Hardware-Aware Beamspace Precoding for All-Digital mmWave Massive MU-MIMO

Emre Gönultaş*, Sueda Taner*, Alexandra Gallyas-Sanhueza, Seyed Hadi Mirfarshbafan, and Christoph Studer

***Abstract*—Massive multi-user multiple-input multiple-output (MU-MIMO) wireless systems operating at millimeter-wave (mmWave) frequencies enable simultaneous wideband data transmission to a large number of users. In order to reduce the complexity of MU precoding in all-digital basestation architectures that equip each antenna element with a pair of data converters, we propose a two-stage precoding architecture which first generates a sparse precoding matrix in the beamspace domain, followed by an inverse fast Fourier transform that converts the result to the antenna domain. The sparse precoding matrix requires a small amount of multipliers and enables regular hardware architectures, which allows the design of hardware-efficient all-digital precoders. Simulation results demonstrate that our methods approach the error-rate performance of conventional Wiener filter precoding with more than 2× lower complexity.**

I. Introduction

Massive multi-user (MU) multiple-input multiple-output (MIMO) systems operating at millimeter-wave (mmWave) frequencies enable simultaneous, wideband wireless transmission to a large number of user equipments (UEs) [1], [2]. While the large contiguous bandwidths available at mmWave enable high per-UE data rates, the strong atmospheric absorption necessitates MU precoding to provide sufficiently high signal-to-noise ratios (SNRs) at the UE side. Since massive MU-MIMO equips the infrastructure basestations (BSs) with a large number of antennas, fine-grained beamforming and simultaneous data transmission to multiple UEs via spatial multiplexing is possible. Hybrid analog-digital beamforming architectures for mmWave systems have been proposed in the past [3], [4]. However, the trend is towards all-digital architectures [5], [6] that enable superior beamforming and spatial multiplexing capabilities and achieve comparable system costs and power consumption by deploying low-precision data converters at each antenna element. In order to successfully deploy all-digital architectures in practice, novel hardware-efficient baseband processing techniques for channel estimation, data detection, and MU precoding are necessary.

An emerging approach towards low-complexity baseband processing algorithms and simpler hardware architectures for all-digital BSs is to exploit beamspace sparsity [7]–[13]. Since mmWave propagation is highly directional, the UE signals arrive at the BS from only a few incident angles [2]. By taking a spatial discrete Fourier transform (DFT) across the antenna array (e.g., a uniform linear array), the received signal is transformed from the antenna domain to the beamspace domain, which concisely reveals the underlying angular sparsity [3], [14], [15]. The sparse nature of the received beamspace signals can then be exploited to design low-complexity baseband algorithms and simpler hardware architectures [7]–[13]. In the uplink, beamspace data detectors have been proposed in [12], [13] and beamspace channel estimators in [10], [16]–[18]. In the downlink, MU beamspace precoders have been proposed only recently in [9], [11], [19], [20].

1) Contributions: We propose two-stage beamspace precoding algorithms for all-digital mmWave massive MU-MIMO systems. Our algorithms rely on orthogonal matching pursuit (OMP) to compute sparse linear precoding matrices in the beamspace domain, which can result in lower precoding complexity than conventional, linear antenna-domain precoders that perform a dense matrix-vector product. The precoded output is then converted to the antenna domain using an inverse fast Fourier transform (IFFT). We use simulations for line-of-sight (LoS) and non-LoS mmWave channels to demonstrate that our algorithms approach the bit error-rate (BER) performance of conventional, antenna-domain Wiener filter (WF) precoding, while enabling more than 2× lower complexity.

2) Notation: Boldface lowercase and uppercase letters represent vectors and matrices, respectively. For a vector $\mathbf{a}$, the kth entry is $a_k = [\mathbf{a}]_k$. For a matrix $\mathbf{A}$, the transpose is $\mathbf{A}^\mathrm{T}$ and the conjugate transpose is $\mathbf{A}^\mathrm{H}$; the kth column is $\mathbf{a}_k = [\mathbf{A}]_k$ and the kth row is $\underline{\mathbf{a}}_k = [\mathbf{A}^T]_k^T$. For an index set Ω, $\mathbf{A}_\Omega$ refers to the submatrix of $\mathbf{A}$ with columns taken from Ω. The ℓ_2-norm of $\mathbf{a}$ is $\|\mathbf{a}\|$, the number of nonzero entries of $\mathbf{a}$ is denoted by $\|\mathbf{a}\|_0$, and the Frobenius norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F$. The $N \times N$ identity matrix is $\mathbf{I}_N$ and the $N \times M$ all-zeros matrix is $\mathbf{0}_{N\times M}$. The $N\times N$ unitary discrete Fourier transform (DFT) matrix is $\mathbf{F}_N$. The unit vector $\mathbf{e}_n$ contains a 1 in the nth entry and zeros otherwise. Vectors and matrices in the antenna domain are denoted with a bar, e.g., $\bar{\mathbf{a}}$ and $\bar{\mathbf{A}}$. The set of integers $\{1,\ldots,N\}$ is $[\![N]\!]$. The multivariate complex-valued circularly-symmetric Gaussian probability density function (PDF) with covariance matrix Σ is denoted by $\mathcal{CN}(\mathbf{0}_{N\times 1}, \mathbf{\Sigma}_N)$.

