**.

B. Stage Two: Sparse Hybrid Precoding Design

Let $\mathbf{u}^{\mathrm{opt}}_{k,s}$ be the optimum solution of problem (29), which is the outcome of stage one of Approach B. To deal with the sparsity constraint (6), we again employ the adaptive re-weighted ℓ1-norm method as in Section IV-A. In particular, the weight parameters are updated as

$$\hat{t}_{k,s} = \nabla f^{(k,s)}_{\mathsf{apx}}(w)|_{w=\|\mathbf{A}\tilde{\mathbf{w}}_{k,s}\|_2^2}, \tag{31}$$

and the constraint (6) can be approximated by

$$\sum_{(k,s)} \hat{t}_{k,s}\left\|\mathbf{A}\tilde{\mathbf{w}}_{k,s}\right\|_2^2 \le T, \tag{32}$$

where $T = \bar{D} + \sum_{(k,s)} f^{(k,s)}_{\mathsf{cnj}}(\hat{t}_{k,s})$. For a given AP matrix $\mathbf{A}$, the DP vector can then be derived such that the resulting HP is as close as possible to the digital one. In particular, we restate problem (12) as

$$\min_{\{\tilde{\mathbf{w}}_{k,s}\}} \sum_{\forall(k,s)} \left\|\mathbf{u}^{\mathrm{opt}}_{k,s} - \mathbf{A}\tilde{\mathbf{w}}_{k,s}\right\|_2^2 \quad \text{s. t. (4c) and (32).} \tag{33}$$

This problem is a QCQP; hence, it can be solved by employing optimization solver tools CVX. After achieving the sparse DP vectors $\tilde{\mathbf{w}}_{k,s}$'s, we can find the AP matrix based on Algorithm 2 or 3 as in Paragraph IV-B.2. Finally, the sparse HP design using Approach B is summarized in Algorithm 5 in which problems (33) and (24) are solved alternatively. For ease of referencing, if Algorithm 2 is used in step 6 of Algorithm 5, the overall algorithm is referred to as Algorithm B-2. Otherwise, if Algorithm 2 is used, the overall algorithm is referred to as Algorithm B-3. It is worth noting that we can analyse the convergence of the modified version of Algorithm 3 for optimizing AP matrix by simply combining the proofs of Proposition 2 and 5. Similar to Remark 4, the objective function of problem (33) will monotonically decrease after each

Algorithm 5 Approach B: Sparse-HP-based Design

```
1:  Initialize:
    Obtain $\mathbf{u}^{\mathrm{opt}}_{k,s}$'s by solving problem (8).
    Setting $\tilde{\mathbf{w}}^{(0)}_{k,s} = \theta\mathbf{1}_{N_{\mathrm{RF}}\times 1}$ for all $(k,s) \in \mathcal{K}\times\mathcal{S}$, where $\theta$ ($> 0$) is small enough to satisfy the constraint (4c).
    Select any $\mathbf{A}^{(0)} \in \mathcal{A}$, set $l_{\mathrm{o}} = l_{\mathrm{i}} = 0$.
2:  repeat
3:      Calculate $\hat{t}^{(l_{\mathrm{o}})}_{k,s}$ as in (31) and $B^{(l_{\mathrm{o}})}$.
4:      repeat
5:          Fix $\mathbf{A}^{(l_{\mathrm{i}})}$, update $\{\tilde{\mathbf{w}}^{(l_{\mathrm{i}})}_{k,s}\}$ by solving problem (33).
6:          Fix $\tilde{\mathbf{w}}^{(l_{\mathrm{i}})}_{k,s}$'s, update $\mathbf{A}^{(l_{\mathrm{i}}+1)}$ (Algorithm 2 or 3).
7:          Update $l_{\mathrm{i}} = l_{\mathrm{i}} + 1$.
8:      until Convergence or a stopping criterion trigger.
9:      Update $l_{\mathrm{o}} = l_{\mathrm{o}} + 1$.
10: until Convergence.
11: Return $\mathbf{A}^{\star}$ and $\tilde{\mathbf{w}}^{\star}_{k,s}$'s.
12: Define $a^{\star}_{k,s}$'s and $\mathbf{w}^{\star}_{k,s}$'s based on $\tilde{\mathbf{w}}^{\star}_{k,s}$'s as in (7).
```

TABLE I

COMPLEXITIES OF THE PROPOSED ALGORITHMS.

A-2	$L_1^{\rm FDP} X_{\sf Q}(KSN_{\rm T},2)+L_1^{\rm HP}(N_{\rm T}^2+X_2)$
A-3	$L_1^{\rm FDP} X_{\sf Q}(KSN_{\rm T},2)+L_1^{\rm HP}(N_{\rm T}^2+X_3)$
B-2	$L_2^{\rm FDP} X_{\sf Q}(KSN_{\rm T},1)+L_2^{\rm HP}(X_{\sf Q}(KSN_{\rm RF},1)+X_2)$
B-3	$L_2^{\rm FDP} X_{\sf Q}(KSN_{\rm T},1)+L_2^{\rm HP}(X_{\sf Q}(KSN_{\rm RF},1)+X_3)$

iteration of Algorithm B-3 which confirms the convergence of this algorithm.

VI. COMPLEXITY ANALYSIS AND OTHER SOLUTION APPROACHES

A. Complexity Analysis

In this section, we investigate the complexities of our two proposed solution approaches integrated with Algorithms 2 and 3, which are denoted as A-2, A-3, B-2, and B-3.

It is observed that stage one of the two approaches includes solving the QCQP problem (20) multiple times. The number of computations of the second stage depends on the complexity in solving Algorithm 2 or Algorithm 3 for both approaches. In addition, the second approach also involves the QCQP in stage two through dealing with problem (33). As given in [35], the computation number for solving QCQP is summarized as

$$X_{\sf Q}(m,n)=\mathcal{O}(\max(m,n)^4 m^{1/2}\log(\zeta_{\sf QCQP}^{-1})). \tag{34}$$

where m and n are the number of variables and constraints, and $\zeta_{\sf QCQP}$ is the solution accuracy.

Algorithm 2 involves solving problem (28) to achieve solution $\mathbf{V}_r^{\rm opt}$, calculating Λ and Σ, and comparing M random vector $\boldsymbol{\varsigma}_m$. Employing the results in [35] one more time, we can estimate the complexity of Algorithm 2 as follows:

$$\begin{aligned} X_2 &= \mathcal{O}\left(N_{\rm T}\left[X_{\sf Q}(N_{\rm RF},N_{\rm RF})+M(N_{\rm RF}+1)^2\right]\right)\\ &= \mathcal{O}\left(N_{\rm T}\left(N_{\rm RF}^{4.5}\zeta_2^{-1}+N_{\rm RF}^2 M\right)\right), \end{aligned} \tag{35}$$

where ζ_2 is the solution accuracy for solving QCQP problem in Algorithm 2.

