IEEE Access
The journal for rapid open access publishing

IEEE Xplore Digital Library
http://ieeexplore.ieee.org/Xplore

Hybrid MMSE Precoding and Combining Designs for mmWave Multiuser Systems

Duy H. N. Nguyen, *Member, IEEE*, Long B. Le, *Senior Member, IEEE*, Tho Le-Ngoc, *Fellow, IEEE*, and Robert W. Heath Jr., *Fellow, IEEE*

***Abstract*—Hybrid analog/digital precoding architectures are a low-complexity alternative for fully-digital precoding in millimeter-wave (mmWave) MIMO wireless systems. This is motivated by the reduction in the number of radio frequency and mixed signal hardware components. Hybrid precoding involves a combination of analog and digital processing that enable both beamforming and spatial multiplexing gains in mmWave systems. This paper develops hybrid analog/digital precoding and combining designs for mmWave multiuser systems, based on the mean-squared error (MSE) criteria. In the first design with the analog combiners being determined at the users, the proposed hybrid minimum mean-squared error precoder (MMSE) is realized by minimizing the sum-MSE of the data streams intended for the users. In the second design, both the hybrid precoder and combiners are jointly designed in an iterative manner to minimize a weighted sum-MSE cost function. By leveraging the sparse structure of mmWave channels, the MMSE precoding/combining design problems are then formulated as sparse reconstruction problems. An orthogonal matching pursuit-based algorithm is then developed to determine the MMSE precoder and combiners. Simulation results show the performance advantages of the proposed precoding/combining designs in various system settings.**

***Index Terms*—Millimeter wave, multiple-input multiple-output (MIMO), antenna arrays, beamforming, precoding, sparse reconstruction, minimum mean squared-error (MMSE).**

I. Introduction

Millimeter-wave (mmWave) communications has emerged as one of the most promising candidates for future cellular systems due to the large and underexploited mmWave band [1]–[5]. MmWave systems need large antenna arrays to provide array gain and achieve reasonable link margin [6]. Interestingly, packing a large number of antenna elements in a sizable space in mmWave systems is possible due to the band's short wavelength. The large antenna arrays also enable MIMO communication strategies, *e.g.*, multiuser MIMO [7], [8].

Multiuser precoding involves assigning the weight vectors for different mobile-stations (MS) before transmitting through the multiple antennas of the base-station (BS). Proper selection of weight vectors enables spatial separation among the users and thus supports multiplexing multiple data streams. In conventional MIMO system at lower frequencies, precoding is performed at baseband by a digital signal processing unit. Such a design requires a dedicated RF chain for each antenna element [6]. Unfortunately, the high cost and power consumption of current mmWave mixed-signal hardware technologies make fully digital transceiver architectures impractical [9]. Therefore, mmWave systems need suitable MIMO architectures and signal processing algorithms [10]. Recent work in precoding/combining designs for mmWave systems has advocated the use of hybrid analog/digital precoders/combiners [6]–[8], [11]–[13]. In this hybrid structure, the analog precoder/combiner allows the beamforming gain, while the digital precoder/combiner provides the multiplexing gain. Analog beamforming/combining can be implemented with phase-shifters [2], [6], [14], switches [9], or even a lens [15], [16]. With phase-shifters, the relative phases of the RF signals are changed to steer the transmit/receive beams in the desired directions [17]. The phase changes might be digitally controlled and thus have only quantized values [8].

Hybrid precoding/combining for single-user mmWave systems has been investigated in [6], [12], [18], [19]. These papers exploit the limited scattering nature of the mmWave channel to develop low-complexity hybrid precoding algorithms. It was shown that hybrid precoding/combining with RF beamsteering codebook is capable of achieving a near-optimal performance to the fully digital design [6]. An algorithmic solution, based on the orthogonal matching pursuit (OMP) algorithm [20], was proposed in [6] to reconstruct such hybrid precoder/combiner. Further improvements to the algorithm in terms of complexity were addressed in [18], [19]. Additional performance gain in hybrid precoding/combining can be obtained by freely optimizing the RF precoder and combiner without restricting them to pre-defined codebooks [21]. Such an approach, though, comes with higher computational complexity due to the treatment of nonconvex constraints on the RF precoder and combiner.

Hybrid precoding/combining was also studied for multiuser mmWave systems [7], [8], [22], [23]. In [7], [8], a two-stage hybrid precoding design was proposed. In the first stage, the MS and the BS jointly select a "best" combination of RF

This material is based upon work supported in part by a startup fund from San Diego State University, in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship Program, in part by the NSERC Grants including NSERC–CRD project no. CRDPJ 461894–13, and in part by the National Science Foundation under Grant No. 1702800.

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

Long B. Le is with INRS-EMT, Université du Québec, Montréal, QC, Canada, H5A 1K6 (e-mail: long.le@emt.inrs.ca).

Tho Le-Ngoc is with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada, H3A 0E9 (e-mail: tho.le-ngoc@mcgill.ca).

Robert W. Heath Jr. is with the Wireless Networking and Communications Group, Department of Electrical and Computer Engineering, University of Texas, Austin, TX, USA 78712 (e-mail: rheath@utexas.edu).

Digital Object Identifier: 10.1109/ACCESS.2017.2754979

combiner and RF beamformer to maximize the channel gain to that particular MS. The baseband digital precoder is then derived as a zero-forcing (ZF) precoder by inverting the effective channel. In [22], an iterative hybrid precoding/combining algorithm exploiting channel reciprocity was proposed for multiuser systems with single-stream transmission for each user. The work in [23] established the required number of RF chains and phase-shifters such that hybrid precoding achieves the same performance as that of the digital precoding. The study in [23], however, is limited to the case of single-antenna users.

In this paper, we examine a multiuser mmWave system similar to that in [7], [8], albeit under a more general setting where each user is equipped with multiple RF chains. Such a general setting was investigated in recent papers [24]–[26]. The work [24] proposed a three-step hybrid precoding/combining design: finding the "best" RF combiner at each MS, finding the RF precoder by co-phasing the effective channel to each MS, and imposing a baseband block-diagonalization (BD) precoder. In [25], hybrid precoding was designed to minimize its performance gap to the fully digital precoding. The work in [26] showed that hybrid precoding structure can realize any digital precoder exactly if the number of RF chains is twice the total number of data streams. For cases with fewer number of RF chains, a heuristic hybrid precoding design that obtains a close performance to that of digital precoding design was proposed in [26]. Note that the near-exact or exact performance obtained by hybrid precoding designs in [24], [26], [27] may require infinite resolution phase shifters, which is challenging to realize in practice. Hybrid precoding/combining was studied in [28] for interference channel with multiple pairs of transmitters and receivers. By first deriving optimal MMSE transmit and receive filters, [28] then tried to approximate the hybrid precoders and combiners, respectively, with predefined RF codebooks. We note that recent work in [23], [29]–[31] has proposed hybrid precoding/combining designs for wideband mmWave systems. Nevertheless, hybrid precoding/combining in narrowband mmWave systems has still attracted a majority of research attention [6]–[8], [12], [18], [19], [21], [22], [24]–[28]. In this paper, we focus our study on narrowband mmWave systems, where minimum mean-squared error (MMSE)-based hybrid precoding/combining designs are being proposed.

In a early version of this paper [32], we proposed a hybrid MMSE precoding design for multiuser mmWave systems where each MS is equipped with a single RF chain. In this paper, we propose two new hybrid MMSE precoding/combining designs for multiuser mmWave systems in a more general setting where each MS can be equipped with multiple RF chains. The proposed designs assume finite RF beamforming codebooks at both the BS and MSs and the sparse structure of mmWave channels. The main contributions of paper can be summarized as follows:

- Proposing a hybrid MMSE precoding algorithm for multiuser mmWave system without prior knowledge of a fully digital precoder as in [6], [28]. In this algorithm, the RF combiners are determined separately at the MSs, whereas the digital baseband and analog RF precoders are jointly optimized to minimize the sum-MSE of the data streams intended for the MSs. We show that this sum-MSE minimization problem resembles the problem of sparse signal recovery with multiple measurement vectors [33]. We thus develop an algorithmic hybrid MMSE precoding solution, which is based on the concept of OMP [20]. The key difference in our proposed algorithm to the OMP-based algorithm in [6] is that the baseband precoder is derived as a closed-form MMSE solution instead of a least-square solution. In addition, a joint design of the RF precoder for all users with an MMSE baseband precoder potentially offers better sum-rate than the two-stage hybrid ZF precoder in [8].
- Proposing a hybrid weighted MMSE (WMMSE) precoding/combining algorithm where the hybrid precoder and hybrid combiners are jointly optimized. In this algorithm, the hybrid precoder and hybrid combiners are updated in an iterative manner to minimize a weighted sum-MSE cost function. Due to equivalence between the weighted sum-MSE and sum-rate functions [34], the proposed algorithm can attain a locally maximum system sum-rate as well. The iterative nature of this algorithms also enable the hybrid combiners at the MSs to match better with the hybrid precoder at the BS and vice versa, compared to the proposed hybrid MMSE precoder. Therefore, it provides much better performance than the first proposed algorithm, albeit with higher complexity.

Simulation results show significant sum-rate performance advantages of the proposed hybrid precoding/combining designs over previous designs in various system settings. In addition, the two proposed designs show a similar performance growth as obtained by fully digital precoding/combining designs, either with the signal-to-noise ratio (SNR) or with the number of users in the system.

The rest of the paper is organized as follows. In Section II, the channel and multiuser system models are described. In Section III, a two-stage hybrid BD is presented as a generalization of two-stage hybrid ZF precoding [8] to the system with multiple-RF-chain MSs. The two-stage hybrid designs serve as a benchmark for comparison to the proposed MMSE-based precoding/combining designed presented in Section IV and Section V, respectively. Simulation results are presented in Section VII before concluding the paper in Section VIII.

Notation: $(\mathbf{X})^\top$, $(\mathbf{X})^*$, and $\mathbf{X}^\dagger$ denote transpose, conjugate transpose (Hermitian operator), and pseudo-inverse of a matrix $\mathbf{X}$; $\|\mathbf{x}\|_0$ and $\|\mathbf{x}\|$ denote norm-0 and Euclidean norm of a vector x, whereas $\|\mathbf{X}\|_F$ denotes the Frobenius norm of $\mathbf{X}$; $\mathrm{Tr}\{\mathbf{X}\}$, $|\mathbf{X}|$ denote the trace and determinant of $\mathbf{X}$; $\mathrm{diag}(\mathbf{A})$ is a vector formed by the diagonal elements of $\mathbf{A}$ whereas $\mathrm{diag}(a_1, \ldots, a_L)$ is an $L \times L$ diagonal matrix with the diagonal elements $(a_1, \ldots, a_L)$; $\mathrm{blkdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_K)$ is a block diagonal matrix formed by matrices $(\mathbf{A}_1, \ldots, \mathbf{A}_K)$; $\mathbf{I}_L$ denotes an $L \times L$ identity matrix; $\mathbb{E}[\cdot]$ denotes the expectation; $\mathcal{N}(\mathbf{m}, \mathbf{R})$ is a complex Gaussian random vector with mean $\mathbf{m}$ and covariance $\mathbf{R}$.

II. System Model

A. MmWave Channel Model

One of the main characteristics of the mmWave channel is the limited number of scatterers in its propagation path. Measurement results [3], [35], [36] have shown that there are fewer dominant scatterers are mmWave due to the significance of blockage and the reduced effects of diffraction. Due to the short wavelength in mmWave, a transmitted signal does not reflect well with surrounding environment. As the signal disperses due to roughness of the reflecting surface, it loses power. In this paper, we adopt the extended Saleh-Valenzuela geometric model for the considered mmWave channel [6], [37], [38]. Let N_{t} and N_{r} be the numbers of transmit antennas at the BS and receive antennas at a MS, respectively. The channel $\mathbf{H}_i \in \mathbb{C}^{N_{\mathrm{r}} \times N_{\mathrm{t}}}$ from the BS to MS-i can be modeled as

$$\mathbf{H}_i = \sqrt{\frac{N_{\mathrm{r}} N_{\mathrm{t}}}{L_i}} \sum_{l=1}^{L_i} \alpha_{i,l} \mathbf{a}_{\mathrm{r}}(\phi_{i,l}^{\mathrm{r}}, \theta_{i,l}^{\mathrm{r}}) \mathbf{a}_{\mathrm{t}}^*(\phi_{i,l}^{\mathrm{t}}, \theta_{i,l}^{\mathrm{t}}) \quad (1)$$

where L_i is the number of propagation paths, $\alpha_{i,l}$ is the complex gain of the lth path, including the path loss with $\alpha_{i,l} \sim \mathcal{N}(0, \sigma_{\alpha,i}^2)$, and $(\phi_{i,l}^{\mathrm{r}}, \theta_{i,l}^{\mathrm{r}})$ and $(\phi_{i,l}^{\mathrm{t}}, \theta_{i,l}^{\mathrm{t}})$ are its (azimuth, elevation) angles of arrival and departure. Then, the vectors $\mathbf{a}_{\mathrm{r}}(\phi_{i,l}^{\mathrm{r}}, \theta_{i,l}^{\mathrm{r}})$ and $\mathbf{a}_{\mathrm{t}}(\phi_{i,l}^{\mathrm{t}}, \theta_{i,l}^{\mathrm{t}})$ represent the unit-norm receive and transmit array response vectors at (azimuth, elevation) angles of $(\phi_{i,l}^{\mathrm{r}}, \theta_{i,l}^{\mathrm{r}})$ and $(\phi_{i,l}^{\mathrm{t}}, \theta_{i,l}^{\mathrm{t}})$. Denote $\mathbf{A}_{\mathrm{r},i} = \left[\mathbf{a}_{\mathrm{r}}(\phi_{i,1}^{\mathrm{r}}, \theta_{i,1}^{\mathrm{r}}), \ldots, \mathbf{a}_{\mathrm{r}}(\phi_{i,L_i}^{\mathrm{r}}, \theta_{i,L_i}^{\mathrm{r}})\right]$ and $\mathbf{A}_{\mathrm{t},i} = \left[\mathbf{a}_{\mathrm{t}}(\phi_{i,1}^{\mathrm{t}}, \theta_{i,1}^{\mathrm{t}}), \ldots, \mathbf{a}_{\mathrm{t}}(\phi_{i,L_i}^{\mathrm{t}}, \theta_{i,L_i}^{\mathrm{t}})\right]$. Since mmWave channels are expected to have limited scattering [36], a small number of propagation paths L_i is assumed, relatively to the number of transmit antennas N_{t}.

