Decentralized Likelihood Ascent Search-Aided Detection For Distributed Large-Scale MIMO Systems

Qiqiang Chen, Zheng Wang, *Senior Member, IEEE*, Chenhao Qi, *Senior Member, IEEE*, Zhen Gao, *Member, IEEE*, Yongming Huang, *Fellow, IEEE*, and Dusit Niyato, *Fellow, IEEE*

***Abstract*—In this paper, we propose the decentralized likelihood ascent search (DLAS)-aided detection for the distributed large-scale multiple-input multiple-output (MIMO) systems to achieve more remarkable performance gains. With the help of DLAS, traditional distributed iterative methods are able to achieve better performance than the linear detection schemes such as ZF and MMSE. According to analysis, we derive the equivalent noise and the post-processing SNR for DLAS. More importantly, based on them, we demonstrate that the proposed DLAS-aided detection achieves the full received diversity. To further facilitate its implementation in practice, we design the decentralized effective ring (DER) architecture with significantly reduced bandwidth requirement and better parallel computation. Finally, simulation results demonstrate that the proposed DLAS-aided detection attains the same received diversity as ML detection while surpassing state-of-the-art decentralized schemes in terms of BER performance, with reduced complexity and bandwidth costs.**

***Index Terms*—Large-scale MIMO, distributed MIMO detection, decentralized signal detection, likelihood ascent search.**

I. Introduction

THE large-scale multiple-input multiple-output (MIMO) system has become a core technology for enabling beyond fifth-generation (5G) and sixth-generation (6G) wireless communications, due to its promising capacity, ultra-fast data rate, and high energy efficiency [1]–[3]. The fundamental concept of large-scale MIMO involves equipping base station (BS) with hundreds or even thousands of antennas, which provides an effective approach to simultaneously serve a large number of users within the same time-frequency resource [4]. However, the rapid increase in the number of antennas also introduces some pressing challenges in transferring extensive volumes of raw data for signal processing, despite state-of-the-art hardware capabilities [5]. For instance, a large-scale MIMO system with 256 antennas, an 80MHz sampling rate, and a 12-bit analog-to-digital converter (ADC) generates raw baseband data rates in excess of 1Tbit/s, which surpasses the bandwidth capabilities of the existing high-speed interconnects like the common public radio interface (CPRI) [6]. Additionally, escalating computational complexity and storage requirements further strain single computing fabrics, making them inadequate for practical deployment. To address these challenges, numerous of distributed detection schemes for large-scale MIMO systems have been proposed [6]–[18].

In particular, the decentralized baseband processing (DBP) architecture was firstly introduced in [6] to alleviate the bottlenecks in centralized detection. DBP divides the BS antennas into C individual distributed units (DUs), where each DU contains B antennas and is equipped with an independent computing fabric. Based on DBP, techniques such as conjugate gradient (CG) [6] and alternating direction method of multipliers (ADMM) [7] are applicable for distributed detection in large-scale MIMO systems. Along iterations, these methods gradually approach the performance of centralized minimum mean square error (MMSE) detection with reduced bandwidth overhead. For complexity reduction, the decentralized Newton (DN) [8] and coordinate descent (CD) [9] algorithms have been proposed. Compared to the parallel implementation in DBP, daisy-chain architecture provides an alternative solution for the distributed detection with a pipelined design [10]–[12]. Based on it, the stochastic gradient descent (SGD) [10], averaged stochastic gradient descent (ASGD) [11], and general recursive least square (GRLS) [12] algorithms can be implemented in a distributed manner. However, all these algorithms only achieve performance comparable to that of centralized MMSE detection and are therefore limited by its performance constraints [12]. In fact, it is well known that the optimal maximum likelihood (ML) detection performance can only be approximated by MMSE when the number of antennas at a BS is sufficiently greater than that at the transmitter side, i.e., $N_r \gg N_t$. Unfortunately, such a scenario is not always met in practice, resulting in a considerable performance gap in distributed detection [13], [14]. For better performance, decentralized nonlinear detection schemes such as large-MIMO approximate message passing (LAMA) [15], Gaussian message passing (GMP) [16], and expectation propagation algorithm (EPA) [17] have been proposed at the expense of computational complexity. However, the comprehensive performance analysis with respect to these schemes is not given, which is important to evaluate their performance gains.

On the other hand, the neighborhood search algorithms have

This work was supported in part by the National Natural Science Foundation of China under Grant 62371124, 62225107, 61720106003, 62071116, 62471036, U2233216, in part by the Fundamental Research Funds for the Central Universities under Grant 2242022k60002, in part by Shandong Province Natural Science Foundation under Grant ZR2022YQ62, in part by Beijing Nova Program, in part by Beijing Natural Science Foundation under Grant L242011. (Corresponding author: Zheng Wang, Yongming Huang.)

Qiqiang Chen, Zheng Wang, Chenhao Qi and Yongming Huang are with School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: q.chen@seu.edu.cn; wznuaa@gmail.com; qch@seu.edu.cn; huangym@seu.edu.cn).

Zhen Gao is with the Advanced Technology Research Institute, BIT (Jinan), Jinan 250307 China, also with the BIT (Zhuhai), Zhuhai 519088, and also with the Yangtze Delta Region Academy, BIT (Jiaxing), Jiaxing 314019, China (e-mail: gaozhen16@bit.edu.cn).

Dusit Niyato is with the Nanyang Technological University, Singapore (e-mail: dniyato@ntu.edu.sg).

TABLE I
A BRIEF COMPARISONS OF THE RELATED LITERATURE OF DECENTRALIZED ALGORITHMS

Decentralized Algorithms	Architecture	Computational Complexity	Data Bandwidth	Nonlinear Performance	Diversity Order
CG [6]	DBP	$O(K^2T_{\max})$	$O(KT_{\max})$		
ADMM [7]	DBP	$O(K^2T_{\max})$	$O(KT_{\max})$		
DN [8]	DBP and Ring[1]	$O(K^2T_{\max})$	$O(KT_{\max})$		
CD [9]	DBP	$O(K^2I_{\max})$	$O(K)$		
SGD [10]	Daisy-chain	$O(KN)$	$O(K)$		
ASGD [11]	Daisy-chain	$O(KN)$	$O(K)$		
GRLS [12]	Daisy-chain	$O(KN)$	$O(K)$		
LAMA [15]	FD[2]	$O(KN\mathcal{O}I_{\max})$	$O(K)$	✓	
GMP [16]	FD	$O(KN\mathcal{O}I_{\max})$	$O(K)$	✓	
EPA [17]	DBP	$O(KN\mathcal{O}I_{\max}T_{\max})$	$O(KT_{\max})$	✓	
DLAS (This work)	DER	$O(K^2)$	$O(K)$	✓	✓

[1] The Ring architecture extends the daisy-chain structure by connecting the last DU back to the first, forming a closed loop.
[2] The fully decentralized (FD) architecture builds on DBP but maintains only an unidirectional link from DUs to the CPU.

emerged as an effective way to provide significant gains over the linear MMSE detection with low complexity cost [20]–[30]. Among these, likelihood ascent search (LAS) [20] is the first reported algorithm that starts with an initial estimate and then searches for the optimal solution in the neighboring set by minimizing the ML cost. The one symbol LAS (1-LAS) in [21] identifies the optimal vector in a neighborhood set that differs from the previous solution by only one symbol. For better BER performance, methods such as multistage LAS (MLAS) [21], reactive tabu search (RTS) [22], layered tabu search (LTS) [23], multiple initial vectors LAS (MIV-LAS) [24], grouped genetic algorithm LAS (GGA-LAS) [25], and unconstrained LAS (ULAS) [26] have been proposed to avoid the obstacles of local minimum. To further reduce the complexity, channel hardening [27], hopfield neural network (HNN) [28], and reduced neighborhood [29], [30] techniques have also been incorporated into LAS. However, all of these neighborhood search algorithms persist in their search until the ML cost is no longer reduced, making it difficult to measure the number of iterations and thus to perform the convergence analysis. In addition, all these neighborhood search algorithms are inherently limited to centralized detection without further decentralized implementations.

