THz-band, Tbps MIMO Communications: A Joint Data Detection and Decoding Framework

Hakim Jemaa, *Student Member, IEEE,* Hadi Sarieddeen, *Member, IEEE,* Mohamed-Slim Alouini, *Fellow, IEEE,* and Tareq Y. Al-Naffouri, *Senior Member, IEEE*

***Abstract*—Efficient data detection and decoding are addressed under terahertz (THz)-band channel conditions and terabit-per-second (Tbps) baseband processing constraints. We investigate the performance and complexity tradeoffs of candidate data detectors in correlated ultra-massive multiple-input multiple-output (UM-MIMO) THz channels. Under high correlation, channel-matrix puncturing in subspace detectors can significantly reduce computational complexity and introduce much-needed parallelizability. Simulation results demonstrate that subspace detectors outperform conventional detectors in typical line-of-sight-dominated THz channel conditions. We advocate for a joint data detection and decoding framework that does not parallelize channel-code decoders to satisfy stringent Tbps baseband constraints but parallelizes data sources through channel puncturing and adopts short codes instead.**

***Index Terms*—THz communications, ultra-massive MIMO, subspace detection, short codes, terabit-per-second.**

I. Introduction

The THz band between 0.1 THz and 10 THz is a critical spectrum enabler for sixth-generation (6G) wireless communication systems and beyond. Compared to millimeter-wave (mmWave) communications, THz communications promise higher data rates (Tbps), reduced latency, enhanced security, and better localization and sensing capabilities [1], [2]. Recent advances in electronic, photonic, and plasmonic (graphene-based) solutions are closing the gap in THz devices, emphasizing high electron tunability in reconfigurable THz systems. However, major challenges remain in designing efficient THz-specific signal processing techniques [3], [4]. With the fifth-generation (5G) only promising peak data rates of 20 gigabit-per-second, baseband signal processing should be hardware-friendly and low-complexity to bridge the gap between promised Tbps rates and state-of-the-art processor clock speeds. THz-band signal processing should account for constraints germane to THz device technologies, hardware impairments, and channel and noise modeling.

In this work, we consider using UM-MIMO technology as an infrastructure enabler that combats the large THz path and molecular absorption losses and extends the communication range [5], [6]. We focus on the corresponding UM-MIMO data detection and decoding problems that are most computationally demanding, exploring solutions that account for both THz channel characteristics and baseband processing limitations. We study the applicability of various data detectors for different THz channel conditions, highlighting the performance and complexity tradeoffs and the importance of parallelizability in Tbps operations. In particular, subspace detection algorithms [7], [8] can introduce favorable parallelizability alongside performance gains. We examine short-code-length channel codes to reduce the channel-code decoding complexity, leveraging the parallelizability introduced by channel puncturing. Lower case, bold lower case, and bold upper case letters correspond to scalars, vectors, and matrices, respectively. Scalar norms, vector norms, transpose, and conjugate transpose are denoted by $|\cdot|$, $\|\cdot\|$, $(\cdot)^T$ and $(\cdot)^H$, respectively.

Hakim Jemaa, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri are with the Department of Computer, Electrical and Mathematical Sciences and Engineering (CEMSE), King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia, 23955-6900 (hakim.jemaa@kaust.edu.sa; slim.alouini@kaust.edu.sa; tareq.alnaffouri@kaust.edu.sa). Hadi Sarieddeen is with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA (hadisari@mit.edu). The work was supported by the KAUST Office of Sponsored Research (OSR) under Award ORA-CRG2021-4695.

II. System and Channel Models

We consider a 3D UM-MIMO system with an array-of-subarrays (AoSA) architecture [3] consisting of $M^{(\mathrm{t})}\times N^{(\mathrm{t})}$ and $M^{(\mathrm{r})}\times N^{(\mathrm{r})}$ subarrays (SAs) at the transmitter (Tx) and receiver (Rx), respectively. Each SA consists of $Q\times Q$ antenna elements (AEs). The distances separating two Tx (or Rx) SAs is Δ. We assume beamforming at the AE level, where each SA generates a single beam. We consider a single-carrier frequency-flat channel for line-of-sight (LoS)-dominant THz systems, favoring low-complexity designs and reduced peak-to-average power ratios [3]. The complex baseband input-output relation is $\mathbf{y}=\mathbf{H}\mathbf{x}+\mathbf{n}$, where $\mathbf{x}=[x_1 x_2 \cdots x_{M_t N_t}]^T \in \mathcal{X}^{M_t N_t \times 1}$ is a vector of information symbols, each drawn from a quadrature amplitude modulation (QAM) constellation, $\mathcal{X}$, $\mathbf{y}\in C^{M_r N_r}$ is the received symbol vector, $\mathbf{H}=[\mathbf{h}_1\mathbf{h}_2\cdots\mathbf{h}_{M_t N_t}] \in C^{M_r N_r \times M_t N_t}$ is the channel matrix, and $\mathbf{n}\in C^{M_r N_r\times 1}$ is the additive white Gaussian noise of variance σ^2. We assume perfect channel state information at the Rx, which is reasonable in low-mobility THz scenarios. We adapt our TeraMIMO channel simulator [9], which accurately simulates stochastic THz channels from measurement-based models. The frequency-domain channel between the $q^{(\mathrm{t})}$th Tx and $q^{(\mathrm{r})}$th Rx SAs is