The array response vectors $\mathbf{a}_{\mathrm{r}}(\phi_{i,l}^{\mathrm{r}}, \theta_{i,l}^{\mathrm{r}})$ and $\mathbf{a}_{\mathrm{t}}(\phi_{i,l}^{\mathrm{t}}, \theta_{i,l}^{\mathrm{t}})$ only depend on the transmit and receive antenna array structures. Two commonly-used antenna array structures are the uniform linear array (ULA) and the uniform planar array (UPA). While the following algorithms and results presented in this work are applicable to any antenna arrays, we use UPAs in the simulations of Section VII.

B. Multiuser MIMO Signal Model

Consider a downlink MIMO multiuser system shown in Fig. 1 in which a BS is communicating with K independent remote MSs. In this system, the BS is equipped with $N_{\mathrm{t}}^{\mathrm{RF}}$ RF chains, whereas each MS is equipped $N_{\mathrm{r}}^{\mathrm{RF}}$ RF chains. Hence, each MS can support up to $N_{\mathrm{r}}^{\mathrm{RF}}$ data streams. We assume that $KN_{\mathrm{r}}^{\mathrm{RF}} \leq N_{\mathrm{t}}^{\mathrm{RF}}$ so that the BS is capable of multiplexing $KN_{\mathrm{r}}^{\mathrm{RF}}$ concurrent data streams of the K MSs.

In this paper, we consider a narrowband block-fading channel model as in [6], [8], [15]. At first, the BS applies a $N_{\mathrm{t}}^{\mathrm{RF}} \times KN_{\mathrm{r}}^{\mathrm{RF}}$ baseband precoder $\mathbf{F}_{\mathrm{B}} = [\mathbf{F}_{\mathrm{B}_1}, \ldots, \mathbf{F}_{\mathrm{B}_K}]$, where $\mathbf{F}_{\mathrm{B}_i} \in \mathbb{C}^{N_{\mathrm{t}}^{\mathrm{RF}} \times N_{\mathrm{r}}^{\mathrm{RF}}}$ is the baseband precoding matrix applied to the information symbol vector intended for MS-i, $\mathbf{s}_i \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{RF}}}$. Following the baseband precoding and RF processing steps through the $N_{\mathrm{t}}^{\mathrm{RF}}$ RF chains, the BS then applies an $N_{\mathrm{t}} \times N_{\mathrm{t}}^{\mathrm{RF}}$ RF precoding matrix $\mathbf{F}_{\mathrm{R}}$ on its transmitted signal. Given $\mathbf{F}_i = \mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}_i}$ as the combined BS precoding matrix for MS-i, the transmitted signal at the BS can be modeled as

$$\mathbf{x} = \sum_{i=1}^{K} \mathbf{F}_i \mathbf{s}_i = \mathbf{F}\mathbf{s} \quad (2)$$

where $\mathbf{F} = [\mathbf{F}_1, \ldots, \mathbf{F}_K] \in \mathbb{C}^{N_{\mathrm{t}} \times KN_{\mathrm{r}}^{\mathrm{RF}}}$ and $\mathbf{s} = [\mathbf{s}_1^\top, \ldots, \mathbf{s}_K^\top]^\top$. It is assumed that the information symbols are independent for each MS and have unit power, *i.e.*, $\mathbb{E}[\mathbf{s}_i \mathbf{s}_i^*] = \mathbf{I}_{R_{\mathrm{r}}^{\mathrm{RF}}}$, and $\mathbb{E}[\mathbf{s}_i \mathbf{s}_j^*] = \mathbf{0}$ for $j \neq i$. To meet the total power budget of P at the BS, the precoding matrix $\mathbf{F}_i$'s are constrained by $\sum_{i=1}^{K} \mathrm{Tr}\{\mathbf{F}_i \mathbf{F}_i^*\} \leq P$.

Given $\mathbf{H}_i \in \mathbb{C}^{N_{\mathrm{r}} \times N_{\mathrm{t}}}$ as the downlink channel from the BS to MS-i, the received signal at MS-i can be modeled as

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x} + \mathbf{z}_i = \mathbf{H}_i \mathbf{F}_i \mathbf{s}_i + \mathbf{H}_i \sum_{j \neq i}^{K} \mathbf{F}_j \mathbf{s}_j + \mathbf{z}_i \quad (3)$$

where the noise $\mathbf{z}_i$ is assumed to be $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_{\mathrm{r}}})$.

Let $\mathbf{W}_{\mathrm{R}_i}^* \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{RF}} \times N_{\mathrm{r}}}$ and $\mathbf{W}_{\mathrm{B}_i}^* \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{RF}} \times N_{\mathrm{r}}^{\mathrm{RF}}}$ be the RF combiner and baseband equalizer, respectively, at MS-i. Denote $\mathbf{W}_i^* = \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^*$ as the combined receive filter to process the received signal $\mathbf{y}_i$, which results in the estimated signal vector at MS-i

$$\hat{\mathbf{s}}_i = \mathbf{W}_i^* \mathbf{H}_i \mathbf{F}_i \mathbf{s}_i + \mathbf{W}_i^* \mathbf{H}_i \sum_{j \neq i}^{K} \mathbf{F}_j \mathbf{s}_j + \mathbf{W}_i^* \mathbf{z}_i. \quad (4)$$

We assume that $\mathbf{F}_{\mathrm{R}}$ and $\mathbf{W}_{\mathrm{R}_i}$'s are implemented using analog phase shifters. Thus, their entries are of constant modulus. We normalize these entries to satisfy $\left|\left[\mathbf{F}_{\mathrm{R}}\right]_{m,t}\right| = \frac{1}{\sqrt{N_{\mathrm{t}}}}$ and $\left|\left[\mathbf{W}_{\mathrm{R}_i}\right]_{n,r}\right| = \frac{1}{\sqrt{N_{\mathrm{r}}}}, \forall i$. Due to the constraints on RF hardware, the analog beamforming/combining vectors can only take on certain quantized angles of the phase-shifters. Therefore, these vectors should be selected from finite-size codebooks. Inspired by the good performance of single user precoding algorithms [6] and multiuser precoding algorithms [7], [8] which relied on RF beamsteering vectors, we will adopt the the beamsteering codebooks for the analog beamforming/combining vectors. In this case, the analog beamforming/combining vectors have the same form of the array response vector and thus can be parameterized by quantized azimuth and elevation angles [6].

Let $\mathcal{F}$ represent the analog beamforming codebook, with cardinality $|\mathcal{F}| = 2^{N_\phi N_\theta}$, where N_ϕ and N_θ are the numbers of azimuth and elevation quantizing bits, respectively. Then, $\mathcal{F}$ consists of vectors in the form $\mathbf{a}_{\mathrm{t}}\left(\frac{2\pi k_\phi}{2^{N_\phi}}, \frac{2\pi k_\theta}{2^{N_\theta}}\right)$, for the variables k_ϕ taking on values $0, 1, \ldots, 2^{N_\phi} - 1$ and the variables k_θ taking on values $0, 1, \ldots, 2^{N_\theta} - 1$. The analog combining codebook $\mathcal{W}_i$ can be similarly defined for MS-i.

Denote $\mathbf{R}_i = \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \big(\mathbf{H}_i \sum_{j \neq i}^{K} \mathbf{F}_{\mathrm{R}} \mathbf{F}_{\mathrm{B}_j} \mathbf{F}_{\mathrm{B}_j}^* \mathbf{F}_{\mathrm{R}}^* \mathbf{H}_i^* + \sigma^2 \mathbf{I}_{N_{\mathrm{r}}}\big) \mathbf{W}_{\mathrm{R}_i} \mathbf{W}_{\mathrm{B}_i}$ as the covariance matrix of the total interuser interference plus noise at MS-i. Under Gaussian signaling, the achievable rate for the transmission to MS-i is

$$R_i = \log \big| \mathbf{I}_{N_{\mathrm{r}}^{\mathrm{RF}}} + \mathbf{R}_i^{-1} \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{H}_i \mathbf{F}_{\mathrm{R}} \mathbf{F}_{\mathrm{B}_i} \times \mathbf{F}_{\mathrm{B}_i}^* \mathbf{F}_{\mathrm{R}}^* \mathbf{H}_i^* \mathbf{W}_{\mathrm{R}_i} \mathbf{W}_{\mathrm{B}_i} \big|. \quad (5)$$

In this paper, we are interested in jointly designing the baseband precoder, RF precoder, RF combiner and baseband

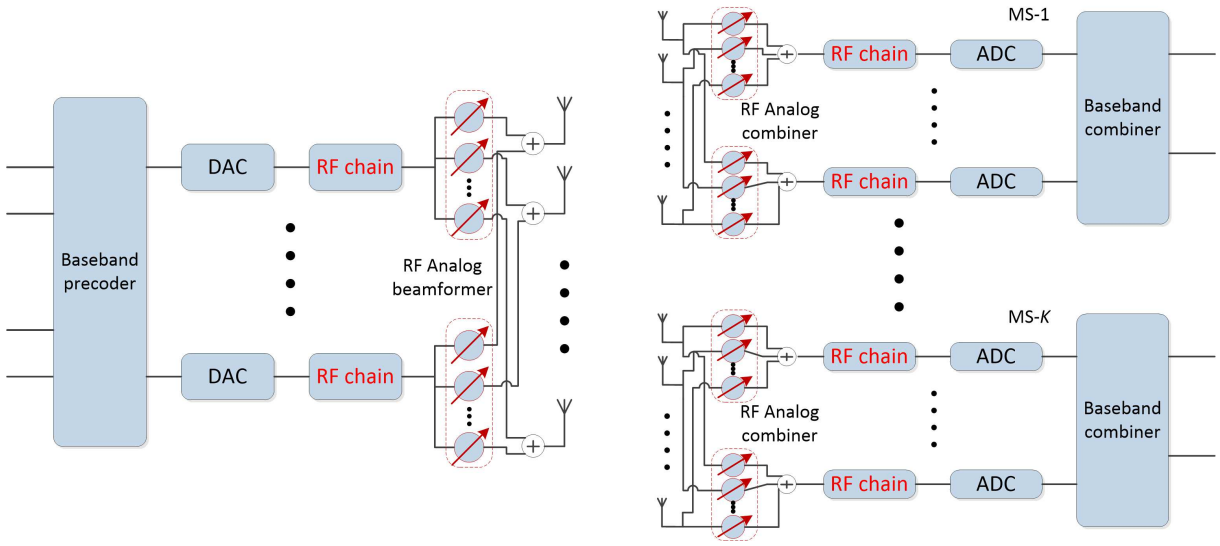

Fig. 1. Diagram of a mmWave multiuser system with hybrid analog/digital precoding and combining.

equalizer to maximize the system sum-rate. This optimization can be stated as

$$\begin{aligned} \underset{\substack{\mathbf{F}_{\mathrm{R}},\mathbf{F}_{\mathrm{B}_1},\ldots,\mathbf{F}_{\mathrm{B}_K}\\ \mathbf{W}_{\mathrm{B}_1},\ldots,\mathbf{W}_{\mathrm{B}_K}}}{\text{maximize}} \quad & \sum_{i=1}^{K} \log\left|\mathbf{I}_{N_{\mathrm{r}}^{\mathrm{RF}}} + \mathbf{R}_i^{-1}\mathbf{W}_{\mathrm{B}_i}^{*}\mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{H}_i\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}_i}\right. \\ & \qquad \left.\times\mathbf{F}_{\mathrm{B}_i}^{*}\mathbf{F}_{\mathrm{R}}^{*}\mathbf{H}_i^{*}\mathbf{W}_{\mathrm{R}_i}\mathbf{W}_{\mathrm{B}_i}\right| \quad (6) \\ \text{subject to} \quad & \mathbf{F}_{\mathrm{R}} \in \mathcal{F},\ \mathbf{W}_{\mathrm{R}_i} \in \mathcal{W}_i, \forall i \\ & \sum_{i=1}^{K} \mathrm{Tr}\left\{\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}_i}\mathbf{F}_{\mathrm{B}_i}^{*}\mathbf{F}_{\mathrm{R}}^{*}\right\} \le P. \end{aligned}$$

In general, the optimization in (6) is a nonconvex problem due to the presence of the variables $\{\mathbf{F}_{\mathrm{R}}, \mathbf{F}_{\mathrm{B}_i}\}$ and $\{\mathbf{W}_{\mathrm{R}_i}, \mathbf{W}_{\mathrm{B}_i}\}$ in the interference plus noise term $\mathbf{R}_i$ and the multiplication of the variables. In addition, the nonconvex constraints on $\mathbf{F}_{\mathrm{R}}$ and $\mathbf{W}_{\mathrm{R}_i}$'s make the optimization problem even more challenging to solve. Thus, obtaining a globally optimal solution of problem (6) is not only complicated, but also intractable for practical implementation. Instead, by taking advantage of the sparse-scattering property in the mmWave propagation environment, we then develop low-complexity, yet efficient hybrid analog/digital precoding and combining algorithms for sum-rate maximization problem (6).

III. Two-stage Hybrid Block-Diagonalization Precoding/Combining

In recent papers [7], [8], hybrid multiuser ZF precoding has been developed for multiuser mmWave systems, where each MS is equipped with *one* RF chain. In this section, we generalize the two-stage multiuser hybrid precoding design algorithm proposed in [7], [8] to the system with multiple-RF-chain MSs. In the first stage of the algorithm, $R_{\mathrm{r}}^{\mathrm{RF}}$ "best" RF single-user RF precoding/combining vector pairs for each MS-i are selected to maximize the signal power for user-i while overlooking the inter-user interference, *i.e.*,

$$(\mathbf{f}_{\mathrm{R}_i}^{(1)}, \mathbf{w}_{\mathrm{R}_i}^{(1)}) = \arg \max_{\substack{\mathbf{w}_{\mathrm{R}_i}\in\mathcal{W}_i\\ \mathbf{f}_{\mathrm{R}_i}\in\mathcal{F}}} \left|\mathbf{w}_{\mathrm{R}_i}^{*}\mathbf{H}_i\mathbf{f}_{\mathrm{R}_i}\right| \quad (7)$$

$$\vdots$$

$$(\mathbf{f}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}})}, \mathbf{w}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}})}) = \arg \max_{\substack{\mathbf{w}_{\mathrm{R}_i}\in\mathcal{W}_i\setminus\left\{\mathbf{w}_{\mathrm{R}_i}^{(1)},\ldots,\mathbf{w}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}}-1)}\right\}\\ \mathbf{f}_{\mathrm{R}_i}\in\mathcal{F}\setminus\left\{\mathbf{f}_{\mathrm{R}_i}^{(1)},\ldots,\mathbf{f}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}}-1)}\right\}}} \left|\mathbf{w}_{\mathrm{R}_i}^{*}\mathbf{H}_i\mathbf{f}_{\mathrm{R}_i}\right|.$$

MS-i then sets $\mathbf{W}_{\mathrm{R}_i} = [\mathbf{w}_{\mathrm{R}_i}^{(1)}, \ldots, \mathbf{w}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}})}]$ as its RF combiner, whereas the BS forms its RF precoding matrix as $\mathbf{F}_{\mathrm{R}} = [\mathbf{F}_{\mathrm{R}_1}, \ldots, \mathbf{F}_{\mathrm{R}_K}]$ with $\mathbf{F}_{\mathrm{R}_i} = [\mathbf{f}_{\mathrm{R}_i}^{(1)}, \ldots, \mathbf{f}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}})}]$. Effectively, $\bar{\mathbf{H}}_i \triangleq \mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{H}_i\mathbf{F}_{\mathrm{R}}$ can be regarded as the downlink channel to MS-i. In the second stage of the algorithm, a baseband BD precoder [39] can be derived, based on the effective downlink channels to all the users.