To address the related issues, the novel contributions of this paper are as follows:

- For distributed detection and parallel execution, we propose the decentralized likelihood ascent search (DLAS) mechanism. We then introduce it into the traditional distributed iterative algorithms, such as CG, ADMM, and DN, resulting in the proposed DLAS-aided detection scheme, which surpasses the linear MMSE performance with reduced complexity and bandwidth costs.
- We derive the equivalent noise, the post-processing signal-to-noise ratio (SNR), and the diversity order for D-LAS. Our analysis demonstrates that the proposed DLAS-aided detection is able to achieve the same received diversity as ML detection, which implies the remarkable performance gain compared to the traditional linear detection. To the authors' knowledge, this is the first comprehensive performance evaluation of the neighborhood search algorithms.
- For better implementation in practice, we design the decentralized effective ring (DER) architecture that allows parallel processing with the significant reduction in bandwidth overhead.

To sum up, we present a clear comparison of our novel contributions to the existing literature in Table I. Here, $I_{\max}$ and $T_{\max}$ represent the maximum numbers of inner and outer iterations, respectively. $\mathcal{O}$ refers to the constellation set. $K = 2N_t$ and $N = 2N_r$, where N_t and N_r denote the number of antennas at the user side and BS side, respectively.

In the context of signal detection for distributed large-scale MIMO systems, the exploration of efficient methods in diverse channel environments is critically important [5]. Among various channel models, the Rayleigh fading channel is widely adopted in related research for its ability to effectively capture the complex fading phenomena encountered in practical communication environments [6]–[19]. Therefore, it serves as a fundamental basis for the in-depth investigation of detection methods presented in this paper.

The rest of this paper is organized as follows. Section II briefly introduces the traditional detection for uplink large-scale MIMO systems in both centralized and decentralized architectures. In Section III, the proposed DLAS mechanism is described, and its complexity as well as bandwidth analysis is also given. Section IV derives the equivalent noise, the post-processing SNR, and the diversity order of DLAS to guarantee its performance gain. In Section V, the DER architecture is proposed for reduced bandwidth cost and parallel computation. After that, Section VI presents simulations of the proposed DLAS-aided detection for uplink large-scale MIMO systems. Finally, Section VII concludes the paper.

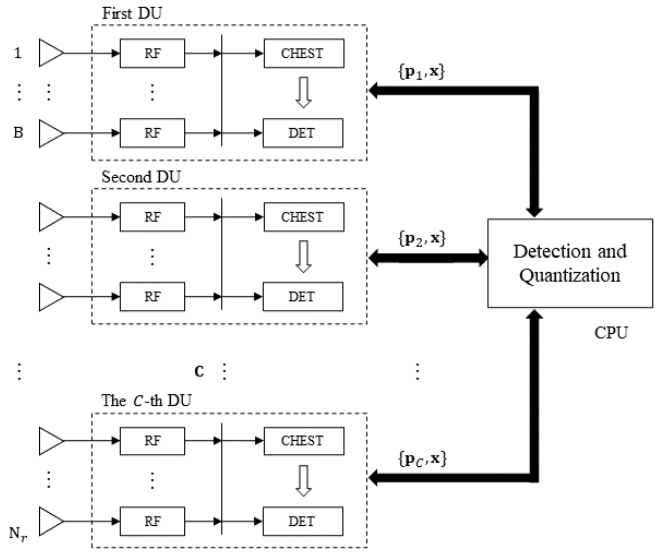


Fig. 1. Illustration of the decentralized baseband processing (DBP) architecture with C DUs. Each DU is equipped with $B = N_r/C$ antennas and an independent computing fabric for channel estimation (CHEST) and detection (DET), while the CPU performs signal detection and finally quantizes ($\mathcal{Q}$) the outputs.

Algorithm 1: Decentralized ADMM Algorithm

Input : $\mathbf{y}_c$, $\mathbf{H}_c$, $c = 1, 2, \dots C$, ρ, γ, σ_n^2
Output : estimated transmit signal $\widehat{\mathbf{x}}$

Pre-processing: $\mathbf{W}_c = \mathbf{G}_c + \rho \mathbf{I}_K$, $\mathbf{G}_c = \mathbf{H}_c^T \mathbf{H}_c$, $\mathbf{y}_c^{\text{MRC}} = \mathbf{W}_c^{-1} \mathbf{y}_c^{\text{MF}}$, $\mathbf{y}_c^{\text{MF}} = \mathbf{H}_c^T \mathbf{y}_c$
Initialization: $\boldsymbol{\theta}_c^1 = \mathbf{y}_c^{\text{MRC}}$, $\boldsymbol{\lambda}_c^1 = \mathbf{0}$, $\mathbf{x}^1 = \frac{\rho}{\sigma_n^2 + C\rho} \sum_{c=1}^{C} \boldsymbol{\theta}_c^1$
for $t = 2, \cdots, T_{\max}$ **do**
  // Decentralized processing in each DU:
  $\boldsymbol{\theta}_c^t = \mathbf{y}_c^{\text{MRC}} + \rho \mathbf{W}_c^{-1} (\mathbf{x}^{t-1} - \boldsymbol{\lambda}_c^{t-1})$
  $\boldsymbol{\lambda}_c^t = \boldsymbol{\lambda}_c^{t-1} - \gamma \left(\mathbf{x}^{t-1} - \boldsymbol{\theta}_c^t \right)$
  $\mathbf{p}_c^t = \boldsymbol{\theta}_c^t + \boldsymbol{\lambda}_c^t$
  // Centralized processing in CPU:
  $\mathbf{x}^t = \frac{\rho}{\sigma_n^2 + C\rho} \sum_{c=1}^{C} \mathbf{p}_c^t$
end for
output $\widehat{\mathbf{x}} = \lceil \mathbf{x}^{T_{\max}} \rfloor_{\mathcal{Q}} \in \mathcal{X}^K$

Notation: Matrices and column vectors are denoted by upper and lowercase boldface letters, and the transpose, inverse of a matrix $\mathbf{B}$ by $\mathbf{B}^T$ and $\mathbf{B}^{-1}$, respectively. We use $\mathbf{b}_i$ for the i-th column of the matrix $\mathbf{B}$, $b_{i,j}$ for the entry in the i-th row and j-th column of the matrix $\mathbf{B}$. The vector inverse $\mathbf{b}^{-1}$ is computed by taking the reciprocal of each individual element, and $\odot$ represent the Hadamard product, which performs element-wise multiplication between two vectors. Additionally, $\text{Tr}(\cdot)$ and $E[\cdot]$ denote the trace of the matrix and expectation, respectively. $\text{diag}(\mathbf{B})$ extracts the diagonal elements of the square matrix $\mathbf{B}$, and $\lceil \cdot \rfloor$ rounds to the closest integer. Finally, $\Re(\cdot)$ and $\Im(\cdot)$ indicate the real and imaginary components, respectively.