II. mmWave Massive MU-MIMO Downlink

A. Downlink Channel and System Model

We consider the mmWave massive MU-MIMO downlink, in which a BS with a B-antenna uniform linear array (ULA)

*EG and ST contributed equally to this work.

EG, ST, and AGS are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853; e-mail: eg566@cornell.edu, st939@cornell.edu, ag753@cornell.edu.

SHM and CS are with the Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland; e-mail: mirfarshbafan@iis.ee.ethz.ch, studer@ethz.ch.

The work of ST, AGS, and SHM was supported by ComSenTer, one of six centers in JUMP, an SRC program sponsored by DARPA. The work of EG and CS was supported by Xilinx, Inc. and by the US NSF under grants ECCS-1408006, CCF-1535897, CCF-1652065, CNS-1717559, and ECCS-1824379.

The authors thank O. Castañeda for discussions on computational complexity.

transmits data to U single-antenna UEs. We assume that the UEs are in the far-field and mmWave propagation conditions [2]. For illustrative purposes, we can model wave propagation from the BS to UE u with the standard plane-wave approximation [21] $\underline{\bar{\mathbf{h}}}_u = \sum_{\ell=0}^{L-1} \alpha_\ell \underline{\mathbf{a}}(\phi_\ell)$, where L refers to the number of transmission paths between UE u and the BS antenna array (including a possible LoS path), $\alpha_l \in \mathbb{C}$ is the complex-valued channel gain of the ℓth transmission path, and

$$\underline{\mathbf{a}}(\phi_\ell) = \left[1, e^{j\phi_\ell}, e^{j2\phi_\ell}, \ldots, e^{j(B-1)\phi_\ell}\right], \tag{1}$$

where ϕ_ℓ is the spatial frequency determined by the ℓth path's incident angle to the ULA. The downlink channel matrix $\bar{\mathbf{H}} \in \mathbb{C}^{U\times B}$ comprises the rows $\underline{\bar{\mathbf{h}}}_u$ for $u \in [\![U]\!]$. In Sec. IV, we show simulation results with more realistic mmWave channel vectors generated from the mmMAGIC QuaDRiGa model [22].

We consider a block-fading frequency-flat channel, in which the channel stays constant over a block of T time slots. We model the downlink input-output relation as follows:

$$\bar{\mathbf{y}} = \bar{\mathbf{H}}\bar{\mathbf{x}} + \bar{\mathbf{n}}. \tag{2}$$

Here, the U-dimensional vector $\bar{\mathbf{y}} \in \mathbb{C}^U$ comprises the signals received at all U UEs and the noise vector is $\bar{\mathbf{n}} \sim \mathcal{CN}(\mathbf{0}_{U\times 1}, N_0\mathbf{I}_U)$, where N_0 is assumed to be known at the BS. To mitigate MU interference, the BS must precode the transmit symbols. To this end, a B-dimensional antenna-domain precoded vector $\bar{\mathbf{x}}$ is formed according to

$$\bar{\mathbf{x}} = \mathcal{P}(\mathbf{s}, \bar{\mathbf{H}}, N_0, \rho^2), \tag{3}$$

where the transmit vector $\mathbf{s} \in \mathcal{O}^U$ contains the U data symbols to be transmitted to the UEs, $\mathcal{O}$ is the constellation set (e.g., 16-QAM), the transmit signals are assumed to be i.i.d. zero-mean and normalized so that $\mathbb{E}\left[|s_u|^2\right] = E_s$ for all $u \in [\![U]\!]$ and ρ^2 is the average power constraint so that $\mathbb{E}_{\mathbf{s}}\left[\|\bar{\mathbf{x}}\|^2\right] \leqslant \rho^2$. As in [23], we define the SNR as $\mathit{SNR} \triangleq \rho^2/N_0$.

