Tiled Beamspace Processing for Scaling mmWave Massive MU-MIMO

Jiyoon Han*†, Canan Cebeci*§, Wei Tang†, Zhengya Zhang† and Upamanyu Madhow§
†University of Michigan, Ann Arbor, MI, U.S.A. §University of California, Santa Barbara, CA, U.S.A.
{hanjyoon, weitang, zhengya}@umich.edu {ccebeci, madhow}@ucsb.edu

2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall) | 979-8-3315-1778-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/VTC2024-Fall63153.2024.10757813

***Abstract*—Recent progress in millimeter wave (mmWave) silicon technologies has given rise to a new possibility: digital beamforming for truly massive multiuser (MU-) MIMO. However, there are two key challenges in scaling: packaging a vast array of antennas with the corresponding radio frequency integrated circuits (RFICs), and controlling the complexity of digital signal processing (DSP) for MU detection. In this paper, we show that modular tiled architectures, which simplify the task of RF packaging, also enable significant reduction of the communication and computational burden of DSP for MU-MIMO by utilizing beamspace techniques that take advantage of the sparsity of the mmWave channel. Specifically, we propose and investigate Linear Minimum Mean Squared Error (LMMSE) adaptive MU detection via novel tiled beamspace architectures in which the bulk of the DSP occurs in-place at each tile. The dimensionality reduction and parallelization enabled by such architectures not only reduce the computational burden of inference and training relative to a traditional "full array" baseline, but also significantly reduce the length of the required training period. We consider three different training strategies with differing requirements for computation and inter-tile communication: independent training for each tile, coordinated training across tiles, and hierarchical training based on independent training as a first stage. Simulation results show that these approaches can actually outperform the full array baseline when we limit the length of the training period.**

***Index Terms*—Beamspace processing, Tiled architecture, mmWave, MU-MIMO, Linear MMSE**

I. INTRODUCTION

MmWave communication [1] and massive MU-MIMO [2] are key enablers of vastly increased data rates and network capacity in next-generation wireless communication systems. While the first generation of mmWave RFICs [3] utilize RF beamforming, recent work suggests that large-scale fully digital beamforming in these bands, using tiled architectures to alleviate packaging difficulties [4], is on the cusp of feasibility. In this paper, we show that such tiled architectures also alleviate the computational bottleneck of scaling DSP for MU-MIMO to a massive number of antennas and users, by parallelizing computation and exploiting the sparsity of the mmWave channel.

It is well known that increasing the number of base station (BS) antenna elements positioned at higher locations leads to a narrower angular spread and improved spatial resolution within the beamspace domain [5]. Larger losses due to scattering and blockage in mmWave channels further increases the sparsity, leading to well-defined channels characterized by a small number of dominant paths [1]. Channel matrices of this nature are sparse in the "beamspace" domain. For a linear array, taking an FFT across the antennas concentrates the signal energy of any given user into a small number of angular frequency bins, and recent work has utilized dimensionality reduction in beamspace to reduce the complexity of MU-MIMO [6]–[8].

In the tiled beamspace architectures considered here, we exploit the benefits of dimensionality reduction in beamspace while utilizing in-place computation at each tile and minimizing inter-tile communication. We consider adaptive LMMSE multiuser detection, which enables parallelization of computation across users as well. Our key contributions are summarized as follows:

- We model tiled uniform linear array (ULA) architectures for a mmWave MU-MIMO uplink, and show how to subdivide multiuser detection processing across tiles to achieve lightweight parallel computation within each tile with minimal inter-tile communication overhead. Sparsity in beamspace is utilized in each tile by using a small window of observations for each user, thus reducing both the complexity of multiuser detection and the required training overhead. Inference using LMMSE detection is parallelized across users and across tiles, and requires minimal inter-tile communication.
- We introduce and evaluate three innovative approaches for LMMSE adaptation for our tiled windowed beamspace architecture with different requirements for computation and inter-tile communication: independent, coordinated, and hierarchical training. We provide numerical results illustrating the tradeoffs between training complexity and BER performance for these schemes, all of which have exactly the same (low) inference complexity.
- Our numerical results illustrate the efficacy of dimensionality reduction on reducing training overhead, and show that our tiled windowed beamspace schemes significantly outperform a conventional full-array LMMSE detector when we limit the number of training symbols.