Remark 1: While being simple to implement, ZF or BD precoding performs poorly at low signal-to-noise ratio (SNR) regime, since the multiuser interference suppression comes with the expense of noise enhancement. In addition, the performance of ZF or precoding does not grow linearly in fully loaded systems where the number of users is the same as the number of transmit antennas [40], [41]. In the two-stage algorithm proposed in [8], a ZF baseband precoder is designed to serve K users by using only K RF chains. Similarly, $KN_{\mathrm{r}}^{\mathrm{RF}}$ RF chains will be used in this two-stage hybrid BD precoder even if $N_{\mathrm{t}}^{\mathrm{RF}} > KN_{\mathrm{r}}^{\mathrm{RF}}$. Thus, the drawbacks of the baseband ZF and BD precoders may become the limiting factor to the system sum-rate performance with increasing number of users K.

IV. Hybrid MMSE Precoding Design with Pre-determined RF Combiners

In this section, we are interested in proposing a multiuser hybrid precoding design when the RF combiner at each MS is pre-determined. Unlike the approach mentioned in Section III,

where the RF beamformer/combiner is obtained independently for each BS-MS link, the proposed approach allows a joint design of RF and baseband precoders for all the MSs.

In the first stage, each MS-i independently decides its RF combining matrix $\mathbf{W}_{\mathrm{R}_i}$ to maximize its downlink channel gains while ignoring the inter-user interferences. Specifically, the $R_{\mathrm{r}}^{\mathrm{RF}}$ RF combining vectors at MS-i are selected as follows:

$$\begin{aligned}\mathbf{w}_{\mathrm{R}_i}^{(1)} &= \arg \max_{\mathbf{w}_{\mathrm{R}_i}\in\mathcal{W}_i} \left\|\mathbf{w}_{\mathrm{R}_i}^{*}\mathbf{H}_i\right\| \\ &\vdots \\ \mathbf{w}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}})} &= \arg \max_{\mathbf{w}_{\mathrm{R}_i}\in\mathcal{W}_i\setminus\left\{\mathbf{w}_{\mathrm{R}_i}^{(1)},\ldots,\mathbf{w}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}}-1)}\right\}} \left\|\mathbf{w}_{\mathrm{R}_i}^{*}\mathbf{H}_i\right\|. \quad (8)\end{aligned}$$

MS-i then sets $\mathbf{W}_{\mathrm{R}_i} = [\mathbf{w}_{\mathrm{R}_i}^{(1)},\ldots,\mathbf{w}_{\mathrm{R}_i}^{(R_{\mathrm{r}}^{\mathrm{RF}})}]$ as its RF combiner. Denote $\tilde{\mathbf{H}}_i \triangleq \mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{H}_i \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{RF}}\times N_{\mathrm{t}}}$ as the effective channel from the BS to MS-i and $\tilde{\mathbf{z}}_i \triangleq \mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{z}_i$ as the colored noise at MS-i. Also denote $\tilde{\mathbf{R}}_i = \mathbf{W}_{\mathrm{B}_i}^{*}\Big(\tilde{\mathbf{H}}_i\sum_{j\neq i}^{K}\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}_j}\mathbf{F}_{\mathrm{B}_j}^{*}\mathbf{F}_{\mathrm{R}}^{*}\tilde{\mathbf{H}}_i^{*}+\sigma^2\mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{W}_{\mathrm{R}_i}\Big)\mathbf{W}_{\mathrm{B}_i}$.

In the second stage, the optimal hybrid precoder and the baseband equalizer by solving problem (6) with known RF combiners $\mathbf{W}_{\mathrm{R}_i}$. Even so, the sum-rate maximization is still a nonconvex problem/ In this section, iwe first examine a suboptimal and non-iterative fully-digital MMSE precoder and its spatially sparse approximation. We subsequently propose our hybrid MMSE precoding design counterpart.

A. A Fully Digital MMSE Multiuser Precoding Design

To address the drawbacks of BD precoding as mentioned in Remark 1, several MMSE-based multiuser precoding techniques have been proposed, including a joint iterative method [42] and non-iterative methods [41], [43]. The aim of MMSE precoding is to generate the transmit precoder which results in the received signal vector $\hat{\mathbf{s}} = [\hat{\mathbf{s}}_1,\ldots,\hat{\mathbf{s}}_K]^{\top}$ as close as possible to the original signal vector $\mathbf{s}$. With a fully digital MMSE precoder, an $N_{\mathrm{t}}\times N_{\mathrm{r}}^{\mathrm{RF}}$ precoder $\mathbf{F}_i^{\mathrm{MMSE}}$ is employed in place of $\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}_i}$. Let $\mathbf{T} = [\mathbf{T}_1,\ldots,\mathbf{T}_K]$ be an *unnormalized* MMSE precoder at the BS and γ be the power gain factor to meet the power constraint at the BS. Next, we define $\mathbf{U}_i^{*}$ and $\mathbf{V}_i$ as the unitary matrices employed at MS-i and the BS in order to decompose the block channel $\mathbf{H}_i\mathbf{T}_i$ into parallel subchannels. Finally, the fully digital MMSE precoder $\mathbf{F}_i^{\mathrm{MMSE}}$ for MS-i is constructed as $\sqrt{1/\gamma}\mathbf{T}_i\mathbf{V}_i$, whereas the baseband equalizer $\mathbf{W}_{\mathrm{B}_i}$ at MS-i is given by $\sqrt{\gamma}\mathbf{U}_i$.

Substitute $\mathbf{F}_i^{\mathrm{MMSE}} = \sqrt{1/\gamma}\mathbf{T}_i\mathbf{V}_i$ and $\mathbf{W}_{\mathrm{B}_i} = \sqrt{\gamma}\mathbf{U}_i$ into the received signal model in (4), the estimated symbol vector at MS-i is given by

$$\hat{\mathbf{s}}_i = \mathbf{U}_i^{*}\tilde{\mathbf{H}}_i\mathbf{T}_i\mathbf{V}_i\mathbf{s}_i + \mathbf{U}_i^{*}\tilde{\mathbf{H}}_i\sum_{j\neq i}^{K}\mathbf{T}_j\mathbf{V}_j\mathbf{s}_j + \sqrt{\gamma}\mathbf{U}_i^{*}\tilde{\mathbf{z}}_i. \quad (9)$$

The estimated signal vectors for all MSs given (9) can be stacked into one vector:

$$\hat{\mathbf{s}} = \mathbf{U}^{*}\tilde{\mathbf{H}}\mathbf{T}\mathbf{V}\mathbf{s} + \sqrt{\gamma}\mathbf{U}^{*}\tilde{\mathbf{z}} \quad (10)$$

where $\tilde{\mathbf{H}} = [\tilde{\mathbf{H}}_1^{\top},\ldots,\tilde{\mathbf{H}}_K^{\top}]^{\top}$, $\mathbf{T} = [\mathbf{T}_1,\ldots,\mathbf{T}_K]$, $\mathbf{V} = \mathrm{blkdiag}(\mathbf{V}_1,\ldots,\mathbf{V}_K)$, $\mathbf{U} = \mathrm{blkdiag}(\mathbf{U}_1,\ldots,\mathbf{U}_K)$, and $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1^{\top},\ldots,\tilde{\mathbf{z}}_K^{\top}]^{\top}$. An MSE-based optimization for designing the precoder and combiner is to minimize the MSE cost function $\mathbb{E}\left[\left\|\mathbf{U}_i^{*}\mathbf{V}_i\mathbf{s}_i - \hat{\mathbf{s}}_i\right\|^2\right]$. This cost function can be expanded as

$$\begin{aligned}&\sum_{i=1}^{K}\mathbb{E}\left[\left\|\mathbf{U}_i^{*}\mathbf{V}_i\mathbf{s}_i - \hat{\mathbf{s}}_i\right\|^2\right] \\ &= \left\|\mathbf{U}^{*}\mathbf{V} - \mathbf{U}^{*}\tilde{\mathbf{H}}\mathbf{T}\mathbf{V}\right\|_F^2 + \gamma\sigma^2\sum_{i=1}^{K}\mathrm{Tr}\{\mathbf{U}_i^{*}\mathbf{U}_i\} \\ &= \left\|\mathbf{I}_{KN_{\mathrm{r}}^{\mathrm{RF}}} - \tilde{\mathbf{H}}\mathbf{T}\right\|_F^2 + \gamma K N_{\mathrm{r}}^{\mathrm{RF}}\sigma^2. \quad (11)\end{aligned}$$

Thus, the MSE minimization problem can be stated as

$$\begin{aligned}\underset{\mathbf{T}_i\gamma}{\text{minimize}} \quad & \left\|\mathbf{I}_{KN_{\mathrm{r}}^{\mathrm{RF}}} - \tilde{\mathbf{H}}\mathbf{T}\right\|_F^2 + \gamma K N_{\mathrm{r}}^{\mathrm{RF}}\sigma^2 \quad (12)\\ \text{subject to} \quad & \mathrm{Tr}\{\mathbf{T}\mathbf{T}^{*}\} \leq \gamma P.\end{aligned}$$

The optimization in (12) is a convex quadratic problem. Its optimal solution, obtained via standard optimization techniques, can be stated in closed-form as follows [44], [45]:

$$\mathbf{T} = \left(\tilde{\mathbf{H}}^{*}\tilde{\mathbf{H}} + \frac{KR_{\mathrm{r}}^{\mathrm{RF}}\sigma^2}{P}\mathbf{I}_{N_{\mathrm{t}}}\right)^{-1}\tilde{\mathbf{H}}^{*} \quad (13)$$

whereas the optimal scaling factor γ is given by $\sum_{i=1}^{K}\mathrm{Tr}\{\mathbf{T}_i\mathbf{T}_i^{*}\}/P$. Finally, $\mathbf{U}_i$ and $\mathbf{V}_i$ are obtained through the singular-value decomposition of $\tilde{\mathbf{H}}_i\mathbf{T}_i$, *i.e.*, $\tilde{\mathbf{H}}_i\mathbf{T}_i = \mathbf{U}_i\mathbf{D}_i\mathbf{V}_i^{*}$ where $\mathbf{D}_i$ is a diagonal matrix.

B. Spatially Sparse MMSE Precoding

Given a fully digital MMSE precoder $\mathbf{F}_{\mathrm{MMSE}} = [\mathbf{F}_1^{\mathrm{MMSE}},\ldots,\mathbf{F}_K^{\mathrm{MMSE}}]$, a hybrid RF and baseband precoder can be reconstructed such that $\mathbf{F}_{\mathrm{MMSE}} \approx \mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}}$ with $\mathbf{F}_{\mathrm{R}} \in \mathcal{F}$. Based on the sparse approximation framework, an OMP algorithm was proposed in [6] for jointly designing hybrid analog/digital precoder $(\mathbf{F}_{\mathrm{R}}, \mathbf{F}_{\mathrm{B}})$.

For ease of referencing, the pseudo-code of the OMP algorithm [6] is presented in Algorithm 1. The algorithm finds a solution to the sparsity constrained matrix reconstruction problem

$$\begin{aligned}\underset{\tilde{\mathbf{X}}_{\mathrm{B}}}{\text{minimize}} \quad & \left\|\mathbf{B} - \mathbf{C}\mathbf{A}\tilde{\mathbf{X}}_{\mathrm{B}}\right\|_F \quad (14)\\ \text{subject to} \quad & \left\|\mathrm{diag}(\tilde{\mathbf{X}}_{\mathrm{B}}\tilde{\mathbf{X}}_{\mathrm{B}}^{*})\right\|_0 = N^{\mathrm{RF}} \\ & \left\|\mathbf{A}\tilde{\mathbf{X}}_{\mathrm{B}}\right\|_F^2 = P \quad (\textit{if applicable}).\end{aligned}$$

Depending on the parameters $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, Algorithm 1 can be used for sparsely approximating a digital precoder or a digital combiner. Herein, $\mathbf{A} = [\mathbf{a}_1,\ldots,\mathbf{a}_L]$ contains predefined RF beamforming vectors as its columns, and N^{RF} is the number of RF chains at the transmitter or receiver whichever applicable. For a given RF precoder/combiner $\mathbf{X}_{\mathrm{R}}$, the baseband precoder/combiner $\mathbf{X}_{\mathrm{B}}$ obtained in step **9** of Algorithm 1 is the solution to the unconstrained least-square minimization $\left\|\mathbf{B} - \mathbf{C}\mathbf{X}_{\mathrm{R}}\mathbf{X}_{\mathrm{B}}\right\|_F$. For spatially reconstructing $\mathbf{F}_{\mathrm{MMSE}}$ into $\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}}$, the inputs to be passed into Algorithm 1 include: $\mathbf{A}_{\mathcal{F}}$ formed by stacking column-wise all the RF beamforming vectors in $\mathcal{F}$, the digital MMSE precoder $\mathbf{F}_{\mathrm{MMSE}}$ and P as the transmit power at the BS. We then obtain $(\mathbf{F}_{\mathrm{R}}, \mathbf{F}_{\mathrm{B}}) = \mathrm{Sparse\_TX\_RX}(\mathbf{A}_{\mathcal{F}}, \mathbf{F}_{\mathrm{MMSE}}, \mathbf{I}_{N_{\mathrm{t}}}, T, P)$. Due

Algorithm 1: Spatially Sparse Precoding/Combining Design via OMP

Function: $\text{Sparse\_TX\_RX}(\mathbf{A}, \mathbf{B}, \mathbf{C}, N^{\text{RF}}, P)$;
Inputs: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, N^{RF}, and P (*if applicable*);
Outputs: $\mathbf{X}_{\text{R}}$, $\mathbf{X}_{\text{B}}$;
$\mathbf{X}_{\text{res}} = \mathbf{B}$;
$\mathbf{X}_{\text{R}} = \text{Empty}$;
for $r \leq N^{\text{RF}}$ **do**
  $\mathbf{\Phi} = \mathbf{A}^* \mathbf{C}^* \mathbf{X}_{\text{res}}$;
  $k = \arg\max_l \left[\mathbf{\Phi}\mathbf{\Phi}^*\right]_{l,l}$;
  $\mathbf{X}_{\text{R}} = \left[\mathbf{X}_{\text{R}} \,|\, \mathbf{A}^{(k)}\right]$;
  $\mathbf{X}_{\text{B}} = \left(\mathbf{X}_{\text{R}}^* \mathbf{C}^* \mathbf{C} \mathbf{X}_{\text{R}}\right)^{-1} \mathbf{X}_{\text{R}}^* \mathbf{C}^* \mathbf{B}$;
  $\mathbf{X}_{\text{res}} = \frac{\mathbf{B} - \mathbf{C}\mathbf{X}_{\text{R}}\mathbf{X}_{\text{B}}}{\|\mathbf{B} - \mathbf{C}\mathbf{X}_{\text{R}}\mathbf{X}_{\text{B}}\|_F}$;
Normalize $\mathbf{X}_{\text{B}} = \sqrt{P} \frac{\mathbf{X}_{\text{B}}}{\|\mathbf{X}_{\text{R}}\mathbf{X}_{\text{B}}\|_F}$, if the power constraint is active.

to the sparse reconstruction of the hybrid precoder from a given MMSE precoder $\mathbf{F}_{\text{MMSE}}$, this precoding design will be referred to as the "Two-stage Sparse MMSE Precoding".