B. MSE-Optimal Linear Precoding

To minimize the precoding complexity, we focus on linear precoders for which the function in (3) is linear, i.e.,

$$\bar{\mathbf{x}} = \mathcal{P}(\mathbf{s}, \bar{\mathbf{H}}, N_0, \rho^2) = \mathbf{P}\mathbf{s}, \tag{4}$$

where $\mathbf{P} \in \mathbb{C}^{B\times U}$ is a precoding matrix. Since multi-antenna transmission causes an array gain, each UE u performs scalar equalization of the received signal y_u with a precoding factor $\beta_u \in \mathbb{C}$ according to $\hat{s}_u = \beta_u y_u$, $u = 1, \ldots, U$. As in [24], we consider pilot-based estimation of the precoding factors: In the first time slot, the BS transmits U pilots with energy E_s, which are then used at each UE to estimate β_u. We focus on linear precoders that minimize the UE-side mean-square error (MSE) for a common $\beta \in \mathbb{C}$ so that

$$\mathit{MSE} \triangleq \mathbb{E}_{\mathbf{s},\mathbf{n}}\left[\|\mathbf{s} - \hat{\mathbf{s}}\|^2\right] = \mathbb{E}_{\mathbf{s},\mathbf{n}}\left[\|\mathbf{s} - \beta\mathbf{y}\|^2\right] \tag{5}$$

$$= \mathbb{E}_{\mathbf{s}}\left[\|\mathbf{s} - \beta\bar{\mathbf{H}}\mathbf{x}\|^2\right] + |\beta|^2 U N_0 \tag{6}$$

is minimized. The MSE-optimal linear precoder is known as the Wiener filter (WF) precoder [25], where the precoding matrix $\mathbf{P}^{\text{WF}} = \frac{1}{\beta(\bar{\mathbf{Q}}^{\text{WF}})}\bar{\mathbf{Q}}^{\text{WF}}$ is given by

$$\bar{\mathbf{Q}}^{\text{WF}} = \left(\bar{\mathbf{H}}^H\bar{\mathbf{H}} + \kappa^{\text{WF}}\mathbf{I}_B\right)^{-1}\bar{\mathbf{H}}^H. \tag{7}$$

Here, $\kappa^{\text{WF}} = UN_0/\rho^2$, and $\beta : \mathbb{C}^{N\times N} \to \mathbb{R}$ is a function that computes a pre-factor to satisfy the power constraint:

$$\beta(\mathbf{Q}) = \sqrt{\operatorname{tr}\left(\bar{\mathbf{Q}}^H\bar{\mathbf{Q}}\right)E_s/\rho^2}. \tag{8}$$

As it will become useful later, one can alternatively obtain the (unnormalized) WF precoding matrix $\bar{\mathbf{Q}}^{\text{WF}}$ in (7) by solving the following unconstrained optimization problem [26]:

$$\bar{\mathbf{Q}}^{\text{WF}} = \underset{\bar{\mathbf{Q}}\in\mathbb{C}^{B\times U}}{\arg\min}\ \|\bar{\mathbf{H}}\bar{\mathbf{Q}} - \mathbf{I}_U\|_F^2 + \kappa^{\text{WF}}\|\bar{\mathbf{Q}}\|_F^2. \tag{9}$$

C. Linear Precoding in the Beamspace Domain

In order to reduce the complexity of conventional, antenna-domain WF precoding $\bar{\mathbf{x}} = \mathbf{P}^{\text{WF}}\mathbf{s}$, one can perform linear precoding in the beamspace domain [11]. The key idea is to deploy linear precoders of the form

$$\bar{\mathbf{x}} = \mathcal{P}(\mathbf{s}, \bar{\mathbf{H}}, N_0, \rho^2) = \mathbf{F}_B^{\text{H}}\mathbf{P}\mathbf{s}, \tag{10}$$

where the inverse DFT matrix $\mathbf{F}_B^H$ converts the beamspace domain precoding vector $\mathbf{x} = \mathbf{P}\mathbf{s}$ into the antenna domain. Such two-stage precoders are able to exploit the sparse nature of the rows of the channel matrix $\bar{\mathbf{H}}$ in the beamspace domain, because the rows consist of a superposition of a few complex-valued sinusoids as in (1), i.e., the rows of the beamspace-domain mmWave MIMO channel matrix $\mathbf{H} = \bar{\mathbf{H}}\mathbf{F}_B$ are typically sparse [3], [15], [27], [28]; this property allows for the design of precoding matrices $\mathbf{P}$ whose columns are also sparsely populated [11]. For such sparse precoding matrices, computing (10), which is carried out at symbol rate, requires lower complexity than antenna-domain precoding as in (4).

III. Sparse Beamspace Precoding Algorithms

We now propose algorithms to compute sparse precoding matrices that are suitable for beamspace precoding as in (10). We start by an OMP-based algorithm, and then propose alternative algorithms with additional structure on the sparse matrix $\mathbf{P}$, which simplify corresponding hardware architectures.

