IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 11, 2021, accepted November 29, 2021, date of publication December 13, 2021,
date of current version December 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3135003

QoS-Aware Hybrid Beamforming With Minimal Power in mmWave Massive MIMO Systems

MOHSEN TAJALLIFAR[1], AHMAD R. SHARAFAT[1], (Life Senior Member, IEEE), AND HALIM YANIKOMEROGLU[2], (Fellow, IEEE)

[1]Faculty of Electrical and Computer Engineering, Tarbiat Modares University (TMU), Tehran 14117-13116, Iran
[2]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

Corresponding author: Ahmad R. Sharafat (sharafat@ieee.org)

ABSTRACT Hybrid beamforming is used to leverage the benefits of both massive multiple input and multiple output (MIMO) systems and millimeter waves for significantly increasing the capacity of wireless networks. Existing schemes for hybrid beamforming in multiple radio frequency (RF) chains optimize a global measure of performance and ignore quality of service (QoS) *per data stream*. In this paper, we propose a novel scheme for hybrid beamforming to minimize the transmit power while satisfying QoS defined as the mean square error (MSE) *per data stream*. We propose a two-stage cascade structure for the baseband precoder and combiner, and obtain their respective matrices in both single- and multi-user systems. We also propose a simplified scheme for designing hybrid beamformers with no nested loops in the alternating optimization method. Simulations show less than 2 dBW optimality gap for the multi-user scenario, with significantly less transmit power as compared to other existing schemes. Our simplified scheme consumes a negligible amount of 0.1 dBW more transmit power than our scheme with nested loops.

INDEX TERMS Hybrid beamforming, massive MIMO, mmWave, precoder/combiner, QoS-aware.

I. INTRODUCTION

Mobile data is experiencing a phenomenal growth, which necessitates a paradigm shift in the allocation, utilization, and management of resources. Massive multiple input and multiple output (MIMO) is a key enabler for increasing spectral efficiency, can facilitate multi-stream transmission, and increases the gain of the effective channel, resulting in higher gains in spatial multiplexing and better signal-to-noise ratio, both at the same time. Besides, millimeter waves (mmWaves) provide more bandwidth but suffer from very high attenuation. Beamforming via precoding at the transmitter and combining at the receiver, compensates for high attenuation and is practical in mmWaves as antenna arrays occupy a small area due to small wavelengths [1].

In fully digital beamforming, there is one antenna element per each radio frequency (RF) chain and beamforming is performed in the baseband, where it is possible to apply control over the signal amplitude and phase, resulting in multiple simultaneous beams [2]–[4]. In practice, the number of RF chains is much less than the number of antenna elements, which renders fully digital beamforming impractical. Instead, hybrid (digital-analog) beamforming is performed in baseband and in RF. Digital signal processing is employed in baseband to eliminate interference, and phase shifters are used in RF to steer the beam to the desired direction [5], [6].

Hybrid beamforming has been the subject of extensive research, and various schemes have been proposed to achieve different objectives such as maximizing spectral or energy efficiencies, or minimizing the mean squared error (MSE) in both single-user [5], [7]–[15] or multi-user systems [16]–[25]. A widely used technique is to decouple the transmitter design from the receiver design [5], [7]–[13], [16].

In general, RF precoder/combiner design is a non-convex Euclidean distance minimization problem; and heuristic methods are used to obtain sub-optimal RF phase shifts. Some methods obtain a subset of array response vectors (or variants of such vectors) that minimize the above mentioned distance. Other methods leverage phase information in the channel matrix. Moreover, baseband precoders are mainly least-squares solutions for the distance minimization problem. The orthogonal matching pursuit algorithm (OMP) [5], [26] finds a subset of array response vectors with maximum projection on the optimal fully digital precoder to obtain the hybrid RF precoder.

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

When the optimal fully digital precoder matrix is multiplied by a unitary matrix, the result is a variant of the former that preserves optimality. This notion was used in [8] to minimize the Euclidean distance between the hybrid precoder matrix and the optimal fully digital precoder matrix. In [9], [27], the alternating minimization method is used to obtain the baseband and RF precoder matrices by minimizing the Euclidean distance between the hybrid precoder matrix and the optimal one. Other methods directly find a hybrid precoder matrix that maximizes mutual information between transmitted symbols and received signals [16]. Another scheme finds the RF precoder and combiner matrices that maximize the capacity of the effective channel [14], [17].

The way in which RF chains are connected to antenna elements impacts the performance. In fully connected structures (FCS), each RF chain is connected to all antenna elements, resulting in high array gains [11], [14]–[21], [23], [24], [28]. But N_{RF} RF chains and N antenna elements require $N_{\mathrm{RF}} \times N$ phase shifters, which are costly and consume energy. In contrast, in partially connected structures (PCS), each RF chain is connected to a subset of antenna elements, i.e., forming subarrays [7], [10]–[13], [22]. PCSs can be realized in any of the overlapped [22], non-overlapped [7], or hybridly-connected forms [12], each with a different performance-complexity trade-off. FCS is more spectrally efficient as the array gain is high, and PCS is less costly and more energy efficient as it needs less phase shifters [7], [13].

Quality of service (QoS) is of paramount importance in future networks. A user may simultaneously request several services (e.g., web browsing, video streaming, messaging, etc.) or different users may request various services, each with a different QoS requirement. In QoS-aware beamforming, it is very desirable to reduce transmit power levels to the extent possible. Existing works mostly optimize a global measure of performance in multiple RF chain receivers for the aggregate of all data streams without considering the QoS for individual data streams when the transmit power is minimized. In contrast, spectral efficiency of each data stream is constrained to an acceptable value in QoS-aware systems when the transmit power is minimized. QoS-aware designs in the literature are mostly confined to single RF chain receivers [29]–[32] and there is no QoS-aware design when the transmit power is minimized for multiple RF chain receivers, which is the target of this paper.

In [29], the transmit power is minimized for a given signal-to-interference-plus-noise ratio (SINR) for each receiver via the semi-definite relaxation method. In [30], transmit power is minimized for a given data rate for each receiver via the zero-forcing method. In doing so, each column of the RF precoder matrix is chosen from a discrete Fourier transform (DFT)-based codebook that matches eigenmodes of multi-user channels. A scheduling problem for beam and sub-band selection is considered in [31] to maintain a given data rate for each receiver, but the precoder is not optimized. In [32], the optimal hybrid precoder is designed, but the approach requires other structures that are complex and costly.

In the literature, when the receiver has multiple RF chains, precoder design is decoupled from combiner design, and the optimal minimum MSE (MMSE) combiner is assumed at the receiver. With this assumption, spectral efficiency is equal to mutual information between transmitted symbols and received signals. Hybrid precoder is an approximation of the fully digital one, and in its problem formulation, mutual information is considered instead of spectral efficiency [5], [9], [11]. In QoS-aware hybrid beamforming, optimal MMSE combiner cannot be assumed, and hence, spectral efficiency may not be achieved with (near)-optimal transmit power. In such cases, a different approach is needed for designing *per data stream* QoS-aware hybrid beamformers.

A *per data stream* QoS-aware massive MIMO system is designed by minimizing the transmit power while satisfying the *per data stream* MSE constraints. When the baseband precoder performs channel diagonalization and the baseband combiner is MMSE combiner, data is transmitted over multiple sub-channels chosen from the effective channel matrix [33]. In this case, MSE in different streams are independent and inversely proportional to both the sub-channel gains and transmit power levels. The power allocated to each stream is minimized when its sub-channel gain, i.e., the eigenvalue of the effective channel matrix, is maximized, which can be achieved by eigen-beamforming at both the transmitter and the receiver, i.e., by adopting the right and left eigenvectors of the effective channel matrix as the transmit and the receive beamforming vectors, respectively.

In the conventional single-stage beamformer design in [18], [19], [22], eigen-beamforming is applied in the RF beamformer, whose matrices include only exponential entries with unit amplitude that represent phase shifts. As such, the RF beamforming matrices are approximations of the right and left eigenvectors of the effective channel matrix, resulting in small gains for sub-channels. To compensate for low gains, the transmit power needs to be increased, which means that the QoS may be satisfied but at higher transmit power levels.

We deviate from the above mentioned conventional design, and propose a two-stage cascade structure for both the baseband precoder and combiner, where the entries in the matrices of our RF beamformers are complex values in the form of linear combinations of unit-amplitude exponential entries. In this way, the RF beamforming matrices are better approximations of the right and left eigenvectors of the effective channel matrix. Specifically, our RF precoder is composed of the second stage of baseband precoder and the conventional RF precoder, and our RF combiner is composed of the second stage of baseband combiner and the conventional RF combiner. In this way, our RF beamformers maximize sub-channel gain by better approximating the right and left eigenvectors of the effective channel gain matrix as explained in Sections III-IV. The first stage of our baseband precoder diagonalizes the effective channel matrix, and the first stage of our baseband combiner is MMSE combiner. As we will show, our design is QoS-aware *per data stream* with less transmit power as compared to the single-stage structure.

A. CONTRIBUTION AND TECHNICAL NOVELTY

We develop a *per data stream* QoS-aware hybrid beamforming scheme for mmWave massive MIMO systems. Similar to [14], [17], we decouple RF beamforming from baseband beamforming; and use the alternating optimization method. However, unlike [14], [17] that maximize spectral efficiency (which may lead to outage in *per data stream* QoS-sensitive cases), we minimize the transmit power by minimizing cross-channel interference in different data streams while maintaining the required QoS for each stream. We achieve this by splitting the baseband precoder and combiner into two cascaded stages. Our main contributions and achievements are as follows:

- We propose a two-stage cascade structure for both baseband precoder and combiner, and obtain their respective matrices by minimizing the transmit power subject to QoS constraints for each data stream via nested loops.
- We also propose a simplified scheme for designing hybrid beamformers by using eigenvectors of channel matrix (instead of eigenvectors of effective channel matrix) in our two-stage cascade structure.
- We minimize the Euclidean distance between the hybrid precoder and the fully digital precoder by alternating between baseband and RF precoders. In contrast to the existing methods, we find the phase of each component in the RF precoder by directly minimizing the Euclidean distance. In contrast to [7], [9], [10], [27], our proposed method is applicable on both FCS and PCS.
- We apply our schemes to single-user and multi-user scenarios. Simulations show less than 2 dBW optimality gap for the multi-user scenario, with significantly less transmit power as compared to other existing schemes. Transmit power in our simplified scheme is negligibly more compared to our scheme with nested loops.

B. ORGANIZATION

The rest of this paper is organized as follows. Section II contains system and channel models, followed by problem formulation and our design framework in Section III. Hybrid beamforming for our two-stage structure via the alternating distance minimization, and our simplified hybrid beamforming are in Section IV. Multi-user and wideband scenarios are in Sections V and VI, respectively. Simulations and conclusion are in Sections VII and VIII, respectively.

C. NOTATIONS

The following notations are used in this paper: $\mathbf{A}$, $\mathbf{a}$ and a denote a matrix, a vector and a scalar, respectively; $\mathbb{E}$ denotes statistical expectation; and $\mathbb{C}^{n\times m}$ is a $n \times m$ matrix with complex elements. The operators $\|\mathbf{A}\|_{\mathrm{F}}$, $(\mathbf{A})^{\mathrm{H}}$, $(\mathbf{A})^{-1}$ and $\mathrm{Tr}(\mathbf{A})$ show Frobenius norm, Hermitian, inverse and trace of $\mathbf{A}$, respectively; $[\mathbf{A}]_{1:n}$, $[\mathbf{A}]_{n,m}$, and $\boldsymbol{\lambda}(\mathbf{A})$ denote the first n columns, (n, m) entry, and the vector of eigenvalues of $\mathbf{A}$, respectively; $\mathrm{diag}\{\mathbf{a}\}$ is a diagonal matrix with diagonal elements $\mathbf{a}$; and $\mathrm{blkdiag}\,(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ is a block diagonal matrix whose diagonal blocks are $\mathbf{A}_1, \ldots, \mathbf{A}_n$.

II. SYSTEM AND CHANNEL MODELS

A. SYSTEM MODEL

Consider a system comprised of one transmitter and one receiver, each with N_{t} and N_{r} antennas, respectively. The transmitter has baseband and RF precoders and the receiver has baseband and RF combiners. As in Fig. 1, a vector $\mathbf{s} \in \mathbb{C}^{N_{\mathrm{s}}\times 1}$ of independent data streams with covariance matrix $\mathbb{E}\{\mathbf{s}\mathbf{s}^{\mathrm{H}}\} = \mathbf{I}_{N_{\mathrm{s}}}$ is processed by the baseband precoder matrix $\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{t}}^{\mathrm{RF}}\times N_{\mathrm{s}}}$, fed into $N_{\mathrm{t}}^{\mathrm{RF}}$ RF chains, and passed through the RF precoder comprising a set of phase shifters whose gain multipliers are $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{t}}^{\mathrm{RF}}}$. Since at least N_{s} independent linear equations should be solved for signal recovery, we assume $N_{\mathrm{t}}^{\mathrm{RF}} \geq N_{\mathrm{s}}$. We also assume RF chains are fewer than antennas, i.e., $N_{\mathrm{t}}^{\mathrm{RF}} \leq N_{\mathrm{t}}$. Similar assumptions are made for the combiner.