Throughout the paper, lowercase and uppercase boldface letters refer to column vectors and matrices, respectively. $*$, $\cdot^*$, $\cdot^\top$, and $\cdot^H$ are the convolution, conjugate, transpose, and conjugate transpose operators, respectively. $\mathbf{a}[r_1 : r_2]$ is a vector consisting of the r_1-th through the r_2-th elements of $\mathbf{a}$.

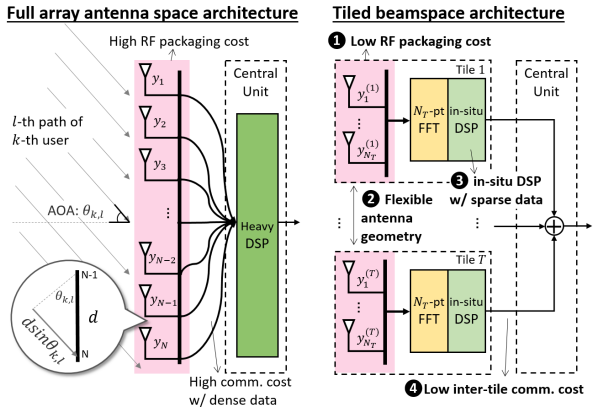
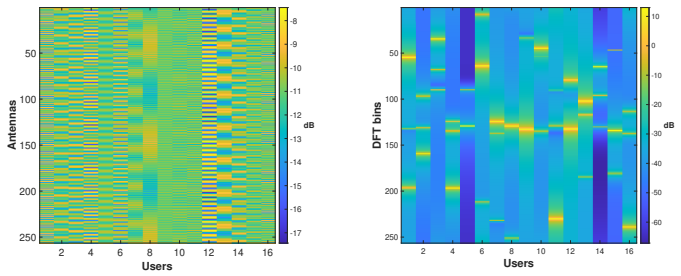


Fig. 1: Conventional full array antenna space architecture (left) and tiled beamspace architecture (right) for a mmWave MU-MIMO uplink with a tiled ULA at the base station.



(a) Antenna space channel (b) Beamspace channel

Fig. 2: Energy of a mmWave channel in (a) antenna space and (b) beamspace for 256 BS antennas and 16 users. The channel is generated at 28 GHz, incorporating 3 paths per user. The channel matrix in beamspace is sparse.

II. SYSTEM MODEL

We consider a massive MU-MIMO uplink as depicted in Fig. 1. The BS is equipped with an N-antenna ULA with inter-element spacing d, and communicates simultaneously with K users. We consider a narrowband symbol-synchronous model with singlecarrier modulation in order to focus on fundamental aspects of the comparison between standard architectures and tiled beamspace architecture. The received signal $\mathbf{y}$ at the BS is represented as

$$\mathbf{y}[n] = \sum_{k=1}^{K} (b_k * \mathbf{h}_k)[n] + \mathbf{n}[n], \tag{1}$$

where $\{b_k[n]\}$ and $\mathbf{h}_k[n]$ represent the symbol stream and vector channel response, respectively, for user k, and $\mathbf{n}[n]$ is a complex AWGN process with independent and identically distributed $\mathcal{CN}(0, 2\sigma^2)$ entries across space and time.