A. Sparse Beamspace Precoding (SBP)

In order to design SBP matrices, we modify the optimization problem in (9) to deliver sparse matrices. As a first method, we propose to solve the following optimization problem

$$\mathbf{Q}^{\text{SBP}} = \begin{cases} \underset{\mathbf{Q}\in\mathbb{C}^{B\times U}}{\text{minimize}} & \|\mathbf{H}\mathbf{Q} - \mathbf{I}_U\|_F^2 + \kappa^{\text{WF}}\|\mathbf{Q}\|_F^2 \\ \text{subject to} & \mathbf{Q} \in \mathfrak{S}_{\text{SBP}} \end{cases} \tag{11}$$

where we impose a constraint that ensures each column of $\mathbf{Q}$ to have exactly K entries, i.e.,

$$\mathfrak{S}_{\text{SBP}} \triangleq \{\mathbf{Q} \in \mathbb{C}^{B\times U} : \|\mathbf{q}_u\|_0 = K, u = 1, \ldots, U\}. \tag{12}$$

We then normalize the matrix $\mathbf{Q}^{\text{SBP}}$ to obtain the SBP matrix $\mathbf{P}^{\text{SBP}} = \frac{1}{\beta(\mathbf{Q}^{\text{SBP}})}\mathbf{Q}^{\text{SBP}}$, where $\beta(\mathbf{Q}^{\text{SBP}})$ was defined in (8). It is important to realize that one can solve the problem in (11) on a per-column basis, i.e., we can solve

$$\mathbf{q}_u^{\text{SBP}} = \begin{cases} \underset{\mathbf{q}\in\mathbb{C}^{B}}{\text{minimize}} & \|\mathbf{H}\mathbf{q} - \mathbf{e}_u\|_2^2 + \kappa^{\text{WF}}\|\mathbf{q}\|_2^2 \\ \text{subject to} & \|\mathbf{q}\|_0 = K, \end{cases} \tag{13}$$

for $u = 1, \ldots, U$. Unfortunately, this sparse approximation problem is NP-hard [29] and thus must be solved using approximate methods. We propose to compute an approximate solution to (13) using OMP [30], as detailed next.

Let $\mathbf{q}_u^{(k)} \in \mathbb{C}^k$ be the vector computed after the kth OMP iteration, and $\mathbf{r}_u^{(k)}$ the associated residual. Let $\mho_u^{(k)}$ be the set of indices of the k nonzero entries of $\mathbf{q}_u$, and let $\Omega_u^{(k)}$ be the set of available indices for the new nonzero entry in the $(k+1)$th iteration. Here, $\Omega_u^{(k)} = [\![B]\!] \backslash \mho_u^{(k)}, \forall k$. We initialize the available and already-selected indices $\Omega_u^{(0)} = [\![B]\!]$, $\mho_u^{(0)} = \varnothing$, and the residual $\mathbf{r}_u^{(0)} = \mathbf{e}_u$. Then, repeat the following three steps for iterations $k = 1, \ldots, K$: (i) Identify the next best beam index by correlating the residual with the columns of $\mathbf{H}$,

$$b_u^{(k)} = \underset{b \in \Omega_u^{(k-1)}}{\arg\max}\, |\mathbf{h}_b^{\mathrm{H}} \mathbf{r}_u^{(k-1)}|, \tag{14}$$

and augment the support set, $\mho_u^{(k)} = \mho_u^{(k-1)} \cup \{b_u^{(k)}\}$. By definition, $b_u^{(k)}$ is unavailable for selection in subsequent iterations and we use $\Omega_u^{(k)} = \Omega_u^{(k-1)} \backslash \{b_u^{(k)}\}$. (ii) Update the SBP vector as for the WF precoder,

$$\mathbf{q}_u^{(k)} = (\mathbf{H}_{\mho_u^{(k)}}^{\mathrm{H}} \mathbf{H}_{\mho_u^{(k)}} + \kappa^{\mathrm{WF}} \mathbf{I}_k)^{-1} \mathbf{H}_{\mho_u^{(k)}}^{\mathrm{H}} \mathbf{e}_u. \tag{15}$$

(iii) Update the residual,

$$\mathbf{r}_u^{(k)} = \mathbf{e}_u - \mathbf{H}_{\mho_u^{(k)}} \mathbf{q}_u^{(k)}. \tag{16}$$

After K iterations, $\mathbf{q}_u^{(K)}$ gives the nonzero entries $[\mathbf{q}_u]_b, b \in \mho^{(K)}$, of the SBP column $\mathbf{q}_u$; and this procedure is repeated for all columns $\mathbf{q}_u$, $u \in [\![U]\!]$, of the unnormalized SBP matrix $\mathbf{Q}^{\mathrm{SBP}}$. We then normalize the sparse matrix $\mathbf{Q}^{\mathrm{SBP}}$ to obtain the SBP matrix $\mathbf{P}^{\mathrm{SBP}} = \mathbf{Q}^{\mathrm{SBP}}/\beta(\mathbf{Q}^{\mathrm{SBP}})$, where the precoding factor is calculated according to (8). The resulting SBP matrix $\mathbf{P}^{\mathrm{SBP}}$ contains—as desired—exactly KU nonzero entries.