The received vector at the receiver antenna array is

$$\mathbf{x} = \mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{H} \in \mathbb{C}^{N_{\mathrm{r}}\times N_{\mathrm{t}}}$ is the channel matrix and $\mathbf{n} \in \mathbb{C}^{N_{\mathrm{r}}\times 1}$ is the noise vector with covariance matrix $\mathbb{E}\{\mathbf{n}\mathbf{n}^{\mathrm{H}}\} = \sigma_{\mathrm{n}}^2\mathbf{I}_{N_{\mathbf{r}}}$. The received vector is phase shifted by the RF combiner comprising a set of phase shifters $\mathbf{W}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{r}}\times N_{\mathrm{r}}^{\mathrm{RF}}}$, passed through $N_{\mathrm{r}}^{\mathrm{RF}}$ RF chains, and fed into the baseband combiner $\mathbf{W}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{RF}}\times N_{\mathrm{s}}}$, resulting in

$$\hat{\mathbf{s}} = \mathbf{W}_{\mathrm{BB}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{s} + \mathbf{W}_{\mathrm{BB}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{n}. \tag{2}$$

We consider both FCS and PCS. In FCS, the (k, l) entry in $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{RF}}$ is in the form of $e^{j\varphi_{kl}}$ where φ_{kl} is the phase shift applied to the l^{th} RF chain connected to the k^{th} antenna element. The sets of matrices for RF precoders and combiners in FCS are $\mathcal{F}_{\mathrm{FC}}$ and $\mathcal{W}_{\mathrm{FC}}$, respectively. In PCS, each RF chain is connected to a subset of antennas. Although our mathematical derivations and proposed methods are applicable to both overlapped and non-overlapped PCSs, we only consider non-overlapped PCS for brevity, for which

$$\mathbf{F}_{\mathrm{RF}} = \mathrm{blkdiag}\left(\mathbf{f}_{\mathrm{RF},1}, \ldots, \mathbf{f}_{\mathrm{RF},N_{\mathrm{t}}^{\mathrm{RF}}}\right), \tag{3}$$

where $\mathbf{f}_{\mathrm{RF},k} \in \mathbb{C}^{M\times 1}$ is a vector of M phase shifts for the antenna elements in the k^{th} sub-array. Without loss of generality, we assume that all sub-arrays have the same number of antenna elements, i.e., $M = N_{\mathrm{t}}/N_{\mathrm{t}}^{\mathrm{RF}}$. The same is true for the receiving partially connected RF combiner $\mathbf{W}_{\mathrm{RF}}$ in the receiver. The sets of matrices for RF precoders and combiners in PCS are $\mathcal{F}_{\mathrm{PC}}$ and $\mathcal{W}_{\mathrm{PC}}$, respectively.

B. CHANNEL MODEL

Our approach is independent of channel model. Without loss of generality, we begin by considering the widely adopted slow-fading narrow-band clustered channel model given by

$$\mathbf{H} = \sqrt{\frac{N_{\mathrm{t}}N_{\mathrm{r}}}{N_{\mathrm{cl}}N_{\mathrm{ray}}}}\sum_{i=1}^{N_{\mathrm{cl}}}\sum_{l=1}^{N_{\mathrm{ray}}}\alpha_{l,i}\mathbf{a}_{\mathrm{r}}(\phi_{l,i}^{\mathrm{r}}, \theta_{l,i}^{\mathrm{r}})\mathbf{a}_{\mathrm{t}}(\phi_{l,i}^{\mathrm{t}}, \theta_{l,i}^{\mathrm{t}})^{\mathrm{H}}, \tag{4}$$

where N_{cl} and N_{ray} are the number of clusters and rays in each cluster, respectively. The complex gain of l^{th} ray in i^{th}

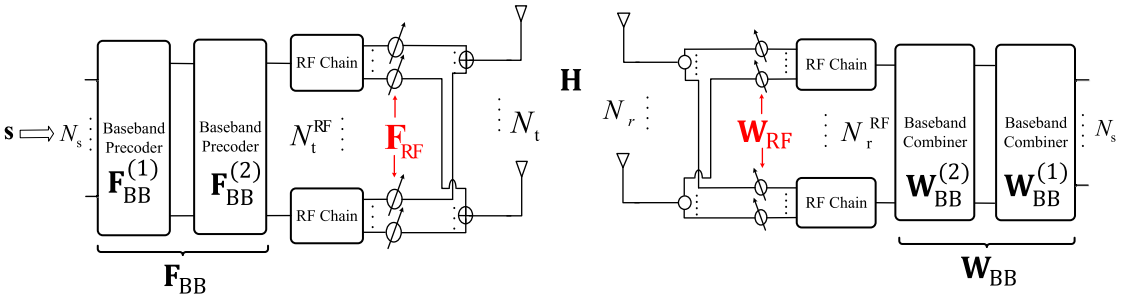

FIGURE 1. Block diagram of hybrid beamformers at the transmitter and receiver with fully connected arrays.

cluster is $\alpha_{l,i}$. The azimuth angle-of-arrival (AoA) and angle-of-departure (AoD) of l^{th} ray in i^{th} cluster are $\phi_{l,i}^{\text{r}}$ and $\phi_{l,i}^{\text{t}}$, respectively; and the elevation AoAs and AoDs are $\theta_{l,i}^{\text{r}}$ and $\theta_{l,i}^{\text{t}}$, respectively. Moreover, $\mathbf{a}_{\text{r}}(\phi_{l,i}^{\text{r}}, \theta_{l,i}^{\text{r}})$ and $\mathbf{a}_{\text{t}}(\phi_{l,i}^{\text{t}}, \theta_{l,i}^{\text{t}})$ are the receiver's and transmitter's normalized array responses to AoA and AoD of l^{th} ray in i^{th} cluster, respectively. Although our approach is independent of array types, we assume uniform planar arrays to be consistent with the literature. The array has N_{H} and N_{V} equally spaced antenna elements with inter-element spacing d in horizontal and vertical axes, respectively. The array response vector is

$$\mathbf{a}(\phi,\theta) = \frac{1}{\sqrt{N_{\text{H}}N_{\text{V}}}}[1,\ldots,e^{j\frac{2\pi d}{\lambda}(m\sin\theta\cos\phi+n\cos\theta)},\ldots,e^{j\frac{2\pi d}{\lambda}((N_{\text{H}}-1)\sin\theta\cos\phi+(N_{\text{V}}-1)\cos\theta)}], \quad (5)$$

where $0 \le m \le N_{\text{H}}$ and $0 \le n \le N_{\text{V}}$, and λ is the wavelength of operating frequency.

III. PROBLEM FORMULATION AND DESIGN FRAMEWORK

We now formulate the problem of minimizing the transmit power subject to a given QoS for each data stream and other constraints. Without loss of generality, we consider *per data stream* MSE as the measure of QoS. However, as stated in [33], other measures such as signal-to-noise ratio, spectral efficiency, and bit-error-rate can also be considered in formulating the problem. The MSE matrix is

$$\mathbf{E} = \mathbb{E}\{(\hat{\mathbf{s}}-\mathbf{s})(\hat{\mathbf{s}}-\mathbf{s})^{\text{H}}\}, \quad (6)$$

whose k^{th} diagonal entry $[\mathbf{E}]_{k,k}$ is the MSE of the estimate of k^{th} data stream. From (2), the MSE matrix is (7), as shown at the bottom of the page. The total consumed power P_{FC} is the sum of consumed power by all RF chains, phase shifters, and amplifiers. For FCS, we have

$$P_{\text{FC}} = (N_{\text{t}}^{\text{RF}}N_{\text{t}} + N_{\text{r}}^{\text{RF}}N_{\text{r}})P_{\text{PS}} + (N_{\text{t}}^{\text{RF}} + N_{\text{r}}^{\text{RF}})P_{\text{RF}} + P_{\text{T}}, \quad (8)$$

where P_{PS} is the power consumed by each phase shifter, P_{RF} is the power consumed by each RF chain, and P_{T} is the transmit power, i.e., the power consumed by all amplifiers,

$$P_{\text{T}} = \text{Tr}(\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^{\text{H}}\mathbf{F}_{\text{RF}}^{\text{H}}). \quad (9)$$

We wish to minimize P_{T} by optimizing the baseband and RF precoders and combiners subject to QoS constraints in each data stream, written as

$$\min_{\mathbf{W}_{\text{RF}},\mathbf{W}_{\text{BB}},\mathbf{F}_{\text{RF}},\mathbf{F}_{\text{BB}}} \quad \text{Tr}(\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}\mathbf{F}_{\text{BB}}^{\text{H}}\mathbf{F}_{\text{RF}}^{\text{H}})$$
$$\text{subject to} \quad \begin{cases} [\mathbf{E}]_{k,k} \le \rho_k, & \forall k \\ \mathbf{F}_{\text{RF}} \in \mathcal{F}_{\text{FC}}, & \mathbf{W}_{\text{RF}} \in \mathcal{W}_{\text{FC}}, \end{cases} \quad (10)$$

where ρ_k is the maximum tolerable MSE of k^{th} data stream. Our scheme is QoS-aware for all data streams, each with its own QoS requirement. When QoS requirements of all data streams are the same, a QoS-aware system is characterized by its maximum tolerable MSE denoted by $\rho_k = \rho, \forall k$, i.e., QoS constraints in (10) are satisfied when MSEs in all data streams are less than ρ. In other words, in *per data stream* QoS-aware systems with $\rho_k = \rho, \forall k$, the complementary cumulative distribution function CCDF $= \Pr\{\text{MSE} > r\}$ for all $r > \rho$ is zero.

We now develop a framework for solving (10). As in Fig. 1, we split both the baseband precoder and combiner matrices into two cascaded matrices $\mathbf{F}_{\text{BB}} = \mathbf{F}_{\text{BB}}^{(2)}\mathbf{F}_{\text{BB}}^{(1)}$ and $\mathbf{W}_{\text{BB}} = \mathbf{W}_{\text{BB}}^{(2)}\mathbf{W}_{\text{BB}}^{(1)}$, where $\mathbf{F}_{\text{BB}}^{(1)} \in \mathbb{C}^{N_{\text{s}}\times N_{\text{s}}}$, $\mathbf{F}_{\text{BB}}^{(2)} \in \mathbb{C}^{N_{\text{t}}^{\text{RF}}\times N_{\text{s}}}$, $\mathbf{W}_{\text{BB}}^{(1)} \in \mathbb{C}^{N_{\text{s}}\times N_{\text{s}}}$, and $\mathbf{W}_{\text{BB}}^{(2)} \in \mathbb{C}^{N_{\text{r}}^{\text{RF}}\times N_{\text{s}}}$. The effective channel matrix is $\mathbf{H}_{\text{eff}} = \mathbf{W}_{\text{BB}}^{(2)\,\text{H}}\mathbf{W}_{\text{RF}}^{\text{H}}\mathbf{H}\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}^{(2)}$, and (2) becomes

$$\hat{\mathbf{s}} = \mathbf{W}_{\text{BB}}^{(1)\,\text{H}}\mathbf{H}_{\text{eff}}\mathbf{F}_{\text{BB}}^{(1)}\mathbf{s} + \mathbf{W}_{\text{BB}}^{(1)\,\text{H}}\tilde{\mathbf{n}}, \quad (11)$$

where $\tilde{\mathbf{n}} = \mathbf{W}_{\text{BB}}^{(2)\,\text{H}}\mathbf{W}_{\text{RF}}^{\text{H}}\mathbf{n}$ is the effective noise vector. The estimate (11) is identical to the estimate in the conventional (non-hybrid) single-user MIMO systems when $\mathbf{H}_{\text{eff}}$ and $\tilde{\mathbf{n}}$ are given. Hence, $\mathbf{F}_{\text{BB}}^{(1)}$ and $\mathbf{W}_{\text{BB}}^{(1)}$ in (11) are the same as those in fully digital systems that are QoS-aware. As a result, our hybrid design with the same $\mathbf{F}_{\text{BB}}^{(1)}$ and $\mathbf{W}_{\text{BB}}^{(1)}$ is also QoS-aware.

$$\mathbf{E} = \left(\mathbf{W}_{\text{BB}}^{\text{H}}\mathbf{W}_{\text{RF}}^{\text{H}}\mathbf{H}\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}} - \mathbf{I}_{N_s}\right)\left(\mathbf{W}_{\text{BB}}^{\text{H}}\mathbf{W}_{\text{RF}}^{\text{H}}\mathbf{H}\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}} - \mathbf{I}_{N_s}\right)^{\text{H}} + \sigma_{\text{n}}^2\mathbf{W}_{\text{BB}}^{\text{H}}\mathbf{W}_{\text{RF}}^{\text{H}}\mathbf{W}_{\text{RF}}\mathbf{W}_{\text{BB}}. \quad (7)$$

TABLE 1. Different methods for the optimization of hybrid beamforming matrices.