$$\begin{aligned} h_{q^{(\mathrm{r})},q^{(\mathrm{t})}} &= \alpha^{\mathrm{LoS}} G^{(\mathrm{t})}(\mathbf{\Phi}^{(\mathrm{t})}) G^{(\mathrm{r})}(\mathbf{\Phi}^{(\mathrm{r})}) \\ &\quad \times \mathcal{A}^{(\mathrm{r})}_{eq}(\mathbf{\Phi}^{(\mathrm{r})}, \mathbf{\Phi}^{(\mathrm{r})}_0) \mathcal{A}^{(\mathrm{t})^{\mathrm{Tr}}}_{eq}(\mathbf{\Phi}^{(\mathrm{t})}, \mathbf{\Phi}^{(\mathrm{t})}_0) e^{-j2\pi f \tau^{\mathrm{LoS}}} \\ &\quad + \sum_{c=1}^{N_{\mathrm{clu}}} \sum_{\ell=1}^{N^c_{\mathrm{ray}}} \alpha^{\mathrm{NLoS}}_{c,\ell} G^{(\mathrm{t})}(\mathbf{\Phi}^{(\mathrm{t})}_{c,\ell}) G^{(\mathrm{r})}(\mathbf{\Phi}^{(\mathrm{r})}_{c,\ell}) \\ &\quad \times \mathcal{A}^{(\mathrm{r})}_{eq}(\mathbf{\Phi}^{(\mathrm{r})}_{c,\ell}, \mathbf{\Phi}^{(\mathrm{r})}_0) \mathcal{A}^{(\mathrm{t})^{\mathrm{Tr}}}_{eq}(\mathbf{\Phi}^{(\mathrm{t})}_{c,\ell}, \mathbf{\Phi}^{(\mathrm{t})}_0) e^{-j2\pi f \tau^{\mathrm{NLoS}}_{c,\ell}}, \end{aligned} \tag{1}$$

where N_{clu} and N_{ray}^{c} are the number of clusters and number of rays in the cth cluster, $\alpha_{q^{(\text{r})},q^{(\text{t})}}^{\text{LoS}}$ and $\alpha_{c,\ell}^{\text{NLoS}}$ are the LoS and NLoS path gains (including molecular absorption), $G^{(\text{t})}$ and $G^{(\text{r})}$ are the Tx and Rx antenna gains, and $\mathbf{a}^{(\text{t})} \in C^{Q^{(\text{t})}\times 1}$ and $\mathbf{a}^{(\text{r})} \in C^{Q^{(\text{r})}\times 1}$ are the Tx and Rx beamsteering vectors. The vectors $\mathbf{\Phi}^{(\text{t})} = [\phi^{(\text{t})}, \theta^{(\text{t})}]^{\text{Tr}}$/$\mathbf{\Phi}^{(\text{r})} = [\phi^{(\text{r})}, \theta^{(\text{r})}]^{\text{Tr}}$ and $\mathbf{\Phi}_{c,\ell}^{(\text{t})} = [\phi_{c,\ell}^{(\text{t})}, \theta_{c,\ell}^{(\text{t})}]^{\text{Tr}}$/$\mathbf{\Phi}_{c,\ell}^{(\text{r})} = [\phi_{c,\ell}^{(\text{r})}, \theta_{c,\ell}^{(\text{r})}]^{\text{Tr}}$ are the LoS and ℓth-ray cth-cluster NLoS angles of departure/arrival, and $\tau^{\text{LoS}} = d/c$ and $\tau_{c,\ell}^{\text{NLoS}}$ are the the LoS and NLoS ray times of arrival; f is the operating frequency and c is the speed of light in vacuum. $\mathcal{A}_{eq}^{(\text{t})}$ and $\mathcal{A}_{eq}^{(\text{r})}$ are the equivalent array responses [9].