Typical mmWave channels consist of a dominant path, normally the Line of Sight (LoS) path, along with a few numbers of weaker paths which get further attenuated when beamforming along the dominant path [1]. Here, we consider L paths arriving in the far field from each user. The vector channel for user k, $1 \le k \le K$ is given by

$$\mathbf{h}_k[n] = \sum_{l=1}^{L} \mathbf{h}_{k,l}[n]\delta(n - \tau_{k,l}), \tag{2}$$

$$\mathbf{h}_{k,l}[n] = \alpha_{k,l}\mathbf{a}(\Omega_{k,l}), \tag{3}$$

$$\mathbf{a}(\Omega_{k,l}) = [e^{j0\Omega_{k,l}}\ e^{j1\Omega_{k,l}}\ e^{j2\Omega_{k,l}} \cdots e^{j(N-1)\Omega_{k,l}}]^\top. \tag{4}$$

where $\alpha_{k,l}$ is the path gain, $\Omega_{k,l} = \frac{2\pi d \sin\theta_{k,l}}{\lambda}$ is the spatial frequency, $\theta_{k,l}$ is the Angle of Arrival (AoA), and $\tau_{k,l}$ is the delay for the l^{th} path, quantized to the nearest sample to avoid detailed modeling of transmit and receive filters. Here λ denotes the carrier wavelength, and we set the inter-element spacing $d = \lambda/2$ in our numerical results.

A. Beamspace Representation

To convert from antenna space to beamspace, we utilize the Discrete Fourier Transform (DFT) at the BS. Denoting the N-point DFT matrix by $\mathbf{D}_N$, the beamspace representation for a user's channel is given by

$$\tilde{\mathbf{h}}_k[n] = \mathbf{D}_N \mathbf{h}_k[n]. \tag{5}$$

Fig. 2 highlights the sparsity of the beamspace channels. The energy of each user's channel vector is concentrated within a limited number of bins concentrated around its spatial frequency.

B. Tiled Architecture

We now consider the tiled architecture depicted in Fig. 1, where the array is partitioned into T tiles, each with N_T antennas at $\lambda/2$ spacing. Each tile has local RF chains and an FFT block responsible for transforming signals into beamspace. Such an architecture offers several advantages. From a DSP perspective, it allows for the distribution of high-complexity computations across multiple tiles, with the flexibility for different levels of inter-tile coordination. From a RF hardware design perspective, it alleviates packaging challenges associated with scaling array size at the traditional $\lambda/2$ pitch, while accommodating the associated electronics for controlling the array. From a systems perspective, it allows flexible design of antenna geometries to utilize the entire available aperture in order to optimize spatial resolution and reuse.

The received signal for tile t, $1 \le t \le T$, can be written as

$$\mathbf{y}^{(t)} = \sum_{k=1}^{K} \mathbf{b}_k * \mathbf{h}_k^{(t)} + \mathbf{n}^{(t)}. \tag{6}$$

In this paper, we focus on comparing tiled DSP with DSP for the full array, and assume that the antennas for the tiles are simply a partition of an $N = N_T T$ element $\lambda/2$-spaced linear array into T contiguous N_T-element arrays. In this case, the channels for k-th user and l-th path across tiles are related as follows:

$$\mathbf{h}_{k,l}^{(t)}[n] = e^{j\Omega_{k,l}N_T(t-1)}\mathbf{h}_{k,l}^{(1)}[n] \tag{7}$$

where $\mathbf{h}_{k,l}^{(1)}[n] = [e^{j0\Omega_{k,l}}\ e^{j1\Omega_{k,l}}\ e^{j2\Omega_{k,l}} \cdots e^{j(N_T-1)\Omega_{k,l}}]^\top$.

Note that each tile receives a superposition of signals from all users while having fewer antennas than the full array. Thus, coordination among tiles is required if we hope to approach the interference suppression capability of the full array.

III. MULTIUSER DETECTION

We compare LMMSE detection across different receiver architectures. We begin with a quick review of LMMSE reception for a generic model, and then map each architecture to this model. The D-dimensional received vector corresponding to the nth symbol $b[n]$ transmitted by a "desired" user k is given by

$$\mathbf{r}[n] = b_k[n]\mathbf{u}_k + \mathbf{I}_{\backslash k}[n] \tag{8}$$

where $\mathbf{u}_k$ denotes the "desired signal vector," and $\mathbf{I}_{\backslash k}[n]$ denotes the "interference vector" comprising noise and multiuser interference. Assuming that $b_k[n]$ and $\mathbf{I}_{\backslash k}[n]$ are each stationary (or at least wide-sense stationary), the LMMSE correlator $\mathbf{c}$ minimizes the MSE $E[|\mathbf{c}^H\mathbf{r}[n] - b_k[n]|^2]$, the solution to which is well known:

$$\mathbf{c}_{MMSE} = \mathbf{R}^{-1}\mathbf{p} \tag{9}$$

where $\mathbf{R} = E[\mathbf{r}[n]\mathbf{r}^H[n]]$ and $\mathbf{p} = E[b_k^*[n]\mathbf{r}[n]]$. In least squares adaptive implementations, these statistical averages are replaced by empirical averages over a sequence of, say, L_t training symbols, where it suffices to choose L_t to be a small multiple of the dimension D [9]. If there are multiple users being demodulated simultaneously based on the same received vector, then the least squares estimate of $\mathbf{R}$ is common for all users, while the estimates of $\mathbf{p}$ are separately computed based on the training sequence for each user. These training sequences are chosen to be orthogonal or quasi-orthogonal.

We investigate the performance of LMMSE reception for the system (1) for four receiver architectures detailed in the following subsections, where we define the received vector $\mathbf{r}$ with an arbitrarily chosen symbol, dropping the symbol number n, in order to map to the generic model (8). LMMSE detection, which may be viewed as a compromise between matched filtering and zero-forcing detection, typically yields good performance when the dimension D of the underlying signal space is larger than the number of strong interferers, which makes it likely that projecting orthogonal to the subspace spanned by these interferers does not lead to significant reduction in the energy of the desired signal. The tiled windowed beamspace architectures considered here are effective because they reduce dimension in a manner that preserves most of the desired signal's energy while reducing the number of strong interferers. Such dimension reduction strategies are particularly attractive for adaptive implementations: the computation of the inverse in (9) scales as $O(D^3)$, while the training overhead scales linearly with D.

A. Full Array Antenna Space Processing

We consider the baseline benchmark associated with the full array in antenna space ($D = N$), for which the received vector is the $N \times 1$ vector $\mathbf{r} = \mathbf{y}$. As we scale the number of antennas N, both the computational complexity and overhead for training becomes unattractive.

B. Tiled Windowed Beamspace (TWB) Processing

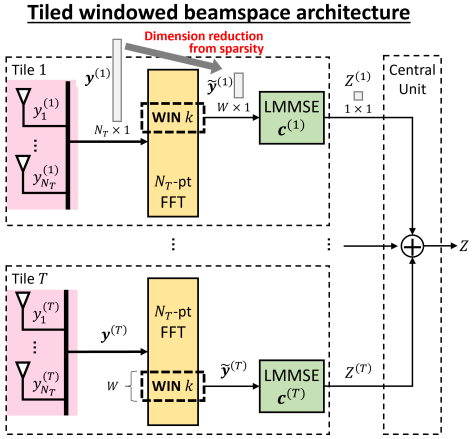


Fig. 3: Tiled windowed beamspace inference for each user consists of linear correlation with the received signal in a selected beamspace window in parallel for each tile. The final decision statistic is a sum of the tile-level decision statistics.

Here, we perform separate N_T-point DFTs on the tile-level received signals to obtain $\tilde{\mathbf{y}}^{(t)} = \mathbf{D}_{N_T}\mathbf{y}^{(t)}$, $1 \le t \le T$. For each tile t, we create a window around the highest energy bin $i_k^{(t)}$ for user k to obtain $W \times 1$ tile-level windowed beamspace signals:

$$\mathbf{r}^{(t)} = \tilde{\mathbf{y}}^{(t)}[i_k^{(t)} - W/2 + 1 : i_k^{(t)} + W/2]\ , 1 \le t \le T, \text{ for user } k$$