Matrices	Optimization methods
$\mathbf{W}_{\mathrm{BB}}^{(1)}$	MSE minimization
$\mathbf{F}_{\mathrm{BB}}^{(1)}$	Zero forcing, maximum ratio precoding, eigen-beamforming
$\mathbf{P}$	Global measure of performance: Water-filling; Individual QoS constraints: Tightening QoS constraints
$\mathbf{F}_{\mathrm{BB}}^{(2)}, \mathbf{F}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}}^{(2)}, \mathbf{W}_{\mathrm{RF}}$	Minimizing a function of MSEs: Sum of MSEs, Max. of MSEs; Distance minimization: via eigenvectors of $\mathbf{H}$; Maximizing a function of eigenvalues: Sum of eigenvalues, Min. of eigenvalues

This is in contrast to the existing hybrid designs that use approximations of their respective QoS measures.

The optimal $\mathbf{W}_{\mathrm{BB}}^{(1)}$ is MMSE combiner. Different types of precoders can be used for $\mathbf{F}_{\mathrm{BB}}^{(1)}$, e.g., zero-forcing, maximum ratio precoding, or eigen-beamforming. The power allocation matrix $\mathbf{P}$ in our per data stream QoS-aware design is embedded in $\mathbf{F}_{\mathrm{BB}}^{(1)}$ as shown in (14) below.

Next we obtain $\mathbf{F}_{\mathrm{BB}}^{(2)}$, $\mathbf{F}_{\mathrm{RF}}$, $\mathbf{W}_{\mathrm{BB}}^{(2)}$, and $\mathbf{W}_{\mathrm{RF}}$ to maximize the sum of eigenvalues of $\mathbf{H}_{\mathrm{eff}}$. There can be other objectives such as minimizing a function of MSEs (e.g., sum of MSEs, maximum of MSEs, etc.), minimizing the distance with a target fully digital design (e.g., eigenvectors of $\mathbf{H}$), etc. Table 1 shows various optimization methods for different matrices. In contrast to the single-stage structure, in our two-stage cascade structure, a linear combination of exponential entries in $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{RF}}$ are utilized to increase the effective channel gain $\mathbf{H}_{\mathrm{eff}}$, resulting in less transmit power.

Since it is not easy to simultaneously find all optimal matrices, sub-optimal matrices can be obtained by searching for a specific matrix when other matrices are fixed. Fig. 2 shows the flowchart of our *per data stream* QoS-aware hybrid beamforming scheme. We first find $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}^{(2)}$ via the alternating minimization method. After convergence, we find $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}^{(2)}$ via the same method. Next, $\mathbf{H}_{\mathrm{eff}}$ is obtained to check whether eigenvalues have converged. Finally, $\mathbf{F}_{\mathrm{BB}}^{(1)}$ and $\mathbf{W}_{\mathrm{BB}}^{(1)}$ are computed.

IV. DESIGNING PRECODERS AND COMBINERS

We now use our framework to obtain matrix values for hybrid beamforming. As stated in Section III, we use MMSE combiner. From (11), MMSE combiner can be written as

$$\mathbf{W}_{\mathrm{BB}}^{(1)} = \left(\mathbf{H}_{\mathrm{eff}}\mathbf{F}_{\mathrm{BB}}^{(1)}\mathbf{F}_{\mathrm{BB}}^{(1)\mathrm{H}}\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}} + \mathbf{R}_{\tilde{\mathrm{n}}}\right)^{-1}\mathbf{H}_{\mathrm{eff}}\mathbf{F}_{\mathrm{BB}}^{(1)}, \quad (12)$$

where $\mathbf{R}_{\tilde{\mathrm{n}}} = \sigma_{\mathrm{n}}^2\mathbf{W}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)}$ is the covariance matrix of the effective noise vector. The MMSE matrix is

$$\bar{\mathbf{E}} = \left(\mathbf{F}_{\mathrm{BB}}^{(1)\mathrm{H}}\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}}\mathbf{F}_{\mathrm{BB}}^{(1)} + \mathbf{I}_{N_{\mathrm{s}}}\right)^{-1}. \quad (13)$$

To obtain $\mathbf{F}_{\mathrm{BB}}^{(1)}$, we use eigen-beamforming in Section III, as it eliminates mutual interference in data streams by using orthogonal sub-channels, i.e., by diagonalizing $\bar{\mathbf{E}}$. When

$$\mathbf{F}_{\mathrm{BB}}^{(1)} = \mathbf{V}_{\mathrm{eff}}\mathbf{P}^{\frac{1}{2}}, \quad (14)$$

MMSE matrix is diagonal, where $\mathbf{P} = \mathrm{diag}\{p_1, \ldots, p_{N_{\mathrm{s}}}\} \in \mathbb{R}^{N_{\mathrm{s}}\times N_{\mathrm{s}}}$ is the diagonal transmit power matrix, and

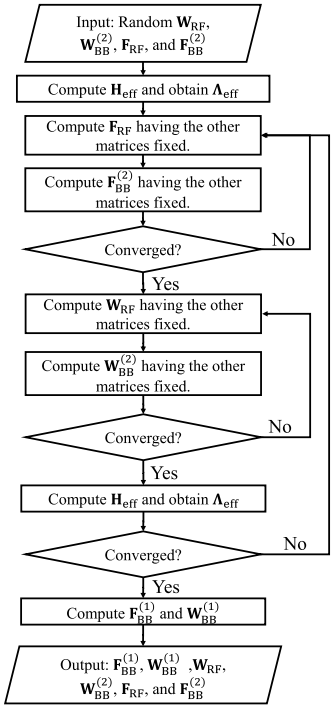

FIGURE 2. Flowchart of our scheme for hybrid beamforming.

$\mathbf{V}_{\mathrm{eff}}$ is obtained via eigenvalue decomposition (EVD) of $\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}}$ as

$$\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}} = \mathbf{V}_{\mathrm{eff}}\mathbf{\Lambda}_{\mathrm{eff}}\mathbf{V}_{\mathrm{eff}}^{\mathrm{H}}, \quad (15)$$

in which eigenvalues in $\mathbf{\Lambda}_{\mathrm{eff}}$ are in descending order. This beamforming matrix has N_{s} sub-channels whose gains correspond to the first N_{s} eigenvalues in $\mathbf{\Lambda}_{\mathrm{eff}}$.

By applying $\mathbf{F}_{\mathrm{BB}}^{(1)}$ to the MMSE matrix, we get

$$\bar{\mathbf{E}} = \left(\mathbf{P}^{\frac{\mathrm{H}}{2}}\mathbf{V}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{V}_{\mathrm{eff}}\mathbf{\Lambda}_{\mathrm{eff}}\mathbf{V}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{V}_{\mathrm{eff}}\mathbf{P}^{\frac{1}{2}} + \mathbf{I}_{N_{\mathrm{s}}}\right)^{-1} = \left(\mathbf{P}^{\frac{\mathrm{H}}{2}}\mathbf{\Lambda}_{\mathrm{eff}}\mathbf{P}^{\frac{1}{2}} + \mathbf{I}_{N_{\mathrm{s}}}\right)^{-1}. \quad (16)$$

MMSE matrix is diagonal with elements $[\bar{\mathbf{E}}]_{k,k} = \frac{1}{1+p_k\lambda_{\mathrm{eff},k}}$, where $\lambda_{\mathrm{eff},k}$ is the k^{th} largest eigenvalue of $\mathbf{\Lambda}_{\mathrm{eff}}$. To minimize P_{T}, we need $\mathbf{W}_{\mathrm{RF}}$, $\mathbf{W}_{\mathrm{BB}}^{(2)}$, $\mathbf{F}_{\mathrm{RF}}$, and $\mathbf{F}_{\mathrm{BB}}^{(2)}$ for highest values of $\lambda_{\mathrm{eff},k}$. We maximize the sum of eigenvalues, i.e., $\sum_{k=1}^{N_s}\lambda_{\mathrm{eff},k}$, which is equal to the effective channel gain $\mathrm{Tr}(\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}})$.

Let

$$\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}} = \mathbf{F}_{\mathrm{BB}}^{\mathrm{H}(2)}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}_{\mathrm{W}}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} \quad (17)$$

where $\mathbf{H}_{\mathrm{W}} = \mathbf{H}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{W}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}$. When $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}$ is the set of N_{s} eigenvectors corresponding to the highest eigenvalues of $\mathbf{H}_{\mathrm{W}}$, the effective channel gain is maximized. The EVD of $\mathbf{H}_{\mathrm{W}}$ is $\mathbf{H}_{\mathrm{W}} = \mathbf{V}_{\mathrm{W}}\mathbf{\Lambda}_{\mathrm{W}}\mathbf{V}_{\mathrm{W}}^{\mathrm{H}}$, and $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}$ is equal to $\widetilde{\mathbf{V}}_{\mathrm{W}} = [\mathbf{V}_{\mathrm{W}}]_{1:N_{\mathrm{s}}}$. However, since $\mathbf{F}_{\mathrm{RF}}$ belongs to $\mathcal{F}_{\mathrm{FC}}$, we obtain $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}$ that is as close to $\widetilde{\mathbf{V}}_{\mathrm{W}}$ as possible in the Euclidean sense. Thus,

$$\begin{aligned} \min_{\mathbf{F}_{\mathrm{RF}},\mathbf{F}_{\mathrm{BB}}^{(2)}} \quad & \|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \mathbf{F}_{\mathrm{RF}} \in \mathcal{F}_{\mathrm{FC}}. \end{aligned} \tag{18}$$

We divide (18) into two subproblems and use the alternating minimization method to obtain $\mathbf{F}_{\mathrm{RF}}$ for a fixed $\mathbf{F}_{\mathrm{BB}}^{(2)}$ and vice versa. Inspired by [16], we optimize each phase shift at a time having the others fixed. Note that each phase shift in [16] is obtained for sum-rate maximization, but we obtain phase shifts by minimizing the Euclidean distance in (18). Theorem 1 finds the $(k,l)^{\text{th}}$ phase shift in $\mathbf{F}_{\mathrm{RF}}$.

Theorem 1: For a given $\mathbf{F}_{BB}^{(2)}$, the $(k,l)^{th}$ phase shift that minimizes (18) *in FCS is $\varphi_{k,l}^{\star} = -\angle\delta_{k,l}$, where $\delta_{k,l} = \sum_{n=1}^{N_t^{RF}} \gamma_{k,n}^{*}[\mathbf{F}_{BB}^{(2)}]_{l,n}$, and $\gamma_{k,n} = [\widetilde{\mathbf{V}}_W]_{k,n} - \sum_{i\neq l} e^{j\varphi_{k,i}}[\mathbf{F}_{BB}^{(2)}]_{i,n}$.*

Proof: The optimal phase shift $\varphi_{k,l}$ is obtained by differentiating the objective function with respect to $\varphi_{k,l}$. To do so, we expand $\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}$ as

$$\begin{aligned} &\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}^2 \\ &\quad = \sum_{m=1}^{N_{\mathrm{t}}}\sum_{n=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} \left| [\widetilde{\mathbf{V}}_{\mathrm{W}}]_{m,n} - \sum_{i=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} e^{j\varphi_{m,i}}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{i,n} \right|^2. \end{aligned} \tag{19}$$

The phase shift $\varphi_{k,l}$ contributes to the objective function only when $m = k$. Therefore,

$$\begin{aligned} &\frac{\partial}{\partial\varphi_{k,l}}\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}^2 \\ &\quad = \frac{\partial}{\partial\varphi_{k,l}} \sum_{n=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} \left| [\widetilde{\mathbf{V}}_{\mathrm{W}}]_{k,n} - \sum_{i=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} e^{j\varphi_{k,i}}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{i,n} \right|^2. \end{aligned} \tag{20}$$

Hence,

$$\begin{aligned} \frac{\partial}{\partial\varphi_{k,l}}\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}^2 &= \sum_{n=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} \frac{\partial}{\partial\varphi_{k,l}} \\ &\times \left| [\widetilde{\mathbf{V}}_{\mathrm{W}}]_{k,n} - e^{j\varphi_{k,l}}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{l,n} + \sum_{i=1,i\neq l}^{N_{\mathrm{t}}^{\mathrm{RF}}} e^{j\varphi_{k,i}}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{i,n} \right|^2. \end{aligned}$$

For $\gamma_{k,n} = [\widetilde{\mathbf{V}}_{\mathrm{W}}]_{k,n} - \sum_{i\neq l} e^{j\varphi_{k,i}}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{i,n}$, we have

$$\begin{aligned} &\frac{\partial}{\partial\varphi_{k,l}}\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}^2 \\ &= \sum_{n=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} \frac{\partial}{\partial\varphi_{k,l}} \left| \gamma_{k,n} - e^{j\varphi_{k,l}}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{l,n} \right|^2 \\ &= \sum_{n=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} \frac{\partial}{\partial\varphi_{k,l}} \left(|\gamma_{k,n}|^2 + |[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{l,n}|^2 - 2\mathrm{Re}(\gamma_{k,n}^{*}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{l,n}e^{j\varphi_{k,l}}) \right) \\ &= \sum_{n=1}^{N_{\mathrm{t}}^{\mathrm{RF}}} \left(-j\gamma_{k,n}^{*}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{l,n}e^{j\varphi_{k,l}} + j\gamma_{k,n}[\mathbf{F}_{\mathrm{BB}}^{(2)}]_{l,n}^{*}e^{-j\varphi_{k,l}} \right) \\ &= -j\delta_{k,l}e^{j\varphi_{k,l}} + j\delta_{k,l}^{*}e^{-j\varphi_{k,l}} = 2|\delta_{k,l}|\sin(\varphi_{k,l} + \angle\delta_{k,l}). \end{aligned} \tag{21}$$

When the derivative of Euclidean distance is zero, we have

$$\varphi_{k,l} = \begin{cases} -\angle\delta_{k,l}, \\ \pi - \angle\delta_{k,l}. \end{cases} \tag{22}$$

Note that one solution minimizes and another maximizes the Euclidean distance. However, $\varphi_{k,l} = -\angle\delta_{k,l}$ leads to $\frac{\partial^2}{\partial\varphi_{k,l}^2}\|\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}^2 = 2|\delta_{k,l}|\cos(\varphi_{k,l} + \angle\delta_{k,l}) \geq 0$. Thus, the Euclidean distance minimizer is $\varphi_{k,l}^{\star} = -\angle\delta_{k,l}$. □

Having obtained $\mathbf{F}_{\mathrm{RF}}$, the least squares solution for $\mathbf{F}_{\mathrm{BB}}^{(2)}$ in (18) is obtained as

$$\mathbf{F}_{\mathrm{BB}}^{(2)} = \left(\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{F}_{\mathrm{RF}}\right)^{-1}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\widetilde{\mathbf{V}}_{\mathrm{W}}. \tag{23}$$

Note that the least squares solution is also adopted by several works such as [5], [9], where the least squares solution approximates the hybrid precoder to the fully digital one. However, the least squares solution in (23) approximates $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}$ to the matrix of the right eigenvectors $\widetilde{\mathbf{V}}_{\mathrm{W}}$.