C. Proposed Hybrid MMSE Precoder

In this section, we propose a new hybrid MMSE precoding structure with the objective of jointly minimizing the sum-MSE of all data streams. Instead of approximating a hybrid precoder to a known fully digital precoder $\mathbf{F}_{\text{MMSE}}$ using Algorithm 1, the proposed hybrid precoder can bypass the step of deriving $\mathbf{F}_{\text{MMSE}}$. Denote $\mathbf{T}_{\text{B}_i}$ as an unnormalized baseband precoder for MS-i that is used for inter-user interference mitigation. Replacing $\mathbf{T}_i = \mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}_i}$ to the problem (12), we arrive at a new optimization

$$\begin{aligned} \underset{\mathbf{F}_{\text{R}}, \mathbf{T}_{\text{B}}, \gamma}{\text{minimize}} \;& \left\|\mathbf{I}_{KN_{\text{r}}^{\text{RF}}} - \tilde{\mathbf{H}}\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}}\right\|_F^2 + KN_{\text{r}}^{\text{RF}}\gamma\sigma^2 \qquad (15) \\ \text{subject to} \;& \mathbf{F}_{\text{R}} \in \mathcal{F} \\ & \text{Tr}\left\{\mathbf{F}_{\text{R}}^*\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}}\mathbf{T}_{\text{B}}^*\right\} \leq \gamma P. \end{aligned}$$

Herein, $\mathbf{F}_{\text{R}} \in \mathcal{F}$ is added to constrain the RF beamformers in the codebook $\mathcal{F}$.

We note that problem (15) is nonconvex due to the multiplication of the variables $\mathbf{F}_{\text{R}}$ and $\mathbf{T}_{\text{B}}$ and the nonconvex constraint on $\mathbf{F}_{\text{R}}$. However, for a known RF precoder, problem (15) becomes the convex problem (12) its optimal solution can be derived in closed-form [46]:

$$\mathbf{T}_{\text{B}} = \left(\mathbf{F}_{\text{R}}^*\tilde{\mathbf{H}}^*\tilde{\mathbf{H}}\mathbf{F}_{\text{R}} + \frac{KN_{\text{r}}^{\text{RF}}\sigma^2}{P}\mathbf{F}_{\text{R}}^*\mathbf{F}_{\text{R}}\right)^{\dagger} \mathbf{F}_{\text{R}}^*\tilde{\mathbf{H}}^*. \quad (16)$$

whereas the optimal scaling factor is $\gamma = \|\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}}\|_F^2/P$. Let the SVD of the effective channel $\tilde{\mathbf{H}}_i\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}_i}$ be $\mathbf{U}_i\mathbf{D}_i\mathbf{V}_i^*$. The optimal baseband precoder $\mathbf{F}_{\text{B}_i}$ is then given by $\sqrt{1/\gamma}\mathbf{T}_{\text{B}_i}\mathbf{V}_i$, whereas the baseband combiner is set as $\mathbf{W}_{\text{B}_i} = \mathbf{U}_i$.

In order to find the RF precoder $\mathbf{F}_{\text{R}}$, we take a similar approach as in [6] by restricting its search within a set of predetermined basis vectors in $\mathcal{F}$. Our proposed hybrid precoder is obtained from solving the optimization

$$\begin{aligned} \underset{\tilde{\mathbf{T}}_{\text{B}}, \gamma}{\text{minimize}} \;& \left\|\mathbf{I}_{KN_{\text{r}}^{\text{RF}}} - \tilde{\mathbf{H}}\mathbf{A}_{\mathcal{F}}\tilde{\mathbf{T}}_{\text{B}}\right\|_F^2 + KN_{\text{r}}^{\text{RF}}\gamma\sigma^2 \qquad (17) \\ \text{subject to} \;& \left\|\text{diag}\{\tilde{\mathbf{T}}_{\text{B}}\tilde{\mathbf{T}}_{\text{B}}^*\}\right\|_0 = N_{\text{t}}^{\text{RF}} \\ & \text{Tr}\left\{\mathbf{A}_{\mathcal{F}}^*\mathbf{A}_{\mathcal{F}}\tilde{\mathbf{T}}_{\text{B}}\tilde{\mathbf{T}}_{\text{B}}^*\right\} \leq \gamma P \end{aligned}$$

Algorithm 2: Proposed Hybrid MMSE Precoding via OMP

Inputs: $\tilde{\mathbf{H}}$, $\mathbf{A}_{\mathcal{F}}$, N_{t}^{RF}, P;
Outputs: $\mathbf{F}_{\text{R}}$, $\mathbf{F}_{\text{B}}$, $\mathbf{W}_{\text{B}}$;
$\mathbf{V}_{\text{res}} = \mathbf{I}$;
$\mathbf{F}_{\text{R}} = \text{Empty}$;
for $r \leq N_{\text{t}}^{\text{RF}}$ **do**
  $\mathbf{\Phi} = \mathbf{A}_{\mathcal{F}}^H\tilde{\mathbf{H}}^*\mathbf{T}_{\text{res}}$;
  $k = \arg\max_l \left[\mathbf{\Phi}\mathbf{\Phi}^*\right]_{l,l}$;
  $\mathbf{F}_{\text{R}} = [\mathbf{F}_{\text{R}}|\mathbf{A}_{\mathcal{F}}^{(k)}]$;
  $\mathbf{T}_{\text{B}} = \left(\mathbf{F}_{\text{R}}^*\tilde{\mathbf{H}}^*\tilde{\mathbf{H}}\mathbf{F}_{\text{R}} + \frac{KN_{\text{r}}^{\text{RF}}\sigma^2}{P}\mathbf{F}_{\text{R}}^*\mathbf{F}_{\text{R}}\right)^{\dagger} \mathbf{F}_{\text{R}}^*\tilde{\mathbf{H}}^*$;
  $\mathbf{T}_{\text{res}} = \frac{\mathbf{I}_{KN_{\text{r}}^{\text{RF}}} - \tilde{\mathbf{H}}\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}}}{\|\mathbf{I}_{KN_{\text{r}}^{\text{RF}}} - \tilde{\mathbf{H}}\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}}\|_F}$;
$\gamma = \frac{\|\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}}\|_F^2}{P}$;
for $i \leq K$ **do**
  Perform the SVD: $\tilde{\mathbf{H}}_i\mathbf{F}_{\text{R}}\mathbf{T}_{\text{B}_i} = \mathbf{U}_i\mathbf{D}_i\mathbf{V}_i^*$;
  Set $\mathbf{W}_{\text{B}_i} = \mathbf{U}_i$;
  $\mathbf{F}_{\text{B}_i} = \frac{1}{\sqrt{\gamma}}\mathbf{T}_{\text{B}_i}\mathbf{V}_i$;

where the constraint $\mathbf{F}_{\text{R}} \in \mathcal{F}$ is embedded into the objective function. Thanks to the sparsity constraint $\left\|\text{diag}\{\tilde{\mathbf{T}}_{\text{B}}\tilde{\mathbf{T}}_{\text{B}}^*\}\right\|_0 = N_{\text{t}}^{\text{RF}}$, no more than N_{t}^{RF} rows of $\tilde{\mathbf{T}}_{\text{B}}$ are non-zero. These N_{t}^{RF} non-zero rows are selected to form the baseband precoder $\mathbf{F}_{\text{B}}$ subject to a power scaling step, whereas the corresponding N_{t}^{RF} columns of $\mathbf{A}_{\mathcal{F}}$ are selected to form the RF precoder $\mathbf{F}_{\text{R}}$. Since problem (17) resembles optimization problems usually encountered in sparse signal recovery, extensive literature on this topic can be readily used to solve it. Here, we present an algorithmic solution based on the concept of OMP [20], [47]. The pseudo-code for the hybrid MMSE precoder solution is given in Algorithm 2. Our choice of an OMP-based algorithm is to provide a direct comparison the the OMP-based spatially precoding/combining solution in Algorithm 1. We note that step **9** of Algorithm 2 utilizes the solution (16) of the MSE minimization problem (17) as the baseband precoder. This is the key difference to step **9** in Algorithm 1.

V. Hybrid WMMSE Precoding/Combining Design

In this section, we seek to design hybrid mmWave precoders $(\mathbf{F}_{\text{R}}, \mathbf{F}_{\text{B}_i})$ and hybrid mmWave combiners $(\mathbf{W}_{\text{R}_i}, \mathbf{F}_{\text{B}_i})$ to maximize the total spectral efficiency. To this end, we first revisit an iterative algorithm that is capable of achieving a locally optimal digital precoder/combiner to the sum-rate maximization problem. We then propose an iterative algorithm to obtain the hybrid mmWave precoding/combining $(\mathbf{F}_{\text{R}}, \mathbf{F}_{\text{B}_i}, \mathbf{W}_{\text{R}_i}, \mathbf{F}_{\text{B}_i})$ counterpart.

A. A Locally Optimal Digital Precoding/Combining Design

When abstracting the receiver operation and dropping the constraints on $\mathbf{F}_{\text{R}}$ and $\mathbf{W}_{\text{R}_i}$, a suboptimal solution to the sum-

rate maximization problem (6) can be found based on the MSE criterion [34]:

$$\begin{aligned} \underset{\mathbf{W}_i,\mathbf{\Omega}_i,\mathbf{F}_i}{\text{minimize}} \quad & \sum_{i=1}^{K} \left[\mathrm{Tr}\left\{ \mathbf{\Omega}_i \mathbf{E}_i^{\mathrm{D}} \right\} - \log |\mathbf{\Omega}_i| \right] \\ \text{subject to} \quad & \sum_{i=1}^{K} \mathrm{Tr}\left\{ \mathbf{F}_i \mathbf{F}_i^* \right\} \leq P. \end{aligned} \tag{18}$$

Herein, the MSE covariance matrix is defined as

$$\mathbf{E}_i^{\mathrm{D}} = \mathbb{E}\left[\left(\mathbf{s}_i - \mathbf{W}_i^* \mathbf{y}_i \right) \left(\mathbf{s}_i - \mathbf{W}_i^* \mathbf{y}_i \right)^* \right]. \tag{19}$$

and $\mathbf{\Omega}_i$ is the weight matrix associated with $\mathbf{E}_i^{\mathrm{D}}$. Problem (18) is usually referred to as the minimization of weighted mean squared error (WMMSE) problem in literature [34], [48]. Although Problem (18) is still not a jointly convex problem, it is convex over each set of variables $\mathbf{W}_i$, $\mathbf{\Omega}_i$ and $\mathbf{F}_i$. A three-step iterative algorithm, namely WMMSE, was then proposed in [34] to sequentially update $\mathbf{W}_i$, $\mathbf{\Omega}_i$ and $\mathbf{F}_i$ until convergence to a locally optimal solution. We denote such solution as $(\mathbf{F}_i^{\mathrm{opt}}, \mathbf{W}_i^{\mathrm{opt}})$. Since the WMMSE algorithm can only converge to a locally optimal solution, its performance depends on its starting points. To address this issue, multiple random initializations can be used and the best performing solution is then chosen [48].