We wish to perform linear correlation on these signals to obtain a decision statistic Z for the symbol transmitted by user k. The linear correlator operating on the selected beamspace window for tile t is denoted by $\mathbf{c}^{(t)}$, and the decision statistic for user k's symbol is given by

$$Z = \sum_{t=1}^{T} \left(\mathbf{c}^{(t)}\right)^H \mathbf{r}^{(t)} = \mathbf{c}^H\mathbf{r} \tag{10}$$

where $\mathbf{c}$, $\mathbf{r}$ are $WT \times 1$ vectors obtained by concatenating $\mathbf{c}^{(t)}$, $1 \le t \le T$, and $\mathbf{r}^{(t)}$, $1 \le t \le T$, respectively. Note that the decision statistic Z is simply a sum of decision statistics $Z^{(t)} = \left(\mathbf{c}^{(t)}\right)^H \mathbf{r}^{(t)}$ that can be computed separately at each tile, yielding a highly parallelized architecture for inference, as depicted in Fig. 3.

Note that, while Fig. 3 depicts a general situation in which the window centers $i_k^{(t)}$ for user k can vary across tiles, in our setting, since each tile has exactly the same geometry with respect to user k, we would design the window selection algorithm to select a common window center $i_k^{(t)}$ across t.

While the correlator $\mathbf{c}$ in (10) operates over a signal space of dimension $D = WT$, the *effective* dimension of the signal space available for interference suppression depends on the

strategy for adapting it. We now consider three options for the latter, corresponding to different training complexity.

Coordinated training: Here we train based on the $WT \times 1$ vector $\mathbf{r}$ obtained by concatenating across tiles, minimizing the MSE $E[|Z - b_k|^2]$, where b denotes a generic symbol for a given user. Referring to the generic LMMSE solution (9), $D = WT$, computation of $\mathbf{R}$ requires coordination across tiles, while computation of $\mathbf{p}$ can be parallelized across tiles.

Independent training: Here we train the correlators $\mathbf{c}^{(t)}$ for each tile separately, minimizing the MSE $E[|TZ^{(t)} - b_k|^2]$ for each tile t, where the factor T is inserted to account for summation of tile-level decision statistics during inference. Referring to the generic solution (9), we compute T LMMSE solutions in parallel, with T entirely parallel computations of $W \times W$ matrices $\mathbf{R}^{(t)}$, and $W \times 1$ vectors $\mathbf{p}^{(t)}$, in each tile. This scheme is clearly suboptimal, since the tile-level correlators operate in a lower-dimensional signal space of dimension $D = W$, while seeing the same number of interfering users.

Hierarchical training: Here we use the tile-level correlators based on independent training as a first stage of interference suppression with signal space dimension $D = W$, and add a second training stage to exploit the correlations between the decision statistics produced by the tiles. In the second stage, we have signal space dimension $D = T$, using the received vector $\mathbf{r} = [Z^{(1)}, Z^{(2)}, ..., Z^{(T)}]$ from the first stage. The inter-tile communication overhead for training is as small as for inference, since we only require that the decision statistics produced by the tile-level correlators reach a centralized location. Given the limited signal space dimension available in the first stage, we regularize the LMMSE correlator to prevent excessive noise enhancement by adding diagonal loading to the correlation matrix prior to inversion. Specifically, in the first stage, we compute the LMMSE correlator at each tile as $\left(\mathbf{R} + \gamma \boldsymbol{I}_{W \times W}\right)^{-1} \mathbf{p}$ for each tile. In our simulation results, we find that $\gamma = 0.1 \sum_i \mathbf{R}_{ii}$, which effectively introduces "artificial noise" which is 10 dB weaker than the sum of the signal plus noise powers falling in the window, works well across the settings considered. Further improvements could potentially be obtained by adapting γ to each setting.

IV. PERFORMANCE EVALUATION

For the system in Fig. 1, we set the number of BS antennas to $N = 256$. The users' AoAs are uniformly distributed in $[-\pi/3, \pi/3]$, while enforcing a minimum separation in spatial frequency of $\frac{2.782}{N}$ [10] between users (we assume that higher layer resource allocation strategies would schedule users with smaller spatial frequency separations in different slots). For every channel realization, each user receives 3 paths, with the first path always being a Line of Sight (LoS). The amplitude of the second and third paths diminish exponentially as a function of their respective path delays, measured in a symbol unit. These path delays are uniformly distributed across a range from 1 to 100.