II. Preliminary

This section introduces the signal detection in both centralized and decentralized uplink large-scale MIMO systems, along with the ADMM serving as an example of the distributed iterative detection schemes.

A. Uplink Signal Detection

Considering a large-scale MIMO system with N_r receive antennas at a BS, serving different N_t single antenna users ($N_r \geq N_t$), the input-output relationship of the system can be represented as

$$\bar{\mathbf{y}} = \bar{\mathbf{H}} \bar{\mathbf{x}} + \bar{\mathbf{n}}. \tag{1}$$

Here, $\bar{\mathbf{y}} \in \mathbb{C}^{N_r}$ denotes the received vector, and $\bar{\mathbf{x}} \in \mathcal{O}^{N_t}$ represents the transmitted vector from the discrete complex M-quadrature amplitude modulation (QAM) constellation set $\mathcal{O}^{N_t}$. $\bar{\mathbf{H}} \in \mathbb{C}^{N_r \times N_t}$ represents the Rayleigh fading channel matrix whose entries follow $\mathcal{CN}(0, 1)$ and $\bar{\mathbf{n}} \in \mathbb{C}^{N_r}$ denotes the addictive white Gaussian noise (AWGN) with zero mean and covariance matrix $\sigma_{\bar{n}}^2 \mathbf{I}_{N_r}$.

The complex-valued model in (1) can be equivalently translated into a real-valued system of dimensions $2N_r \times 2N_t$ as follows:

$$\begin{bmatrix} \Re(\bar{\mathbf{y}}) \\ \Im(\bar{\mathbf{y}}) \end{bmatrix} = \begin{bmatrix} \Re(\bar{\mathbf{H}}) & -\Im(\bar{\mathbf{H}}) \\ \Im(\bar{\mathbf{H}}) & \Re(\bar{\mathbf{H}}) \end{bmatrix} \begin{bmatrix} \Re(\bar{\mathbf{x}}) \\ \Im(\bar{\mathbf{x}}) \end{bmatrix} + \begin{bmatrix} \Re(\bar{\mathbf{n}}) \\ \Im(\bar{\mathbf{n}}) \end{bmatrix}, \tag{2}$$

which can be succinctly expressed by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{3}$$

For simplicity of notation, from this point onward, let N and K denote $2N_r$ and $2N_t$, respectively. In this way, the complex constellation $\mathcal{O}^{N_t}$ is transformed into a real-valued $\sqrt{M}$-amplitude-shift keying (ASK) constellation set $\mathcal{X}^K$, defined as $\mathcal{X} = \{\pm 1, \pm 3, \dots, \pm(\sqrt{M} - 1)\}$. Considering the equivalent channel matrix $\mathbf{H} \in \mathbb{R}^{N \times K}$ with entries distributed as $\mathcal{N}\left(0, \frac{1}{2}\right)$, the optimal maximum likelihood (ML) detection aims to recover the transmitted signal vector $\mathbf{x}$ from the received vector $\mathbf{y}$ by

$$\widehat{\mathbf{x}}_{\text{ML}} = \underset{\mathbf{x} \in \mathcal{X}^K}{\arg\min} \frac{1}{2} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2, \tag{4}$$

which is an NP-hard problem in principle [31], [32]. As the suboptimal detection schemes, with $\sigma_n^2 = \frac{1}{2}\sigma_{\bar{n}}^2$, the traditional linear detection methods, i.e., ZF and MMSE, follow

$$\mathbf{x}_{\text{ZF}} = \left(\mathbf{H}^T \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{y}, \tag{5}$$

$$\mathbf{x}_{\text{MMSE}} = \left(\mathbf{H}^T \mathbf{H} + \sigma_n^2 \mathbf{I}_K\right)^{-1} \mathbf{H}^T \mathbf{y}, \tag{6}$$

where the final detection output $\widehat{\mathbf{x}}$ is obtained by quantizing to the nearest constellation point, resulting in $\widehat{\mathbf{x}}_{\text{ZF}} = \lceil \mathbf{x}_{\text{ZF}} \rfloor_{\mathcal{Q}} \in \mathcal{X}^K$ or $\widehat{\mathbf{x}}_{\text{MMSE}} = \lceil \mathbf{x}_{\text{MMSE}} \rfloor_{\mathcal{Q}} \in \mathcal{X}^K$.

However, all these methods aggregate data from BS antennas to the central processing unit (CPU) for signal detection, where scaling to hundreds or even thousands of antennas challenges the hardware with increased complexity and bandwidth demands [5]. Therefore, a number of decentralized architectures and corresponding distributed detection schemes

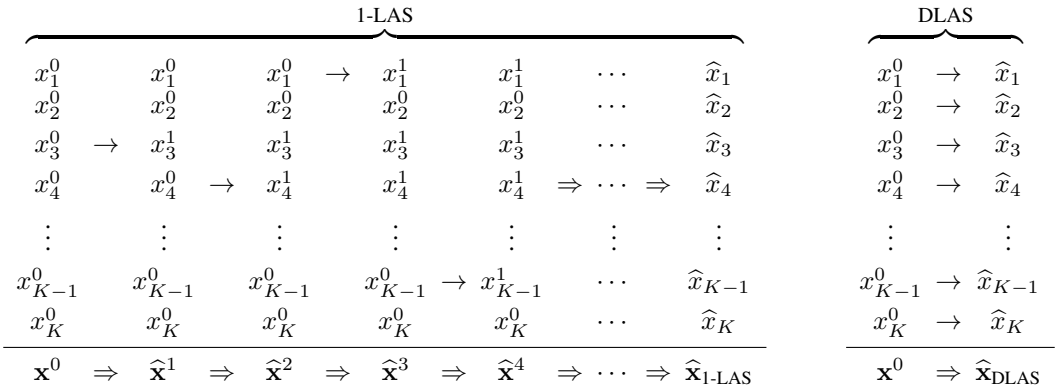


Fig. 2. Comparison of the 1-LAS and DLAS update processes. On the left, in 1-LAS, only one symbol is updated in each iteration, namely the one that leads to the largest reduction in $\mathcal{F}_{\text{ML}}(\rho_i)$. This process continues until convergence. Typically, the number of iterations required is several times greater than K. On the right, the proposed DLAS mechanism sequentially updates all elements in $\mathbf{x}^0$, reaching the final solution within a single iteration.

have been proposed [6]–[18].