B. Spatially Sparse WMMSE Precoding/Combining

Given the fully digital WMMSE precoders and combiners $(\mathbf{F}_i^{\mathrm{opt}}, \mathbf{W}_i^{\mathrm{opt}})$, Algorithm 1 can be recalled to find the sparse hybrid precoder $(\mathbf{F}_{\mathrm{R}}, \mathbf{F}_{\mathrm{B}_i})$ and hybrid combiners $(\mathbf{W}_{\mathrm{R}_i}, \mathbf{W}_{\mathrm{B}_i})$, respectively. For the precoder, we want to jointly approximate $\mathbf{F}_{\mathrm{opt}} \triangleq [\mathbf{F}_1^{\mathrm{opt}}, \ldots, \mathbf{F}_K^{\mathrm{opt}}] \approx \mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}}$ by $(\mathbf{F}_{\mathrm{R}}, \mathbf{F}_{\mathrm{B}}) = \mathrm{Sparse\_TX\_RX}(\mathbf{A}_{\mathcal{F}}, \mathbf{F}_{\mathrm{opt}}, \mathbf{I}_{N_{\mathrm{t}}}, T, P)$. For the combiner at each MS, finding a sparse hybrid combining structure is more involved. A hybrid combiner $(\mathbf{W}_{\mathrm{R}_i}, \mathbf{W}_{\mathrm{B}_i})$ that minimizes the MSE of the transmitted and received signals at MS-i can be found by

$$\begin{aligned} \underset{\mathbf{W}_{\mathrm{R}_i},\mathbf{W}_{\mathrm{B}_i}}{\text{minimize}} \quad & \mathbb{E}\left[\left\| \mathbf{s}_i - \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{y}_i \right\|^2 \right] \\ \text{subject to} \quad & \mathbf{W}_{\mathrm{R}_i} \in \mathcal{W}_i. \end{aligned} \tag{20}$$

This problem is shown to equivalent to [6], [28]:

$$\begin{aligned} \underset{\mathbf{W}_{\mathrm{R}_i},\mathbf{W}_{\mathrm{B}_i}}{\text{minimize}} \quad & \left\| \mathbb{E}[\mathbf{y}_i \mathbf{y}_i^*]^{\frac{1}{2}} \left(\mathbf{W}_i^{\mathrm{opt}} - \mathbf{W}_{\mathrm{R}_i} \mathbf{W}_{\mathrm{B}_i} \right) \right\|_F \\ \text{subject to} \quad & \mathbf{W}_{\mathrm{R}_i} \in \mathcal{W}_i \end{aligned} \tag{21}$$

where $\mathbb{E}[\mathbf{y}_i \mathbf{y}_i^*] = \mathbf{H}_i \mathbf{F}_{\mathrm{opt}} \mathbf{F}_{\mathrm{opt}}^* \mathbf{H}_i^* + \sigma^2 \mathbf{I}_{N_{\mathrm{r}}}$. Algorithm 1 can be readily applied to find the sparse combiner at MS-i as $(\mathbf{W}_{\mathrm{R}_i}, \mathbf{W}_{\mathrm{B}_i}) = \mathrm{Sparse\_TX\_RX}\big(\mathbf{A}_{\mathcal{W}_i}, \mathbb{E}[\mathbf{y}_i \mathbf{y}_i^*]^{\frac{1}{2}} \mathbf{W}_i^{\mathrm{opt}}, \mathbb{E}[\mathbf{y}_i \mathbf{y}_i^*]^{\frac{1}{2}}, N_{\mathrm{r}}^{\mathrm{RF}}\big)$, where $\mathbf{A}_{\mathcal{W}_i}$ contains the RF receive beamsteering codebook vectors in $\mathcal{W}_i$. Hereafter, the hybrid precoder/combiner obtained from the sparse reconstruction of the WMMSE precoder/combiner is referred to as "Two-stage Sparse WMMSE Precoding/Combining".

C. Proposed Hybrid WMMSE Precoder/Combiner

Based on the algorithmic approach to obtain the digital WMMSE precoder/combiner [34] and the proposed hybrid MMSE precoder in Section IV-C, we seek to design a hybrid WMMSE precoding/combining counterpart. We note that a prior digital WMMSE precoder/combiner is not needed. Instead, all the variables $(\mathbf{F}_{\mathrm{R}}, \mathbf{F}_{\mathrm{B}_i}, \mathbf{W}_{\mathrm{R}_i}, \mathbf{F}_{\mathrm{B}_i})$ are jointly designed aiming to minimize a weighted sum-MSE cost function as given in (18), *i.e.*,

$$\begin{aligned} \underset{\substack{\mathbf{W}_{\mathrm{R}_i},\mathbf{W}_{\mathrm{B}_i},\mathbf{\Omega}_i \\ \mathbf{F}_{\mathrm{R}},\mathbf{F}_{\mathrm{B}_i}}}{\text{minimize}} \quad & \sum_{i=1}^{K} \left[\mathrm{Tr}\left\{ \mathbf{\Omega}_i \mathbf{E}_i^{\mathrm{HB}} \right\} - \log |\mathbf{\Omega}_i| \right] \\ \text{subject to} \quad & \mathbf{F} \in \mathcal{F},\ \mathbf{W}_{\mathrm{R}_i} \in \mathcal{W}_i, \forall i \\ & \sum_{i=1}^{K} \mathrm{Tr}\left\{ \mathbf{F}_{\mathrm{R}}^* \mathbf{F}_{\mathrm{R}} \mathbf{F}_{\mathrm{B}_i} \mathbf{F}_{\mathrm{B}_i}^* \right\} \leq P \end{aligned} \tag{22}$$

where the MSE covariance matrix is given by

$$\begin{aligned} &\mathbf{E}_i^{\mathrm{HB}} \\ &= \mathbb{E}\left[\left(\mathbf{s}_i - \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{y}_i \right) \left(\mathbf{s}_i - \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{y}_i \right)^* \right] \\ &= \left(\mathbf{I}_{N_{\mathrm{r}}^{\mathrm{RF}}} - \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{H}_i \mathbf{F}_{\mathrm{R}} \mathbf{F}_{\mathrm{B}_i} \right) \left(\mathbf{I}_{N_{\mathrm{r}}^{\mathrm{RF}}} - \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{H}_i \mathbf{F}_{\mathrm{R}} \mathbf{F}_{\mathrm{B}_i} \right)^* \\ &\quad + \sum_{j \neq i}^{K} \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{H}_i \mathbf{F}_{\mathrm{R}} \mathbf{F}_{\mathrm{B}_j} \mathbf{F}_{\mathrm{B}_j}^* \mathbf{F}_{\mathrm{R}}^* \mathbf{H}_i^* \mathbf{W}_{\mathrm{R}_i} \mathbf{W}_{\mathrm{B}_i} \\ &\quad + \sigma^2 \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{W}_{\mathrm{R}_i} \mathbf{W}_{\mathrm{B}_i}. \end{aligned} \tag{23}$$

Similar to the three-step iterative algorithm in devising the digital WMMSE precoder/combiner, the proposed hybrid WMMSE precoder/combiner is also obtained through three-step iterative updates. At time-n, the iterative procedure is as follows:

1) Update the hybrid combiners: Fixing the hybrid precoder and the weight matrices at time-$(n-1)$ at $(\check{\mathbf{F}}_{\mathrm{R}}, \check{\mathbf{F}}_{\mathrm{B}_i})$ and $\check{\mathbf{\Omega}}_i$, the hybrid combiner $(\mathbf{W}_{\mathrm{R}_i}^{(n)}, \mathbf{W}_{\mathrm{B}_i}^{(n)})$ is given as the solution of the optimization

$$\begin{aligned} \underset{\mathbf{W}_{\mathrm{R}_i},\mathbf{W}_{\mathrm{B}_i}}{\text{minimize}} \quad & \mathbb{E}\left[\left\| \mathbf{s}_i - \mathbf{W}_{\mathrm{B}_i}^* \mathbf{W}_{\mathrm{R}_i}^* \mathbf{y}_i \right\|^2 \right] \\ \text{subject to} \quad & \mathbf{W}_{\mathrm{R}_i} \in \mathcal{W}_i. \end{aligned} \tag{24}$$

Following similar derivation steps to [6], it can be shown that problem (24) is equivalent to

$$\begin{aligned} \underset{\mathbf{W}_{\mathrm{R}_i},\mathbf{W}_{\mathrm{B}_i}}{\text{minimize}} \quad & \left\| \mathbb{E}[\check{\mathbf{y}}_i \check{\mathbf{y}}_i^*]^{\frac{1}{2}} \left(\mathbf{W}_i^{(n)} - \mathbf{W}_{\mathrm{R}_i} \mathbf{W}_{\mathrm{B}_i} \right) \right\|_F \\ \text{subject to} \quad & \mathbf{W}_{\mathrm{R}_i} \in \mathcal{W}_i \end{aligned} \tag{25}$$

where $\mathbb{E}[\check{\mathbf{y}}_i \check{\mathbf{y}}_i^*] = \sum_{j=1}^{K} \mathbf{H}_i \check{\mathbf{F}}_{\mathrm{R}} \check{\mathbf{F}}_{\mathrm{B}_j} \check{\mathbf{F}}_{\mathrm{B}_j}^* \check{\mathbf{F}}_{\mathrm{R}}^* \mathbf{H}_i^* + \sigma^2 \mathbf{I}_{N_{\mathrm{r}}}$, and $\mathbf{W}_i^{(n)} = \mathbb{E}[\check{\mathbf{y}}_i \check{\mathbf{y}}_i^*]^{-1} \mathbf{H}_i \check{\mathbf{F}}_{\mathrm{R}} \check{\mathbf{F}}_{\mathrm{B}_i}$. Algorithm 1 is then applied to find the sparse combiner at MS-i in which

$$\begin{aligned} &(\mathbf{W}_{\mathrm{R}_i}^{(n)}, \mathbf{W}_{\mathrm{B}_i}^{(n)}) \\ &= \mathrm{Sparse\_TX\_RX}\big(\mathbf{A}_{\mathcal{W}_i}, \mathbb{E}[\check{\mathbf{y}}_i \check{\mathbf{y}}_i^*]^{\frac{1}{2}} \mathbf{W}_i^{(n)}, \mathbb{E}[\check{\mathbf{y}}_i \check{\mathbf{y}}_i^*]^{\frac{1}{2}}, N_{\mathrm{r}}^{\mathrm{RF}}\big). \end{aligned} \tag{26}$$

2) Update the weighted matrix: While fixing the hybrid precoders at $(\check{\mathbf{F}}_{\mathrm{R}}, \check{\mathbf{F}}_{\mathrm{B}_i})$ and hybrid combiners at $(\check{\mathbf{W}}_{\mathrm{R}_i}, \check{\mathbf{W}}_{\mathrm{B}_i})$, the weight matrix at time-n is updated as

$$\mathbf{\Omega}_i^{(n)} = \left(\mathbf{I}_{N_{\mathrm{r}}} - \check{\mathbf{W}}_{\mathrm{B}_i}^* \check{\mathbf{W}}_{\mathrm{R}_i}^* \mathbf{H}_i \check{\mathbf{F}}_{\mathrm{R}} \check{\mathbf{F}}_{\mathrm{B}_i} \right)^{-1}. \tag{27}$$

Algorithm 3: Hybrid WMMSE Precoding via OMP

Inputs: $\check{\mathbf{H}}$, $\mathbf{A}_{\mathrm{t}}$, $N_{\mathrm{t}}^{\mathrm{RF}}$, P, $\check{\mathbf{\Omega}}$;
Outputs: $\check{\mathbf{F}}_{\mathrm{R}}$, $\check{\mathbf{F}}_{\mathrm{B}_i}$;
$\mathbf{V}_{\mathrm{res}} = \mathbf{I}$;
$\mathbf{F}_{\mathrm{R}} = \mathrm{Empty}$;
for $r \leq N_{\mathrm{t}}^{\mathrm{RF}}$ **do**
  $\mathbf{\Phi} = \mathbf{A}_{\mathrm{t}}^{H}\check{\mathbf{H}}^{*}\mathbf{T}_{\mathrm{res}}$;
  $k = \arg\max_l [\mathbf{\Phi}\mathbf{\Phi}^{*}]_{l,l}$;
  $\mathbf{F}_{\mathrm{R}} = [\mathbf{F}_{\mathrm{R}}|\mathbf{A}_{\mathrm{t}}^{(k)}]$;
  $\check{\mathbf{T}}_{\mathrm{B}}$ as given in (29);
  $\mathbf{T}_{\mathrm{res}} = \frac{\check{\mathbf{\Omega}}^{1/2}\left(\mathbf{I}_{KN_{\mathrm{r}}^{\mathrm{RF}}} - \check{\mathbf{H}}\check{\mathbf{F}}_{\mathrm{R}}\check{\mathbf{T}}_{\mathrm{B}}\right)}{\left\|\check{\mathbf{\Omega}}^{1/2}\left(\mathbf{I}_{KN_{\mathrm{r}}^{\mathrm{RF}}} - \check{\mathbf{H}}\check{\mathbf{F}}_{\mathrm{R}}\check{\mathbf{T}}_{\mathrm{B}}\right)\right\|_F}$;
$\gamma = \frac{\|\check{\mathbf{F}}_{\mathrm{R}}\check{\mathbf{T}}_{\mathrm{B}}\|_F^2}{P}$;
$\mathbf{F}_{\mathrm{B}_i} = \sqrt{1/\gamma}\check{\mathbf{T}}_{\mathrm{B}_i}$;

3) Update the hybrid precoder: Finally, we keep the hybrid combiners fixed at $(\check{\mathbf{W}}_{\mathrm{R}_i}, \check{\mathbf{W}}_{\mathrm{B}_i})$, the weighted matrices fixed at $\check{\mathbf{\Omega}}_i$ and update the hybrid precoder $(\mathbf{F}_{\mathrm{R}_i}^{(n)}, \mathbf{F}_{\mathrm{B}_i}^{(n)})$. Let us denote $\check{\mathbf{H}}_i = \mathbf{W}_{\mathrm{B}_i}^{*}\mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{H}_i$ as the effective channel and $\check{\mathbf{z}}_i = \mathbf{W}_{\mathrm{B}_i}^{*}\mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{z}_i$ as the effective colored noise at MS-i. Then, the hybrid precoder to minimize the weighted sum-MSE $\sum_{i=1}^{K}\mathrm{Tr}\left\{\check{\mathbf{\Omega}}_i\mathbf{E}_i^{\mathrm{HB}}\right\}$ can be obtained in a similar fashion as in Section IV-C.