B. Row-Select Sparse Beamspace Precoding (RS)

Although the above approach results in a sparse precoding matrix with KU nonzero entries, the unstructured nature of the nonzero entries in $\mathbf{P}$ prevents efficient hardware architectures that perform the sparse matrix-vector multiplication at high rates. To overcome this issue, we propose to enforce *structured* sparsity in the matrix $\mathbf{P}$ such that its rows have either all (U) non-zero entries or all zeros, so we can only store the non-zero rows and use efficient hardware for the sparse matrix-vector multiplication. Concretely, we aim to solve the sparse beamspace precoding problem in (11) with the constraint set

$$\mathfrak{S}_{\mathrm{RS}} \triangleq \Bigg\{ \mathbf{Q} \in \mathbb{C}^{B \times U} : \|\underline{\mathbf{q}}_b\|_0 = \begin{cases} U, & \text{if } b \text{ is selected} \\ 0, & \text{otherwise} \end{cases}, \quad \|\mathbf{q}_u\|_0 = K, u = 1, \ldots, U \Bigg\}, \tag{17}$$

which requires us to find K non-zero rows of the unnormalized precoding matrix $\mathbf{Q}$ with each having U non-zero entries. This problem resembles a multiple measurement vector (MMV) problem [31] and we use an OMP-MMV-like algorithm; we call the method *Row-Select SBP*, simply denoted by RS.

Let $\mho^{(k)}$ denote the rows of $\mathbf{Q}$ that are selected as nonzero in the first k iterations, and $\Omega^{(k)} = [\![B]\!] \backslash \mho^{(k)}$ the remaining ones, i.e., rows available for selection in the $(k+1)$th iteration. Let $\mathbf{Q}^{(k)} \in \mathbb{C}^{k \times U}$ denote a submatrix of the precoding matrix computed at the kth iteration, and $\mathbf{R}^{(k)}$ the residual. We initialize the set of selected nonzero rows $\mho^{(0)} = \varnothing$ and the residual $\mathbf{R}^{(0)} = \mathbf{I}_U$. We repeat the following steps for iterations $k = 1, \ldots, K$: (i) Identify the next best beam index,

$$\hat{b}^{(k)} = \underset{b \in \Omega^{(k-1)}}{\arg\max}\, \|\mathbf{h}_b^H \mathbf{R}^{(k-1)}\|_2, \tag{18}$$

and add this index to the support set $\mho^{(k)} = \mho^{(k-1)} \cup \{\hat{b}\}$. By definition, $\Omega^{(k)} = \Omega^{(k-1)} \backslash \{\hat{b}\}$. (ii) Update the submatrix of the precoding matrix,

$$\mathbf{Q}^{(k)} = (\mathbf{H}_{\mho^{(k)}}^{\mathrm{H}} \mathbf{H}_{\mho^{(k)}} + \kappa^{\mathrm{WF}} \mathbf{I}_k)^{-1} \mathbf{H}_{\mho^{(k)}}^{\mathrm{H}}. \tag{19}$$

(iii) Update the residual,

$$\mathbf{R}^{(k)} = \mathbf{I}_U - \mathbf{H}_{\mho^{(k)}} \mathbf{Q}^{(k)}. \tag{20}$$

After K iterations, the rows of $\mathbf{Q}^{(K)}$ deliver the nonzero rows $\underline{\mathbf{q}}_b, b \in \mho^{(K)}$, of the unnormalized RS matrix $\mathbf{Q}^{\mathrm{RS}}$, which has exactly KU nonzero entries with $\underline{\mathbf{q}}_b$ containing exactly U nonzeros. The RS matrix is obtained by $\mathbf{P}^{\mathrm{RS}} = \mathbf{Q}^{\mathrm{RS}}/\beta\big(\mathbf{Q}^{\mathrm{RS}}\big)$ with the pre-factor in (8).

C. Simplified One-Shot SBP Algorithms

All of the above methods require K iterations to construct K-sparse beam vectors for each UE. To further reduce the preprocessing complexity, we propose simplified methods that require only one iteration. For the counterpart of SBP, we construct the support set Ω_u per user u by selecting K beam indices that maximize the criterion in (14). For the counterpart of RS, we construct the support set of nonzero rows by selecting the K beam indices maximizing (18). We call each of these methods One-Shot SBP (1S-SBP) and One-Shot RS (1S-RS).

IV. Results

A. Simulation Setup

We simulate LoS and non-LoS channel conditions using the QuaDRiGa mmMAGIC UMi model [22] at a carrier frequency of 60 GHz with $\lambda/2$-spaced antennas arranged as a ULA. We generate channel matrices for a mmWave massive MIMO system with $B = 128$ antennas, and for $U = 16$ and $U = 32$ UEs. The UEs are placed randomly in a 120° circular sector around the BS between a distance of 10 m and 110 m, and we assume a minimum UE separation of 1°. We add BS-side power control so that the UE with highest received power has at most 6 dB more than the weakest UE. In order to account for channel estimation errors, we assume that the BS has access to a noisy version of $\mathbf{H}$ modeled as $\hat{\mathbf{H}} = \sqrt{1-\epsilon}\mathbf{H} + \sqrt{\epsilon}\mathbf{Z}$ as in [23]. Here, $\mathbf{Z} \sim \mathcal{CN}(\mathbf{0}_{U \times B}, \mathbf{I}_N)$ models the error for pilot-based channel estimation and we set $\epsilon = 0.0099$ so that the error corresponds to operating the system at 20 dB SNR.