We consider two configurations: (1) $N_T = 64$, $T = 4$ serving $K = 16$ users; and (2) $N_T = 16$, $T = 16$ serving $K = 8$ users. For full array processing, the load factor is $1/16$ for the first configuration and $1/32$ for the second, both of which are small enough to expect negligible noise enhancement due to linear interference suppression. On the other hand, the load factor per tile is higher, equaling $1/4$ and $1/2$ respectively. Our simulation results (Fig. 4) show that the tiled beamspace algorithms work well even under these high load factor environment. We set the window for beamspace processing to $W = 4$ for each tile, since this captures more than 90% of the desired signal energy. For all tiled beamspace algorithms, we set the training sequence length to $L_t = 256$ (divided into two segments of length L_{t1} and L_{t2} for hierarchical training). We note that the overhead can be reduced even further with negligible impact on performance. On the other hand, for full-array processing with dimension $N = 256$, $L_t = N = 256$ is barely enough to obtain an invertible correlation matrix, hence we also present results for $L_t = 2N = 512$.

TABLE I: Complexity of different architectures and trainings.

	Full array antenna space	**Tiled windowed beamspace**		
FFT	0	$T\,N_T \log_2(N_T) = N \log_2(N_T)$		
Inference	KN	KTW		
Training		**Independent**	**Coordinated**	**Hierarchical** (1st stage, 2nd stage)
$\mathbf{R}$	$N^2 L_t$	$KTW^2 L_t$	$K(TW)^2 L_t$	$KTW^2 L_{t1},\ KT^2 L_{t2}$
$\mathbf{P}$	KNL_t	$KTWL_t$	$KTWL_t$	$KTWL_{t1},\ KTL_{t2}$
$\mathbf{R}^{-1}$	$h(N)$	$KT\,h(W)$	$K\,h(TW)$	$KT\,h(W),\ K\,h(T)$
$\mathbf{C}$	KN^2	KTW^2	$K(TW)^2$	$KTW^2,\ KT^2$

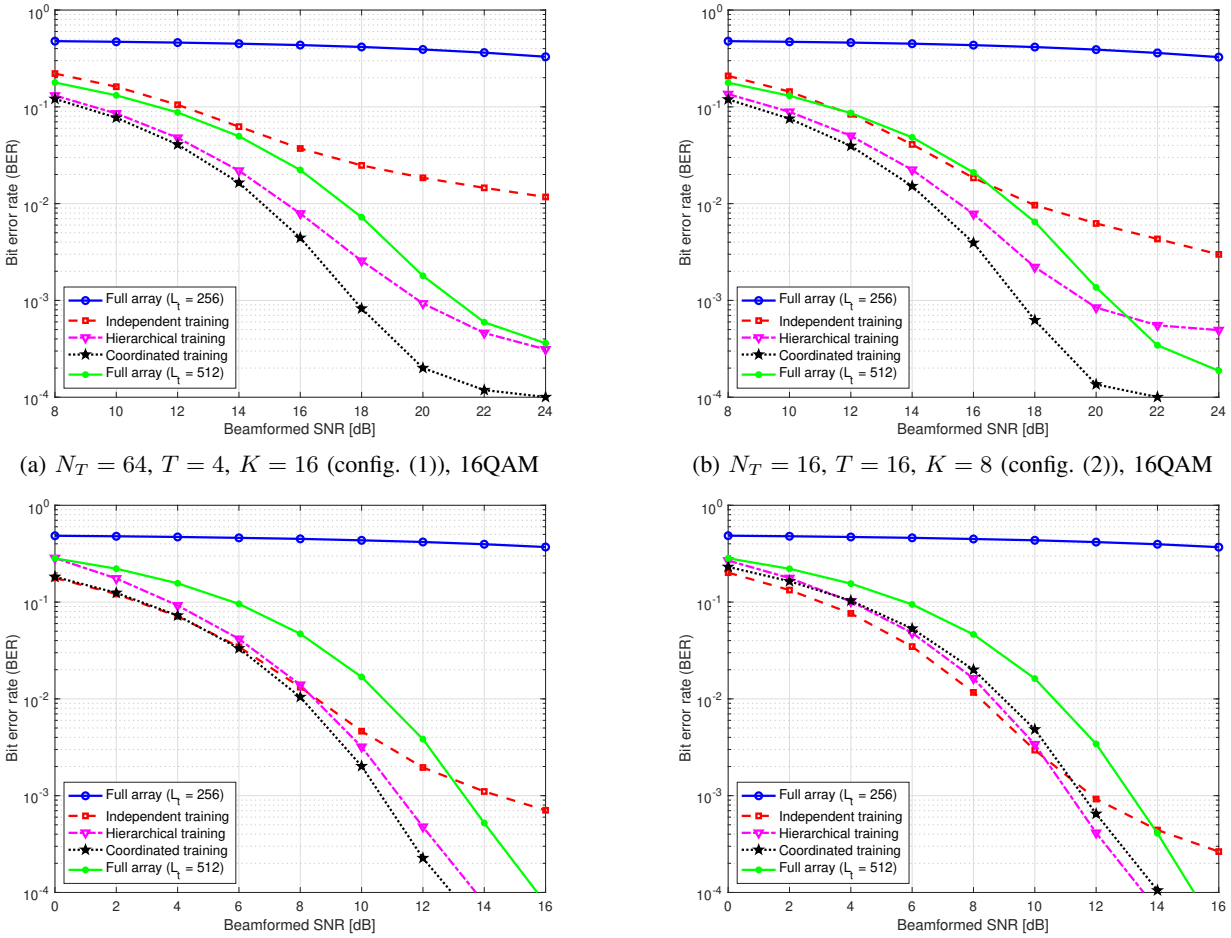
* Assuming LDL-based inversion, $h(N) = \frac{1}{2}N^3 + \frac{1}{2}N^2 - \frac{1}{6}N$.