Denote $\check{\mathbf{T}}_{\mathrm{B}_i}$ as an unnormalized baseband precoder and γ as the power scaling factor to meet the power constraint. Under the assumption that all combiners are rescaled by $\sqrt{\gamma}$, the weighted sum-MSE can be reorganized into:

$$\begin{aligned}\sum_{i=1}^{K}\mathrm{Tr}\left\{\check{\mathbf{\Omega}}_i\mathbf{E}_i^{\mathrm{HB}}\right\} &= \sum_{i=1}^{K}\mathbb{E}\left[\left\|\check{\mathbf{\Omega}}_i^{\frac{1}{2}}\left(\mathbf{s}_i - \sqrt{\gamma}\mathbf{W}_{\mathrm{B}_i}^{*}\mathbf{W}_{\mathrm{R}_i}^{*}\mathbf{y}_i\right)\right\|_F^2\right] \\ &= \left\|\check{\mathbf{\Omega}}^{\frac{1}{2}}\left(\mathbf{I}_{KN_{\mathrm{r}}^{\mathrm{RF}}} - \check{\mathbf{H}}_i\check{\mathbf{F}}_{\mathrm{R}}\check{\mathbf{T}}_{\mathrm{B}}\right)\right\|_F^2 \\ &\quad + \gamma\sigma^2\sum_{i=1}^{K}\left\|\check{\mathbf{\Omega}}_i^{\frac{1}{2}}\mathbf{W}_{\mathrm{B}_i}^{*}\mathbf{W}_{\mathrm{R}_i}^{*}\right\|_F^2 \end{aligned} \quad (28)$$

where $\check{\mathbf{H}} = [\check{\mathbf{H}}_1^{\top}, \ldots, \check{\mathbf{H}}_K^{\top}]^{\top}$, $\check{\mathbf{T}}_{\mathrm{B}} = [\check{\mathbf{T}}_{\mathrm{B}_1}, \ldots, \check{\mathbf{T}}_{\mathrm{B}_K}]$, and $\check{\mathbf{\Omega}} = \mathrm{blkdiag}(\check{\mathbf{\Omega}}_1, \ldots, \check{\mathbf{\Omega}}_K)$. Given an RF precoder $\mathbf{F}_{\mathrm{R}}$, the baseband precoder $\mathbf{T}_{\mathrm{B}}$ is then obtained by minimizing the weighted MSE-cost function subject to the power constraint $\mathrm{Tr}\left\{\check{\mathbf{F}}_{\mathrm{R}}^{*}\check{\mathbf{F}}_{\mathrm{R}}\check{\mathbf{T}}_{\mathrm{B}}\check{\mathbf{T}}_{\mathrm{B}}^{*}\right\} \leq \gamma P$. Similar to solving problem (17) such an optimal baseband precoder can be derived in closed-form solution:

$$\check{\mathbf{T}}_{\mathrm{B}} = \left(\check{\mathbf{F}}_{\mathrm{R}}^{*}\check{\mathbf{H}}^{*}\check{\mathbf{\Omega}}\check{\mathbf{H}}\check{\mathbf{F}}_{\mathrm{R}} + \lambda\check{\mathbf{F}}_{\mathrm{R}}^{*}\check{\mathbf{F}}_{\mathrm{R}}\right)^{\dagger}\check{\mathbf{F}}_{\mathrm{R}}^{*}\check{\mathbf{H}}^{*}\check{\mathbf{\Omega}} \quad (29)$$

where $\lambda = \frac{\sigma^2\sum_{i=1}^{K}\left\|\mathbf{\Omega}_i^{\frac{1}{2}}\mathbf{W}_{\mathrm{B}_i}^{*}\mathbf{W}_{\mathrm{R}_i}^{*}\right\|_F^2}{P}$, and $\gamma = \frac{\|\check{\mathbf{F}}_{\mathrm{R}}\check{\mathbf{T}}_{\mathrm{B}}\|_F^2}{P}$. Since the RF precoder is selected from $N_{\mathrm{t}}^{\mathrm{RF}}$ columns of $\mathbf{A}_{\mathcal{F}}$, a modified version of Algorithm 2, presented in Algorithm 3, can be applied to construct $\mathbf{F}_{\mathrm{R}}$ and $\mathbf{F}_{\mathrm{B}}$. The overall algorithm to obtain the proposed hybrid WMMSE precoding/combining design is presented in Algorithm 4.

Remark 2: The iterative nature of Algorithm 4 allows the hybrid combiners at the MSs to match better with the hybrid precoder at the BS and vice versa, compared to the hybrid MMSE precoder obtained from Algorithm 2. Similar to the digital WMMSE precoder/combiner [34], the performance of the proposed hybrid WMMSE precoder/combiner also depends on its starting points. Thus, multiple random initializations should be used and the best performing solution is then chosen. However, convergence of Algorithm 4 is not guaranteed. This is because steps 1 and 3 of the algorithm only yield approximate solutions to the sum-MSE minimization problem. Hence, a maximum number of iterations should be set in the algorithm as a stopping criterion.

Algorithm 4: Proposed Hybrid WMMSE Precoding/Combining Design

Inputs: $\mathbf{H}$, $\mathbf{A}_{\mathrm{t}}$, $\mathbf{A}_{\mathrm{r},i}$, $N_{\mathrm{t}}^{\mathrm{RF}}$, $N_{\mathrm{r}}^{\mathrm{RF}}$, P;
Outputs: $\mathbf{F}_{\mathrm{R}}$, $\mathbf{F}_{\mathrm{B}_i}$, $\mathbf{W}_{\mathrm{R}_i}$, and $\mathbf{W}_{\mathrm{B}_i}$;
Initialize: $\mathbf{F}_{\mathrm{R}}$ and $\mathbf{F}_{\mathrm{B}_i}$ such that $\sum_{i=1}^{K}\|\mathbf{F}_{\mathrm{R}}\mathbf{F}_{\mathrm{B}_i}\|_F^2 = P$;
repeat
  Set $\check{\mathbf{F}}_{\mathrm{R}} \leftarrow \mathbf{F}_{\mathrm{R}}$, $\check{\mathbf{F}}_{\mathrm{B}_i} \leftarrow \mathbf{F}_{\mathrm{B}_i}$;
  Update $\mathbf{W}_{\mathrm{R}_i}$ and $\mathbf{W}_{\mathrm{B}_i}$ as in (26), and set $\check{\mathbf{W}}_{\mathrm{R}_i} \leftarrow \mathbf{W}_{\mathrm{R}_i}$, $\check{\mathbf{W}}_{\mathrm{B}_i} \leftarrow \mathbf{W}_{\mathrm{B}_i}$;
  Update $\mathbf{\Omega}_i$ as in (27), and set $\check{\mathbf{\Omega}}_i \leftarrow \mathbf{\Omega}_i$;
  Update $\mathbf{F}_{\mathrm{R}}$ and $\mathbf{F}_{\mathrm{B}_i}$ as in Algorithm 3;
until *convergence or reaching the maximum number of iterations*;

Remark 3: As an alternative to Algorithm 4, we propose a modified version of the algorithm by setting a small limit on the number of updates on the RF beamformer $\mathbf{F}_{\mathrm{R}}$ and RF combiners $\mathbf{F}_{\mathrm{B}_i}$'s. This adjustment step allows the BS and the MSs to steer the signal beams in strongest transmission paths. After freezing the RF beamformers and combiners, the WMMSE algorithm can be called to optimize the baseband precoder and combiners. The modified algorithm comes with the benefits of guaranteed convergence by the WMMSE algorithm [34], [48] and lower complexity due to the fewer number of approximation steps via the OMP algorithm.

VI. Complexity Analysis of The Proposed Algorithm

In this section, we analyze the complexity in implementing the proposed hybrid BD, WMMSE and WMMSE precoding/combining algorithms. We also present the comparison with the complexity in implementing fully digital precoding/combining algorithms. To simplify the complexity analysis, let us denote $N = \max\left\{N_{\mathrm{t}}, N_{\mathrm{r}}\right\}$, $N_{\mathrm{RF}} = \max\left\{N_{\mathrm{t}}^{\mathrm{RF}}, N_{\mathrm{r}}^{\mathrm{RF}}\right\}$, and $L = \max\left\{|\mathcal{F}|, |\mathcal{W}_i|\right\}$ as the length of the RF codebooks. In Table I, we enumerate the complexity in undertaking major computational steps (by computing the listed variables) and the total complexity of each algorithm. For ease of presentation, we make an assumption $L \leq N$, which is true for AoA/AoD codebooks where the number of transmission paths is much smaller than the number of antennas. For example, the complexity of the OMP-based procedure in Algorithms 1 and 2 is simplified from $\mathcal{O}(LNN_{\mathrm{RF}}^2) + \mathcal{O}(L^3N_{\mathrm{RF}}) + \mathcal{O}(N^3N_{\mathrm{RF}}^2)$ to $\mathcal{O}(N^3N_{\mathrm{RF}}^2)$.

In two-stage hybrid BD/ZF precoding/combining, the primary factor in the algorithm's complexity is in the selection of RF beamformers and combiners (7). This process, which involves an exhaustive search for the codebooks at both the

TABLE I
COMPLEXITY OF THE PROPOSED ALGORITHMIC SOLUTIONS

Algorithm	Operation	Complexity	Number of Operations	Total Complexity
Hybrid ZF/BD	Select RF beamformers and combiners (7)	$\mathcal{O}(L^2N^2)$	K	$\mathcal{O}(KL^2N^2)$
	Compute baseband precoder and equalizers	$\mathcal{O}(N_{\mathrm{RF}}^3)$	K	$\mathcal{O}(KN_{\mathrm{RF}}^3)$
				$\mathcal{O}(KL^2N^2)+\mathcal{O}(KN_{\mathrm{RF}}^3)$
MMSE-based	Select RF combiners (8)	$\mathcal{O}(LN^2)$	K	$\mathcal{O}(KLN^2)$
Fully Digital	Compute digital precoder (13)	$\mathcal{O}(N^3)$	1	$\mathcal{O}(N^3)$
				$\mathcal{O}(KN^3)$
Two-stage Sparse	Compute precoder with OMP	$\mathcal{O}(N^3N_{\mathrm{RF}})$	1	$\mathcal{O}(N^3N_{\mathrm{RF}})$
				$\mathcal{O}(KN^3)+\mathcal{O}(N^3N_{\mathrm{RF}})$
Proposed Hybrid	Compute precoder with Alg. 2	$\mathcal{O}(N^3N_{\mathrm{RF}})$	1	$\mathcal{O}(N^3N_{\mathrm{RF}})$
				$\mathcal{O}(KN^3)+\mathcal{O}(N^3N_{\mathrm{RF}})$
WMMSE-based				
Fully Digital	Compute combiner $\mathbf{W}_i$	$\mathcal{O}(N^3)$	K	$\mathcal{O}(KN^3)$
	Compute weight matrix $\mathbf{\Omega}_i$	$\mathcal{O}(N^3)$	K	$\mathcal{O}(KN^3)$
	Compute precoder $\mathbf{F}_i$	$\mathcal{O}(N^3)$	K	$\mathcal{O}(KN^3)$
				$\mathcal{O}(KN^3)$
Two-stage Sparse	Compute precoder with OMP	$\mathcal{O}(N^3N_{\mathrm{RF}})$	1	$\mathcal{O}(N^3N_{\mathrm{RF}})$
	Compute combiner with OMP	$\mathcal{O}(N^3N_{\mathrm{RF}})$	K	$\mathcal{O}(KN^3N_{\mathrm{RF}})$
				$\mathcal{O}(KN^3N_{\mathrm{RF}})$
Proposed Hybrid	Compute precoder with OMP	$\mathcal{O}(N^3N_{\mathrm{RF}})$	1	$\mathcal{O}(N^3N_{\mathrm{RF}})$
	Compute weight matrix $\mathbf{\Omega}_i$	$\mathcal{O}(N^3)$	K	$\mathcal{O}(KN^3)$
	Compute precoder with OMP	$\mathcal{O}(N^3N_{\mathrm{RF}})$	K	$\mathcal{O}(KN^3N_{\mathrm{RF}})$
				$\mathcal{O}(KN^3N_{\mathrm{RF}})$
OMP	Calculate $\mathbf{\Phi}$	$\mathcal{O}(LNN_{\mathrm{RF}})$	N_{RF}	$\mathcal{O}(LNN_{\mathrm{RF}}^2)$
	Find k	$\mathcal{O}(L^3)$	N_{RF}	$\mathcal{O}(L^3N_{\mathrm{RF}})$
	Calculate $\mathbf{X}_{\mathrm{B}}$	$\mathcal{O}(N^3)$	N_{RF}	$\mathcal{O}(N^3N_{\mathrm{RF}})$
	Calculate $\mathbf{X}_{\mathrm{res}}$	$\mathcal{O}(N^3)$	N_{RF}	$\mathcal{O}(N^3N_{\mathrm{RF}})$
				$\mathcal{O}(N^3N_{\mathrm{RF}})$

BS and the MSs, yields the complexity order of $\mathcal{O}(KL^2N^2)$. For all other hybrid precoding/combining designs, the primary factor in their complexity comes from the OMP-based procedure. With MMSE-based algorithms, the selection of RF combiners at the MSs has the complexity of $\mathcal{O}(KLN^2)$. The two-stage sparse MMSE precoding and the proposed hybrid MMSE precoding designs then include the OMP-based algorithms, which yield the same total complexity order $\mathcal{O}(KN^3)+\mathcal{O}(N^3N_{\mathrm{RF}})$.

With WMMSE-based algorithms, while the complexity of the fully digital design is $\mathcal{O}(KN^3)$, the complexity of the two-stage sparse hybrid WMMSE precoder/combiner is increased to $\mathcal{O}(KN^3N_{\mathrm{RF}})$. With the proposed hybrid WMMSE precoding/combining design, the complexity at each iteration is also $\mathcal{O}(KN^3N_{\mathrm{RF}})$. However, the complexity of the whole hybrid WMMSE algorithm is magnified by the number of required iterations until convergence. In the modified WMMSE algorithm, after a small number of updates on the RF beamformers and combiners, the WMMSE algorithm then proceeds with the optimization of the baseband precoder/combiner only. This modified algorithm can significantly reduce the complexity due to the OMP-based algorithm. The WMMSE algorithm for optimizing the baseband precoder/combiner in a "simplified" K users $N_{\mathrm{t}}^{\mathrm{RF}}\times N_{\mathrm{r}}^{\mathrm{RF}}$ then yields the complexity $\mathcal{O}(KN_{\mathrm{RF}}^3)\ll\mathcal{O}(KN^3)$. It is noted that the complexity analysis of iterative WMMSE-based algorithms presented in Table I is applicable to per-iteration basis. Certainly, the number of iterations has the impact on the overall complexity of the algorithm. Thus, to keep the complexity of the proposed algorithms manageable, we limit the number of iterations to 100 in the simulations.

VII. SIMULATION RESULTS

In this section, we present numerical results to illustrate the performance advantages of the proposed hybrid MMSE-based precoding/combing designs. We compare our hybrid designs to two fully digital precoding/combining designs (the MMSE precoder and the WMMSE precoder/combiner) and two-stage sparse designs obtained by approximating these digital ones. Also presented are the performances of the two-stage hybrid ZF/BD precoding. With WMMSE-based algorithms (for both digital and hybrid designs), 10 randomized initializations are generated and the best performing solution is chosen. In some of the figures, we also plot the multiuser capacity limit attained by dirty paper coding (DPC) [49]. To provide a fair comparison with other fully digital and hybrid precoding/combining design, the number of data streams transmitted to a MS using DPC must be also set at $N_{\mathrm{r}}^{\mathrm{RF}}$. This constraint sets a limit on the rank of transmit covariance for each user and requires an alteration to the DPC optimization problem. We refer the interested readers to [50] for algorithmic solutions to optimize DPC with rank constraints.