Algorithm 1 solves (18) via the alternating minimization method. In each iteration q, we obtain $\mathbf{F}_{\mathrm{RF}}^{(q)}$ as per Theorem 1. The solution to $\mathbf{F}_{\mathrm{BB}}^{(2)(q)}$ can then be found by (23). The alternating calculations of $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}^{(2)}$ are continued N_{it} times or until the reduction in distance in (18) is less than ϵ_{F}.

Algorithm 1 Solving (18)

Input: Random $\mathbf{F}_{\mathrm{BB}}^{(2)(0)}$, ϵ_{F}, N_{it}, $q = 0$
Compute $\mathbf{H}_{\mathrm{W}} = \mathbf{H}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{W}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}$
Compute EVD of $\mathbf{H}_{\mathrm{W}}$ and return $\widetilde{\mathbf{V}}_{\mathrm{W}}$
repeat
  $q = q + 1$
  Compute $\mathbf{F}_{\mathrm{RF}}^{(q)}$ using Theorem 1
  Compute $\mathbf{F}_{\mathrm{BB}}^{(2)(q)} = \left(\mathbf{F}_{\mathrm{RF}}^{(q)\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{(q)}\right)^{-1}\mathbf{F}_{\mathrm{RF}}^{(q)\mathrm{H}}\widetilde{\mathbf{V}}_{\mathrm{W}}$
until
$\|\mathbf{F}_{\mathrm{RF}}^{(q-1)}\mathbf{F}_{\mathrm{BB}}^{(2)(q-1)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}} - \|\mathbf{F}_{\mathrm{RF}}^{(q)}\mathbf{F}_{\mathrm{BB}}^{(2)(q)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}} \leq \epsilon_{\mathrm{F}}$ or $q \geq N_{\mathrm{it}}$
Output: $\mathbf{F}_{\mathrm{BB}}^{(2)}$ and $\mathbf{F}_{\mathrm{RF}}$

To optimize $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}^{(2)}$, we maximize the effective channel gain. Similar to $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}$, the matrix $\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)}$ is obtained as an approximation of a semi-unitary matrix, i.e., $\mathbf{W}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)} \approx \mathbf{I}_{N_{\mathrm{s}}}$, for large N_{r}. The accuracy of approximation is shown through computer simulations. In this case, $\mathbf{R}_{\tilde{\mathrm{n}}} \approx \sigma_{\mathrm{n}}^2\mathbf{I}_{N_{\mathrm{s}}}$, we get

$$\begin{aligned} &\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}} \\ &\approx \sigma_{\mathrm{n}}^{-2}\mathbf{F}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)}\mathbf{W}_{\mathrm{BB}}^{(2)\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}. \end{aligned} \tag{24}$$

Since $\mathrm{Tr}(\mathbf{A}^{\mathrm{H}}\mathbf{A}) = \mathrm{Tr}(\mathbf{A}\mathbf{A}^{\mathrm{H}})$, we have

$$\mathrm{Tr}\left(\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}}\right) \approx \mathrm{Tr}\left(\sigma_{\mathrm{n}}^{-2}\mathbf{W}_{\mathrm{BB}}^{(2)\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}_{\mathrm{F}}\mathbf{W}_{\mathrm{RF}}^{(2)}\mathbf{W}_{\mathrm{BB}}^{(2)}\right), \quad (25)$$

where $\mathbf{H}_{\mathrm{F}} = \mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}\mathbf{F}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}$. Therefore, instead of maximizing $\mathrm{Tr}\left(\mathbf{H}_{\mathrm{eff}}^{\mathrm{H}}\mathbf{R}_{\tilde{\mathrm{n}}}^{-1}\mathbf{H}_{\mathrm{eff}}\right)$, we maximize the right-hand side of (25).

Let $\mathbf{H}_{\mathrm{F}} = \mathbf{V}_{\mathrm{F}}\mathbf{\Lambda}_{\mathrm{F}}\mathbf{V}_{\mathrm{F}}^{\mathrm{H}}$. The eigenvalues of $\mathbf{\Lambda}_{\mathrm{eff}}$ are maximized by minimizing the Euclidean distance $\|\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{F}}\|_{\mathrm{F}}$, where $\widetilde{\mathbf{V}}_{\mathrm{F}} = [\mathbf{V}_{\mathrm{F}}]_{1:N_{\mathrm{s}}}$. The problem for RF and baseband combiners is

$$\min_{\mathbf{W}_{\mathrm{RF}}, W_{\mathrm{BB}}^{(2)}} \quad \|\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)} - \widetilde{\mathbf{V}}_{\mathrm{F}}\|_{\mathrm{F}}^2 \quad \text{subject to} \quad \mathbf{W}_{\mathrm{RF}} \in \mathcal{W}_{\mathrm{FC}}. \quad (26)$$

Theorem 2 finds the $(k, l)^{\mathrm{th}}$ phase shift for our RF combiner.

Theorem 2: For a given $\mathbf{W}_{BB}^{(2)}$, the $(k, l)^{th}$ phase shift that minimizes (26) *in FCS is $\varphi_{k,l}^{\star} = -\angle\beta_{k,l}$, where $\beta_{k,l} = \sum_{n=1}^{N_r^{RF}} \xi_{k,n}^{*}[\mathbf{W}_{BB}^{(2)}]_{l,n}$, when $\xi_{k,n} = [\widetilde{\mathbf{V}}_F]_{k,n} - \sum_{i\neq l} e^{j\varphi_{k,i}}[\mathbf{W}_{BB}^{(2)}]_{i,n}$.*

Proof: The proof is similar to the proof of Theorem 1. □

The solution to (26) for $\mathbf{W}_{\mathrm{BB}}^{(2)}$ is obtained by the least squares method, resulting in

$$\mathbf{W}_{\mathrm{BB}}^{(2)} = \left(\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{W}_{\mathrm{RF}}\right)^{-1}\mathbf{W}_{\mathrm{RF}}^{\mathrm{H}}\widetilde{\mathbf{V}}_{\mathrm{F}}. \quad (27)$$

Algorithm 2 solves (26) via the alternating minimization method. In each iteration r, we obtain $\mathbf{W}_{\mathrm{RF}}^{(r)}$ as per Theorem 2. The solution to $\mathbf{W}_{\mathrm{BB}}^{(2)(r)}$ is then found by (27). The alternating calculation of $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}^{(2)}$ is continued N_{it} times or until the reduction of distance in (26) is less than ϵ_{W}.

Having obtained $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}$ and $\mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}}^{(2)}$, solving (10) reduces to finding the transmit power P_{T}, which can be stated in terms of scalars in $\{p_1, \ldots, p_{N_{\mathrm{s}}}\}$ subject to per data stream QoS constraints. The solution to (18) suggests that $\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}$ is an approximation of the semi-unitary matrix $\widetilde{\mathbf{V}}_{\mathrm{W}}$, which can also be shown via computer simulations. Thus, for large N_{t}, we have

$$\mathbf{F}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)} \approx \mathbf{I}_{N_{\mathrm{s}}}, \quad (28)$$

From the above, the transmit power is approximated as (29), shown at the bottom of the page, where in (a) we use $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$, in (b) we use (28), and in (c) we use (14).

Algorithm 2 Solving (26)

Input: Random $\mathbf{W}_{\mathrm{BB}}^{(2)(0)}$, ϵ_{W}, N_{it}, $r = 0$
Compute $\mathbf{H}_{\mathrm{F}} = \mathbf{H}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}\mathbf{F}_{\mathrm{BB}}^{(2)\,\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}$
Compute EVD of $\mathbf{H}_{\mathrm{F}}$ and return $\widetilde{\mathbf{V}}_{\mathrm{F}}$
repeat
  $r = r + 1$
  Compute $\mathbf{W}_{\mathrm{RF}}^{(r)}$ using Theorem 2
  Compute $\mathbf{W}_{\mathrm{BB}}^{(2)(r)} = \left(\mathbf{W}_{\mathrm{RF}}^{(r)\mathrm{H}}\mathbf{W}_{\mathrm{RF}}^{(r)}\right)^{-1}\mathbf{W}_{\mathrm{RF}}^{(r)\mathrm{H}}\widetilde{\mathbf{V}}_{\mathrm{F}}$
until
$\|\mathbf{W}_{\mathrm{RF}}^{(r-1)}\mathbf{W}_{\mathrm{BB}}^{(2)(r-1)} - \widetilde{\mathbf{V}}_{\mathrm{F}}\|_{\mathrm{F}} - \|\mathbf{W}_{\mathrm{RF}}^{(r)}\mathbf{W}_{\mathrm{BB}}^{(2)(r)} - \widetilde{\mathbf{V}}_{\mathrm{F}}\|_{\mathrm{F}} \leq \epsilon_{\mathrm{W}}$ or $r \geq N_{\mathrm{it}}$
Output: $\mathbf{W}_{\mathrm{BB}}^{(2)}$ and $\mathbf{W}_{\mathrm{RF}}$

Having obtained P_{T}, the optimization problem (10) can be written as

$$\min_{p_1,\ldots,p_{N_{\mathrm{s}}}} \quad \sum_{k=1}^{N_{\mathrm{s}}} p_k \quad \text{subject to} \quad \begin{cases} \dfrac{1}{1 + p_k\lambda_{\mathrm{eff},k}} \leq \rho_k, & \forall k \\ p_k \geq 0, & \forall k, \end{cases} \quad (30)$$

whose optimal solution is

$$p_k^{\star} = \lambda_{\mathrm{eff},k}^{-1}(\rho_k^{-1} - 1). \quad (31)$$

Our procedure for designing hybrid precoders and combiners is as follows. In each iteration s of Algorithm 3, we obtain $\mathbf{F}_{\mathrm{RF}}^{(s)}$ and $\mathbf{F}_{\mathrm{BB}}^{(2)(s)}$ via Algorithm 1. Similarly, $\mathbf{W}_{\mathrm{RF}}^{(s)}$ and $\mathbf{W}_{\mathrm{BB}}^{(2)(s)}$ are obtained via Algorithm 2. The procedure continues N_{it} times or until the increase in the effective channel gain is less than ϵ. Then $\mathbf{H}_{\mathrm{eff}}$, $\mathbf{R}_{\tilde{\mathrm{n}}}$, $\mathbf{\Lambda}_{\mathrm{eff}}$ and $\mathbf{V}_{\mathrm{eff}}$ are obtained. Next, the transmit power P_{T} is obtained via (31), baseband precoder $\mathbf{F}_{\mathrm{BB}}^{(1)}$ is obtained via (14), and baseband combiner $\mathbf{W}_{\mathrm{BB}}^{(1)}$ is obtained via (12).