A. Bit Error Rates

Fig. 4 shows BER plots for adaptive implementations of the 4 multiuser detection strategies described in Section III for 16QAM and QPSK modulation for the two tiled configurations. As expected, adaptation of the full-array with $L_t = N = 256$ training symbols yields unacceptable performance, but even the number of training symbols for full-array is doubled to $L_t = 2N = 512$, it performs worse than coordinated and hierarchical training using $L_t = 256$ symbols. The performance of hierarchical training is competitive with that of coordinated training despite its significantly lower training complexity, but it does incur a larger error floor for 16QAM due to residual interference. Both coordinated and hierarchical training significantly outperform independent training: the reduced dimension available for interference suppression causes residual interference which is large enough to incur error floors even for the smaller QPSK constellation. It is interesting to note, however, that there is one setting in which independent training performs better: QPSK at low beamformed SNR in configuration 2 (Fig. 4(d)), where noise averaging becomes more important than interference suppression.

B. Computational Complexity

Table I shows the computational complexity in terms of the number of multiply–accumulate (MAC) operations required for each scheme to support K simultaneous users. Assuming



(a) $N_T = 64$, $T = 4$, $K = 16$ (config. (1)), 16QAM

(b) $N_T = 16$, $T = 16$, $K = 8$ (config. (2)), 16QAM

(c) $N_T = 64$, $T = 4$, $K = 16$ (config. (1)), QPSK

(d) $N_T = 16$, $T = 16$, $K = 8$ (config. (2)), QPSK

Fig. 4: Uncoded Bit Error Rate (BER) plots. For tiled beamspace processing approaches, we use $L_t = 256$. For hierarchical training, we use $L_{t1} = L_{t2} = 128$, and $L_{t1} = 32, L_{t2} = 224$ for config. (1) and (2), respectively. We can see that tiled beamspace approaches have lower BER than the full array with longer L_t.

LDL decomposition-based inversion [11] - [12], $h(\cdot)$ can be defined as follows, $h(N) = \frac{1}{2}N^3 + \frac{1}{2}N^2 - \frac{1}{6}N$, which is used for inversion of $\mathbf{R}$.