We consider a MIMO system where the BS is equipped with 8×8 UPA ($M=64$) and each MS is equipped with 4×4 UPA ($N=16$). There are $K=8$ MSs, unless stated otherwise. The channel to each user contains 3 clusters with 6 rays per cluster ($L_i=18,\forall i$). The azimuth and evaluation

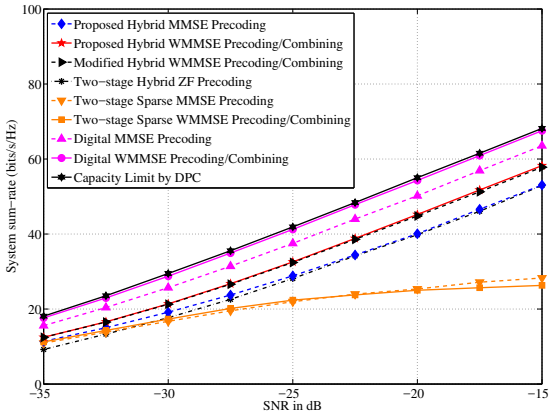

Fig. 2. Achievable system sum-rate *versus* SNR with AoA/AoD codebooks, $N_{\mathrm{r}}^{\mathrm{RF}} = 1$. This figure shows that the proposed hybrid (W)MMSE-based precoding/combining designs provides large gain over the two-stage hybrid precoding/combining designs.

angles of arrival and departure of the rays within a cluster are assumed to be randomly Laplacian distributed with an angle spread of 5°, unless stated otherwise. The AoA/AoD azimuths of the cluster means are assumed to be uniformly distributed in $[0; 2\pi]$, while their AoA/AoD elevations are uniformly distributed in $[-\frac{\pi}{2}; \frac{\pi}{2}]$. For simplicity of exposition, all channel path gains $\alpha_{i,l}$'s are assumed to be Gaussian distribution with the same variance σ_α^2, unless stated otherwise. The noise variance σ^2 is set at 1. The SNR in the plots is defined as $\mathrm{SNR} = \frac{P\sigma_\alpha^2}{K}$. In all simulations, the digital WMMSE precoder/combiner establishes the upper-bound benchmark for the hybrid WMMSE-based precoders. the digital MMSE precoder establishes the upper-bound benchmark for the hybrid MMSE-based precoders. It is also observed that the proposed hybrid WMMSE precoding/combining design always outperforms the hybrid MMSE precoding one. This superior performance is due to the better matching between the hybrid precoder and combiners in the WMMSE-based design.

In the first set of simulation results being presented in Figs. 2, 3, 4, 5 and 6, we consider a scenario where each MS is equipped with one RF chain. The number of RF chains at the BS and the number of MSs are set as $N_{\mathrm{t}}^{\mathrm{RF}} = K = 8$, except for the simulation results in Fig. 3. In Figs. 2 and 3, AoD/AoA beamforming codebooks are assumed for hybrid precoding/combining designs. Specifically, the BS sets $\mathbf{A}_{\mathcal{F}} = [\mathbf{A}_{\mathrm{t},1}, \ldots, \mathbf{A}_{\mathrm{t},K}]$, whereas MS-$i$ sets $\mathbf{A}_{\mathcal{W}_i} = \mathbf{A}_{\mathrm{r},i}$. In Figs. 4, 5 and 6, we utilize quantized RF beamforming/combining codebooks. Specifically, we use uniform quantization of the azimuth angle and uniform quantization of the elevation angle at the BS and each MS. The interested readers are referred to Equation (26) in [6] for the formulation of the RF beamforming/combining codebooks.

Fig. 2 illustrates the achievable system sum-rate with different digital and hybrid precoding/combining designs versus the SNR. As observed from the figure, DPC achieves the highest performance. The fully digital WMMSE precoder/combiner achieves a near-capacity performance, which serves as the benchmark as a linear precoding technique for comparison with hybrid precoding/combining designs. The proposed hybrid (W)MMSE-based precoding/combining designs significantly outperform two-stage hybrid precoding/combining ones, especially at high SNR. This is because the two-stage hybrid MMSE combiner and WMMSE precoder/combiner are obtained by one-shot approximations as in Algorithm 1. While Algorithm 1 does generate near-optimal solutions in single-user systems, it does not necessarily perform well in multiuser systems in the presence of inter-user interference. Fig. 2 shows that the modified hybrid WMMSE precoding/combining design can obtain a slightly inferior performance to the hybrid design in Algorithm 4. In the modified WMMSE design, we set the number of RF beamformer/combiner updates at 3. It is worth mentioning that a better performing design (WMMSE $>$ MMSE $>$ BD) comes at the cost of higher computational complexity with the joint optimization of RF and baseband precoder/combiner.

In Fig. 3, we compare the system sum-rates versus the number of users K (and the number of RF chains $N_{\mathrm{r}}^{\mathrm{RF}}$ with $N_{\mathrm{t}}^{\mathrm{RF}} = K$). The SNR is set at $-25\,\mathrm{dB}$. As displayed in the figure, the proposed hybrid (W)MMSE-based precoding/combining designs significantly outperform the two-stage hybrid MMSE-based ones, especially with high K, where the latter's performance tends to saturate. While performing comparably to the proposed hybrid MMSE precoding with low K, the performance of the two-stage hybrid ZF precoding also tends to saturate at high K. This poor performance is due to the baseband ZF precoder which is devised for an equivalent $N_{\mathrm{t}}^{\mathrm{RF}}$ transmit antennas and K users system. In a system where the number of transmit antennas equals to the number of users ($N_{\mathrm{t}}^{\mathrm{NF}} = K$ as in this case), the performance of ZF precoding does not scale with the number of users [40]. In contrast, the performances of proposed hybrid MMSE-based precoding/combining designs scale almost linearly with the number of users in the system. The reasons for the superior performance of the proposed hybrid designs compared to the two-stage hybrid ZF precoder are two-fold. First, MMSE precoding usually outperforms ZF precoding [40], [44]. Second, the proposed hybrid designs allow a joint decision of the RF precoder, instead of independently selecting each columns of the RF precoder as in the two-stage hybrid ZF precoder.

Fig. 4 compares the sum-rate performances of different hybrid precoding designs versus SNR with quantized RF beamforming/combining codebooks. Herein, we use 3-bit uniform quantization of the azimuth angle and 3-bit uniform quantization of the elevation angle at the BS and each MS. Fig. 4 shows a significant performance advantage of the proposed hybrid MMSE-based designs. The differences in the achievable sum-rates between hybrid designs are even more noticeable than the results in Fig. 2. At high SNR, the two-stage hybrid precoding designs tend to saturate. On the contrary, the performances of the proposed hybrid (W)MMSE-based designs grow linearly with SNR.

Fig. 5 illustrates the sum-rate performances of hybrid precoding designs when the number of quantized bits is varied. The SNR is set at $-20\,\mathrm{dB}$. Herein, we assume uniform quantization of each azimuth/evaluation angle at the the BS and

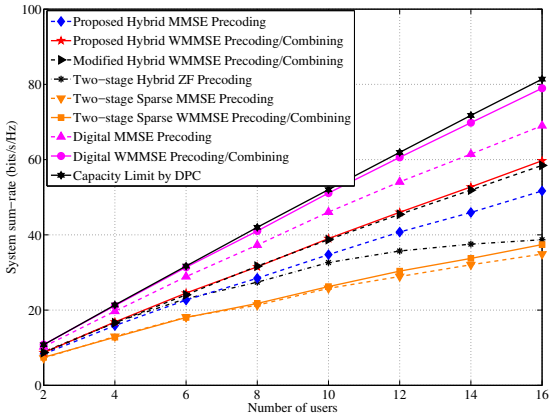

Fig. 3. Achievable system sum-rate *versus* number of user K with $N_{\mathrm{t}}^{\mathrm{RF}} = K$ and $N_{\mathrm{r}}^{\mathrm{RF}} = 1$. This figure demonstrates a linear growth in the performances of the proposed hybrid (W)MMSE-based designs. The figure also indicates that performance of the two-stage hybrid ZF precoding degrades at very high K.

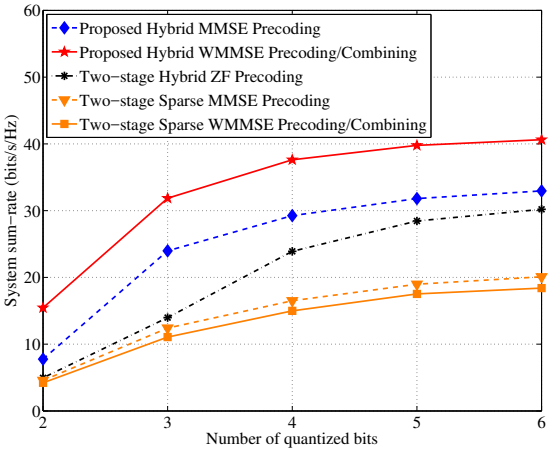

Fig. 5. Achievable system sum-rate *versus* number of quantized bits per azimuth/elevation angle. The figure indicates the saturated performances of all hybrid precoding/combining designs when the number of quantized bits approaches 6. The figure also shows the significant performance advantage of the proposed WMMSE precoding/combining designs.

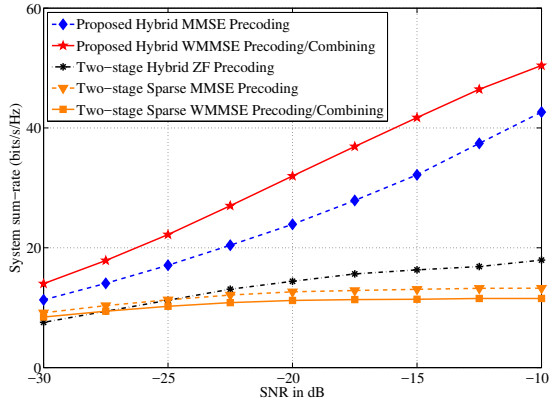

Fig. 4. Achievable system sum-rate *versus* SNR with quantized RF codebooks, $N_{\mathrm{r}}^{\mathrm{RF}} = 1$. The figure indicates the linear growth in the performances of the proposed hybrid (W)MMSE-based precoding/combining designs. The figure also indicates the saturated performances of two-stage hybrid precoding designs at high SNR.

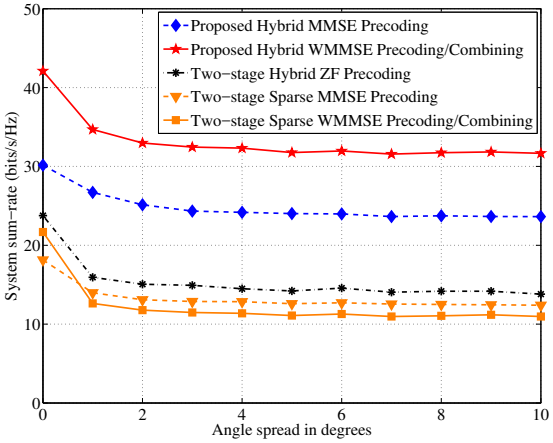

Fig. 6. Achievable system sum-rate *versus* angle spread with quantized RF codebooks, $N_{\mathrm{r}}^{\mathrm{RF}} = 1$. It shows a significant performance advantage of the proposed hybrid MMSE-based designs over the two-stage hybrid designs at all levels of scattering.

each MS with the same number of quantized bits. Fig. 5 shows the saturated performances of all hybrid precoding/combining designs when the number of quantized bits approaches 6. In other words, there are little benefit to use larger quantized codebooks. The figure also shows the significant performance advantage of the proposed WMMSE precoding/combining designs.

Fig. 6 illustrates the system sum-rate versus the angle spread with quantized RF beamforming/combining codebooks. The SNR is set at $-20\,\mathrm{dB}$. When the angle spread increases, the levels of scattering become richer, the performances of all designs degrade. Fig. 6 shows a significant performance advantage of the proposed hybrid MMSE-based designs.

In the second set of simulations with the results presented in Figs. 7, 8 and 9, we consider the setting where each MS is equipped with two RF chains and there are $K = 4$ MSs. With two RF chains at the MS, the two-stage hybrid BD precoding design is utilized to support 2 data streams for each MS. Fig. 7 illustrates the achievable system sum-rates with different digital and hybrid precoding/combining designs when the SNR is varied. In this figure, we observe the same tendency as in Fig. 2 with $N_{\mathrm{r}}^{\mathrm{RF}} = 1$. In particular, the proposed hybrid WMMSE-based precoding/combining design noticeably outperforms all other hybrid precoding designs.

In Figs. 8 and 9, the system sum-rates are displayed at $\mathrm{SNR} = -20\,\mathrm{dB}$ when varying the numbers of BS antennas and MS antennas. The figure shows a considerable performance gain by the proposed hybrid WMMSE precoding/combining over the spatially sparse precoding obtained from Algorithm 1. In fact, the proposed hybrid WMMSE scheme can maintain a constant performance gap with the fully digital WMMSE scheme, whose performance closely matches

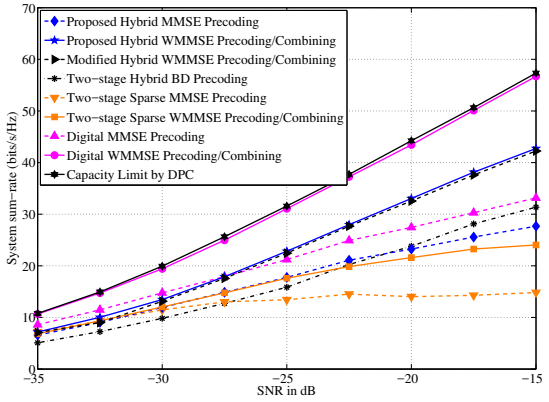

Fig. 7. Achievable system sum-rate *versus* SNR with AoA/AoD codebooks, $N_{\mathrm{r}}^{\mathrm{RF}} = 2$. This figure shows that the proposed hybrid (W)MMSE-based precoding/combining designs provides large gain over the two-stage hybrid precoding/combining designs.