A. SIMPLIFIED HYBRID BEAMFORMING

Algorithm 3 contains nested *while* loops which is time consuming. The scheme can be simplified if obtaining $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}^{(2)}$ is decoupled from obtaining $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}^{(2)}$. In this case, there is no need for nested loops, and precoder and combiner matrices solely depend on $\mathbf{H}$ (not $\mathbf{H}_{\mathrm{W}}$ and $\mathbf{H}_{\mathrm{F}}$). When the number of antenna elements at both the transmitter and receiver is high, the precoder and combiner matrices can increase the effective channel's gain, which is maximized when beams are in the direction of channel eigenvectors.

$$P_{\mathrm{T}} = \mathrm{Tr}(\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}\mathbf{F}_{\mathrm{BB}}^{\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}) = \mathrm{Tr}(\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}\mathbf{F}_{\mathrm{BB}}^{(1)}\mathbf{F}_{\mathrm{BB}}^{(1)\mathrm{H}}\mathbf{F}_{\mathrm{BB}}^{(2)\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}) \overset{(a)}{=} \mathrm{Tr}(\mathbf{F}_{\mathrm{BB}}^{(2)\mathrm{H}}\mathbf{F}_{\mathrm{RF}}^{\mathrm{H}}\mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}^{(2)}\mathbf{F}_{\mathrm{BB}}^{(1)}\mathbf{F}_{\mathrm{BB}}^{(1)\mathrm{H}})$$
$$\overset{(b)}{\approx} \mathrm{Tr}(\mathbf{F}_{\mathrm{BB}}^{(1)}\mathbf{F}_{\mathrm{BB}}^{(1)\mathrm{H}}) \overset{(c)}{=} \mathrm{Tr}(\widetilde{\mathbf{V}}_{\mathrm{eff}}\mathbf{P}^{\frac{1}{2}}\mathbf{P}^{\frac{\mathrm{H}}{2}}\widetilde{\mathbf{V}}_{\mathrm{eff}}^{\mathrm{H}}) \overset{(a)}{=} \mathrm{Tr}(\widetilde{\mathbf{V}}_{\mathrm{eff}}^{\mathrm{H}}\widetilde{\mathbf{V}}_{\mathrm{eff}}\mathbf{P}^{\frac{1}{2}}\mathbf{P}^{\frac{\mathrm{H}}{2}}) = \mathrm{Tr}(\mathbf{P}) = \sum_{k=1}^{N_{\mathrm{s}}} p_k. \quad (29)$$

Algorithm 3 Solving (10)

Input: Random $\mathbf{W}_{\mathrm{RF}}^{(0)} \in \mathcal{W}_{\mathrm{FC}}$, $\mathbf{W}_{\mathrm{BB}}^{(2)(0)}$, $\mathbf{F}_{\mathrm{RF}}^{(0)} \in \mathcal{F}_{\mathrm{FC}}$, $\mathbf{F}_{\mathrm{BB}}^{(2)(0)}$, and $s = 0$

repeat

$s = s + 1$

Compute $\mathbf{F}_{\mathrm{RF}}^{(s)}$ and $\mathbf{F}_{\mathrm{BB}}^{(2)(s)}$ using Algorithm 1

Compute $\mathbf{W}_{\mathrm{RF}}^{(s)}$ and $\mathbf{W}_{\mathrm{BB}}^{(2)(s)}$ using Algorithm 2

$\mathbf{H}_{\mathrm{eff}}^{(s)} = \mathbf{W}_{\mathrm{BB}}^{(2)(s)\mathrm{H}} \mathbf{W}_{\mathrm{RF}}^{(s)\mathrm{H}} \mathbf{H} \mathbf{F}_{\mathrm{RF}}^{(s)} \mathbf{F}_{\mathrm{BB}}^{(2)(s)}$

until $\mathrm{Tr}(\mathbf{H}_{\mathrm{eff}}^{(s)\mathrm{H}} \mathbf{R}_{\tilde{\mathrm{n}}}^{-1} \mathbf{H}_{\mathrm{eff}}^{(s)}) - \mathrm{Tr}(\mathbf{H}_{\mathrm{eff}}^{(s-1)\mathrm{H}} \mathbf{R}_{\tilde{\mathrm{n}}}^{-1} \mathbf{H}_{\mathrm{eff}}^{(s-1)}) \leq \epsilon$ or $s \geq N_{\mathrm{it}}$

Compute $\mathbf{P}$ via (31).

$\mathbf{F}_{\mathrm{BB}}^{(1)} = \widetilde{\mathbf{V}}_{\mathrm{eff}} \mathbf{P}^{\frac{1}{2}}$

Compute $\mathbf{W}_{\mathrm{BB}}^{(1)}$ using (12).

Output: $\mathbf{F}_{\mathrm{BB}}^{(1)}$, $\mathbf{F}_{\mathrm{BB}}^{(2)}$, $\mathbf{F}_{\mathrm{RF}}$, $\mathbf{W}_{\mathrm{BB}}^{(1)}$, $\mathbf{W}_{\mathrm{BB}}^{(2)}$, $\mathbf{W}_{\mathrm{RF}}$

In the simplified Algorithm 4, we begin with singular value decomposition (SVD) of $\mathbf{H}$, i.e., $\mathbf{H} = \mathbf{U}_{\mathrm{H}} \mathbf{\Sigma}_{\mathrm{H}} \mathbf{V}_{\mathrm{H}}^{\mathrm{H}}$, and obtain precoder and combiner matrices by solving (18) and (26), respectively via Algorithms 1 - 2 and replacing $\mathbf{V}_{\mathrm{W}}$ with $\mathbf{V}_{\mathrm{H}}$ and $\mathbf{V}_{\mathrm{F}}$ with $\mathbf{U}_{\mathrm{H}}$. Theorems 1 - 2 are valid in this case as well. In contrast to Algorithm 3, there is no nested *while* loop in Algorithm 4 for obtaining $\mathbf{F}_{\mathrm{BB}}^{(1)}$, $\mathbf{P}$, and $\mathbf{W}_{\mathrm{BB}}^{(1)}$.

B. HYBRID BEAMFORMING IN PCS

In PCS, the matrices $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{RF}}$ are as in (3). The total number of phase shifters at the transmitter and receiver are N_{t} and N_{r}, respectively. Hence, the total consumed power is

$$P_{\mathrm{PC}} = (N_{\mathrm{t}} + N_{\mathrm{r}}) P_{\mathrm{PS}} + (N_{\mathrm{t}}^{\mathrm{RF}} + N_{\mathrm{r}}^{\mathrm{RF}}) P_{\mathrm{RF}} + P_{\mathrm{T}}. \quad (32)$$

The sets $\mathcal{F}_{\mathrm{FC}}$ and $\mathcal{W}_{\mathrm{FC}}$ in (10), (18), and (26) should be substituted with $\mathcal{F}_{\mathrm{PC}}$ and $\mathcal{W}_{\mathrm{PC}}$. All steps in Algorithms 1-4 can also be used for PCS, but references to Theorems 1 and 2 should be replaced with references to Theorems 3 and 4, respectively. This is not the case in [7], [9], [10], [27], which require different approaches for FCS and PCS.

Theorem 3: For a given $\mathbf{F}_{BB}^{(2)}$*, the* $(k, l)^{th}$ *phase shift that minimizes* (18) *in PCS is* $\varphi_{k,l}^{\star} = -\angle \delta_{k,l}$*, where* $\delta_{k,l} = \sum_{n=1}^{N_t^{RF}} [\widetilde{\mathbf{V}}_W]_{k,n}^{*} [\mathbf{F}_{BB}^{(2)}]_{l,n}$.

Proof: The proof is similar to the proof of Theorem 1, but not straightforward. □

Theorem 4: For a given $\mathbf{W}_{BB}^{(2)}$*, the* $(k, l)^{th}$ *phase shift that minimizes* (26) *in PCS is* $\varphi_{k,l}^{\star} = -\angle \beta_{k,l}$*, where* $\beta_{k,l} = \sum_{n=1}^{N_r^{RF}} [\widetilde{\mathbf{V}}_F]_{k,n}^{*} [\mathbf{W}_{BB}^{(2)}]_{l,n}$.

Proof: See the proof of Theorem 3. □

C. CONVERGENCE

In iteration q of Algorithm 1, the objective function in (18) is minimized with respect to $\mathbf{F}_{\mathrm{RF}}^{(q)}$ and $\mathbf{F}_{\mathrm{BB}}^{(2)(q)}$. Hence, for iterations q and $q+1$, we have $\|\mathbf{F}_{\mathrm{RF}}^{(q+1)} \mathbf{F}_{\mathrm{BB}}^{(2)(q+1)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}} \leq \|\mathbf{F}_{\mathrm{RF}}^{(q)} \mathbf{F}_{\mathrm{BB}}^{(2)(q)} - \widetilde{\mathbf{V}}_{\mathrm{W}}\|_{\mathrm{F}}$, i.e., the objective function is non-increasing. On the other hand, the objective function is lower bounded by zero. Thus, Algorithm 1 converges after sufficient iterations. The same is true for Algorithm 2.

Algorithm 3 aims to maximize the effective channel gain, which may decrease in some iterations due to minimizing the Euclidean distance instead of maximizing the effective channel gain, while using approximation (25). To resolve this, when eigenvalues decrease, Algorithm 3 is initialized by another set of random matrices, as shown in Section VII. Algorithm 4 solves (10) since Algorithms 1 and 2 converge and other steps include matrix derivations.

Algorithm 4 Simplified Scheme for Solving (10)

Input: Random $\mathbf{W}_{\mathrm{RF}}^{(0)} \in \mathcal{W}_{\mathrm{FC}}$, $\mathbf{W}_{\mathrm{BB}}^{(2)(0)}$, $\mathbf{F}_{\mathrm{RF}}^{(0)} \in \mathcal{F}_{\mathrm{FC}}$, and $\mathbf{F}_{\mathrm{BB}}^{(2)(0)}$

Compute SVD of $\mathbf{H}$, return $\mathbf{U}_{\mathrm{H}} \mathbf{\Sigma}_{\mathrm{H}} \mathbf{V}_{\mathrm{H}}^{\mathrm{H}}$

Compute $\mathbf{F}_{\mathrm{RF}}$ and $\mathbf{F}_{\mathrm{BB}}^{(2)}$ using Algorithm 1

Compute $\mathbf{W}_{\mathrm{RF}}$ and $\mathbf{W}_{\mathrm{BB}}^{(2)}$ using Algorithm 2

$\mathbf{H}_{\mathrm{eff}} = \mathbf{W}_{\mathrm{BB}}^{(2)\mathrm{H}} \mathbf{W}_{\mathrm{RF}}^{\mathrm{H}} \mathbf{H} \mathbf{F}_{\mathrm{RF}} \mathbf{F}_{\mathrm{BB}}^{(2)}$

Compute EVD of $\mathbf{H}_{\mathrm{eff}}$, return $\mathbf{V}_{\mathrm{eff}}$ and $\mathbf{\Lambda}_{\mathrm{eff}}$

Compute $\mathbf{P}$ via (31).

$\mathbf{F}_{\mathrm{BB}}^{(1)} = \widetilde{\mathbf{V}}_{\mathrm{eff}} \mathbf{P}^{\frac{1}{2}}$

Compute $\mathbf{W}_{\mathrm{BB}}^{(1)}$ using (12).

Output: $\mathbf{F}_{\mathrm{BB}}^{(1)}$, $\mathbf{F}_{\mathrm{BB}}^{(2)}$, $\mathbf{F}_{\mathrm{RF}}$, $\mathbf{W}_{\mathrm{BB}}^{(1)}$, $\mathbf{W}_{\mathrm{BB}}^{(2)}$, $\mathbf{W}_{\mathrm{RF}}$

D. COMPLEXITY

Here we analyze the computational complexity (CC) of our algorithms. Assume N_{t} and N_{r} are of the same order, i.e., $N = \mathcal{O}(N_{\mathrm{t}}) = \mathcal{O}(N_{\mathrm{r}})$, and the same assumption holds for the number of RF chains, i.e., $N_{\mathrm{RF}} = \mathcal{O}(N_{\mathrm{t}}^{\mathrm{RF}}) = \mathcal{O}(N_{\mathrm{r}}^{\mathrm{RF}})$. CC of Algorithms 1 and 2 is dominated by CC of line 5 therein. In FCS, CC of line 5 is $\mathcal{O}(NN_{\mathrm{RF}}^3)$. Hence CC of Algorithms 1 and 2 is $\mathcal{O}(N_{\mathrm{it}} N N_{\mathrm{RF}}^3)$, and CC of algorithms 3 and 4 are $\mathcal{O}(N_{\mathrm{it}}^2 N N_{\mathrm{RF}}^3)$ and $\mathcal{O}(N_{\mathrm{it}} N N_{\mathrm{RF}}^3)$, respectively.

In PCS, CC of line 5 is $\mathcal{O}(NN_{\mathrm{RF}})$. Hence, CC of Algorithms 1 and 2 is $\mathcal{O}(N_{\mathrm{it}} N N_{\mathrm{RF}})$, and CC of Algorithms 3 and 4 are $\mathcal{O}(N_{\mathrm{it}}^2 N N_{\mathrm{RF}})$ and $\mathcal{O}(N_{\mathrm{it}} N N_{\mathrm{RF}})$, respectively. Table 2 summarizes CC of our schemes in FCS and PCS.