For inference, tiled windowed beamspace architectures divide and parallelize the bulk of the processing across T tiles, with reduced per-user correlator dimension W at the cost of tile-level FFTs amortized across all users. Tiled windowed beamspace inference therefore offers an advantage over full-array processing when $KN > KTW + N\log_2 N_T$; this holds if K scales faster than $\log N_T$, which is satisfied by the typical linear scaling of K with N.

For training, tiled windowed beamspace yields dramatic reductions in complexity over the full-array benchmark. LMMSE adaptation requires computation of $\mathbf{R}$, $\mathbf{p}$, inversion of $\mathbf{R}$, and computing $\mathbf{c}$, with the matrix inversion step benefiting the most from the dimensionality reduction strategies we apply. Coordinated training requires more communication and computation overhead than independent and hierarchical training. To sum up, hierarchical training strikes a favorable balance between complexity and performance with minimal inter-tile communication. With a computational complexity of $O(TW^3 + T^3)$, hierarchical training is significantly lower in complexity than coordinated training ($O((TW)^3)$), and only marginally higher in complexity than independent training ($O(TW^3)$).

V. CONCLUSION

We have highlighted the promise of tiled beamspace architectures as a means of scaling massive MU-MIMO by exploiting the sparsity of mmWave channels. By utilizing in-place computation at the tiles, we parallelize computation while minimizing the communication overhead for data movement. For LMMSE interference suppression as considered here, the dimensionality reduction provided by our tiled beamspace approaches yield significant advantages in terms of both computational complexity and training overhead. This is illustrated in this initial investigation by considering a single $\lambda/2$-spaced linear array partitioned into tiles in order to provide a direct performance comparison with an antenna space benchmark. In addition to the substantial gains in computational complexity,

we show that our proposed approaches comprehensively outperform the antenna space benchmark when training overhead is limited. An important topic for future work is to explore the performance gains provided by the more flexible antenna geometries permitted by tiling. Another promising area of study involves leveraging the efficiencies inherent in tiled architectures to develop more efficient hardware systems.

ACKNOWLEDGMENT

This work was supported in part by the Center for Ubiquitous Connectivity (CUbiC), sponsored by Semiconductor Research Corporation (SRC) and Defense Advanced Research Projects Agency (DARPA) under the JUMP 2.0 program.

REFERENCES

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

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

[3] B. Sadhu *et al.*, "A 28-ghz 32-element trx phased-array ic with concurrent dual-polarized operation and orthogonal phase and gain control for 5g communications," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 12, pp. 3373–3391, 2017.

[4] A. A. Farid, A. S. H. Ahmed, A. Dhananjay, and M. J. W. Rodwell, "A fully packaged 135-ghz multiuser mimo transmitter array tile for wireless communications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 7, pp. 3396–3405, 2022.

[5] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[6] M. Abdelghany, U. Madhow, and A. Tölli, "Beamspace local lmmse: An efficient digital backend for mmwave massive mimo," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[7] M. Mahdavi, O. Edfors, V. Öwall, and L. Liu, "Angular-domain massive mimo detection: Algorithm, implementation, and design tradeoffs," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 6, pp. 1948–1961, 2020.

[8] T. Takahashi, A. Tölli, S. Ibi, and S. Sampei, "Low-complexity large mimo detection via layered belief propagation in beam domain," *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 234–249, 2021.

[9] J. D. M. Reed, Irving S. and L. E. Brennan, "Rapid convergence rate in adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems 6*, 1974.

[10] C. A. Balanis, *Antenna theory: analysis and design*. John wiley & sons, 2016.

[11] E. Bertilsson, C. Ingemarsson, and O. Gustafsson, "Low-latency parallel hermitian positive-definite matrix inversion for massive mimo," in *2021 IEEE Workshop on Signal Processing Systems (SiPS)*, 2021, pp. 23–28.

[12] W. Tang, C.-H. Chen, and Z. Zhang, "A 2.4-mm2 130-mw mmse-nonbinary ldpc iterative detector decoder for 4 $\times$ 4 256-qam mimo in 65-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 7, pp. 2070–2080, 2019.