As given in the proof of Proposition 5, Algorithm 3 employs the Gradient Descent method for the Lipschitz function. Based on the results in [36], the complexity of Algorithm 3 can be thus expressed as

$$X_3=\mathcal{O}\left(N_{\rm T}N_{\rm RF}^2\zeta_3^{-1}\right), \tag{36}$$

where ζ_3 is the solution accuracy for Algorithm 3. Comparing (35) and (36), the computation of Algorithm 3 is less complex than that of SDR-based Algorithm 2 if $\zeta_3^{-1} < \max(N_{\rm RF}^{2.5}\zeta_2^{-1}, M)$, and vice versa.

We now are ready to analyze the complexities of our proposed algorithms. Let $L_1^{\rm FDP}$, $L_1^{\rm HP}$, $L_2^{\rm FDP}$, and $L_2^{\rm HP}$ as the number of iterations of stage one and stage two due to our two proposed solution approaches, respectively. By considering the number of variables and constraints in problems (20) and (33), the complexities of our proposed algorithms can be estimated as given in Table I.

B. Other Solution Approaches

For comparison purposes, we also consider another approach for solving the optimization problem (4). In this approach, the SA (i.e., $a_{k,s}$'s) is determined first. The HP for a given SA strategy $a_{k,s}$'s is then designed by employing some well-known MU-HP design methods in the literature. Here, there are two ways to indicate the values of $a_{k,s}$'s which are given as follows:

1) Sparse WMMSE-FDP Based Method: The values of $a_{k,s}$'s can be inferred from the Sparse WMMSE-FDP outcome of stage one of the first approach as

$$a_{k,s}^{\star}=\begin{cases}1 & \text{if } \mathbf{u}_{k,s}^{{\rm Spar},H}\mathbf{u}_{k,s}^{\rm Spar}>0,\\ 0 & \text{if } \mathbf{u}_{k,s}^{{\rm Spar},H}\mathbf{u}_{k,s}^{\rm Spar}=0.\end{cases} \tag{37}$$

2) Heuristic Method: In this heuristic algorithm, we start with a uniform power allocation, which means that $\mathbf{w}_{k,s}^H\mathbf{w}_{k,s}=p^{\sf nom}=P_{\rm T}/\bar{D}$, $\forall(k,s)\in\mathcal{K}\times\mathcal{S}$. We assume that the ZF precoding is applied to mitigate the interference over each sub-carrier. Then, the transmission rate $R_{k,s}$ can be upper-bounded as $R_{k,s}\le\log\left(1+\left|\mathbf{h}_{k,s}^H\mathbf{A}\right|^2 p^{\sf nom}/\sigma^2\right)$. For the OPM-HP design method, columns of the AP matrix $\mathbf{A}$ are selected from a pre-determined set of $L_{\rm OMP}$ basis vectors, $\mathcal{V}_{\rm OMP}=\{\mathbf{v}_1,\ldots,\mathbf{v}_{L_{\rm OMP}}\}$. Let $\mathbf{v}_{k,s}^{\star}=\arg\max_{\mathbf{v}\in\mathcal{V}_{\rm OMP}}\left|\mathbf{h}_{k,s}^H\mathbf{v}\right|^2$. Then, we have

$$R_{k,s}\le\bar{R}_{k,s}=\log\left(1+\frac{N_{\rm RF}\left|\mathbf{h}_{k,s}^H\mathbf{v}_{k,s}^{\star}\right|^2 p^{\sf nom}}{\sigma^2}\right) \tag{38}$$

Then, we select the $\bar{D}$ largest values of $\bar{R}_{k,s}$'s and set the $a_{k,s}^{\star}$'s corresponding to those $\bar{D}$ strongest values to ones while keeping others as zeros.

3) Hybrid Precoder Design for Given Sub-Carrier Allocation: After having the SA, we optimize the HP by solving the following problem.

$$\begin{aligned} &\max_{\{\mathbf{w}_{k,s}\},\mathbf{A}}\ \sum_{\forall(k,s)}\log\left(1+\frac{a_{k,s}^{\star}\left|\mathbf{h}_{k,s}^H\mathbf{A}\mathbf{w}_{k,s}\right|^2}{\sum_{j\neq k}a_{j,s}^{\star}\left|\mathbf{h}_{k,s}^H\mathbf{A}\mathbf{w}_{j,s}\right|^2+\sigma^2}\right)\\ &\ \text{s.t.}\ \ \text{(4b), (4c).} \end{aligned} \tag{39}$$

This problem can be considered as a HP design problem for wideband multi-user mm-wave systems which can be solved by employing the HP design method for mm-wave wideband multiuser MIMO-OFDM systems in [40] and OMP-HP method, a well-known MU-HP algorithmic solution in [5] and [7]. Note that, in our channel model, the transmit array vectors are the same for all sub-carriers; hence, the OMP-HP method can be applied to our model without any modification.

VII. SIMULATION RESULTS

In this simulation results section, we illustrate the performance advantages of the proposed OFDMA- mm-wave HP using the two proposed approaches (Algorithms 4 and 5 integrated with either Algorithms 2 or 3, denoted respectively Alg. A-2, A-3, B-2, and B-3) compared to other precoding designs: i.) the Sparse FDP achieved by Algorithm 1 presented in Section IV-A, denoted as "sparse WMMSE-FDP," ii.) the sparse FDP achieved by the modified ZF-WF method given

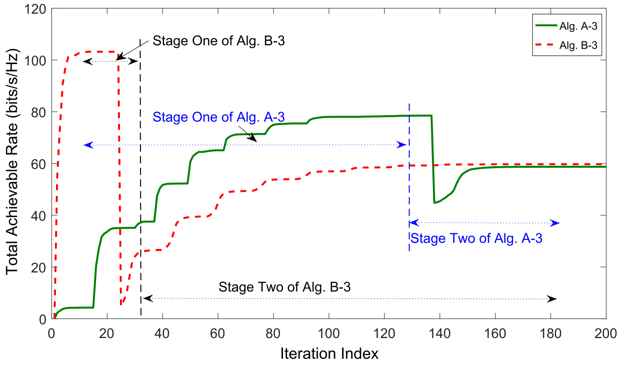

Fig. 3. Total achievable rate *versus* the iteration index.