B. Decentralized System Model

The DBP architecture was introduced in [6], [7] as a solution to reduce both the computational complexity and data bandwidth costs of centralized detection. As illustrated in Fig.1, DBP divides the BS antennas into C individual DUs, each having B_c antennas and an independent computing fabric for the c-th DU to perform local processing tasks. For simplicity, we assume DUs of equal size and set $B = B_c$. As a result, the uplink received vector $\mathbf{y} \in \mathbb{R}^N$ and channel matrix $\mathbf{H} \in \mathbb{R}^{N\times K}$ are divided into

$$\mathbf{y} = \left[\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_C^T\right]^T, \tag{7}$$

$$\mathbf{H} = \left[\mathbf{H}_1^T, \mathbf{H}_2^T, \ldots, \mathbf{H}_C^T\right]^T, \tag{8}$$

where $\mathbf{y}_c \in \mathbb{R}^Q$ and $\mathbf{H}_c \in \mathbb{R}^{Q\times K}$ with $Q = 2B$ represent the local received vector and corresponding channel matrix of the c-th DU, respectively. The noise vector is also divided into $\mathbf{n} = \left[\mathbf{n}_1^T, \mathbf{n}_2^T, \ldots, \mathbf{n}_C^T\right]^T$, allowing the following model for each DU

$$\mathbf{y}_c = \mathbf{H}_c\mathbf{x} + \mathbf{n}_c,\ c = 1, 2, \ldots, C. \tag{9}$$

In other distributed architectures such as ring [8], daisy-chain [10]–[12], and fully decentralized (FD) [13]–[16], DUs employ the same model as in (9), together with the CPU to accomplish signal detection. Based on this, distributed iterative methods such as CG [6], ADMM [7], and DN [8] have been proposed to solve the problem in (6) in a decentralized manner. Their general form can be expressed as the following two equations. Specifically, each DU updates the local information $\mathbf{p}_c^t$ as follows:

$$\mathbf{p}_c^t = f\left(\mathbf{p}_c^{t-1}, \mathbf{x}^{t-1}; \mathbf{y}_c, \mathbf{H}_c\right). \tag{10}$$

Then, the global estimated vector $\mathbf{x}^{t+1}$ is updated by

$$\mathbf{x}^t = g\left(\mathbf{x}^{t-1}, \mathbf{p}_1^t, \mathbf{p}_2^t, \ldots, \mathbf{p}_C^t\right). \tag{11}$$

Here, $f(\cdot)$ and $g(\cdot)$ signify series of operations that vary depending on the specific algorithm used. Attributed to its flexibility and capability for distributed computation, ADMM is widely regarded as a classical and efficient approach for decentralized signal detection [5]. To this end, we take the ADMM approach based on the DBP architecture in [7] as an example. Specifically, each DU updates its local information in parallel as follows:

$$\boldsymbol{\theta}_c^t = \mathbf{y}_c^{\text{MRC}} + \rho\mathbf{W}_c^{-1}(\mathbf{x}^{t-1} - \boldsymbol{\lambda}_c^{t-1}), \tag{12}$$

$$\boldsymbol{\lambda}_c^t = \boldsymbol{\lambda}_c^{t-1} - \gamma\left(\mathbf{x}^{t-1} - \boldsymbol{\theta}_c^t\right), \tag{13}$$

$$\mathbf{p}_c^t = \boldsymbol{\theta}_c^t + \boldsymbol{\lambda}_c^t, \tag{14}$$

based on the initial set up $\boldsymbol{\theta}_c^1 = \mathbf{y}_c^{\text{MRC}}$, $\boldsymbol{\lambda}_c^1 = \mathbf{0}$, and $\mathbf{x}^1 = \frac{\rho}{\sigma_n^2 + C\rho}\sum_{c=1}^C \boldsymbol{\theta}_c^1$. After that, the CPU updates the global estimated vector by

$$\mathbf{x}^t = \frac{\rho}{\sigma_n^2 + C\rho}\sum_{c=1}^C \mathbf{p}_c^t, \tag{15}$$

and then pass it back to each DU for the next iteration. Here $\mathbf{W}_c = \mathbf{G}_c + \rho\mathbf{I}_K$, $\mathbf{G}_c = \mathbf{H}_c^T\mathbf{H}_c$, $\mathbf{y}_c^{\text{MRC}} = \mathbf{W}_c^{-1}\mathbf{y}_c^{\text{MF}}$, and $\mathbf{y}_c^{\text{MF}} = \mathbf{H}_c^T\mathbf{y}_c$ correspond to the local regularized Gram matrix, Gram matrix, maximum-ratio combining (MRC) output, and match filter (MF) output, respectively. ρ and γ denote the fixed penalty parameter and step size, respectively. Through $T_{\max}$ times iterations, the global estimated vector $\mathbf{x}^{T_{\max}}$ is obtained by the CPU to produce the final detection output. More specifically, the decentralized ADMM detection algorithm is summarized in Algorithm 1.

III. The Proposed Distributed Detection Scheme

In this section, we first introduce the decentralized likelihood ascent search (DLAS) mechanism over the traditional distributed iterative algorithms, which leads to the proposed DLAS-aided detection. Then, we analyze the computational complexity and data bandwidth to demonstrate the superiority of DLAS compared to other distributed detection schemes.

A. The Proposed DLAS Mechanism

We now upgrade the one symbol likelihood ascent search (1-LAS) algorithm in [21] to a distributed version with less

complexity cost, and introduce it to the distributed iterative algorithms for more satisfactory performance.

Typically, the objective of DLAS is to minimize the value of ML cost function as follows:

$$\mathcal{F}_{\text{ML}} = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2, \tag{16}$$

which serves a metric to evaluate the estimated vector $\mathbf{x}$. Given that the ML detection in (4) achieves the minimum $\mathcal{F}_{\text{ML}}$, a lower ML cost function value corresponds to a more accurate estimated vector $\mathbf{x}$. Meanwhile, the final output derived from distributed iterative algorithms such as CG, ADMM, and DN is used as an initial solution for the subsequent DLAS mechanism, namely, $\mathbf{x}^0 = \lceil \mathbf{x}^{T_{\max}} \rfloor_{\mathcal{Q}} \in \mathcal{X}^K$.

Given $\mathbf{x}^0$, we update one symbol at a time (specifically, the i-th element for the i-th update, $i = 1, 2, \ldots, K$), and the update rule can be expressed as

$$\tilde{\mathbf{x}} = \mathbf{x}^0 + \rho_i \mathbf{u}_i, \tag{17}$$

where $\mathbf{u}_i \in \mathbb{R}^K$ represents the unit vector with only i-th element set to one, and all other elements set to zero. Since both $\tilde{\mathbf{x}}$ and $\mathbf{x}^0$ must reside within the constellation space $\mathcal{X}^K$, the value of ρ_i is restricted to integers that are multiples of 2. For instance, in 16-QAM modulation with the constellation space $\mathcal{X} = \{-3, -1, 1, 3\}$, the value of ρ_i is limited to the neighboring set $\mathcal{A} = \{-6, -4, -2, 0, 2, 4, 6\}$.