The relation between the THz channel conditions and design compactness is studied in [5]; near-orthogonal LoS channels are retained at an optimal SA separation, $\Delta_{\text{opt}} = \sqrt{zdc/Mf}$, for odd z and symmetric arrays of dimension M. We differentiate between orthogonal and non-orthogonal LoS, depending on whether a spatial tuning of Δ is performed in a reconfigurable design or not. Our system model also accounts for THz-specific characteristics such as spherical wave models, misalignment, and beam split [9]. We consider multiple THz channel scenarios, each incurring different data detection and decoding challenges. In addition to short-distance, LoS links, we consider indoor environments [10] that can sustain multipath components and a data center scenario [11]. High rates, low latency, and high eavesdropping resiliency are favorable THz features for complex high-traffic data-center connectivity.

III. Problem Formulation and Data Detectors

Low-complexity linear detectors perform well in conventional low-frequency massive MIMO systems because of favorable channel hardening resulting from many antennas at the base station and a few at the user equipment. However, the ill-conditioned doubly-massive UM-MIMO THz channels require near-optimal non-linear detection algorithms. The stringent Tbps processing constraints further necessitate low-complexity solutions that consider the underlying hardware implementation. A popular approach is to limit the choice of candidate information symbol vectors in non-linear detectors from a complete lattice to a reduced space.

A. Reference Detectors

The optimal maximum likelihood (ML) detector searches the entire lattice, $\bar{\mathcal{X}} = \mathcal{X}^{M_t N_t}$, for $\mathbf{x}^{\text{ML}} = \arg\min_{\mathbf{x}\in\bar{\mathcal{X}}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$. Furthermore, enhanced bit decoding performance can be achieved by leveraging soft-output (SO) detection information in the form of a posteriori log-likelihood ratios (LLRs) of detected bits. Let $\mathbf{c}$ be the bit representation of $\mathbf{x}$; $c_{i,m}$ is the ith bit of the mth symbol. The LLR, $\gamma_{i,m}$, of $c_{i,m}$ is defined as

$$\gamma\left(c_{i,m}\middle|\,\mathbf{y}\right) \triangleq \gamma_{i,m} \triangleq \ln\left(\text{P}\left[c_{i,m}=+1|\mathbf{y},\mathbf{H}\right]/\text{P}\left[c_{i,m}=-1|\mathbf{y},\mathbf{H}\right]\right), \tag{2}$$

where P is the probability function; $c_{i,m}=0$ and $c_{i,m}=1$ are assumed to be equally likely. Assuming optimality in the max-log sense (because true LLR computations are exhaustive),

$$\gamma_{i,m}^{\text{ML}} = \begin{cases} d^{\text{ML}} - d_{i,m}^{\overline{\text{ML}}} & \text{if } c_{i,m}^{\text{ML}} = 0 \\ d_{i,m}^{\overline{\text{ML}}} - d^{\text{ML}} & \text{if } c_{i,m}^{\text{ML}} = 1, \end{cases} \tag{3}$$

where the ML and counter ML distances are, respectively,

$$d^{\text{ML}} = \|\mathbf{y} - \mathbf{H}\mathbf{x}^{\text{ML}}\|^2 \text{ and } d_{i,m}^{\overline{\text{ML}}} = \min_{\mathbf{x}\in\bar{\mathcal{X}}_{j,b}^{\left(c_{i,m}^{\overline{\text{ML}}}\right)}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2, \tag{4}$$

with $\bar{\mathcal{X}}_{i,m}^{(0)} = \left\{\mathbf{x}\in\bar{\mathcal{X}}\colon c_{i,m}=0\right\}$ and $\bar{\mathcal{X}}_{i,m}^{(1)} = \left\{\mathbf{x}\in\bar{\mathcal{X}}\colon c_{i,m}=1\right\}$; $c_{i,m}^{\overline{\text{ML}}}$ is the counter-ML bit value.

The exhaustive search in ML detection incurs a prohibitive exponential complexity. Conversely, a linear zero-forcing (ZF) detector multiplies the received symbol vector with the pseudo-inverse of the channel, $\mathbf{y}^{\text{ZF}} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{y}$, to decouple symbols and slice them individually on their respective constellations. The ZF SOs are computed as

$$\gamma_{i,m}^{\text{ZF}} = \frac{1}{\sigma^2}\left(\min_{x_i\in\mathcal{X}_{i,m}^{(1)}} \left|\hat{y}_i^{\text{ZF}} - x_i\right|^2 - \min_{x_i\in\mathcal{X}_{i,m}^{(0)}} \left|\hat{y}_i^{\text{ZF}} - x_i\right|^2\right), \tag{5}$$

where $\mathcal{X}_{i,m}^{(0)} = \left\{x\in\mathcal{X}\colon c_{i,m}=0\right\}$ and $\mathcal{X}_{i,m}^{(1)} = \left\{x\in\mathcal{X}\colon c_{i,m}=1\right\}$. Nevertheless, large matrix inversions can still be complex.