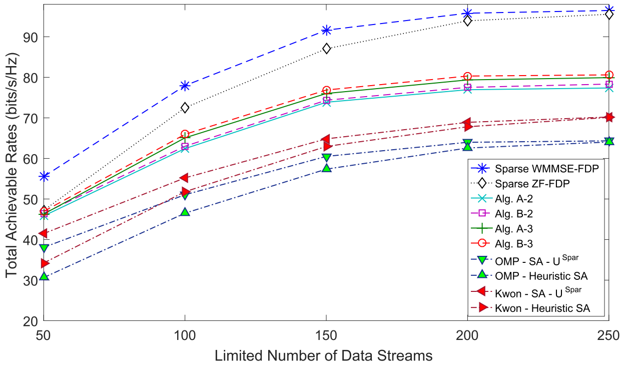

Fig. 4. Total achievable rate *versus* the limited number of data streams.

in Appendix F, denoted as "sparse ZF-FDP," iii.) two OMP algorithms with SA methods given in Section VI-B, denoted OMP – "SA-$\mathbf{U}^{\text{Spar}}$" and "OMP – Heuristic SA," iv) and algorithm given in [40] integrated to two SA methods given in Section VI-B, denoted "Kwon – SA-$\mathbf{U}^{\text{Spar}}$" and "Kwon – Heuristic SA." Here, $\mathbf{U}^{\text{Spar}}$ is the matrix representing all Sparse WMMSE-FDPs $\mathbf{u}_{k,s}^{\text{Spar}}$'s. We consider a MISO system where the BS is equipped with 16×16 UPA ($M = 256$). The channel to each user contains one cluster of 10 paths, i.e., $C = 1$, $L = 10$. 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 $[-\frac{\pi}{2}; \frac{\pi}{2}]$ with an angle spread of $5°$ [10]. The noise variance σ^2 is set at 10^{-13}. In this simulation, we employ the 5G mm-wave path loss model for Austin ($f = 38\ GHz$) as in [37] where all users are randomly located so that their distances to the BS are between 100 m and 200 m. In addition, we set $K = 16$, $N_{RF} = 16$, $\bar{D} = 100$, $P_{\text{T}} = 1.28\ W$, and $S = 32$, unless they are stated otherwise. For implementing Algorithm 2, we choose number of random vectors $\mathbf{t}_m$, M, as 1000. For other algorithms, we choose the solution accuracy as $\zeta_{\mathsf{QCQP}} = \zeta_2 = \zeta_3 = 10^{-3}$.

We illustrate the convergence of our proposed algorithms A-3 and B-3 in Fig. 3 where the variations of total achievable rates achieved by the WMMSE FDP in stage one and the HP in stage two over the iterations are shown. As can be seen, the system achievable rate in each stage of each approach increases over the iterations before reaching its maximum value. Interestingly, algorithm A-3 requires the largest numbers of iterations to deal with the sparsity constraint in stage one and inversely Algorithm B-3 requires the most iterations in stage two. From Table I, we can see that the complexity of each iteration of stage one is much higher than that of stage two. Therefore, the complexity of Approach B is much less than that of Approach A. Due to the fact that the complexity of Algorithm 3 is less than that of Algorithm 2. We can see that Algorithm B-3 has the lowest complexity in comparison to other proposed algorithms.

In Fig. 4, we show the total achievable rate obtained by different schemes versus the limited number of data streams, $\bar{D}$. As can be seen, the "Sparse WMMSE-FDP" is superior the "Sparse ZF-FDP" at all values of $\bar{D}$. This one happens because the directions of FDP vectors according to ZF-WF method are determined and fixed for the scenario that all the users transmit over each sub-carrier, and only the power of the sparse ZF-FDP vectors is updated based on the new value of $\hat{z}_{k,s}$ (Step 3 of Algorithm 1) while all Sparse WMMSE-FDP vectors are updated in each iteration. Hence, the proposed "Sparse WMMSE-FDP" method is better since it enhances the *"degree-of-freedom"* better than "Sparse ZF-FDP" does. However, the gap between two curves representing for these two sparse FDP designing methods becomes smaller when $\bar{D}$ increases. Moreover, the rates achieved by our proposed HP design algorithms are much higher than achieved by the four reference HP algorithms. In addition, the fact that the second approach solution and Algorithm 3 again results in slightly better sum-rate than the other proposed ones confirms their efficiencies. As expected, the system achievable rate increases as limited number of data streams increases. However, in the high regime of $\bar{D}$, the achievable rate of all schemes will be saturated as $\bar{D}$ becomes sufficiently large because of the freedom limitation for designing the DP vector. Additionally, the figure also demonstrates the better performances of HP algorithm in [40] and Sparse WMMSE-FDP based SA method in comparison to OMP method and heuristic SA method, respectively.

Next, we consider the system energy efficiency (SEE) achieved by the various algorithms in Fig. 5. The description of how to define the SEE and the corresponding setting parameters are given in Appendix E. Fig. 5 illustrates the variation of the SEE by different schemes versus the limited number of data streams, $\bar{D}$, where the achievable rates are obtained in Fig. 4 and the SEE is calculated as in (56). As can be observed, the SEE for "Sparse WMMSE-FDP," "Sparse ZF-FDP," "Alg. A-3," "Alg. B-3," "OMP - Heuristic SA," "Kwon - Heuristic SA" schemes increase and then decrease as the number of data stream increases where they achieve their best performances at $\bar{D} = 100$. While the others' SEE results peak at $\bar{D} = 50$, then go down as $\bar{D}$ becomes larger. These results validate the reason why we should limit the number of data streams in the MU-OFDMA mm-wave as we consider in this paper.

Fig. 6 presents the total achievable rate versus the total transmission power of the BS $P_{\max}$. For OMP-HP designs, perfect AoD/AoA codebooks are assumed. As observed from

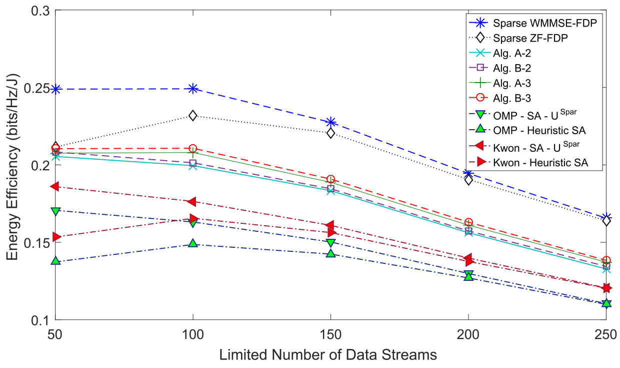

Fig. 5. The system energy efficiency *versus* the limited number of data streams.

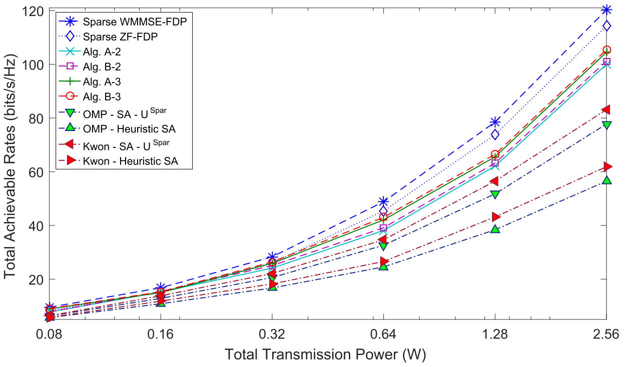

Fig. 6. Total achievable rate *versus* the transmission power.