To determine the best value for ρ_i, which minimizes the ML cost in (16), we consider the following ML cost difference

$$\begin{aligned}
&\triangle\mathcal{F}_{\text{ML}}(\rho_i) = \|\mathbf{y} - \mathbf{H}\tilde{\mathbf{x}}\|_2^2 - \|\mathbf{y} - \mathbf{H}\mathbf{x}^0\|_2^2 \\
&= \rho_i^2 \mathbf{u}_i^T \mathbf{H}^T \mathbf{H} \mathbf{u}_i + \rho_i \mathbf{u}_i^T \mathbf{H}^T \mathbf{H} \mathbf{x}^0 + \rho_i \mathbf{x}^0 \mathbf{H}^T \mathbf{H} \mathbf{u}_i - 2\rho_i \mathbf{y}^T \mathbf{H} \mathbf{u}_i \\
&= \rho_i^2 \left(\mathbf{H}^T \mathbf{H}\right)_{i,i} + 2\rho_i \left(\mathbf{H}^T \mathbf{H} \mathbf{x}^0\right)_i - 2\rho_i \left(\mathbf{H}^T \mathbf{y}\right)_i \\
&= \rho_i^2 g_{i,i} - 2\rho_i \left(\mathbf{H}^T \mathbf{y} - \mathbf{H}^T \mathbf{H} \mathbf{x}^0\right)_i \\
&= \rho_i^2 g_{i,i} - 2\rho_i e_i,
\end{aligned} \tag{18}$$

where

$$\mathbf{G} = \mathbf{H}^T \mathbf{H} = \sum_{c=1}^{C} \mathbf{G}_c \tag{19}$$

and

$$\mathbf{e} = \mathbf{H}^T \left(\mathbf{y} - \mathbf{H}\mathbf{x}^0\right) = \sum_{c=1}^{C} \mathbf{y}_c^{\text{MF}} - \sum_{c=1}^{C} \mathbf{G}_c \mathbf{x}^0 \tag{20}$$

denote the global Gram matrix in $\mathbb{R}^{K \times K}$ and residual vector in $\mathbb{R}^K$, respectively.

To reduce the ML cost function in (16) for a better estimate, the ML cost difference $\triangle\mathcal{F}_{\text{ML}}(\rho_i)$ in (18) should be negative, i.e. $\triangle\mathcal{F}_{\text{ML}}(\rho_i) < 0$. For the case of one symbol update, the maximum reduction in $\triangle\mathcal{F}_{\text{ML}}(\rho_i)$ can be achieved by forcing the gradient of (18) with respect to ρ_i to zero, which leads to a closed-form expression for the optimal ρ_i as follows

$$\rho_i = \mathcal{P}\left(\frac{e_i}{g_{i,i}}\right), \tag{21}$$

where $\mathcal{P}(a) = 2\lceil a/2 \rfloor$ represents the projection of a onto the neighboring set $\mathcal{A}$. Notice that, ρ_i in (21) forces the minimum value of $\triangle\mathcal{F}_{\text{ML}}(\rho_i)$, which is no more than zero [21]. If $\triangle\mathcal{F}_{\text{ML}}(\rho_i) < 0$, ρ_i in (21) indeed reduces the ML cost function and thus improve the performance, otherwise it keeps the original result with no degradation. To facilitate the parallel processing and decentralized design, the entire update process is simplified by updating all elements sequentially just once, i.e., by performing (21) simply from 1 to K in sequence. To be more specific, by defining $\mathbf{D} = \text{diag}(\mathbf{G}) \in \mathbb{R}^K$, the entire update process can be represented as

$$\tilde{\mathbf{x}} = \mathbf{x}^0 + \mathcal{P}\left(\mathbf{D}^{-1} \odot \mathbf{e}\right). \tag{22}$$

Given the possibility for $\tilde{\mathbf{x}}_i$ to exceed $\sqrt{M} - 1$ or to fall below $-\sqrt{M} + 1$, it is necessary to adjust $\tilde{\mathbf{x}}_i$ to the nearest constellation point, which can be expressed as follows

$$\widehat{\mathbf{x}}_i = \begin{cases} \sqrt{M} - 1, & \text{if } \tilde{\mathbf{x}}_i > \sqrt{M} - 1, \\ -(\sqrt{M} - 1), & \text{if } \tilde{\mathbf{x}}_i < -\sqrt{M} + 1, \\ \tilde{\mathbf{x}}_i, & \text{otherwise.} \end{cases} \tag{23}$$

Therefore, we can reformulate the update process represented from (21) to (23) into the following two steps

$$\mathbf{x}_{\text{DLAS}} = \mathbf{x}^0 + \mathbf{D}^{-1} \odot \mathbf{e} \tag{24}$$

and

$$\widehat{\mathbf{x}}_{\text{DLAS}} = \lceil \mathbf{x}_{\text{DLAS}} \rfloor_{\mathcal{Q}} \in \mathcal{X}^K. \tag{25}$$

Finally, the detected signal in (25) is outputted as the detection solution of DLAS. To summarize, the proposed DLAS-aided detection for distributed uplink large-scale MIMO systems, is outlined in Algorithm 2.

Algorithm 2: Decentralized Likelihood Ascent Search (DLAS)-Aided Detection for Distributed MIMO Systems

Input : $\mathbf{y}_c^{\text{MF}}$, $\mathbf{G}_c$, $c = 1, 2, \ldots C$, $T_{\max}$
Output : estimated transmit signal $\widehat{\mathbf{x}}$

Use distributed iterative algorithms to get $\mathbf{x}^{T_{\max}}$
Let initial solution $\mathbf{x}^0 = \lceil \mathbf{x}^{T_{\max}} \rfloor_{\mathcal{Q}} \in \mathcal{X}^K$
// Decentralized processing in each DU:
$\mathbf{D}_c = \text{diag}(\mathbf{G}_c)$
$\mathbf{e}_c = \mathbf{y}_c^{\text{MF}} - \mathbf{G}_c \mathbf{x}^0$
// Centralized processing in CPU:
$\mathbf{D} = \sum_{c=1}^{C} \mathbf{D}_c$
$\mathbf{e} = \sum_{c=1}^{C} \mathbf{e}_c$
$\mathbf{x}_{\text{DLAS}} = \mathbf{x}^0 + \mathbf{D}^{-1} \odot \mathbf{e}$
output $\widehat{\mathbf{x}}_{\text{DLAS}} = \lceil \mathbf{x}_{\text{DLAS}} \rfloor_{\mathcal{Q}} \in \mathcal{X}^K$

DLAS is a distributed and simplified version of 1-LAS in [21], whose update processes are illustrated in Fig. 2 for clarity. In comparison to 1-LAS, the DLAS mechanism exhibits several distinct differences:

1) DLAS updates all elements in $\mathbf{x}^0$ sequentially in a single iteration, which effectively eliminates the need to compute the ML cost difference. In contrast, 1-LAS may require multiple iterations (usually several times than K) to reach stability, where only one symbol chosen for the largest decrease in $\mathcal{F}_{\text{ML}}(\rho_i)$ is updated in each iteration.
2) DLAS is designed for distributed detection and supports parallel execution, while 1-LAS only operates in a centralized and sequential manner.

3) DLAS can be succinctly expressed through vector multiplications, making it more suitable for theoretical analysis. Conversely, the ambiguous number of iterations for convergence renders 1-LAS a heuristic scheme.

Clearly, the proposed DLAS mechanism can be implemented in a decentralized and parallel manner with significantly reduced complexity. Moreover, its performance gain compared to MMSE detection will be guaranteed, which is discussed in Section IV.