We simulate uncoded BER with respect to SNR for the simulation scenarios for $K = U$ and $K = 2U$ using the sparse beamspace precoding methods in Sec. III. We also simulate the performance of WF in Sec. II-B and maximum ratio transmission (MRT) as baseline methods. We further include an algorithm, which we call "local Wiener filter" (local

TABLE I
COMPLEXITY OF VARIOUS PRECODING METHODS.

Algorithm	Preprocessing complexity	Precoding complexity
WF	$2U^3 + 6BU^2 - 2U(U+1) + 1$	$4TBU$
SBP	$4KB(U+2) + 2UK(K+1)$ $+2\sum_{k=1}^{K}(k^3 + 3Uk^2 - (U+1)k + 1)$	$4TKU + 2TB\log_2 B$
1S-SBP	$U(4B(U+2) + 2K^3 + 6UK^2 - 2(U+1)K + 1)$	$4TKU + 2TB\log_2 B$
Local WF	$2U^3 + 6KU^2 - 2U(U+1) + 1$	$4TM + 2TB\log_2 B$
MRT	0	$4TBU$

WF), that follows the idea put forward in [11] by using approximate channel vectors whose entries are equal to the entries of $\mathbf{h}_b, b \in [\![B]\!]$ in the window of $\mathbf{h}_b$ with highest energy and zero otherwise. To enable a fair comparison with this approach, the precoding coefficients are selected to minimize the MSE as in (6), whereas the original objective in [11] maximizes the minimum UE-side SINR. For local WF, we set the window size to K to allow for a fair performance and complexity comparison with our SBP-based methods.

B. Complexity Analysis

We provide a complexity analysis in Tbl. I, where we list the number of real-valued multiplications required during preprocessing (calculating the precoding matrix) and precoding (applying the precoding matrix to T transmit vectors), following the analysis in [26]; as in [12], we assume a complexity of $2B\log_2 B$ for a B-point IFFT. Since RS has the same total complexity as SBP, SBP represents both—the same holds for 1S-SBP and 1S-RS. For local WF, M denotes the average number of nonzero entries in the precoding matrix, where we assume a zero entry if the absolute value is smaller than 10^{-7}. In Fig. 1, we show the speed-up of the algorithms compared to MRT, which we define as the ratio of the complexity required by MRT to that of the algorithm, with respect to the number of transmissions T within a channel coherence interval. For $T \to \infty$, the asymptotic speed-up of our algorithms is $\gamma = \frac{2BU}{B\log_2 B + 4UK}$. Fig. 1 reveals that for small coherence times T, WF is less complex than all of the sparsity-based methods, which renders it the most preferable given that it achieves the smallest MSE. However, since in practical mmWave systems, the coherence time T can be as large as 10^5 [12], we see that already for $T > 10^3$, 1S-SBP is up to 2.91× faster than all of the baseline methods. SBP requires larger T and smaller K than 1S-SBP to outperform the baseline methods.

C. Bit Error-Rate Performance

Fig. 2 and Fig. 3 show the uncoded BER for the scenarios in Sec. IV-A under LoS and non-LoS conditions, respectively, for $U = 16$ users with $K = 16$ (a) and $K = 32$ (b) sparsity levels; and for $U = 32$ with $K = 32$ (c) and $K = 64$ (d). To compare the algorithms, we consider a target BER of 5%, for which all of our algorithms outperform local WF and MRT. For LoS channels with $U = K$, Fig. 2 (a) and (c) demonstrate that the SNR required by SBP and 1S-SBP methods to achieve the target BER is at most 1.5 dB and 2.5 dB higher than WF, respectively. We observe that the performance of RS methods significantly improves when $K = 2U$ in Fig. 2 (b) and (d). Here, 1S-RS requires approximately 1 dB higher SNR than WF, which renders it the most preferable when $T > 10^3$ considering the speed-up and the structure of $\mathbf{P}$ that allows for efficient precoding hardware. For non-LoS channels, we focus on the $K = 2U$ cases in Fig. 3 (b) and (d) for comparable performance to WF, where SBP and 1S-SBP methods require at most 1 dB higher SNR than WF. Considering the speed-up, we find 1S-SBP the most preferable in such scenarios.