V. QoS-AWARE MULTI-USER MIMO HYBRID BEAMFORMING

In this section, we extend the two-stage cascade structure to multi-user MIMO (MU-MIMO) systems. In MU-MIMO, a transmitter transmits $U \times N_{\mathrm{s}}$ data streams to U receivers and each receiver estimates its associated N_{s} data streams. The estimated vector of received symbols at any given instance by receiver u is

$$\hat{\mathbf{s}}_u = \mathbf{W}_{\mathrm{BB},u}^{\mathrm{H}} \mathbf{W}_{\mathrm{RF},u}^{\mathrm{H}} \mathbf{H}_u \mathbf{F}_{\mathrm{RF}} \mathbf{F}_{\mathrm{BB}} \mathbf{s} + \mathbf{W}_{\mathrm{BB},u}^{\mathrm{H}} \mathbf{W}_{\mathrm{RF},u}^{\mathrm{H}} \mathbf{n}_u, \quad (33)$$

where $\hat{\mathbf{s}}_u \in \mathbb{C}^{N_{\mathrm{s}} \times 1}$ is the estimate of $\mathbf{s}_u$ at receiver u, and $\mathbf{s} = [\mathbf{s}_1^{\mathrm{T}}, \mathbf{s}_2^{\mathrm{T}}, \ldots, \mathbf{s}_U^{\mathrm{T}}]^{\mathrm{T}}$ is the transmitted data vector. Moreover, receiver u is equipped with N_{r} antennas and $N_{\mathrm{r}}^{\mathrm{RF}}$ RF chains; and $\mathbf{W}_{\mathrm{BB},u} = \mathbf{W}_{\mathrm{BB},u}^{(2)} \mathbf{W}_{\mathrm{BB},u}^{(1)}$ and $\mathbf{W}_{\mathrm{RF},u} \in \mathbb{C}^{N_{\mathrm{r}} \times N_{\mathrm{r}}^{\mathrm{RF}}}$

TABLE 2. CC of our algorithms.

Structure	Algorithms 1 and 2	Algorithm 3	Algorithm 4
FCS	$\mathcal{O}(N_{\rm it}NN_{\rm RF}^3)$	$\mathcal{O}(N_{\rm it}^2NN_{\rm RF}^3)$	$\mathcal{O}(N_{\rm it}NN_{\rm RF}^3)$
PCS	$\mathcal{O}(N_{\rm it}NN_{\rm RF})$	$\mathcal{O}(N_{\rm it}^2NN_{\rm RF})$	$\mathcal{O}(N_{\rm it}NN_{\rm RF})$

are the baseband and RF combiner of receiver u, respectively. The channel $\mathbf{H}_u$ between transmitter and receiver u is modeled as in (4) and $\mathbf{n}_u \in \mathbb{C}^{N_{\rm r}\times 1}$ is the received noise vector at receiver u. The cascade baseband precoder is $\mathbf{F}_{\rm BB} = \mathbf{F}_{\rm BB}^{(2)}\mathbf{F}_{\rm BB}^{(1)}$, where $\mathbf{F}_{\rm BB}^{(1)} = \left[\mathbf{F}_{\rm BB,1}^{(1)},\ldots,\mathbf{F}_{\rm BB,U}^{(1)}\right]$ and $\mathbf{F}_{\rm BB,u}^{(1)} \in \mathbb{C}^{UN_{\rm s}\times N_{\rm s}}$ is the first stage of the baseband precoder for receiver u.

Our hybrid precoder and combiner minimize the transmit power while per data stream QoS requirements $[\bar{\mathbf{E}}_u]_{k,k} \le \rho_{u,k}$ are satisfied, i.e., the MSE of the $k^{\rm th}$ transmitted data stream to receiver u is less than or equal to $\rho_{u,k}$. We apply the well-known block diagonalization (BD) [34] to the baseband precoder, which gives

$$\mathbf{F}_{\rm BB,u}^{(1)} = \bar{\mathbf{V}}_u^{(0)}\widetilde{\mathbf{V}}_u\mathbf{P}_u^{\frac{1}{2}}, \tag{34}$$

where $\mathbf{P}_u$ is the diagonal transmit power matrix to receiver u. The columns of $\bar{\mathbf{V}}_u^{(0)}$ span the null space of $\bar{\mathbf{H}}_u = [\mathbf{H}_{{\rm eff},1}^{\rm T},\ldots,\mathbf{H}_{{\rm eff},u-1}^{\rm T},\mathbf{H}_{{\rm eff},u+1}^{\rm T},\ldots,\mathbf{H}_{{\rm eff},U}^{\rm T}]^{\rm T}$, where $\mathbf{H}_{{\rm eff},u} = \mathbf{W}_{{\rm BB},u}^{(2){\rm H}}\mathbf{W}_{{\rm RF},u}^{\rm H}\mathbf{H}_u\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}^{(2)}$ is the effective channel matrix of receiver u. The columns of $\bar{\mathbf{V}}_u^{(0)}$ are obtained by the following SVD

$$\bar{\mathbf{H}}_u = \bar{\mathbf{U}}_u[\bar{\mathbf{\Sigma}}_u \quad \mathbf{0}][\bar{\mathbf{V}}_u^{(1)} \quad \bar{\mathbf{V}}_u^{(0)}]^{\rm H}. \tag{35}$$

The inter-user interference is eliminated by embedding $\bar{\mathbf{V}}_u^{(0)}$ in $\mathbf{F}_{{\rm BB},u}^{(1)}$. To eliminate the inter-stream interference in receiver u, the matrix $\widetilde{\mathbf{V}}_u$ in (34) is obtained from the right eigenvectors of $\mathbf{R}_{\tilde{\rm n},u}^{-\frac{1}{2}}\mathbf{H}_u\bar{\mathbf{V}}_u^{(0)} = \mathbf{U}_u\mathbf{\Sigma}_u\mathbf{V}_u^{\rm H}$, where $\mathbf{R}_{\tilde{\rm n},u} = \mathbf{W}_{{\rm BB},u}^{(2){\rm H}}\mathbf{W}_{{\rm RF},u}^{\rm H}\mathbf{W}_{{\rm RF},u}\mathbf{W}_{{\rm BB},u}^{(2)}$ is the covariance matrix of the effective noise vector at receiver u and $\widetilde{\mathbf{V}}_u = [\mathbf{V}_u]_{1:N_{\rm s}}$.

Assuming MMSE combining in $\mathbf{W}_{{\rm BB},u}^{(1)}$ (similar to (12)) and having $\mathbf{F}_{{\rm BB},u}^{(1)}$ as in (34), the MSE matrix of the estimated data streams by receiver u is

$$\bar{\mathbf{E}}_u = \left(\mathbf{P}_u^{\frac{\rm H}{2}}\widetilde{\mathbf{\Lambda}}_{{\rm eff},u}\mathbf{P}_u^{\frac{1}{2}} + \mathbf{I}_{N_{\rm s}}\right)^{-1}, \tag{36}$$

where $\widetilde{\mathbf{\Lambda}}_{{\rm eff},u} = \mathbf{\Sigma}_u^{\rm T}\mathbf{\Sigma}_u$. From (36) and similar to (31), the allocated power to data stream k transmitted to receiver u is

$$p_{u,k} = \lambda_{u,k}^{-1}(\rho_{u,k}^{-1} - 1), \tag{37}$$

where $\lambda_{u,k}$ is the $k^{\rm th}$ largest eigenvalue of $\widetilde{\mathbf{\Lambda}}_{{\rm eff},u}$.

Similar to the single-user system model, $\mathbf{F}_{\rm BB}^{(2)}$, $\mathbf{F}_{\rm RF}$, $\mathbf{W}_{{\rm BB},u}^{(2)}$, and $\mathbf{W}_{{\rm RF},u}$ are obtained to maximize the effective channel gain of receiver u, i.e., ${\rm Tr}\left(\mathbf{H}_{{\rm eff},u}^{\rm H}\mathbf{R}_{\tilde{\rm n},u}^{-1}\mathbf{H}_{{\rm eff},u}\right)$. We extend the simplified Algorithm 4 into Algorithm 5 for multi-user beamforming. Extending Algorithm 3 for multi-user systems is straight forward and is not presented here for brevity.

Algorithm 5 Simplified MU-MIMO Hybrid Beamforming

Input: Random $\mathbf{W}_{{\rm RF},u}^{(0)} \in \mathcal{W}_{\rm FC}$, $\mathbf{W}_{{\rm BB},u}^{(2)(0)}$, $\mathbf{F}_{\rm RF}^{(0)} \in \mathcal{F}_{\rm FC}$, and $\mathbf{F}_{\rm BB}^{(2)(0)}$
for $u = 1 : U$ **do**
  Compute SVD of $\mathbf{H}_u^{\rm H}\mathbf{H}_u$, return $\mathbf{V}_{{\rm H},u}$
  Compute $\mathbf{W}_{{\rm RF},u}$ and $\mathbf{W}_{{\rm BB},u}^{(2)}$ via Algorithm 2
Compute SVD of $\mathbf{H}_{\rm mu}^{\rm H}\mathbf{W}_{\rm mu}\mathbf{W}_{\rm mu}^{\rm H}\mathbf{H}_{\rm mu}$, return $\mathbf{V}_{\rm mu}$
Compute $\mathbf{F}_{\rm RF}$ and $\mathbf{F}_{\rm BB}^{(2)}$ via Algorithm 1
for $u = 1 : U$ **do**
  Compute SVD of $\bar{\mathbf{H}}_u$, return $\mathbf{V}_u^{(0)}$
  Compute SVD of $\mathbf{R}_{\tilde{\rm n},u}^{-\frac{1}{2}}\mathbf{H}_u\bar{\mathbf{V}}_u^{(0)}$, return $\widetilde{\mathbf{V}}_u$
  Compute $\mathbf{P}_u$ via (37)
  $\mathbf{F}_{{\rm BB},u}^{(1)} = \bar{\mathbf{V}}_u^{(0)}\widetilde{\mathbf{V}}_u\mathbf{P}_u$
  Compute $\mathbf{W}_{{\rm BB},u}^{(1)}$ from (12).
Output: $\mathbf{F}_{\rm BB}^{(1)}$, $\mathbf{F}_{\rm BB}^{(2)}$, $\mathbf{F}_{\rm RF}$, $\mathbf{W}_{{\rm BB},u}^{(1)}$, $\mathbf{W}_{{\rm BB},u}^{(2)}$, $\mathbf{W}_{{\rm RF},u}$

The matrices $\mathbf{W}_{{\rm BB},u}^{(2)}$ and $\mathbf{W}_{{\rm RF},u}$ are obtained by minimizing $\|\mathbf{W}_{{\rm RF},u}\mathbf{W}_{{\rm BB},u}^{(2)} - \mathbf{V}_{{\rm H},u}\|_{\rm F}$ via Algorithm 2, where $\mathbf{V}_{{\rm H},u}$ includes the eigenvectors of $\mathbf{H}_u^{\rm H}\mathbf{H}_u$ associated with $N_{\rm s}$ largest eigenvalues.

Next, we obtain $\mathbf{F}_{\rm BB}^{(2)}$ and $\mathbf{F}_{\rm RF}$ by minimizing $\|\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}^{(2)} - \mathbf{V}_{\rm mu}\|_{\rm F}$ via Algorithm 1, where $\mathbf{V}_{\rm mu}$ includes $U \times N_{\rm s}$ eigenvectors of $\mathbf{H}_{\rm mu}^{\rm H}\mathbf{W}_{\rm mu}\mathbf{W}_{\rm mu}^{\rm H}\mathbf{H}_{\rm mu}$ for $U \times N_{\rm s}$ largest eigenvalues, $\mathbf{H}_{\rm mu} = [\mathbf{H}_1^{\rm T},\ldots,\mathbf{H}_U^{\rm T}]^{\rm T}$, and $\mathbf{W}_{\rm mu} = {\rm blkdiag}\left(\mathbf{W}_{{\rm RF},1}\mathbf{W}_{{\rm BB},1}^{(2)},\ldots,\mathbf{W}_{{\rm RF},U}\mathbf{W}_{{\rm BB},U}^{(2)}\right)$. Our MU-MIMO hybrid beamforming scheme is in Algorithm 5.

VI. HYBRID BEAMFORMING IN WIDEBAND SYSTEMS

Our proposed methods developed so far are for narrow-band channels (4). However, since mmWave channels are wideband [35], [36], we extend our approach to wideband channels as well. To do so, we use multi-carrier techniques such as orthogonal frequency division multiplexing (OFDM) to deal with multipath fading in wideband channels.

In MIMO-OFDM systems, per sub-carrier digital precoding is performed first. Then, the signals of all sub-carriers are combined via inverse fast Fourier transform (IFFT). Since a dedicated set of phase shifters for each sub-carrier is not practical, RF precoding is shared among all sub-carriers. The same is true for the receiver. The estimated signal in sub-carrier n at the receiver is

$$\hat{\mathbf{s}}[n] = \mathbf{W}_{\rm BB}^{\rm H}[n]\mathbf{W}_{\rm RF}^{\rm H}\mathbf{H}[n]\mathbf{F}_{\rm RF}\mathbf{F}_{\rm BB}[n]\mathbf{s}[n] + \mathbf{W}_{\rm BB}^{\rm H}[n]\mathbf{W}_{\rm RF}^{\rm H}\mathbf{n}[n], \tag{38}$$

where $\mathbf{H}[n]$ is the channel matrix in sub-carrier $n \in [1, N_{\rm c}]$, and $N_{\rm c}$ is the number of sub-carriers. As in [35], [36], our $\mathbf{H}[n]$ is the modified Saleh and Valenzuela model for mmWave channels.