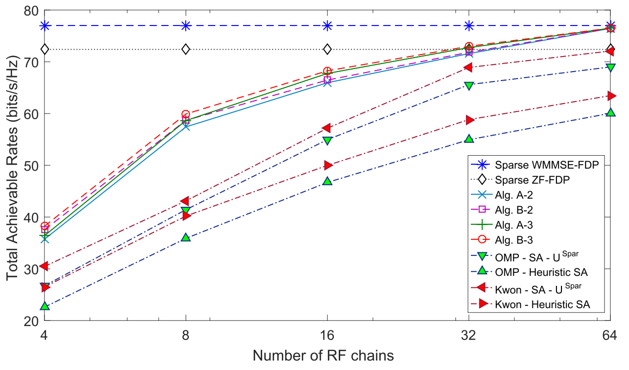

Fig. 7. Total achievable rate *versus* the number of RF chains.

the figure, the achievable rates from all algorithms increase as the transmission power increases. Interestingly, the second design approach surpasses the first approach for both integration with Algorithm 2 or 3, meanwhile Algorithm 3 shows better performance than Algorithm 2 for both proposed approaches. In addition, our proposed HP designs significantly surpass the OMP-HP and Kwon's HP algorithms with the SA achieved from $\mathbf{u}_{k,s}^{\mathrm{Spar}}$'s and heuristic one. Again, the Sparse WMMSE-FDP method shows its superior performance by surpassing the Sparse ZF-FDP at all values of P_{T}.

Next, the impact of the number of RF chains is presented in Fig. 7. Predictably, all schemes except the Sparse WMMSE-FDP upper bound and Sparse ZF-FDP can achieve higher total rate with the increase of the number of RF chains. In addition, our proposed designs again surpass the four reference algorithms. It can also be observed that the total achievable system sum-rate approaches the upper bound as the number of RF chains increases. Moreover, our proposed algorithms can approach the upper bound at the high RF-chain regime for which confirms the excellent performance of our proposed designs.

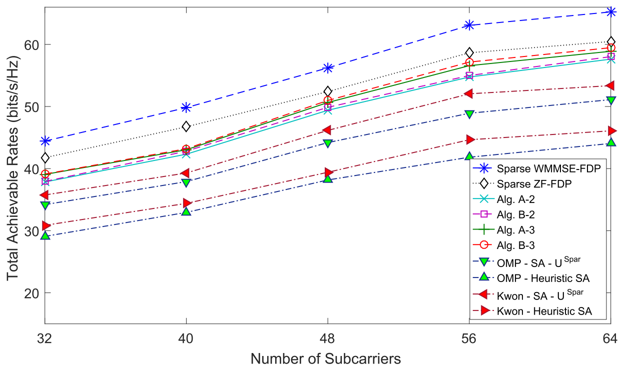

Fig. 8. Total achievable rate *versus* the number of sub-carriers.

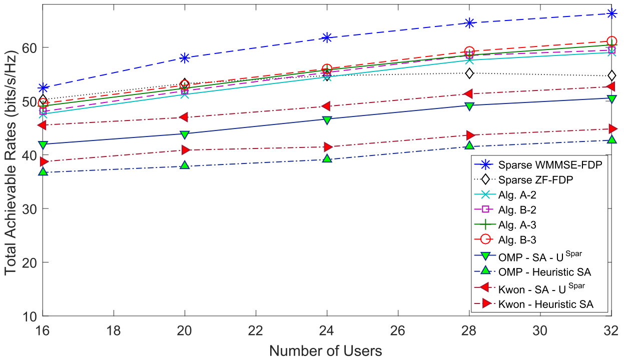

Fig. 9. Total achievable rate *versus* the number of users.

In Figs. 8 and 9, we study the total achievable rates versus the numbers of sub-carriers and users, respectively. As observed from the figures, a larger number of sub-carriers or users results in higher total achievable rate for all schemes except the Sparse ZF-FDP scheme due to the larger number of feasible solutions for sub-carrier-user allocation. Interestingly, when the number of users increases, the rate achieved by the Sparse ZF-FDP scheme increases and then decrease. This can be explained by that the larger number of users results in the higher strictness in designing the "unnormalized" FDP vectors for all users at the first stage of ZF-WF method which can reduce the scheme's performance. Moreover, our proposed algorithms again outperform the other reference algorithms and further confirms their superiority. As can be observed, Algorithm B-3 once again slightly outperforms all other algorithms.

Next, we investigate the network behaviour when the consuming power for computation process of the BS is limited. For illustrative purpose, we assume that the computation power at the BS is 100 W and κ_{pe} is set as 17.009 $GOPS/W$.

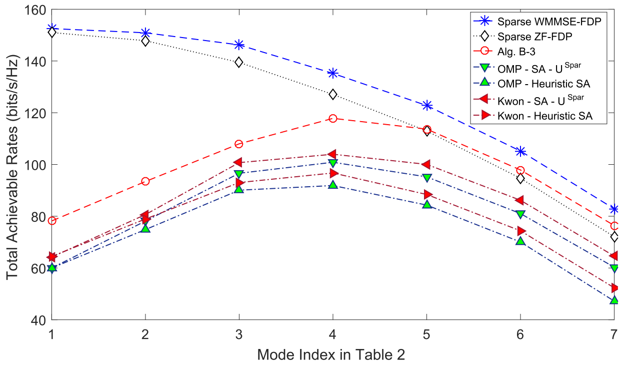

Fig. 10. Total achievable rate *versus* the index of modes in the Table II.

TABLE II

MODE OF $(N_{RF}, \bar{D})$ WHEN $B = 500$ E.G., $\bar{C} = 500 \times 53.328$ GOPS

Mode	1	2	3	4	5	6	7
N_{RF}	4	6	8	10	12	14	16
$\bar{D}$	571	304	189	129	93	70	55

Hence, the total computational capacity is $\bar{C} = 100\times 17.009 = 1700.9$ GOPS. In addition, the number of RF chains can be selected from the set $\{4, 6, 8, 10, 12, 14, 16\}$ and the corresponding limited number of data streams, $\bar{D}$, can be calculated based on the fact that $\bar{D} \times C^{\text{eff}} \le \bar{C}$. Then, all the simulation mode of $(N_{\rm RF}, \bar{D})$ are given in Table II.

Based on the modes given in Table II, we illustrate in Fig. 10 the variation of the total rate achieved by Algorithms Sparse WMMSE-FDP, Sparse ZF-FDP, B-3, and the reference algorithms versus $(N_{\rm RF}, \bar{D})$ represented by the mode index. As can be observed, the total achieved rates for all schemes increase and then decrease as the number of RF chains increases (e.g., the mode index increases). These results can be interpreted as follows. Increasing N_{RF} increases the degrees of freedom for designing the DP vectors; hence, a better performance can be achieved. However, when N_{RF} becomes too large, the number of available data streams must be reduced to avoid violating the constraint of computation capacity of the BS which eventually lead to a degradation of the system achievable rate. The presented trade-off results indicate that one can determine an operating point of the HB system to achieve the best network performance.