Sphere decoders (SDs) transform the ML detection problem into a tree search over which branch pruning reduces complexity. Another approach is lattice reduction (LR)-aided detection. LR transforms the detection problem into a closest vector search problem over an infinite lattice of a near-orthogonal basis. Combined with linear detectors (ZF), LR significantly enhances performance. We adopt the complex lattice reduction (CLLL) algorithm [12] and produce, using Gram-Schmidt orthogonalization, an orthogonal set of vectors that extends the same space as $\{\mathbf{h}_1, \ldots, \mathbf{h}_n\}$. The output LLRs are thus $\gamma_{i,m}^{\text{LR}} = \gamma\left(c_{i,m}\middle|\,\mathbf{y}^{\text{LR}}, \mathbf{H}^{\text{LR}}\right)$. Imperfect basis orthogonalization and nearest neighbor quantization degrade performance; generating SOs following LR is also not straightforward.

B. Subspace Detectors

We study subspace detectors [7] that transform channel matrices into target structures through controlled puncturing. The resultant implementation of search-based detectors over such structures can be parallelized in hardware. An instance of subspace detectors, the layered orthogonal lattice detector (LORD), consists of preprocessing QR decomposition (QRD) and search stages, repeated for different channel column permutations, $\Pi(t)$. For every iteration t, $\mathbf{Q}(t)\mathbf{R}(t) = \mathbf{H}\Pi(t)$, where $\mathbf{Q}$ is a unitary matrix, $\mathbf{R} = [r_{i,j}]$ is upper triangular, and $\mathbf{H}\Pi(t) = \left[\mathbf{h}_1 \ldots \mathbf{h}_{t-1}\mathbf{h}_{t+1} \ldots \mathbf{h}_{N_t M_t}\mathbf{h}_t\right]$. The permutation is also applied to $\mathbf{x}$. We can rewrite the received vector as $\tilde{\mathbf{y}}(t) = \mathbf{Q}(t)^H\mathbf{y} = \mathbf{R}(t)\mathbf{x}\Pi(t) + \mathbf{Q}(t)^H\mathbf{n} = \mathbf{R}(t)\tilde{\mathbf{x}}(t) + \tilde{\mathbf{n}}(t)$. After preprocessing, LORD searches symbols at the root layer of $\tilde{\mathbf{x}}(t)$, $\tilde{x}_{M^{(\text{t})}N^{(\text{t})}}$, and back-substitutes to obtain a candidate $\hat{\mathbf{x}}$:

$$\hat{x}_m = \begin{cases} \tilde{x}_{M_t N_t}, & m = M_t N_t \\ \arg\min\limits_{x\in\mathcal{X}} \left|\tilde{y}_m - r_{m,m}x - \sum_{e=m+1}^{M_t N_t} r_{m,e}\hat{x}_e\right|, & m < M_t N_t. \end{cases} \tag{6}$$

The $M_t N_t$ different sets, $\mathcal{S}_t^{\text{LORD}}$, of $|\mathcal{X}|$ candidate vectors each, form a reduced search set for SO computations:

$$\gamma_{i,m}^{\text{LORD}} = \frac{1}{\sigma^2}\left(\min_{x_i\in\mathcal{S}_{t,i,m}^{(1)}} \|\tilde{\mathbf{y}}(t) - \mathbf{R}(t)\mathbf{x}\|^2 - \min_{x\in\mathcal{S}_{t,i,m}^{(0)}} \|\tilde{\mathbf{y}}(t) - \mathbf{R}(t)\mathbf{x}\|^2\right). \tag{7}$$

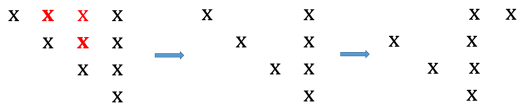


Fig. 1. Puncturing and matrix permutation in a 4x4 MIMO structure.

Populating the mth set (for $t = m$) is crucial for computing the LLRs of bits corresponding to the mth symbol. However, hard-output LORD only requires a single decomposition and set of candidate vectors, where $\mathbf{x}^{\text{LORD}} = \underset{\tilde{\mathbf{x}} \in \mathcal{S}_1^{\text{LORD}}}{\arg\min} \|\mathbf{y} - \mathbf{H}\tilde{\mathbf{x}}\|^2$.