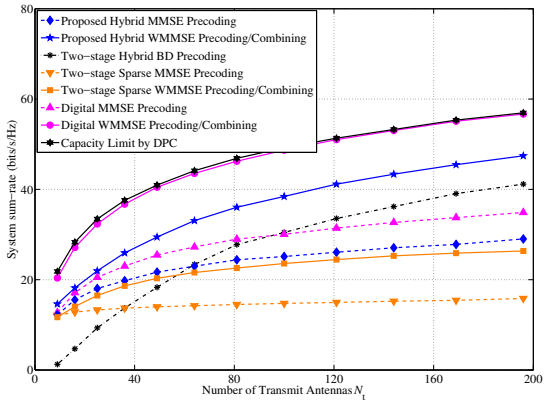

Fig. 8. Achievable system sum-rate *versus* number of transmit antenna N_{t} with AoD/AoA codebooks and $N_{\mathrm{r}}^{\mathrm{RF}} = 2$.

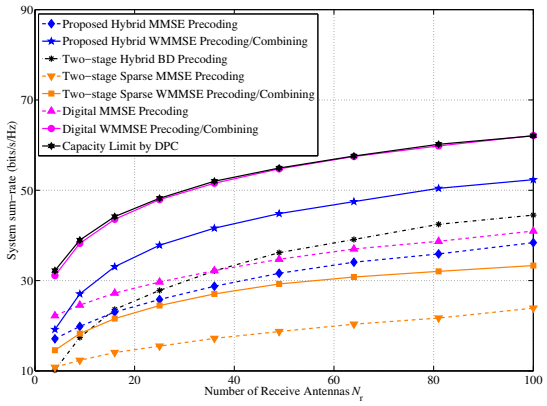

Fig. 9. Achievable system sum-rate *versus* number of receive antenna N_{r} with AoD/AoA codebooks and $N_{\mathrm{r}}^{\mathrm{RF}} = 2$. Both Figs. 8 and 9 demonstrate that superior performance of the proposed hybrid WMMSE precoding/combining design, compared to other hybrid designs.

with the capacity limits by DPC. Interestingly, at high N_{t} or N_{r}, the two-stage hybrid BD precoding starts to surpass the fully digital and hybrid MMSE precoding designs. The reason for this performance is two-fold. First, it is possible to form very narrow beams by RF antenna arrays with high number of transmit and receive antennas. Second, the MMSE-based precoder assumes an independent selection of RF combiners at each MS to maximize its effective downlink channel gain without knowing which direction the BS is beaming its signal. Thus, this RF combiner selection may not be optimal. The MMSE-based precoder is then designed to align its transmit beam to a suboptimal RF combiner. In contrast, the two-stage BD/ZF precoding enables a joint selection of RF beamfomer and combiner between the BS and the MSs at first. With a high number of transmit and receive antennas ($N_{\mathrm{t}} = 100$ and $N_{\mathrm{r}} = 16$), this strategy allows the BS and the MS to form very narrow beams in strongest transmission paths. Thus, the effect of baseband precoding then becomes less prevailing when the RF beams get narrower. Nevertheless, the proposed hybrid WMMSE precoding/combining yields the highest performance out of all hybrid designs.

VIII. Conclusion

This paper has proposed two new hybrid MMSE precoding/combining designs for multiuser mmWave systems. The paper also generalized the two-stage hybrid design to the case of multiple RF chains at the MSs by using a baseband BD precoder. Unlike the two-stage hybrid ZF/BD and the spatially sparse MMSE/WMMSE precoding design, the proposed approach aimed to minimize the sum-MSE in receiving the data streams at the users. By leveraging the sparsity of the mmWave channel, we developed MMSE-based multiuser precoding/combining solutions. Simulation results have shown significant performance advantages of the proposed hybrid precoding/combining designs over two-stage hybrid ZF/BD, MMSE, and WMMSE precoding/combining designs. In addition, the proposed hybrid WMMSE precoder/combiner has been shown to outperform the proposed hybrid MMSE precoder, since the former enables iterative refinements in its precoding and combining designs. Possible extension to the current work might include designing iterative hybrid precoders/combiners that: i.) further narrow the performance gap to the fully digital one and ii.) come with provable convergence analysis.

References

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Select. Areas in Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] 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.

[3] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, J. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[4] T. S. Rappaport, R. W. Heath Jr., J. N. Murdock, and R. C. Daniels, *Millimeter Wave Wireless Communications*. Pearson, 2014.

[5] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[6] 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.

[7] A. Alkhateeb, R. W. Heath Jr., and G. Leus, "Achievable rates of multi-user millimeter wave systems with hybrid precoding," in *Proc. IEEE Int. Conf. Commun.*, London, UK, Jun. 2015.

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

[9] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath Jr., "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, Jan. 2016.

[10] R. W. Heath Jr., N. González-Prelcic, S. Rangan, W. Roh, and A. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *to appear in IEEE J. Sel. Topics in Signal Process.*, 2016.

[11] A. Alkhateeb, J. Mo, N. González-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.

[12] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Select. Topics in Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[13] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[14] P. Sudarshan, N. Mehta, A. Molisch, and J. Zhang, "Channel statistics-based RF pre-processing with antenna selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3501–3511, Dec. 2006.

[15] J. Brady, N. Behdad, and A. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas and Propagation*, vol. 61, no. 7, pp. 3814–3827, Jul. 2013.

[16] Y. Zeng, R. Zhang, and Z. N. Chen, "Electromagnetic lens-focusing antenna enabled massive MIMO: Performance improvement and cost reduction," *IEEE J. Select. Areas in Commun.*, vol. 32, no. 6, pp. 1194–1206, Jun. 2014.

[17] J. Wang, Z. Lan, C. woo Pyo, T. Baykas, C.-S. Sum, M. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, and S. Kato, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE J. Select. Areas in Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009.

[18] C. Rusu, R. Méndez-Rial, N. González-Prelcic, and R. W. Heath Jr., "Low complexity hybrid sparse precoding and combining in millimeter wave MIMO systems," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 1340–1345.

[19] R. Méndez-Rial, C. Rusu, N. González-Prelcic, and R. W. Heath Jr., "Dictionary-free hybrid precoders and combiners for mmwave MIMO systems," in *Proc. IEEE Int. Work. on Signal Process. Advances for Wireless Commun.*, June 2015, pp. 151–155.

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

[21] W. Ni, X. Dong, and W.-S. Lu, "Near-optimal hybrid processing for massive MIMO systems via matrix decomposition," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 3922–3933, Jan. 2017.

[22] R. Méndez-Rial, N. G. Prelcic, and R. W. Heath Jr., "Adaptive hybrid precoding and combining in mmWave multiuser MIMO systems based on compressed covariance estimation," in *Proc. of the IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Cancun, Mexico, Dec. 2015.

[23] 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, Jan. 2016.

[24] W. Ni and X. Dong, "Hybrid block diagonalization for massive multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 201–211, Jan. 2016.

[25] 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.

[26] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Select. Topics in Signal Process.*, 2016.

[27] E. Zhang and C. Huang, "On achieving optimal rate of digital precoder by RF-baseband codesign for MIMO systems," in *Proc. IEEE Veh. Technol. Conf.*, Sep. 2014, pp. 1–5.

[28] M. Kim and Y. Lee, "MSE-based hybrid RF/Baseband processing for millimeter-wave communication systems in MIMO interference channels," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2714–2720, Jun. 2015.

[29] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays," *IEEE J. Select. Areas in Commun.*, vol. 35, no. 7, pp. 1432–1443, Jul. 2017.

[30] 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.

[31] 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.

[32] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, "Hybrid MMSE precoding for mmWave multiuser MIMO systems," in *to appear in IEEE Int. Conf. Commun.*, Kuala Lumpur, Malaysia, May 2016.

[33] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

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

[35] H. Zhao, R. Mayzus, S. Sun, M. Samimi, J. K. Schulz, Y. Azar, K. Wang, G. N. Wong, F. Gutierrez, and T. S. Rappaport, "28 GHz millimeter wave cellular communication measurements for reflection and penetration loss in and around buildings in New York city," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2013, pp. 5163–5167.

[36] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas and Propagation*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.

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

[38] A. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[39] Q. Spencer, A. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 461–471, Feb. 2004.

[40] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communications - Part I: Channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, Jan. 2005.

[41] H. Sung, S.-R. Lee, and I. Lee, "Generalized channel inversion methods for multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3489–3499, Nov. 2009.

[42] J. Joung and Y. H. Lee, "Regularized channel diagonalization for multiuser MIMO downlink using a modified MMSE criterion," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1573–1579, Apr. 2007.

[43] V. Stankovic and M. Haardt, "Generalized design of multi-user MIMO precoding matrices," *IEEE Trans. Wireless Commun.*, vol. 7, no. 3, pp. 953–961, Mar. 2008.

[44] M. Joham, K. Kusume, M. H. Gzara, and W. Utschick, "Transmit Wiener filter for the downlink of TDD DS-CDMA systems," in *IEEE 7th Symp. Spread-Spectrum Technol., Applicat.*, Sep. 2002, pp. 9–13.

[45] D. H. N. Nguyen and T. Le-Ngoc, "MMSE precoding for multiuser MISO downlink transmission with non-homogeneous user SNR conditions," *EURASIP J. Signal Process.*, vol. 2014:85, Jun. 2014.

[46] D. H. N. Nguyen, L. B. Le, and T. Le-Ngoc, "Multiuser MISO precoding for sum-rate maximization under multiple power constraints," in *Proc. IEEE Wireless Commun. and Networking Conf.*, New Orleans, LA, USA, Mar. 2015, pp. 729–734.

[47] L. Rebollo-Neira and D. Lowe, "Optimized orthogonal matching pursuit approach," *IEEE Signal Process. Letters*, vol. 9, no. 4, pp. 137–140, Apr. 2002.

[48] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[49] M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[50] D. H. N. Nguyen and R. W. Heath Jr., "Optimizing dirty paper coding for multiuser MIMO systems with rank constraints," University of Texas at Austin, Austin, TX, USA, Available online: http://engineering.sdsu.edu/~nguyen/downloads/DPC_rank.pdf, 2016.

Duy H. N. Nguyen (S'07–M'14) received the B.Eng. degree (with First Class Honors) from Swinburne University of Technology, Hawthorn, VIC, Australia, in 2005, the M.Sc. degree from University of Saskatchewan, Saskatoon, SK, Canada, in 2009, and the Ph.D. degree from McGill University, Montréal, 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 (INRS), Université du Québec, Montréal, 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 at 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.

Long Bao Le (S'04–M'07–SM'12) received the B.Eng. degree in electrical engineering from Ho Chi Minh City University of Technology, Vietnam, in 1999, the M.Eng. degree in telecommunications from Asian Institute of Technology, Thailand, in 2002, and the Ph.D. degree in electrical engineering from the University of Manitoba, Canada, in 2007. He was a Postdoctoral Researcher at Massachusetts Institute of Technology (2008–2010) and University of Waterloo (2007–2008). Since 2010, he has been with the Institut National de la Recherche Scientifique (INRS), Université du Québec, Montréal, QC, Canada where he is currently an associate professor. His current research interests include 5G wireless technologies, radio resource management, cognitive radio, smartgrids. He is a co-author of the books *Radio Resource Management in Multi-Tier Cellular Wireless Networks* (Wiley, 2013) and *Radio Resource Management in Wireless Networks: An Engineering Approach* (Cambridge University Press, 2017). Dr. Le is a member of the editorial board of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He was in the editorial board for IEEE WIRELESS COMMUNICATIONS LETTERS in 2011–2016. He has served as a technical program committee co-chairs of the Wireless Access track at IEEE VTC 2014-Fall, Wireless Networks track at IEEE VTC 2011-Fall, and the Cognitive Radio and Spectrum Management track at IEEE PIMRC 2011.

Tho Le-Ngoc (F'97) received the B.Eng. degree (Hons.) in electrical engineering and the M.Eng. degree from McGill University, Montréal, QC, Canada, in 1976 and 1978, respectively, and the Ph.D. degree in digital communications from the University of Ottawa, Canada, in 1983. From 1977 to 1982, he was a R&D Senior Engineer with Spar Aerospace Ltd., Sainte-Anne-de-Bellevue, Canada, and involved in the development and design of satellite communications systems. From 1982 to 1985, he was the Engineering Manager of the Radio Group with the Department of Development Engineering of SRTelecom Inc., St.Laurent, Canada, where he developed the new point-to- multipoint DA-TDMA/TDM Subscriber Radio System SR500. From 1985 to 2000, he was a Professor with the Department of Electrical and Computer Engineering, Concordia University, Montréal. Since 2000, he has been with the Department of Electrical and Computer Engineering, McGill University. His research interest is in the area of broadband digital communications. He is a fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada. He is the recipient of the 2004 Canadian Award in Telecommunications Research, and the IEEE Canada Fessenden Award 2005. He holds a Canada Research Chair (Tier I) on Broadband Access Communications.

Robert W. Heath Jr. (S'96–M'01–SM'06–F'11) received the B.S. and M.S. degrees from the University of Virginia, Charlottesville, VA, in 1996 and 1997 respectively, and the Ph.D. from Stanford University, Stanford, CA, in 2002, all in electrical engineering. From 1998 to 2001, he was a Senior Member of the Technical Staff then a Senior Consultant at Iospan Wireless Inc, San Jose, CA where he worked on the design and implementation of the physical and link layers of the first commercial MIMO-OFDM communication system. Since January 2002, he has been with the Department of Electrical and Computer Engineering at The University of Texas at Austin where he is a Cullen Trust for Higher Education Endowed Professor, and is a Member of the Wireless Networking and Communications Group. He is also President and CEO of MIMO Wireless Inc. He authored "Introduction to Wireless Digital Communication" (Prentice Hall, 2017) and "Digital Wireless Communication: Physical Layer Exploration Lab Using the NI USRP" (National Technology and Science Press, 2012), and co-authored "Millimeter Wave Wireless Communications" (Prentice Hall, 2014).

Dr. Heath has been a co-author of fifteen award winning conference and journal papers including the 2010 and 2013 EURASIP Journal on Wireless Communications and Networking best paper awards, the 2012 Signal Processing Magazine best paper award, a 2013 Signal Processing Society best paper award, 2014 EURASIP Journal on Advances in Signal Processing best paper award, the 2014 Journal of Communications and Networks best paper award, the 2016 IEEE Communications Society Fred W. Ellersick Prize, the 2016 IEEE Communications and Information Theory Societies Joint Paper Award, and the 2017 Marconi Prize Paper Award. He received the 2017 EURASIP Technical Achievement award. He was a distinguished lecturer in the IEEE Signal Processing Society and is an ISI Highly Cited Researcher. He is also an elected member of the Board of Governors for the IEEE Signal Processing Society, a licensed Amateur Radio Operator, a Private Pilot, and a registered Professional Engineer in Texas.