The signal model in (38) implies that data streams in each sub-carrier are processed by a dedicated baseband precoder,

and do not interfere with data streams in other sub-carriers. The same is true for the baseband combiners. Hence, baseband precoders and combiners are individually designed for each sub-carrier, similar to the narrow-band systems in section IV, but the RF precoder and combiner are shared among all sub-carriers. Similar to Section IV, we find the RF precoder and combiner such that the sum of channel gains in all sub-carriers is maximized. Hence, we maximize $\sum_{n=1}^{N_c} \text{Tr}\left(\mathbf{H}_{\text{eff}}^{\text{H}}[n]\mathbf{R}_{\tilde{\text{n}}}^{-1}[n]\mathbf{H}_{\text{eff}}[n]\right)$, where $\mathbf{H}_{\text{eff}}[n]$ and $\mathbf{R}_{\tilde{\text{n}}}[n]$ are the effective channel matrix and covariance matrix of the effective noise in sub-carrier n, respectively. From Section IV, maximizing the sum of channel gains is achieved by solving

$$\min_{\mathbf{F}_{\text{RF}},\mathbf{F}_{\text{BB}}^{(2)}[n]} \quad \sum_{n=1}^{N_c} \|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{BB}}^{(2)}[n] - \mathbf{V}[n]\|_{\text{F}}^2 \quad \text{subject to} \quad \mathbf{F}_{\text{RF}} \in \mathcal{F}_{\text{FC}}, \tag{39}$$

where $\mathbf{V}[n]$ is the matrix of right eigenvectors of $\mathbf{H}[n]$. The objective function in (39) is $\|\mathbf{F}_{\text{RF}}\mathbf{F}_{\text{mc}}^{(2)} - \mathbf{V}_{mc}\|_{\text{F}}^2$, where $\mathbf{F}_{\text{mc}}^{(2)} = \left[\mathbf{F}_{\text{BB}}^{(2)}[1], \ldots, \mathbf{F}_{\text{BB}}^{(2)}[N_c]\right]$ and $\mathbf{V}_{\text{mc}} = [\mathbf{V}[1], \ldots, \mathbf{V}[N_c]]$. Hence, $\mathbf{F}_{\text{RF}}$ is obtained from Theorem 1 by substituting $\mathbf{F}_{\text{BB}}^{(2)}$ and $\widetilde{\mathbf{V}}_{\text{W}}$ with $\mathbf{F}_{\text{mc}}^{(2)}$ and $\mathbf{V}_{\text{mc}}$, respectively. Next, baseband precoding matrices are obtained as in Section IV for each subcarrier, which can be done in parallel since they are not the same in different sub-carriers. The same is true for baseband and RF combining matrices.

VII. SIMULATIONS

A. SIMULATION SETUP

Consider square planar antenna arrays ($N_{\text{H}} = N_{\text{V}}$) with half wavelength inter-element spacing ($d = \frac{\lambda}{2}$) at both the transmitter and receiver with $N_{\text{RF}} = N_{\text{t}}^{\text{RF}} = N_{\text{r}}^{\text{RF}}$ RF chains. A channel with $N_{cl} = 5$ clusters and mean values of AoAs and AoDs uniformly distributed between $[0, 2\pi]$ for the azimuth and between $[0, \pi]$ for the elevation, and $N_{ray} = 10$ rays in each cluster are assumed. Each AoA and AoD is drawn from a Laplacian distribution whose average is the corresponding mean value in the cluster and its standard deviation is 10 degrees. The gain of each ray is a complex Gaussian random variable with zero mean and unit variance. We set $P_{\text{RF}} = 100$ mW, $P_{\text{PS}} = 10$ mW [9], $\sigma_{\text{n}} = 1$, and $\rho_k = \rho, \forall k$, and average each point over 1000 channels. As in [37], frequency band is 28 GHz and channel bandwidth is 1 GHz, but normalized values are used in simulations.

B. SIMULATION RESULTS

Fig. 3 shows CCDF of MSE for a 144×64 system (i.e., $N_{\text{t}} = 144$ and $N_{\text{r}} = 64$) with $N_{\text{s}} = N_{\text{RF}} = 4$ and $\rho = 0.2$. OMP [5] is used to find the hybrid beamformer matrices from the fully digital QoS-aware beamformer matrix. The precoder is obtained as $\mathbf{F}_{\text{FD}} = \mathbf{V}_{\text{H}}\mathbf{P}_{\text{FD}}$, where $\mathbf{P}_{\text{FD}}$ is the diagonal power allocation matrix whose k^{th} diagonal entry similar to (31) is

$$p_{\text{FD},k} = \lambda_{\text{H},k}^{-1}(\rho^{-1} - 1), \tag{40}$$

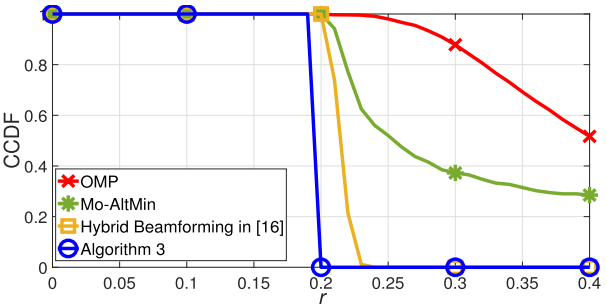

FIGURE 3. CCDF of MSE for a 144×64 system with $N_{\text{S}} = N_{\text{RF}} = 4$.

where $\lambda_{\text{H},k}$ is k^{th} eigenvalue of channel matrix $\mathbf{H}\mathbf{H}^{\text{H}}$. At the receiver, the optimal MMSE combiner is used. In hybrid beamforming, OMP finds approximations of $\mathbf{F}_{\text{FD}}$ and the optimal MMSE combiner as hybrid precoder and hybrid combiner, respectively.

As stated in Section III, in per data stream QoS-aware systems with $\rho_k = \rho, \forall k$, the values of CCDF $= \Pr\{\text{MSE} > r\}$ for all $r > \rho$ should be zero. As shown in Fig. 3, for $\rho = 0.2$, the values of CCDF for OMP is above zero for all $r > 0.2$. For example, CCDF is approximately 0.55 at $r = 0.4$. Hence, the CCDF for OMP and for manifold optimization based on alternating minimization (Mo-AltMin) in [9] is not a sharp step function, i.e., OMP and Mo-AltMin are not QoS-aware per data stream. This is due to the high sensitivity of per data stream MSE to approximation. In other words, by utilizing $\mathbf{F}_{\text{FD}}$ and MMSE combiner, MSE values in all data streams is ρ, but utilizing the approximated hybrid precoder and combiner yield higher MSE values.

Existing schemes for hybrid beamforming achieve various objectives that are incompatible with QoS constraints. For example, in [16], the hybrid precoder maximizes the mutual information. But maximizing mutual information and satisfying QoS constraints are two different objectives that cannot be achieved simultaneously. As such, [16] cannot be made QoS-aware per data stream. In contrast, as shown in Fig. 3, CCDF in our scheme for all $r > 0.2$ (e.g., at $r = 0.4$) is 0, which indicates that our scheme is QoS-aware per data stream with minimized transmit power.

Fig. 4 shows the Euclidean distance in (18) for different schemes. Although beamforming in [5], [9], [26], [27] are not QoS-aware per data stream, they solve a problem similar to (18). Note that the Euclidean distance in our scheme is much less than that for OMP-based distance minimization in [5], [26]. Our Euclidean distance is also slightly less than that in the hybrid design via least-squares relaxation (HD-LSR) in [27] and Mo-AltMin. Moreover, our scheme achieves zero Euclidean distance, i.e., optimal performance, when the number of RF chains is at least twice the number of data streams. Note that HD-LSR and Mo-AltMin are alternating methods whose computational complexity is similar to that of Algorithm 1. For $N_{\text{s}} = 2$ and $N_{RF} = 4$, the run times of HD-LSR, Mo-AltMin, and Algorithm 1 are 1.55, 2.29, and 1.82 seconds, respectively.

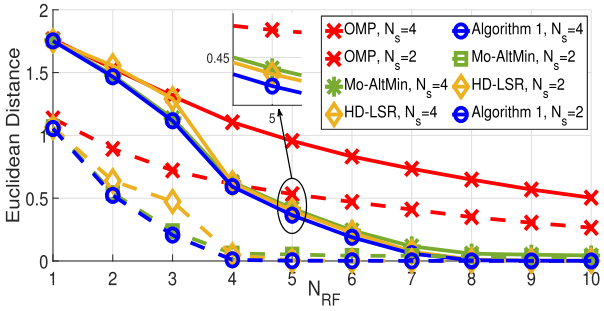

FIGURE 4. Euclidean distance between matrix of hybrid precoder and matrix of eigenvectors for $N_t = 64$ and $N_s = 2, 4$.

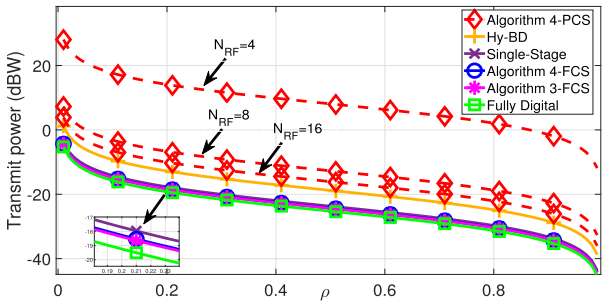

FIGURE 5. Transmit power vs. ρ for a 144 × 64 system with $N_{RF} = 4$ for FCS and $N_{RF} = 4, 8, 16$ for PCS.

Fig. 5 shows transmit power vs. ρ. Fully digital beamforming is per data stream QoS-aware with the lowest transmit power. Note that our Algorithm 4 (simplified scheme) consumes a negligible amount of 0.1 dBW more transmit power than our Algorithm 3 in FCS when $N_s = N_{RF} = 4$ for all $\rho = (0, 1)$. Fig. 5 also compares the performance of our scheme with a two-stage structure with that of a conventional scheme with a single-stage structure in [18] for a 144 × 64 system. As can be seen, our scheme is QoS-aware per data stream with less than 1 dBW optimality gap in the transmit power, whereas other schemes require more transmit power to be QoS-aware. In addition, the hybrid scheme Hy-BD in [25], which chooses RF beamformers from the columns of DFT matrix, requires significantly more transmit power than our scheme. Note that the transmit power in PCS is much higher than those in the fully digital and FCS, but is reduced when N_{RF} is increased.

To demonstrate the efficiency of our two-stage cascade structure, we consider the ratio of transmit power in fully digital beamforming to the same in our scheme, i.e.,

$$\frac{p_k^\star}{p_{\mathrm{FD},k}} = \frac{\lambda_{\mathrm{eff},k}^{-1}(\rho_k^{-1} - 1)}{\lambda_{\mathrm{H},k}^{-1}(\rho_k^{-1} - 1)} = \frac{\lambda_{\mathrm{H},k}}{\lambda_{\mathrm{eff},k}}. \tag{41}$$

Since $\lambda_{\mathrm{eff},k}$ is obtained via the approximated eigenvectors of $\mathbf{H}$ and $\lambda_{\mathrm{H},k}$ is obtained via exact eigenvectors of $\mathbf{H}$, we have $\lambda_{\mathrm{H},k} \geq \lambda_{\mathrm{eff},k}$. Note that energy efficiency of our schemes depends on the accuracy of approximating the eigenvectors of $\mathbf{H}$. Note also that the ratio in (41) is independent of the value of ρ_k.

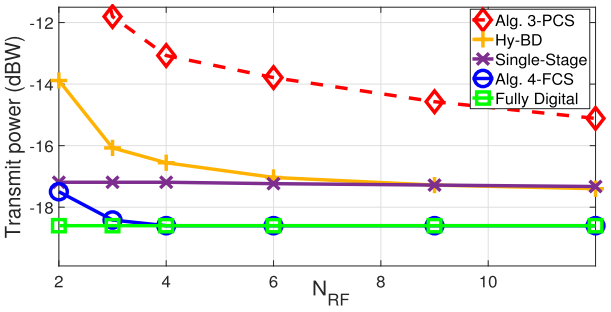

FIGURE 6. Transmit power vs. N_{RF} for a 144 × 64 system with $N_s = 2$ and at $\rho = 0.1$.

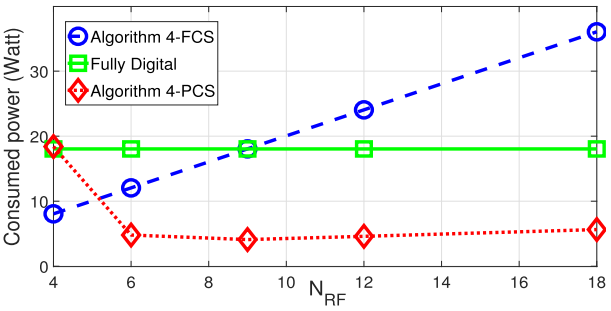

FIGURE 7. Total consumed power vs. N_{RF} for a 144 × 36 system.

Fig. 6 shows the transmit power for different number of RF chains. Algorithm 4 in FCS achieves the performance of fully digital beamforming for $N_{RF} \geq 2N_s$, which means Algorithm 1 and 2 achieve zero Euclidean distances for solving (18) and (26), which is in line with [16]. Moreover, the transmit power in the single-stage hybrid beamforming in [18] remains unchanged when the number of RF chains increases, i.e., Euclidean distances in (18) and (26) do not decrease by increasing the number of RF chains when only the phase information of channel eigenvectors are utilized.