VIII. Conclusion

This paper has proposed a new joint SA and HP design for multi-user OFDMA mm-wave systems under a limited number of data streams. We have proposed two novel two-stage based solution approaches to determine the sparse HP which aims to maximize the total achievable rate of the network. Numerical results have illustrated that our proposed algorithms outperform the reference joint SA and HP design algorithms and confirmed the efficiency of our proposed algorithms. In addition, one has shown that Algorithm B-3 (the second approach with projected gradient descent AP design) can provide the best performance where it achieves the highest sum-rate while requiring the lowest computational complexity.

We have also studied the impacts of various parameters on the system sum-rate and relevant performance tradeoffs between the number of RF chains and that of data streams under the constraint of computation capacity of BS. This may lead into interesting directions for future studies on the HP design.

Appendix A
Proof of Proposition 2

Proof of the first statement: Let Ω_{l_o} be the total system rate which is the out-come Algorithm 1 in iteration l_o. Because $z_{k,s}^{(l_o+1)}$ is calculated as in (15) based on the value of $\mathbf{u}_{k,s}^{l_o}$, we must have $\sum_{(k,s)} z_{k,s}^{(l_o+1)} \mathbf{u}_{k,s}^{l_o H} \mathbf{u}_{k,s}^{l_o} \le B$. Therefore, $\mathbf{u}_{k,s}^{l_o}$ is a feasible solution for the max-sum-rate problem for iteration $l_o{+}1$. On the other hand, similar to the spirit of [27], the alternating minimization process in **Step 4–7** of Algorithm 1 results in a monotonic improvement of the objective function of problem (17). In addition, $\mathbf{u}_{k,s}^{l_o}$ is a feasible solution for inside loop (**Step 4–7**) of Algorithm 1. Therefore, we must have

$$\Omega_l \le \Omega_{l+1}, \quad \forall l > 0. \tag{40}$$

Due to the monotonic convergence of the proposed algorithm, the resultant solution must be a locally optimal solution.

Proof of the second statement: According to the proof of the first statement, we can see that the out-come FDP $\mathbf{u}_{k,s}^{l_o}$ of one iteration will be a feasible solution for the max-sum-rate problem of the next one. In addition, if $\sum_{\forall(k,s)} z_{k,s}^{(l_o)} \tilde{p}_{k,s}^{\mathsf{F}} \le D + \sum_{\forall(k,s)} f_{\mathsf{cnj}}^{(k,s)}(z_{k,s}^{(l_o)})$, we always have

$$\sum_{\forall(k,s)} f_{\mathsf{apx}}^{(k,s)}(\tilde{p}_{k,s}^{\mathsf{F}}) \le D, \tag{41}$$

for any values of $z_{k,s}^{(l_o)}$'s according to the definite of the conjugate function given in Section IV-A. Thus, the outcome solution of problem (16) in any iteration l_o satisfies all constraints of problem (13). Hence, Algorithm 1 returns the solution that satisfies all constraints of problem (13).

Appendix B
Proof of Proposition 3

Note that the objective function of problem (26), denoted $\Phi_r(\mathbf{v})$, can be expressed as

$$\Phi_r(\mathbf{v}) = \begin{bmatrix}\mathbf{v}^H & 1\end{bmatrix} \begin{bmatrix}\boldsymbol{\Pi}_r & -\boldsymbol{\pi}_r \\ -\boldsymbol{\pi}_r^H & 0\end{bmatrix} \begin{bmatrix}\mathbf{v} \\ 1\end{bmatrix}. \tag{42}$$

Upon defining $\boldsymbol{\Omega}_r$ as in Proposition 3, we can equivalently rewritten problem $(\mathcal{P}_r)$ as

$$\min_{\tilde{\mathbf{v}}} \tilde{\mathbf{v}}^H \boldsymbol{\Omega}_r \tilde{\mathbf{v}} \text{ s. t. } |(\tilde{\mathbf{v}})_i| = 1\ \forall i \le N_{\rm RF},\ (\tilde{\mathbf{v}})_{N_{\rm RF}+1} = 1. \tag{43}$$

In addition, $\Phi_r(\mathbf{v})$ can be also expressed as

$$\Phi_r(\mathbf{v}) = \begin{bmatrix}e^{-j\theta}\mathbf{v}^H & e^{-j\theta}\end{bmatrix} \begin{bmatrix}\boldsymbol{\Pi}_r & -\boldsymbol{\pi}_r \\ -\boldsymbol{\pi}_r^H & 0\end{bmatrix} \begin{bmatrix}e^{j\theta}\mathbf{v} \\ e^{j\theta}\end{bmatrix}, \tag{44}$$

for any value of θ. Thus, we $(\mathcal{Q}_r)$ by relaxing the last constraint from $(\tilde{\mathbf{v}})_{N_{\rm RF}+1} = 1$ to $|(\tilde{\mathbf{v}})_{N_{\rm RF}+1}| = 1$. In addition, $\tilde{\mathbf{v}}$ is feasible for $(\mathcal{Q}_r)$ if and only if $(\tilde{\mathbf{v}})'_{N_{\rm RF}+1}\tilde{\mathbf{v}}$ is feasible for problem (43) at the same objective value. Therefore, problem

$(\mathcal{P}_r)$ can be recast as problem $(\mathcal{Q}_r)$ and vice-versa. Then, in stead of solving $(\mathcal{P}_r)$, we can duel with $(\mathcal{Q}_r)$ and return the solution by exploiting the N_{RF} first elements of the vector $(\tilde{\mathbf{v}})'_{N_{\mathrm{RF}}+1}\tilde{\mathbf{v}}$.

Appendix C
Proof of Proposition 4

We can utilize $\bar{\boldsymbol{\Omega}}_r$ in stead of $\boldsymbol{\Omega}_r$ without changing the optimization problem $(\mathcal{Q}_r)$ due to the following result,

$$\begin{aligned}\tilde{\mathbf{v}}^H\bar{\boldsymbol{\Omega}}_r\tilde{\mathbf{v}} &= \tilde{\mathbf{v}}^H\left(\boldsymbol{\Omega}_r - \lambda_r^{\min}\mathbf{I}\right)\tilde{\mathbf{v}}\\ &= \tilde{\mathbf{v}}^H\boldsymbol{\Omega}_r\tilde{\mathbf{v}} - \lambda_r^{\min}\left(N_{\mathrm{RF}}+1\right),\end{aligned} \tag{45}$$

where $\lambda_r^{\min}\left(N_{\mathrm{RF}}+1\right)$ is a constant value. Then, let us define $\mathbf{V} = \tilde{\mathbf{v}}\tilde{\mathbf{v}}^H$. We have $\mathbf{V} \in \mathbb{C}^{(N_{\mathrm{RF}}+1)\times(N_{\mathrm{RF}}+1)}$, which is positive semi-definite and has rank one. Therefore, $(\mathcal{Q}_r)$ can be formulated as the SDP problem as in (28).