B. Complexity and Bandwidth Analysis

Here, the computational complexity is evaluated in terms of the required number of real multiplications, where one complex multiplication is counted as four real multiplications [12]. For example, the complexity required to invert a matrix of dimensions $K \times K$ is quantified as $0.5K^3$ [33].

Specifically, with ADMM initialization as an example, the overall complexity of the proposed DLAS-aided detection consists of the following three stages. During the first stage, namely pre-processing in ADMM, the computational complexity for $\mathbf{G}_c$, $\mathbf{y}_c^{\text{MF}}$ within each DU is K^2Q and KQ, respectively. Moreover, the calculation of $\mathbf{y}_c^{\text{MRC}}$ involves a $K \times K$ matrix inversion and a multiplication between $\mathbf{W}_c^{-1}$ and $\mathbf{y}_c^{\text{MF}}$, resulting in a complexity of $0.5K^3 + K^2$. In the second stage, the ADMM iterations, the complexity of initialization is marginal. Subsequently, the computational burdens for determining $\boldsymbol{\theta}_c^t$, $\boldsymbol{\lambda}_c^t$ and $\mathbf{x}^t$ in each iteration are noted as K^2, K, and K, respectively. As for the final stage, the DLAS mechanism, computing local residual vector $\mathbf{e}_c$ demands a complexity of K^2 in each DU. Subsequently, the calculation of $\mathbf{x}_{\text{DLAS}}$, which involves a vector inversion and multiplication, requires a complexity of $2K$. To summarize, the overall computational complexity for the proposed DLAS mechanism as well as DLAS-aided detection is illustrated in Table II.

On the other hand, as delineated in Table II, the data bandwidth required for interconnection of these distributed large-scale MIMO detection schemes is determined by the total real values transferred on links, where one complex value counts for two real values [8]. Specifically, in the pre-processing stage of ADMM, initializing $\mathbf{x}^1$ involves summing $\boldsymbol{\theta}_c^1$ for each DU, equating to a bandwidth of KC. At every iteration, it is necessary for each DU to obtain the global estimated vector $\mathbf{x}^{t-1}$ from the CPU and then send the local information $\mathbf{p}_c^t$ back to the CPU, leading to the bandwidth of $2KC$. Subsequently, during the DLAS mechanism, each DU obtains the initial vector $\mathbf{x}^0$ from the CPU and then transmits the diagonal elements of local Gram matrix $\mathbf{D}_c$ along with the residual vector $\mathbf{e}_c$ back to the CPU for obtaining the final solution, requiring a bandwidth of $3KC$.

Throughout the context, $T_{\max}$ denotes the number of outer iterations for the CG, ADMM, and DN algorithms. From Table II, the proposed DLAS mechanism exhibits much lower complexity and required bandwidth than other distributed detection schemes. More precisely, the complexity and required bandwidth of the DLAS-aided detection are comparable to the traditional distributed iterative algorithms, making it a better choice for practical hardware implementation.

IV. Detection Performance Analysis of DLAS

Unlike the traditional 1-LAS [21], M-LAS [21], RTS [22], and other neighborhood search algorithms [23]–[30], the proposed DLAS mechanism simplifies the update process by modifying all elements sequentially within a single iteration, making it possible to be analyzed theoretically. By leveraging the properties of the Rayleigh channels, we establish a robust theoretical foundation for the analysis of DLAS. Specifically, our performance analysis shows that the proposed DLAS-aided detection achieves the same received diversity as ML detection. To the best of the authors' knowledge, the forthcoming analysis presents the first comprehensive performance evaluation of the neighborhood search algorithms.

A. Equivalent Noise

The traditional distributed iterative algorithms, such as CG, ADMM, and DN, obtain the estimated vector $\mathbf{x}^{T_{\max}}$ through iterations. Subsequently, this solution is projected onto the constellation set $\mathcal{X}^K$ to serve as an initial input for the following DLAS mechanism, expressed as follows:

$$\begin{aligned}\mathbf{x}^0 &= \lceil \mathbf{x}^{T_{\max}} \rfloor_{\mathcal{Q}} \\ &= \mathbf{x} + \mathbf{n}^0 + \boldsymbol{\delta} \\ &= \mathbf{x} + \mathbf{z}. \end{aligned} \tag{26}$$

Here $\mathbf{x}^0 \in \mathbb{R}^K$ represents the initial input for DLAS, while $\mathbf{n}^0 \in \mathbb{R}^K$ denotes the equivalent noise vector from the distributed iterative algorithm. Additionally, $\mathbf{z} \in \mathcal{A}^K$ is the *distance vector* from the neighboring set, and $\boldsymbol{\delta} = \mathbf{z} - \mathbf{n}^0 \in \mathbb{R}^K$ rounds the solution to the nearest constellation point. Clearly, the elements in $\mathbf{z}$ are distributed as follows:

$$z_i = \begin{cases} 0, & \text{if } |\mathbf{n}_i^0| \le 1, \\ 2, & \text{if } 1 < \mathbf{n}_i^0 \le 3, \\ -2, & \text{if } -3 \le \mathbf{n}_i^0 < -1, \\ \cdots & \cdots \end{cases} \tag{27}$$

for $i = 1, 2, \ldots, K$. Taking (26) into (20), the residual vector can be reformulated as

$$\begin{aligned}\mathbf{e} &= \mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{x}^0) \\ &= \mathbf{H}^T\left(\mathbf{H}\mathbf{x} + \mathbf{n} - \mathbf{H}(\mathbf{x} + \mathbf{z})\right) \\ &= \mathbf{H}^T\mathbf{n} - \mathbf{H}^T\mathbf{H}\mathbf{z}. \end{aligned} \tag{28}$$

Thus, DLAS mechanism updates the initial solution $\mathbf{x}^0$ by

$$\begin{aligned}\mathbf{x}_{\text{DLAS}} &= \mathbf{x}^0 + \mathbf{D}^{-1} \odot \mathbf{e} \\ &= \mathbf{x} + \mathbf{z} + \mathbf{D}^{-1} \odot (\mathbf{H}^T\mathbf{n} - \mathbf{H}^T\mathbf{H}\mathbf{z}) \\ &= \mathbf{x} + \mathbf{z} + \mathbf{D}^{-1} \odot (\mathbf{H}^T\mathbf{n}) - \mathbf{D}^{-1} \odot (\mathbf{H}^T\mathbf{H}\mathbf{z}), \end{aligned} \tag{29}$$

where $\mathbf{x}_{\text{DLAS}} \in \mathbb{R}^K$ will be projected onto the constellation set $\mathcal{X}^K$ as the final detection output. We now investigate the equivalent noise of the proposed DLAS-aided detection.