Fig. 7 shows the impact of N_{RF} on the total consumed power. In FCS, the power consumed by RF chains dominates the transmit power, and a higher N_{RF} leads to a higher total consumed power. For large values of N_{RF}, the number of phase shifters in FCS is high, resulting in excessive power consumption. Hence, in such cases as can be seen in Fig. 7, the total consumed power is more than that of the fully digital beamforming in which no phase shifter is utilized. In PCS, the transmit power is dominant in the total consumed power for small N_{RF}, and a higher N_{RF} (i.e., a higher gain of the effective array) leads to less transmit power, resulting in less total consumed power. However, the increase in gain of the effective array for relatively large N_{RF} is insignificant, and hence, the reduction in transmit power is negligible.

Algorithm 3 minimizes Euclidean distances instead of directly maximizing the effective channel gain. Hence, the effective channel gain may not increase in each iteration, and the convergence of Algorithm 3 is not guaranteed. However, Fig. 8 shows that the effective channel gain on average increases with the number of iterations (each point is averaged over 1000 channels). Note that although the convergence of Algorithm 3 is not guaranteed, in practice, it is convergent.

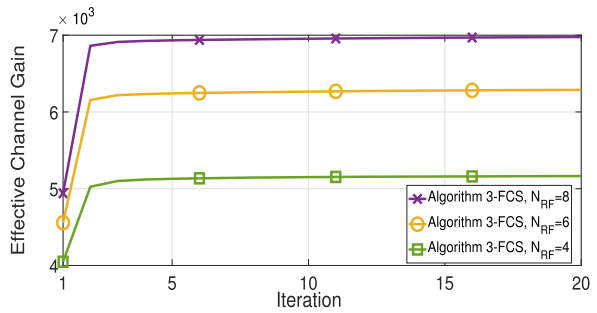

FIGURE 8. The effective channel gain vs. the number of iterations for a 144 × 36 system and $N_{RF} = 4, 6, 8$.

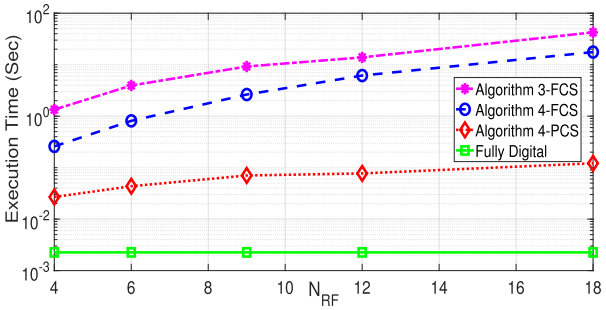

FIGURE 9. Run time of fully digital and hybrid beamforming vs. N_{RF}.

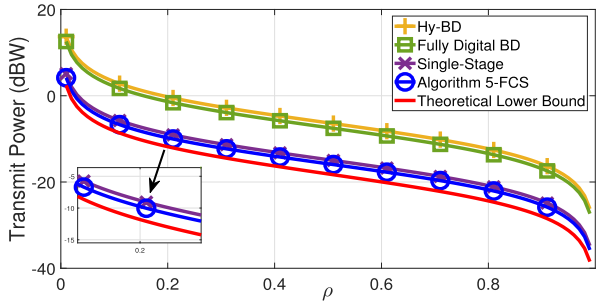

FIGURE 10. Transmit power vs. ρ in the two-stage and single-stage structures for a 144 × 16 system with $N_s = N_r^{RF} = 4$, $N_t^{RF} = UN_r^{RF}$, and $U = 4$.

Fig. 9 shows the run time of hybrid schemes in MATLAB® 2018b on a 3.30 GHz core i5 CPU and 16 GB RAM. Note that run time increases with N_{RF}. Algorithm 4 converges faster than Algorithm 3, and convergence of Algorithm 4 in PCS is much faster than in FCS.

Fig 10 shows that the transmit power of our two-stage cascade structure in a multi-user scenario is within 2 dBW gap of the theoretical lower bound [38] obtained by assuming parallel and non-interfering channels for different users. Similar to the single-user system, the single-stage structure in [18] requires more transmit power be QoS-aware per data stream. Note that both the single-stage and two-stage hybrid beamforming outperform the conventional fully digital BD method in [34].

The transmit power of the multi-user system for different number of RF chains is shown in Fig. 11. The two-stage cascade structure is within 2 dBW gap of the theoretical lower bound when $N_r^{RF} \geq N_s$. In contrast, the transmit power of the single-stage structure and Hy-BD increases with N_r^{RF} due to

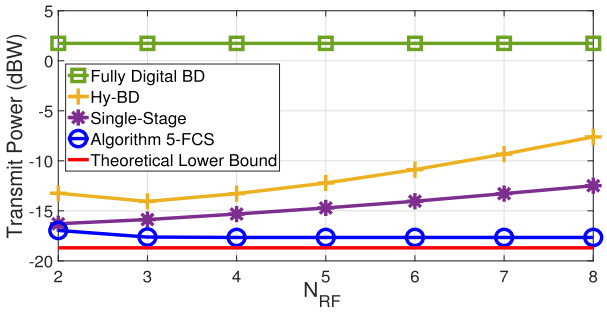

FIGURE 11. Transmit power vs. N_r^{RF} for a 144 × 16 system with $N_s = 2$, $N_t^{RF} = UN_r^{RF}$, and $U = 6$.

low gain of each user's channel on the null space of the interfering users' channels. In other words, the single-stage structure and Hy-BD fail to effectively diagonalize the multi-user channel for larger values of N_r^{RF}.

VIII. CONCLUSION

In this paper, we developed a per data stream QoS-aware hybrid beamforming scheme for mmWave massive MIMO systems that minimizes cross-channel interference and transmit power in different data streams. Specifically, we proposed a two-stage cascade structure for both the baseband precoders and combiners, and obtained their respective matrices by minimizing the transmit power subject to per data stream QoS constraints. Our designs are within 2 dBW optimality gap in transmit power in multiuser scenarios.

REFERENCES

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

[2] K. Xu, Z. Shen, Y. Wang, X. Xia, and D. Zhang, "Hybrid time-switching and power splitting swipt for full-duplex massive MIMO systems: A beam-domain approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7257–7274, Aug. 2018.

[3] Z. Shen, K. Xu, and X. Xia, "Beam-domain anti-jamming transmission for downlink massive MIMO systems: A Stackelberg game perspective," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2727–2742, 2021.

[4] Z. Shen, K. Xu, X. Xia, W. Xie, and D. Zhang, "Spatial sparsity based secure transmission strategy for massive MIMO systems against simultaneous jamming and eavesdropping," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3760–3774, 2020.

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

[6] S. Sun, T. S. Rappaport, R. W. Heath, Jr., A. Nix, and S. Rangan, "MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both?" *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 110–121, Dec. 2014.

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

[8] S. S. Ioushua and Y. C. Eldar, "A family of hybrid analog–digital beamforming methods for massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3243–3257, Jun. 2019.

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

[10] S. Payami, M. Ghoraishi, M. Dianati, and M. Sellathurai, "Hybrid beamforming with a reduced number of phase shifters for massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 4843–4851, Jun. 2018.

[11] J.-C. Chen, "Hybrid beamforming with discrete phase shifters for millimeter-wave massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7604–7608, Aug. 2017.
[12] D. Zhang, Y. Wang, X. Li, and W. Xiang, "Hybridly connected structure for hybrid beamforming in mmWave massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 662–674, Feb. 2018.
[13] C. G. Tsinos, S. Maleki, S. Chatzinotas, and B. Ottersten, "On the energy-efficiency of hybrid analog–digital transceivers for single-and multi-carrier large antenna array systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1980–1995, Sep. 2017.
[14] M. M. Molu, P. Xiao, M. Khalily, K. Cumanan, L. Zhang, and R. Tafazolli, "Low-complexity and robust hybrid beamforming design for multi-antenna communication systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1445–1459, Mar. 2018.
[15] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[16] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, Apr. 2016.
[17] C. Hu, J. Liu, X. Liao, Y. Liu, and J. Wang, "A novel equivalent baseband channel of hybrid beamforming in massive multiuser MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 764–767, Apr. 2018.
[18] X. Wu, D. Liu, and F. Yin, "Hybrid beamforming for multi-user massive MIMO systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3879–3891, Sep. 2018.
[19] N. Song, H. Sun, and T. Yang, "Coordinated hybrid beamforming for millimeter wave multi-user massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.
[20] Z. Wang, M. Li, X. Tian, and Q. Liu, "Iterative hybrid precoder and combiner design for mmWave multiuser MIMO systems," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1581–1584, Jul. 2017.
[21] L. Zhao, D. W. K. Ng, and J. Yuan, "Multi-user precoding and channel estimation for hybrid millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1576–1590, Jul. 2017.
[22] N. Song, T. Yang, and H. Sun, "Overlapped subarray based hybrid beamforming for millimeter wave multiuser massive MIMO," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 550–554, May 2017.
[23] Z. Li, S. Han, S. Sangodoyin, R. Wang, and A. F. Molisch, "Joint optimization of hybrid beamforming for multi-user massive MIMO downlink," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3600–3614, Jun. 2018.
[24] Q. Shi and M. Hong, "Spectral efficiency optimization for millimeter wave multiuser MIMO systems," *IEEE J. Sel. Toptics Signal Process.*, vol. 12, no. 3, pp. 455–468, Jun. 2018.
[25] 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.
[26] D. H. Nguyen, L. B. Le, T. Le-Ngoc, and R. W. Heath, Jr., "Hybrid MMSE precoding and combining designs for mmWave multiuser systems," *IEEE Access*, vol. 5, pp. 19167–19181, 2017.
[27] C. Rusu, R. Mèndez-Rial, N. González-Prelcic, and R. W. Heath, Jr., "Low complexity hybrid precoding strategies for millimeter wave communication systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8380–8393, Dec. 2016.
[28] Y. Han, S. Jin, J. Zhang, J. Zhang, and K. K. Wong, "DFT-based hybrid Beamforming multiuser systems: Rate analysis and beam selection," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 514–528, Jun. 2018.
[29] J. Geng, W. Xiang, Z. Wei, N. Li, and D. Yang, "Multi-user hybrid analogue/digital beamforming for relatively large-scale antenna systems," *IET Commun.*, vol. 8, no. 17, pp. 3038–3049, 2014.
[30] W. Ni, P.-H. Chiang, and S. Dey, "Energy efficient hybrid beamforming in massive MU-MIMO systems via eigenmode selection," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jun. 2017, pp. 400–406.
[31] L.-H. Shen and K.-T. Feng, "Joint beam and subband resource allocation with QoS requirement for millimeter wave MIMO systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.
[32] G. Zang, Y. Cui, H. V. Cheng, F. Yang, L. Ding, and H. Liu, "Optimal hybrid beamforming for multiuser massive MIMO systems with individual SINR constraints," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 532–535, Apr. 2019.
[33] D. P. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Found. Trends Commun. Inf. Theory*, vol. 3, nos. 4–5, pp. 331–551, Nov. 2006.
[34] Q. H. Spencer, A. L. 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.
[35] H. Li, M. Li, Q. Liu, and A. L. Swindlehurst, "Dynamic hybrid beamforming with low-resolution PSs for wideband mmWave MIMO-OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2168–2181, Sep. 2020.
[36] J. Zhang, Y. Huang, J. Wang, and L. Yang, "Hybrid precoding for wideband millimeter-Wave systems with finite resolution phase shifters," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11285–11290, Nov. 2018.
[37] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, and S. Rangan, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.
[38] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1570–1580, Apr. 2005.

MOHSEN TAJALLIFAR received the B.Sc. degree in electrical engineering from Shiraz University, Shiraz, Iran, in 2010, and the M.Sc. degree in electrical engineering from Tarbiat Modares University (TMU), Tehran, Iran, in 2013, where he is currently pursuing the Ph.D. degree with the Faculty of Electrical and Computer Engineering. His research interests include optimization methods, MIMO communication systems, and resource allocation in wireless networks.

AHMAD R. SHARAFAT (Life Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1975, and the M.Sc. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1976 and 1981, respectively. He is currently a Professor of electrical and computer engineering (ECE) at Tarbiat Modares University, Tehran. He has 12 patents, and coauthored four books and more than 160 papers in refereed scholarly journals and professional conferences. His research interests include advanced signal processing, and communications systems and networks. He is a Life Senior Member of Sigma Xi. He is a member of the Iranian Academy of Sciences; the Chairperson of ITU-D Study Group 2 in the International Telecommunication Union (ITU), Geneva, Switzerland; the Past Chairperson of IEEE Iran Section; and an Editor of the *International Journal of Wireless Information Networks* and *Scientia Iranica*.

HALIM YANIKOMEROGLU (Fellow, IEEE) is currently a Professor of SCE at Carleton University, Ottawa, Canada. He has made substantial contributions to wireless technologies, and his current focus is the aerial and satellite networks for 6G and beyond. His extensive collaboration with industry resulted in 38 patents. He is a fellow of the Engineering Institute of Canada (EIC) and the Canadian Academy of Engineering (CAE), and a Distinguished Speaker for the IEEE Communications Society and the IEEE Vehicular Technology Society. He received several awards for his research, teaching, and service.