Appendix D
Proof of Proposition 5

Recall the objective function of $(\mathcal{P}_r)$ as $\Phi_r(\mathbf{v}) = \mathbf{v}^H\boldsymbol{\Pi}_r\mathbf{v} - 2\Re(\boldsymbol{\pi}_r^H\mathbf{v})$. Then, the gradient of $\Phi_r(\mathbf{v})$ can be expressed as $\nabla\Phi_r(\mathbf{v}) = 2\boldsymbol{\Pi}_r\mathbf{v} - 2\boldsymbol{\pi}_r$, which yields

$$\begin{aligned}\|\nabla\Phi_r(\mathbf{v}) - \nabla\Phi_r(\mathbf{t})\|_2^2 &= 4\|\boldsymbol{\Pi}_r(\mathbf{v}-\mathbf{t})\|_2^2\\ &\le 4\|\boldsymbol{\Pi}_r\|_2^2\|\mathbf{v}-\mathbf{t}\|_2^2 \overset{(a)}{=} 4\lambda_r^{\max}(\boldsymbol{\Pi}_r)\|\mathbf{v}-\mathbf{t}\|_2^2,\end{aligned} \tag{46}$$

where (a) follows the fact that the spectral norm of $\boldsymbol{\Pi}_r$ is defined to be $\lambda_r^{\max}(\boldsymbol{\Pi}_r)$. Therefore, if we select the step size satisfying $\alpha_r \le 1/4\lambda_r^{\max}(\boldsymbol{\Pi}_r)$, we can achieve

$$\|\nabla\Phi_r(\mathbf{v}) - \nabla\Phi_r(\mathbf{t})\|_2^2 \le \frac{1}{\alpha_r}\|\mathbf{v}-\mathbf{t}\|_2^2. \tag{47}$$

Hence, $\nabla\Phi_r(\mathbf{v})$ is $(1/\alpha_r)$–Lipschitz continuous. Hence, the quadratic upper bound surrogate function of $\Phi_r(\mathbf{v})$ is

$$\Upsilon_r(\mathbf{v},\mathbf{t}) = \Phi_r(\mathbf{t}) + \nabla\Phi_r(\mathbf{t})^H\left(\mathbf{v}-\mathbf{t}\right) + \frac{1}{2\alpha_r}\|\mathbf{v}-\mathbf{t}\|_2^2 \tag{48}$$

Thanks to (47), we have $\Upsilon_r(\mathbf{v},\mathbf{t}) \ge \Phi_r(\mathbf{v})$ and the equality holds if and only if $\mathbf{t} = \mathbf{v}$. We then can rewrite problem $(\mathcal{P}_r)$ in terms of $\Upsilon_r(\mathbf{v},\mathbf{t})$ as

$$\min_{\mathbf{v},\mathbf{t}} \Upsilon_r(\mathbf{v},\mathbf{t}) \ \text{ s. t. } |(\mathbf{v})_i| = 1 \ \forall i. \tag{49}$$

This problem will now be solved by the following iterative optimization algorithm. In iteration $(n+1)$, we first optimize with respect to $\mathbf{t}$ as

$$\mathbf{t}^{(n+1)} = \arg\min_{\mathbf{t}} \Upsilon_r(\mathbf{v}^{(n)},\mathbf{t}) = \mathbf{v}^{(n)}. \tag{50}$$

Then, we optimize with respect to $\mathbf{v}$ as

$$\begin{aligned}\mathbf{v}^{(n+1)} &= \arg\min_{\mathbf{t}} \Upsilon_r(\mathbf{v},\mathbf{t}^{(n+1)})\\ &= \arg\min_{|(\mathbf{v})_i|=1} \nabla\Phi_r^H(\mathbf{v}^{(n)})\mathbf{v} + \frac{1}{2\alpha_r}\|\mathbf{v}-\mathbf{v}^{(n)}\|_2^2\\ &= \arg\min_{|(\mathbf{v})_i|=1} \|\mathbf{v}-\mathbf{v}^{(n)} + \alpha_r\nabla\Phi_r(\mathbf{v}^{(n)})\|_2^2\\ &= \arg\min_{|(\mathbf{v})_i|=1} \|\mathbf{v}-\boldsymbol{\xi}_r^{(n+1)}\|_2^2 = e^{j\angle(\boldsymbol{\xi}_r^{(n+1)})},\end{aligned} \tag{51}$$

which is exactly the same as the way we update $\mathbf{v}_r^{(n+1)}$ in **Step 5** and **Step 6** of Algorithm 3. Since both of the $\mathbf{v}^{(n+1)}$ and $\mathbf{t}^{(n+1)}$ updates are conditionally optimal, we have

$$\Upsilon_r(\mathbf{v}^{(n)},\mathbf{t}^{(n)}) \ge \Upsilon_r(\mathbf{v}^{(n)},\mathbf{t}^{(n+1)}) \ge \Upsilon_r(\mathbf{v}^{(n+1)},\mathbf{t}^{(n+1)}). \tag{52}$$

Hence, this updating process will converge to a fixed point $(\mathbf{v}^\star,\mathbf{t}^\star)$. In addition, we have $\Upsilon_r(\mathbf{v},\mathbf{t}^\star) \ge \Upsilon_r(\mathbf{v}^\star,\mathbf{t}^\star)$ and $\Upsilon_r(\mathbf{v}^\star,\mathbf{t}) \ge \Upsilon_r(\mathbf{v}^\star,\mathbf{t}^\star)$, which yields that $\mathbf{v}^\star$ and $\mathbf{t}^\star$ are block-wise minimums. Therefore, $(\mathbf{v}^\star,\mathbf{t}^\star)$ is a KKT point of problem (49). We also have $\Phi_r(\mathbf{v}^{(n)}) = \Upsilon_r(\mathbf{v}^{(n)},\mathbf{t}^{(n+1)})$ and $\nabla\Phi_r(\mathbf{v}^{(n)}) = \nabla\Upsilon_r(\mathbf{v}^{(n)},\mathbf{t}^{(n+1)})$. Hence, **Step 4–8** of Algorithm 3 converges to $\mathbf{v}^\star$ which is also a KKT point of the non-convex and NP-hard problem $(\mathcal{P}_r)$.