C. Channel Puncturing

Aiming to realize more efficient and better parallelizabile detectors, QRD is replaced by a more generic WR decomposition (WRD). WRD transforms $\mathbf{H}$ into a punctured upper triangular matrix $\dot{\mathbf{R}}$, where $\mathbf{H} = \mathbf{W}\dot{\mathbf{R}}$, and $\mathbf{W}$ can be efficiently computed via QRD followed by elementary matrix computations [7]. The transformed system model is $\tilde{\mathbf{y}} = \mathbf{W}^H\mathbf{y} = \dot{\mathbf{R}}\mathbf{x} + \mathbf{W}^H\mathbf{n}$. The corresponding subspace detection (SSD) routine then cyclically shifts the columns of $\mathbf{H}$, generating the corresponding punctured $\dot{\mathbf{R}}$ for each of the $M^{(\mathrm{t})} \times N^{(\mathrm{t})}$ steps. Figure 1 illustrates the puncturing and permutation operations. For the tth WRD, the SOs are:

$$\gamma_{i,m}^{\text{SSD}} = \frac{1}{\sigma^2}\left(\min_{x_i \in \mathcal{S}_{t,i,m}^{(1)}} \|\tilde{\mathbf{y}}(t) - \dot{\mathbf{R}}(t)\mathbf{x}\|^2 - \min_{x \in \mathcal{S}_{t,i,m}^{(0)}} \|\tilde{\mathbf{y}}(t) - \dot{\mathbf{R}}(t)\mathbf{x}\|^2 \right). \tag{8}$$

With LORD, $\mathbf{Q}$ is unitary, and the distance computations in (7) are over the original vector space. Therefore, LLRs can be improved by populating global distance metrics across permutations. To the contrary, $\mathbf{W}$ is not unitary, and comparing distances computed from different SSD decompositions is not useful. The LLR computations per symbol can thus be fully parallelized (with an order equal to the UM-MIMO dimension), a critical feature for achieving Tbps rates.

IV. Simulation Results and Discussion

We investigate the bit-error-rate (BER) performance of the studied schemes as a function of the signal-to-noise ratio (SNR). Figure 2 illustrates the performance in data-center and indoor THz environments, with 64x64 MIMO and 4 QAM. Lower modulation orders are favored for complexity reduction in UM-MIMO systems. The data-center channel has fewer multipath components; the distribution of the number of clusters and rays and other parameters are highlighted in Table I. In both the indoor multipath and data-center channels, LORD outperforms LR and ZF (10 dB gain compared to LR). ZF performs badly with high channel correlation.

With spatial tuning in short-distance LoS channels, the channel is orthogonal, and LR and ZF would be equally optimal. However, the LR performance deteriorates with imperfect orthogonalization, especially since the method we adopted for SO computations in LR-aided detectors is not optimal. Therefore, under near-orthogonality in LoS THz channels, ZF is sufficient to decouple the interfering symbols.

TABLE I
Main Channel Model Parameters

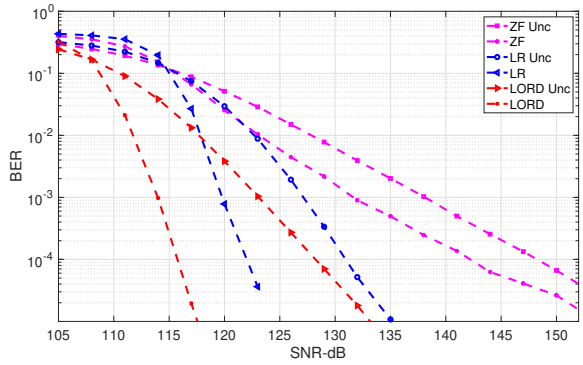
Parameter\channel	Data center	Indoor multipath	LoS
Frequency (THz)	0.3	0.3	1
Bandwidth (THz)	0.02	0.025	0
Distance (m)	2	1.75	2
Clusters	3	Poisson	-
Rays	4	Poisson	-

The very high SNR values in the BER curves' slope region are caused by the very high path loss at THz frequencies. This effect can be remedied by increasing the antenna gains and introducing beamforming gains. However, the antennas' behavior and their architectures at THz frequencies are yet to be perfectly known. Much lower SNR ranges can be achieved by packing thousands of AEs within a SA to achieve a target beamforming gain, as shown in Fig. 3.