Lemma 1. *Given the channel matrix* $\mathbf{H}$ *whose entries follow* $\mathcal{N}\left(0, \frac{1}{2}\right)$*, the distance vector* $\mathbf{z}$ *follows that*

$$\mathbf{z} = \mathbf{D}^{-1} \odot (\mathbf{H}^T\mathbf{H}\mathbf{z}) \tag{30}$$

TABLE II
COMPLEXITY AND BANDWIDTH COMPARISONS OF DECENTRALIZED ALGORITHMS

Algorithm	Number of real multiplications	Total real values transferred on links	
		DBP and Ring	DER
CG [6]	$\left(K^2Q+KQ\right)C+(K^2+7K)CT_{\max}$	$2KC(T_{\max}+2)$	$2K(C-1)(T_{\max}+2)$
ADMM [7]	$\left(K^2Q+KQ+K^2+0.5K^3\right)C+\left((K^2+K)C+K\right)(T_{\max}-1)$	$KC(2T_{\max}-1)$	$K(C-1)(2T_{\max}-1)$
DN [8]	$\left(K^2Q+KQ+2K\right)C+(K^2C+K)T_{\max}$	$4KCT_{\max}$	$4K(C-1)T_{\max}$
DLAS mechanism	K^2C+2K	$3KC$	$3K(C-1)$
DLAS-CG	$\left(K^2Q+KQ+K^2\right)C+(K^2+7K)CT_{\max}+2K$	$KC(2T_{\max}+7)$	$K(C-1)(2T_{\max}+7)$
DLAS-ADMM	$\left(K^2Q+KQ+2K^2+0.5K^3\right)C+\left((K^2+K)C+K\right)(T_{\max}-1)+2K$	$2KC(T_{\max}+1)$	$2K(C-1)(T_{\max}+1)$
DLAS-DN	$\left(K^2Q+KQ+K^2+2K\right)C+(K^2C+K)T_{\max}+2K$	$KC(4T_{\max}+3)$	$K(C-1)(4T_{\max}+3)$

with the increment of SNR.

Proof. In the case of $i=j$, we can find that $h_{i,j}^2$ follows the *Gamma distribution* with the shape parameter $\alpha=\frac{1}{2}$ and the scale parameter $\beta=1$ [34], namely $h_{i,j}^2\sim\Gamma(\frac{1}{2},1)$. Building upon it, $g_{i,i}$, as the summation of $h_{i,j}^2$, obeys the following distribution

$$g_{i,i}=\sum_{l=1}^{N}h_{l,i}^2\sim\Gamma\left(\frac{N}{2},1\right). \tag{31}$$

Subsequently, in the case of $i\neq j$, $h_{l,i}$ and $h_{l,j}$ are independent Gaussian variables with mean zero and variance $\frac{1}{2}$, which leads to

$$g_{i,j}=\sum_{l=1}^{N}h_{l,i}h_{l,j}\sim\mathcal{N}\left(0,\frac{N}{4}\right). \tag{32}$$

Then we define

$$\mathbf{b}=\mathbf{D}^{-1}\odot(\mathbf{H}^H\mathbf{H}\mathbf{z})=(\mathrm{diag}\,(\mathbf{G}))^{-1}\odot(\mathbf{G}\mathbf{z}) \tag{33}$$

with the i-th element

$$\begin{aligned}b_i&=\frac{1}{g_{i,i}}\times(g_{i,1}z_1+g_{i,2}z_2+\cdots+g_{i,K}z_K)\\&=z_i+\frac{1}{g_{i,i}}\sum_{j\neq i}^{K}g_{i,j}z_j.\end{aligned} \tag{34}$$

It is worth noticing that b_i is still a Gaussian variable with the distribution as

$$b_i\sim\mathcal{N}\left(z_i,\frac{N}{4g_{i,i}^2}\sum_{j\neq i}^{K}z_j^2\right). \tag{35}$$

Considering the fact that $g_{i,i}\sim\Gamma\left(\frac{N}{2},1\right)$, its mean and variance can be simply expressed by

$$E\left[g_{i,i}\right]=Var\left[g_{i,i}\right]=\frac{N}{2}, \tag{36}$$

which results in

$$E\left[g_{i,i}^2\right]=E^2\left[g_{i,i}\right]+Var\left[g_{i,i}\right]=\frac{N^2+2N}{4}. \tag{37}$$

Taking the expectation of $g_{i,i}^2$ in (35), the variance of b_i can be further approximated as

$$\sigma_{b_i}^2\approx\frac{K-1}{N+2}E\left[z_j^2\right]<\epsilon, \tag{38}$$

where ϵ denotes an arbitrarily small positive number. This inequality holds under the assumptions of $N\gg K$ and high SNR conditions, but its validity depends on specific system settings and may not always be guaranteed. Therefore, (38) should be interpreted as an approximation rather than an exact result. Under these conditions, $\sigma_{b_i}^2$ becomes sufficiently small, allowing b_i to be treated as a constant equal to its mean value (i.e., $b_i=z_i$ for $i=1,2,\ldots,K$), thus completing the proof. □

Based on Lemma 1, the estimated vector of the proposed DLAS-aided detection in (29) can be further simplified, which leads to the following result.

Theorem 1. *Given the channel matrix* $\mathbf{H}$*, the equivalent noise of proposed DLAS-aided detection is*

$$\mathbf{n}_{\mathrm{DLAS}}=\mathbf{D}^{-1}\odot(\mathbf{H}^T\mathbf{n}). \tag{39}$$

If the noise is independent over each element, the strategy that decodes each component of $\mathbf{x}$ independently is optimal [35]. However, the noise described in Theorem 1 is actually correlated, which leads to a significant degradation in the detection performance. The impact of this degradation on the post-processing SNR of each element is characterized in the following section.

B. The Full Received Diversity Order

In this subsection, we first derive the distribution of the post-processing SNR on the i-th element of DLAS, which in turn allows the characterization of the diversity order.

Lemma 2. *Given the equivalent noise in* (39)*, the post-processing SNR* γ_i *on the* i*-th element of proposed DLAS-aided detection is*