Appendix E
Setting for System Energy Efficiency Simulation

We assume that the BS comprises the most efficient supercomputer *Shoubu system B* [38], which can reportedly achieve a power efficiency $\kappa_{\mathrm{pe}} = 17.009\ GOPS/W$, and C^{eff} is estimated based on a particular model for C^{eff} proposed in [32] as

$$C^{\mathrm{eff}} = \left(3\ N_{\mathrm{RF}} + N_{\mathrm{RF}}^2 + \frac{br_{\mathrm{code}}}{3}\right)/10, \tag{53}$$

where b is the modulation bits per symbol in the data stream and r_{code} is the corresponding coding rate. In addition, the 16-QAM modulation scheme is employed with the code rate at $4/3$; hence, C^{eff} is around $30.5778\ GOPS$. Then, P_{comp} (in Watt) can be calculated based on $\bar{D}$ as

$$P_{\mathrm{comp}} = \frac{30.5778\bar{D}}{17.009} = 1.7977\bar{D}. \tag{54}$$

Now, we will consider the power consumption corresponding to the RF and AP process, denoted as P_{C}. The BB signal, outcome of the IFFT block in Fig. 1, will be up-converted to RF band, multiplied with a phase shifter, amplified before being transmitted at the antennas. In this setting, we consider the full-connected AP system; hence, P_{C} (in Watt) can be calculated as

$$\begin{aligned}P_{\mathrm{C}} &= N_{\mathrm{RF}}(P_{\mathrm{DAC}} + P_{\mathrm{RFC}}) + N_{\mathrm{RF}}N_{\mathrm{T}}P_{\mathrm{PS}} + N_{\mathrm{T}}P_{\mathrm{PA}}\\ &= 131.888(W)\end{aligned} \tag{55}$$

where $P_{\mathrm{DAC}} = 200\ mW$, $P_{\mathrm{RFC}} = 43\ mW$, $P_{\mathrm{PS}} = 30\ mW$, and $P_{\mathrm{PA}} = 20\ mW$ represent for the power of digital-analog converter, RF upconverter, phase shifter, and power amplifier as given in [42], respectively. Let P_{total} be the total power consumption of the system, the SEE then can be calculated as

$$\eta_{\mathrm{SEE}} = \frac{\sum_{\forall(k,s)} R_{k,s}}{P_{\mathrm{total}}} = \frac{\sum_{\forall(k,s)} R_{k,s}}{1.7977\bar{D} + 131.888 + P_{\mathrm{T}}}. \tag{56}$$

Appendix F
Zero-Forcing Water-Filling Method

In this section, we introduce the ZF-WF precoding method to solve the problem (16). In particular, we first define the *unnormalized* ZF precoding vectors, then scale them with a distinct factor to achieve the ZF-WF precoding vectors satisfying the two constraints in this problem. Let $\tilde{\mathbf{u}}_{k,s}$ be the *unnormalized* vector corresponding to $\mathbf{u}_{k,s}$, and we also define $\mathbf{H}_s = [\mathbf{h}_{1,s}, \ldots, \mathbf{h}_{M,s}]$. Over each subcarrier, the ZF precoding vectors corresponding to that subcarrier should be designed such that they do not induce any interference, i.e., $\mathbf{h}_{i,s}^H \tilde{\mathbf{u}}_{k,s} = 0$ for all $i \neq k$. Hence, the *unnormalized* precoding vectors over subcarrier s can be given as

$$\tilde{\mathbf{U}}_{s,\mathrm{ZF}} = \mathbf{H}_s^H (\mathbf{H}_s \mathbf{H}_s^H)^{-1}, \tag{57}$$

where $\tilde{\mathbf{U}}_{s,\mathrm{ZF}} = [\tilde{\mathbf{u}}_{1,s}, \ldots, \tilde{\mathbf{u}}_{M,s}]$. Now, we scale the ZF-WF precoding vextor $\mathbf{u}_{k,s}^{ZF-WF}$ up from the *unnormalized* vector $\tilde{\mathbf{u}}_{k,s}$ with the distinct factor $\sqrt{p_{k,s}^{\mathrm{F}}}$. By replacing $\mathbf{u}_{k,s}^{ZF-WF} = \sqrt{p_{k,s}^{\mathrm{F}}} \tilde{\mathbf{u}}_{k,s}$, the problem (16) can be restated as

$$\begin{aligned} \max_{\{p_{k,s}^{\mathrm{F}} \geq 0\}} & \sum_{\forall (k,s)} \log(1 + \rho_{k,s} p_{k,s}^{\mathrm{F}}) \\ \text{s. t.} & \sum_{\forall (k,s)} \gamma_{k,s} p_{k,s}^{\mathrm{F}} \leq P_{\mathrm{T}}, \\ & \sum_{\forall (k,s)} \hat{z}_{k,s} \gamma_{k,s} p_{k,s}^{\mathrm{F}} \leq Z, \end{aligned} \tag{58}$$

where $\rho_{k,s} = |\mathbf{h}_{k,s}^H \tilde{\mathbf{u}}_{k,s}|^2$ and $\gamma_{k,s} = \tilde{\mathbf{u}}_{k,s}^H \tilde{\mathbf{u}}_{k,s}$ for all (k,s). This problem is defined as a dual-layer constraint power allocation problem which can be solved by employing the iterative water-filling [48] as follows:

$$p_{k,s}^{\mathrm{F}} = \left[\frac{1}{\gamma_{k,s}(\mu_1 + \mu_2 \hat{z}_{k,s})} - \frac{1}{\rho_{k,s}} \right]^+ \tag{59}$$

where $[x]^+ = \max(x, 0)$; and μ_1 and μ_2 are the water level parameters which can be iteratively updated to meet the two constraints [48]. Then, the ZF-WF method can be modified to solve the problem (16) by employing the Algorithm 1 as follows: Calculating all the *"unnormalized"* FDP vectors in the initial step, employing all steps in Algorithm 1, and replacing Step 4-7 by obtaining power as in (59).

REFERENCES

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

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

[3] 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.

[4] 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.

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

[6] E. Castañeda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 239–284, 1st Quart., 2017.

[7] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[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, Jr., "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[10] 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.

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

[12] T. E. Bogale and L. B. Le, "Beamforming for multiuser massive MIMO systems: Digital versus hybrid analog-digital," in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 4066–4071.

[13] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, "Hybrid MMSE precoding for mmWave multiuser MIMO systems," in *Proc. 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, Jr., "Hybrid MMSE preocding and combining designs for mmWave multiuser systems," *IEEE Access*, vol. 5, pp. 19167–19181, Sep. 2017.

[15] C. Kim, T. Kim, and J.-Y. Seol, "Multi-beam transmission diversity with hybrid beamforming for MIMO-OFDM systems," in *Proc. IEEE Globecom Workshops*, Atlanta, GA, USA, Dec. 2013, pp. 61–65.

[16] S. Yoon, T. Jeon, and W. Lee, "Hybrid beam-forming and beam-switching for OFDM based wireless personal area networks," *IEEE J. Select. Areas Commun.*, vol. 27, no. 8, pp. 1425–1432, Oct. 2009.

[17] H.-H. Lee and Y.-C. Ko, "Low complexity codebook-based beamforming for MIMO-OFDM systems in millimeter-wave WPAN," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3607–3612, Nov. 2011.