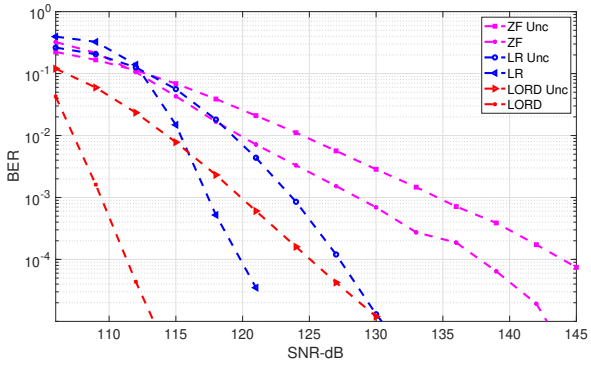
The BER performance using polar and low-density parity-check (LDPC) codes in indoor multipath channels are illustrated in Fig. 3. For Polar codes, we consider a 164-bit code length; for other code lengths, we rate-match the encoded message to a length equal to that of LDPC encoding. We use cyclic redundancy check (CRC)-aided successive cancellation list (SCL) decoding with a list size of 8 and a CRC of 24 bits; polar coding without CRC suffers in performance. For similar code rates, the performance of the studied detectors under polar and LDPC coding is similar to that with turbo coding. However, the results here do not accurately capture the performance of coding schemes, mainly because the code lengths and rates are not perfectly matched. Subspace detection is shown to achieve near-optimal performance at a significantly lower complexity.

As this work motivates introducing high parallelizability through subspace detection, a performance comparison between channel-punctured and non-punctured detection schemes is illustrated in Fig. 4. For a non-orthogonal LoS channel, Fig. 4-a compares SSD, LORD, and LR-aided detection with 16x16 MIMO and 64x64 MIMO, 16 QAM and 64 QAM. The higher the MIMO order, the higher the correlation effect and the more SSD outperforms LORD (an average gain of 5dB). This observation is very promising and highlights a win-win situation. Not only do we profit from complexity reduction and high parallelizability with SSD, but we achieve near-ML performance. At lower frequencies, puncturing in SSD would typically incur a performance degradation cost [7]; with ill-conditioned THz channels, puncturing in SSD reduces channel correlation and enhances performance.

Finally, we investigate how tuning with the coding scheme parameters affects the detectors' performance. Figure 4-b compares SSD and LR for 4x4 MIMO and 16x16 MIMO, 4 QAM and 16 QAM, and code lengths of 256 bits and 64 bits. Using smaller code chunks degrades the performance by 2 dB for 4x4 MIMO and around 10 dB for 16x16 MIMO. However, smaller code lengths allow spatial parallelism and further complexity reduction, which is crucial for Tbps operations.



(a) Data-center channel.



(b) Indoor multipath channel.

Fig. 2. Uncoded and turbo-coded BER performance of the studied detectors in 64x64 MIMO systems using 4QAM.

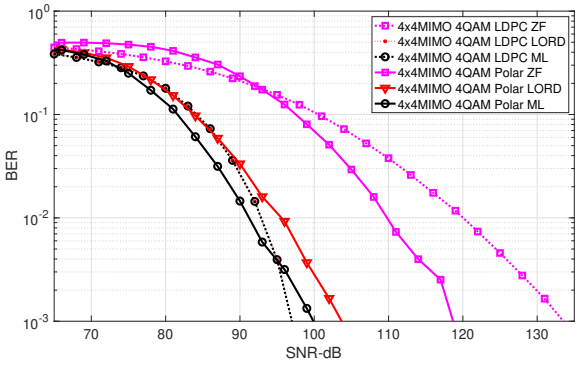


Fig. 3. BER performance with polar and LDPC channel coding in indoor multipath environments (using 100 AEs per SA).

V. Complexity Analysis

A. Data Detection Complexity

ML detection requires an exponential number of Euclidean distance computations, $|\mathcal{X}|^{M_t N_t}$. In contrast, ZF only requires $O(M_t N_t\,|\mathcal{X}|)$ floating-point operations (flops). For SD, the complexity is determined by the number of visited nodes within a tree search. Fixed-radius sphere decoding [13] would be favored in a THz/Tbps scenario because it has a fixed complexity. Nevertheless, parallelizability is crucial for a Tbps baseband, and tree structures are not easily parallelized unless flattened and unrolled. For LR-based detection, the preprocessing stage incurs an $O((M_t N_t)(M_r N_r)^3 \log(M_r N_r))$ complexity on top of ZF. Such preprocessing is not easily parallelized, but needs to be computed once per frame.