V. CONCLUSIONS

We have proposed four different algorithms to perform sparse precoding in the beamspace domain. Our algorithms consist of two stages: The first stage computes a sparse precoding matrix; the second stage converts the precoded vector to the antenna domain using fast Fourier transform. Having a sparse precoding matrix reduces complexity and enables hardware efficient digital precoding architectures. Our simulation results for LoS and non-LoS mmWave channels have demonstrated that our sparse beamspace precoding algorithms enable more than 2× complexity reduction compared to traditional, antenna-domain Wiener filter precoding.

REFERENCES

[1] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

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

[3] 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.

[4] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, Jun. 2016.

[5] P. Skrimponis, S. Dutta, M. Mezzavilla, S. Rangan, S. H. Mirfarshbafan, C. Studer, J. Buckwalter, and M. Rodwell, "Power consumption analysis for mobile mmwave and sub-THz receivers," in *Proc. 2nd 6G Wireless Summit*, Mar. 2020, pp. 1–5.

[6] Z. M. Enciso, S. Hadi Mirfarshbafan, O. Castañeda, C. J. Schaefer, C. Studer, and S. Joshi, "Analog vs. digital spatial transforms: A throughput, power, and area comparison," in *Proc. IEEE Int. Midwest Symp. Circuits Systems*, Aug. 2020, pp. 125–128.

[7] X. Gao, L. Dai, Z. Chen, Z. Wang, and Z. Zhang, "Near-optimal beam selection for beamspace mmWave massive MIMO systems," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1054–1057, Mar. 2016.

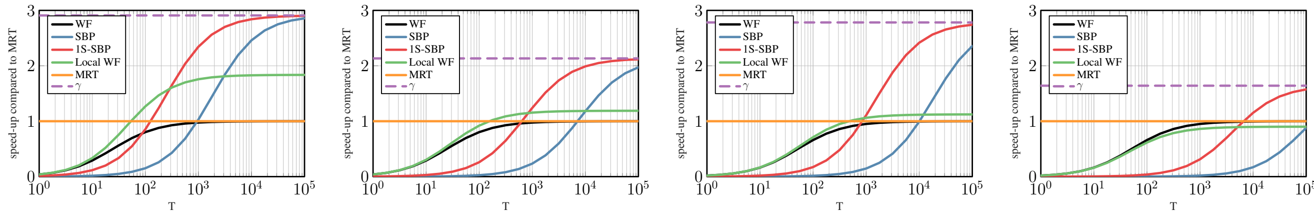
[8] C. Chen, C. Tsai, Y. Liu, W. Hung, and A. Wu, "Compressive sensing (CS) assisted low-complexity beamspace hybrid precoding for millimeter-wave MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1412–1424, Dec. 2017.

[9] A. Sayeed and J. Brady, "Beamspace MIMO for high-dimensional multiuser communication at millimeter-wave frequencies," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Dec. 2013, pp. 3679–3684.

[10] S. H. Mirfarshbafan, A. Gallyas-Sanhueza, R. Ghods, and C. Studer, "Beamspace channel estimation for massive MIMO mmWave systems: Algorithm and VLSI design," *IEEE Trans. Circuits Syst. I*, pp. 1–14, Sep. 2020.

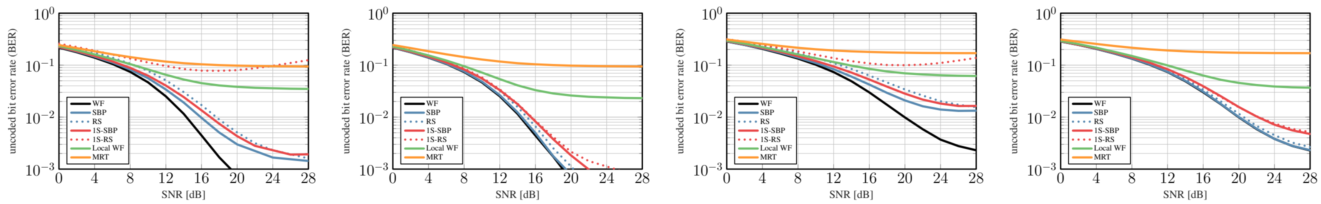
[11] M. Abdelghany, U. Madhow, and A. Tölli, "Efficient beamspace downlink precoding for mmWave massive MIMO," in *Asilomar Conf. Signals, Syst., Comput.*, Nov. 2019, pp. 1459–1464.

[12] S. H. Mirfarshbafan and C. Studer, "Sparse beamspace equalization for massive MU-MIMO mmWave systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2020, pp. 1773–1777.