[18] S. Park, A. Alkhateeb, and R. W. Heath, Jr., "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2907–2920, May 2017.

[19] D. Zhu, B. Li, and P. Liang, "A novel hybrid beamforming algorithm with unified analog beamforming by subspace construction based on partial CSI for massive MIMO-OFDM systems," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 594–607, Feb. 2017.

[20] C.-S. Sum *et al.*, "Virtual time-slot allocation scheme for throughput enhancement in a millimeter-wave multi-Gbps WPAN system," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1379–1389, Oct. 2009.

[21] T. E. Bogale, L. B. Le, and A. Haghighat, "User scheduling for massive MIMO OFDMA systems with hybrid analog-digital beamforming," in *Proc. IEEE Int. Conf. Commun.*, London, U.K., Jun. 2015, pp. 1757–1762.

[22] 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.

[23] I. Ahmed, H. Khammari, and A. Shahid, "Resource allocation for transmit hybrid beamforming in decoupled millimeter wave multiuser-MIMO downlink," *IEEE Access*, vol. 5, pp. 170–182, 2016.

[24] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, nos. 5–6, pp. 877–905, 2008.

[25] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.

[26] R. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.

[27] S. S. Christensen, R. Agarwal, 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.

[28] V. N. Ha, D. H. N. Nguyen, and L. B. Le, "Sparse precoding design for cloud-RANs sum-rate maximization," in *Proc. IEEE Wireless Commun. Netw. Conf.*, New Orleans, LA, USA, Mar. 2015, pp. 1648–1653.

[29] M. Grant, S. Boyd, and Y. Ye. (2009). *CVX: Matlab Software for Disciplined Convex Programming*. [Online]. Available: http://www.stanford.edu/boyd/cvx

[30] X. Zhang, A. F. Molisch, and S.-Y. Kung, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE Trans. Signal Process.*, vol. 53, no. 11, pp. 4091–4103, Nov. 2005.

[31] P. Rost, S. Talarico, and M. C. Valenti, "The complexity–rate tradeoff of centralized radio access networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164–6176, Nov. 2015.

[32] T. Werthmann, H. Grob-Lipski, S. Scholz, and B. Haberland, "Task assignment strategies for pools of baseband computation units in 4G cellular networks," in *Proc. IEEE Int. Conf. Commun. Workshop*, London, U.K., Jun. 2015, pp. 2714–2720.

[33] D. Sabella *et al.*, "Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure," *IEEE Access*, vol. 2, pp. 1586–1597, 2014.

[34] Y. J. A. Zhang and A. M.-C. So, "Optimal spectrum sharing in MIMO cognitive radio networks via semidefinite programming," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 2, pp. 362–373, Feb. 2011.

[35] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[36] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optim.*, vol. 26, no. 3, pp. 1835–1854, Sep. 2016.

[37] G. R. MacCartney, J. Zhang, S. Nie, and T. S. Rappaport, "Path loss models for 5G millimeter wave propagation channels in urban microcells," in *Proc. IEEE Global Commun. Conf.*, Atlanta, GA, USA, Dec. 2013, pp. 3948–3953.

[38] *Green500 List for November 2017*. Accessed: Jul. 3, 2018. [Online]. Available: https://www.top500.org/list/2017/06/?page=1

[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] Y. Kwon, J. Chung, and Y. Sung, "Hybrid beamformer design for mmwave wideband multi-user MIMO-OFDM systems," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun.*, Sapporo, Japan, Jul. 2017, pp. 1–5.

[41] C. Desset *et al.*, "Flexible power modeling of LTE base stations," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Shanghai, China, Apr. 2012, pp. 2858–2862.

[42] W.-T. Li, Y.-C. Chiang, J.-H. Tsai, H.-Y. Yang, J.-H. Cheng, and T.-W. 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.

[43] F. A. P. De Figueiredo, F. A. C. M. Cardoso, I. Moerman, and G. Fraidenraich, "Channel estimation for massive MIMO TDD systems assuming pilot contamination and frequency selective fading," *IEEE Access*, vol. 5, pp. 17733–17741, 2017.

[44] X. Cheng, J. Sun, and S. Li, "Channel estimation for FDD multi-user massive MIMO: A variational Bayesian inference-based approach," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7590–7602, Nov. 2017.

[45] L. N. Tran, M. Juntti, M. Bengtsson, and B. Ottersten, "Beamformer designs for MISO broadcast channels with zero-forcing dirty paper coding," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1173–1185, Mar. 2013.

[46] S. Huang, H. Yin, J. Wu, and V. C. M. Leung, "User selection for multiuser MIMO downlink with zero-forcing beamforming," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3084–3097, Sep. 2013.

[47] D. N. Nguyen and T. Le-Ngoc, "MMSE precoding for multiuser MISO downlink transmission with non-homogeneous user SNR conditions," *EURASIP J. Adv. Signal Process.*, vol. 85, no. 1, pp. 1–12, Dec. 2014.

[48] P. Wang, M. Zhao, L. Xiao, S. Zhou, and J. Wang, "Power allocation in OFDM-based cognitive radio systems," in *Proc. IEEE Global Telecommun. Conf.*, Washington, DC, USA, Nov. 2007, pp. 4061–4065.

Vu Nguyen Ha (S'11–M'17) received the B.Eng. degree from the French Training Program for Excellent Engineers in Vietnam, Ho Chi Minh City University of Technology, Vietnam, the 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, Montreal, 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 is currently a Post-Doctoral Fellow with the École Polytechnique de Montréal, Montreal, QC, Canada. His research interests include radio resource management and emerging enabling technologies for 5G wireless systems with special emphasis on heterogeneous small-cell networks, cloud RAN, and massive MIMO communications. He is currently a recipient of the FRQNT Post-Doctoral Fellowship for International Researcher.

Duy H. N. Nguyen (S'07–M'14) received the B.Eng. degree (Hons.) from the Swinburne University of Technology, Hawthorn, VIC, Australia, in 2005, the M.Sc. degree from the University of Saskatchewan, Saskatoon, SK, Canada, in 2009, and the Ph.D. degree from McGill University, Montreal, QC, Canada, in 2013, all in electrical engineering. From 2013 to 2015, he held a joint appointment as a Research Associate with McGill University and a Post-doctoral Research Fellow with the Institut National de la Recherche Scientifique, Université du Québec, Montreal, QC, Canada. He was a Research Assistant with the University of Houston in 2015 and a Post-Doctoral 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 Post-Doctoral Fellowship, and the NSERC Post-doctoral Fellowship.

Jean-François Frigon (S'95–M'02–SM'12) received the B.Eng. degree from the École Polytechnique de Montréal, Montreal, 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, CA, USA, in 2004. From 2001 to 2003, he was the Director of Wireless Communications Systems at Innovics Wireless. He joined the Electrical Engineering Department, École Polytechnique de Montréal, Montreal, QC, Canada, 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.