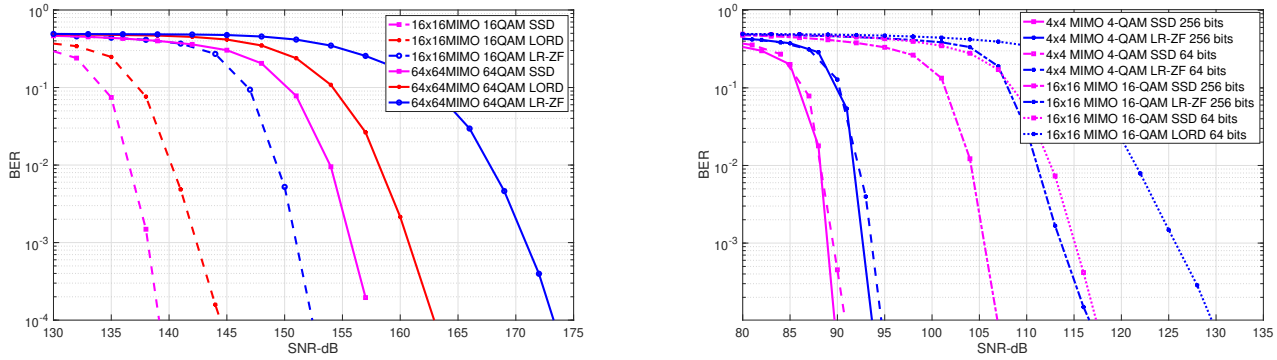
Subspace detectors can be highly parallelized. The complexity of near-optimal LORD is $O((M_r N_r)^2(|\mathcal{X}|))$ (much less than ML). Furthermore, through subspace matrix puncturing, the complexity can be reduced to approach that of ZF. The additional pre-processing cost can be marginal in near-static environments (typical for THz applications), where matrix decompositions can be computed once and stored in memory. In LORD, the procedures following all matrix permutations can be combined to compute the LLRs for the bits of a specific symbol. However, such procedures can run in parallel with SSD [7]. Hence, given a sufficient chip size that can support parallel processing, the latency in a 1000×1000 UM-MIMO SSD can be reduced 1000 times with channel puncturing.

B. Channel Coding Complexity

Communication system design faces implementation bottlenecks in computational complexity, chip area, energy efficiency, and algorithm parallelizability. The EPIC project guidelines [14] for practical deployment of Tbps data rates translates into one pJ/bit energy efficiency, 100 GHz/sec/mm^2 area efficiency, and 0.1 W/mm^2 power density. Improvements in silicon and hardware are insufficient for THz/Tbps communication requirements. Therefore, efficient, parallelizable, and low-complexity data detection schemes should be jointly optimized with efficient channel-code decoding.

High throughput, high power/energy efficiency, and low latency require highly parallelizable architectures that guarantee large locality and minimum control of data flow. Turbo and LDPC schemes are data-flow-based, incorporating limited locality with randomness via interleavers and tanner graphs, respectively. The coding/decoding complexity (in flops) is code-length- and code-rate-dependant [15]. Thus, variable code lengths and rates are crucial to support different THz use cases. Shorter code lengths reduce complexity, lower the energy cost of memory usage, and enhance parallelizability.

Data dependency in iterative decoding hinders parallelism. For instance, the different trellis steps cannot be calculated in parallel in maximum a posteriori turbo decoding because of the structure of interleavers. By decomposing the decoding process into sub-functions, spatial parallelism can be introduced. For example, the variable and check node computations can be treated independently in the belief propagation algorithm. Achieving high throughput is also bounded by locality and regularity, where minimizing power consumption and interconnectivity reduces energy and area consumption (data transfers dominate the belief propagation algorithm). The throughput is affected by multiple practical and theoretical limitations; it can be expressed as a multiparameter variable that depends on factors such as the feasible parallelism capacity in an architecture, the timing overhead due to interconnect and memory access conflicts, and the number of decoding iterations.



(a) Punctured versus non-punctured detectors in LoS channels.

(b) Code length variation with data-center channels.

Fig. 4. Channel-punctured versus channel-unpunctured performance for different code lengths.

The conventional decoding algorithms significantly differ for the three considered coding schemes; thus, the challenges will also be different. Turbo codes require low-complexity decoding algorithms to boost implementation efficiency further, whereas the throughput of LDPC decoding is strongly dominated by the interconnect complexity inherent to the Tanner graph structure. All levels of parallelism must be efficiently exploited to reach Tbps throughput, and pipelining is mandatory to break up data dependencies. Pipelined architectures require a huge number of registers, and the clock tree becomes a bottleneck in clock synchronization and power consumption. This concern is mainly raised for high-throughput polar decoders.