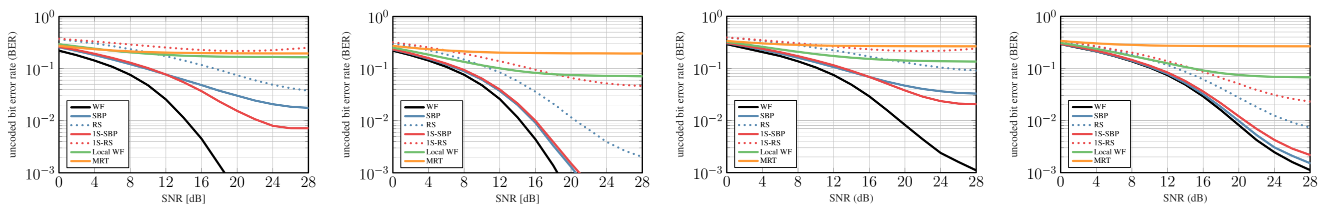
(a) $U = 16,\ K = 16, \gamma = 2.91$ (b) $U = 16,\ K = 32, \gamma = 2.13$ (c) $U = 32,\ K = 32, \gamma = 2.78$ (d) $U = 32,\ K = 64, \gamma = 1.64$

Fig. 1. Speed-up compared to MRT vs the number of transmissions (T) evaluated by the number of real-valued multiplications for $B = 128$ BS antennas, $U = 16$ and $U = 32$ users for sparsity values $K = U$ and $K = 2U$. The proposed sparse beamspace precoding algorithms are up to $2.91\times$ faster than MRT.



(a) $U = 16,\ K = 16$ (b) $U = 16,\ K = 32$ (c) $U = 32,\ K = 32$ (d) $U = 32,\ K = 64$

Fig. 2. BER results of LoS scenario with $B = 128$ BS antennas, $U = 16$ and $U = 32$ users for sparsity values $K = U$ and $K = 2U$. The proposed sparse beamspace precoding algorithms are able to achieve near-WF performance for sparsity levels $K = 2U$ for LoS mmWave channels.



(a) $U = 16,\ K = 16$ (b) $U = 16,\ K = 32$ (c) $U = 32,\ K = 32$ (d) $U = 32,\ K = 64$

Fig. 3. BER results of non-LoS scenario with $B = 128$ BS antennas, $U = 16$ and $U = 32$ users for sparsity values $K = U$ and $K = 2U$. The proposed sparse beamspace precoding algorithms are able to achieve near-WF performance for sparsity levels $K = 2U$ for non-LoS mmWave channels.

[13] M. Mahdavi, O. Edfors, V. Öwall, and L. Liu, "Angular-domain massive MIMO detection: Algorithm, implementation, and design tradeoffs," *IEEE Trans. Circuits Syst.*, vol. 67, no. 6, pp. 1948–1961, Jan. 2020.

[14] J. Mo, P. Schniter, and R. W. Heath Jr., "Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1141–1154, Mar. 2016.

[15] J. Lee, G. Gil, and Y. H. Lee, "Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2370–2386, Jun. 2016.

[16] X. Gao, L. Dai, S. Han, C. I, and X. Wang, "Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6010–6021, Jun. 2017.

[17] H. He, C. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Commun. Lett.*, vol. 7, no. 5, pp. 852–855, May 2018.

[18] L. Dai, X. Gao, S. Han, I. Chih-Lin, and X. Wang, "Beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," in *Int. Conf. Commun. China*, Jul. 2016, pp. 1–6.

[19] R. Pal, K. V. Srinivas, and A. K. Chaitanya, "A beam selection algorithm for millimeter-wave multi-user MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 852–855, Feb. 2018.

[20] M.-F. Tang and B. Su, "Downlink precoding for multiple users in FDD massive MIMO without CSI feedback," *Springer J. Signal Process. Syst.*, vol. 83, no. 2, pp. 151–163, May 2016. [Online]. Available: https://doi.org/10.1007/s11265-015-1079-0

[21] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge Univ. Press, 2005.

[22] S. Jaeckel, L. Raschkowski, K. Börner, L. Thiele, F. Burkhardt, and E. Eberlein, "QuaDRiGa - Quasi Deterministic Radio Channel Generator User Manual and Documentation," Fraunhofer Heinrich Hertz Institute, Tech. Rep. v2.0.0, 2017.

[23] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized precoding for massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[24] K. Li, C. Jeon, J. R. Cavallaro, and C. Studer, "Feedforward architectures for decentralized precoding in massive MU-MIMO systems," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 1659–1665.

[25] M. Joham, W. Utschick, and J. A. Nossek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005.

[26] O. Castañeda, S. Jacobsson, G. Durisi, T. Goldstein, and C. Studer, "Finite-alphabet MMSE equalization for all-digital massive MU-MIMO mmWave communication," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 9, pp. 2128–2141, Jun. 2020.

[27] P. Schniter and A. Sayeed, "Channel estimation and precoder design for millimeter-wave communications: The sparse way," in *Asilomar Conf. Signals, Syst., Comput.*, Nov. 2014, pp. 273–277.

[28] 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.

[29] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995. [Online]. Available: https://doi.org/10.1137/S0097539792240406

[30] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Asilomar Conf. Signals, Syst., Comput.*, Nov. 1993, pp. 40–44 vol.1.

[31] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Nov. 2006.