Instead of parallelizing the decoder, we can parallelize the source and use short codes per source stream. In the particular case of subspace detection, source parallelizability can be introduced in the spatial domain, with each antenna being independently fed short code words. Unconventional universal decoders can complement such source parallelizability. In particular, the recently proposed guessing random additive noise decoding (GRAND) [16], [17] can be used to decode short codes from arbitrary code-book structures. GRAND is particularly efficient with high-rate block-code constructions.

VI. Conclusion

In this paper, we studied the performance and complexity trade-offs in designing UM-MIMO data detection and decoding schemes that account for the THz channels' peculiarities and the baseband constraints if Tbps rates are to be supported. We described channel models for practical THz applications and motivated channel puncturing as a means for near-optimal, low-complexity THz-band data detection. Instead of parallelizing the decoder, we proposed parallelizing the source through puncturing and using short codes. We finally linked our results to a complexity analysis, concluding that the future of THz-band communications will rely on how we approach different trade-offs under power and budget limitations.

References

[1] H. Sarieddeen, N. Saeed, T. Y. Al-Naffouri, and M.-S. Alouini, "Next generation terahertz communications: A rendezvous of sensing, imaging, and localization," *IEEE Commun. Mag.*, May 2020.

[2] I. F. Akyildiz, J. M. Jornet, and C. Han, "Terahertz band: Next frontier for wireless communications," *Physical Communication*, vol. 12, pp. 16–32, Sep. 2014.

[3] H. Sarieddeen, M.-S. Alouini, and T. Y. Al-Naffouri, "An overview of signal processing techniques for terahertz communications," *Proceedings of the IEEE*, vol. 109, no. 10, pp. 1628–1665, 2021.

[4] N. Rajatheva, I. Atzeni, S. Bicais, E. Bjornson, A. Bourdoux, S. Buzzi, C. D'Andrea, J.-B. Dore, S. Erkucuk, M. Fuentes *et al.*, "Scoring the terabit/s goal: Broadband connectivity in 6G," *arXiv preprint arXiv:2008.07220*, 2020.

[5] H. Sarieddeen, M. Alouini, and T. Y. Al-Naffouri, "Terahertz-band ultra-massive spatial modulation MIMO," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 9, pp. 2040–2052, Sep. 2019.

[6] A. Faisal, H. Sarieddeen, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Ultramassive MIMO systems at terahertz bands: Prospects and challenges," *IEEE Vehicular Technology Magazine*, vol. 15, no. 4, pp. 33–42, 2020.

[7] H. Sarieddeen, M. M. Mansour, and A. Chehab, "Large MIMO detection schemes based on channel puncturing: Performance and complexity analysis," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2421–2436, Jun. 2018.

[8] M. M. Mansour, "A near-ML MIMO subspace detection algorithm," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 408–412, Apr. 2015.

[9] S. Tarboush, H. Sarieddeen, H. Chen, M. H. Loukil, H. Jemaa, M.-S. Alouini, and T. Y. Al-Naffouri, "TeraMIMO: A channel simulator for wideband ultra-massive MIMO terahertz communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 12, pp. 12 325–12 341, 2021.

[10] S. Priebe and T. Kurner, "Stochastic modeling of THz indoor radio channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4445–4455, Sep. 2013.

[11] C. Cheng, S. Sangodoyin, and A. Zajić, "THz cluster-based modeling and propagation characterization in a data center environment," *IEEE Access*, vol. 8, pp. 56 544–56 558, 2020.

[12] Y. H. Gan, C. Ling, and W. H. Mow, "Complex lattice reduction algorithm for low-complexity full-diversity mimo detection," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2701–2710, 2009.

[13] L. G. Barbero and J. S. Thompson, "Fixing the complexity of the sphere decoder for MIMO detection," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2131–2142, Jun. 2008.

[14] *Next-Generation Channel Coding Towards Terabit/s Wireless Communications*. Zenodo, Aug. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1346686

[15] A. E. S. Hassan, M. Dessouky, A. Abou Elazm, and M. Shokair, "Evaluation of complexity versus performance for turbo code and LDPC under different code rates," *Proc. SPACOMM*, pp. 93–98, 2012.

[16] K. R. Duffy, J. Li, and M. Médard, "Capacity-achieving guessing random additive noise decoding," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4023–4040, 2019.

[17] H. Sarieddeen, M. Médard, K. Duffy *et al.*, "Grand for fading channels using pseudo-soft information," in *Proc. IEEE Global Conmun. Conf. (GLOBECOM)*, 2022, pp. 1–6.