$$\gamma_i=\gamma_0 g_{i,i}, \tag{40}$$

where $\gamma_0=E[x_i^2]/\sigma_n^2$ *denotes the average SNR on each element.*

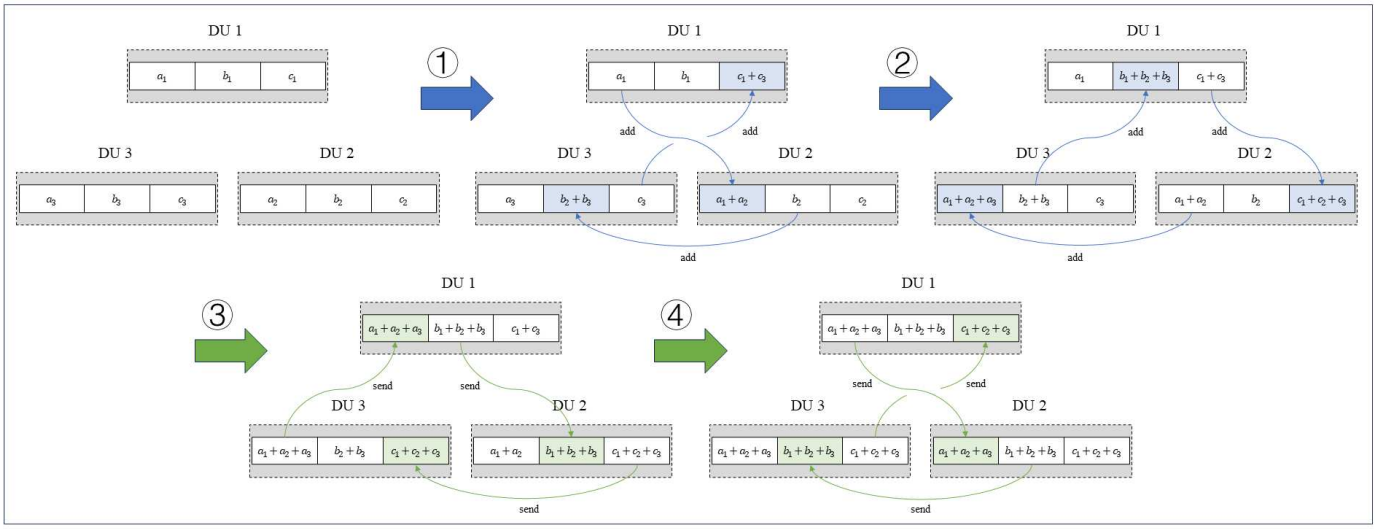


Fig. 3. Illustration of the data exchange approach within the DER architecture. Here, the BS comprises a total of $C = 3$ DUs, each holding three equal-sized data blocks a, b and c to be aggregated. The blue arrows symbolize the data aggregation stage, where data blocks are accumulated, while green arrows represent the data replacement stage, where data synchronization takes place.

Proof. First, we focus on the equivalent noise in (39), with the i-th element

$$n_i = \frac{\mathbf{h}_i^T \mathbf{n}}{g_{i,i}}. \tag{41}$$

In this way, the corresponding noise power can be expressed as

$$\begin{aligned}\sigma_i^2 &= \frac{E\left[\left(\mathbf{h}_i^T \mathbf{n}\right)^2\right]}{g_{i,i}^2} \\ &= \frac{E\left[\mathrm{Tr}(\mathbf{h}_i^T \mathbf{n}\mathbf{n}^T \mathbf{h}_i)\right]}{g_{i,i}^2} \\ &= \frac{\mathrm{Tr}\left(\mathbf{h}_i^T E\left[\mathbf{n}\mathbf{n}^T\right] \mathbf{h}_i\right)}{g_{i,i}^2} \\ &= \frac{\sigma_n^2 \mathrm{Tr}(\mathbf{h}_i^T \mathbf{h}_i)}{g_{i,i}^2} \\ &= \frac{\sigma_n^2}{g_{i,i}}. \end{aligned} \tag{42}$$

Therefore, we can obtain the post-processing SNR on the i-th element as follows:

$$\gamma_i = \frac{E[x_i^2]}{\sigma_i^2} = \gamma_0 g_{i,i}, \tag{43}$$

which completes the proof. □

Theorem 2. *The post-processing SNR on the i-th element of proposed DLAS-aided detection is a weighed Chi-squared variable distributed as*

$$f\left(\gamma_i\right) = \frac{e^{-\frac{\gamma_i}{\gamma_0}}}{\gamma_0 (N_r - 1)!} \left(\frac{\gamma_i}{\gamma_0}\right)^{N_r - 1}. \tag{44}$$

Proof. Considering the result that $g_{i,i}$ follows the *Gamma distribution* as shown in (31), we define a new variable v as follows:

$$v = 2g_{i,i} \sim \chi^2(N), \tag{45}$$

where $\chi^2(N)$ denotes the *Chi-square distribution* with N degrees of freedom. This leads to the following probability density function (PDF) of v

$$f(v) = \frac{e^{-\frac{b}{2}}}{2^{N_r}(N_r - 1)!} b^{N_r - 1}. \tag{46}$$

From (40) and (46), we can derive the exact PDF of the post-processing SNR γ_i, whose value is equal to $\gamma_0 v/2$

$$\begin{aligned} f(\gamma_i) &= f\left(\frac{2\gamma_i}{\gamma_0}\right) \times \frac{2}{\gamma_0} \\ &= \frac{e^{-\frac{\gamma_i}{\gamma_0}}}{2^{N_r}(N_r - 1)!} \left(\frac{2\gamma_i}{\gamma_0}\right)^{N_r - 1} \times \frac{2}{\gamma_0} \\ &= \frac{e^{-\frac{\gamma_i}{\gamma_0}}}{\gamma_0 (N_r - 1)!} \left(\frac{\gamma_i}{\gamma_0}\right)^{N_r - 1}. \end{aligned} \tag{47}$$

□

Clearly, the post-processing SNR γ_i is a weighted *Chi-square distribution* with $2N_r$ degrees of freedom. Based on this result, we can obtain the following theorem, which explains that the proposed DLAS-aided detection can achieve the full received diversity.

Theorem 3. *The diversity order of the proposed DLAS-aided detection is given by*

$$d_{\mathrm{DLAS}} = \lim_{\sigma_n^2 \to 0} \frac{\log\left(\bar{P}_{e,i}\right)}{\log(\sigma_n^2)} = N_r. \tag{48}$$

Proof. Assuming independent ML decoding at the receiver, the corresponding PER on the i-th element, namely $P_{e,i}$ is given by [36]

$$P_{e,i} \approx N_e Q\left(\sqrt{2\gamma_i}\right) \leq \frac{1}{2} N_e e^{-\gamma_i}, \tag{49}$$

where N_e is the average number of nearest neighbors of the constellation on the i-th stream and the second step follows from the *Chernoff bound*. The PER derived above corresponds to a particular realization of the channel as it fades over time due to the correlated N_r links. Therefore, the value of the

diversity order can be analyzed by considering the average PER (i.e., $\bar{P}_{e,i} = E[P_{e,i}]$) as follows

$$\bar{P}_{e,i} \leq \frac{N_e}{2\left(1+\gamma_0\right)^{N_r}}, \tag{50}$$

which is derived from [35]. The diversity order of the proposed DLAS-aided detection is lower bounded according to

$$\begin{aligned} d_{\text{DLAS}} &= \lim_{\sigma_n^2 \to 0} \frac{\log\left(\bar{P}_{e,i}\right)}{\log(\sigma_n^2)} \\ &\geq \lim_{\sigma_n^2 \to 0} \frac{\log\left(\frac{1}{2}N_e\left(1+E[x_i^2]/\sigma_n^2\right)^{-N_r}\right)}{\log(\sigma_n^2)} \\ &\approx \lim_{\sigma_n^2 \to 0} \frac{\log\left(\frac{1}{2}N_e\left(E[x_i^2]/\sigma_n^2\right)^{-N_r}\right)}{\log(\sigma_n^2)} \\ &= \lim_{\sigma_n^2 \to 0} \frac{\log\left(\frac{1}{2}N_e\right) - N_r\log\left(E[x_i^2]\right) + N_r\log(\sigma_n^2)}{\log(\sigma_n^2)} \\ &= N_r. \end{aligned} \tag{51}$$

Given that ML detection represents the optimal detector, the received diversity of DLAS is upper bounded by

$$d_{\text{DLAS}} \leq d_{\text{ML}} = N_r, \tag{52}$$

which coincides with the lower bound in (51), resulting in $d_{\text{DLAS}} = N_r$.

□

As can be seen clearly, after just a single iteration, the proposed DLAS-aided detection achieves the diversity order equivalent to that of ML detection, which demonstrates its ability to asymptotically approach the optimal performance with the increment of SNR.

V. Decentralized Generalization

In this section, a new distributed architecture, termed decentralized effective ring (DER), is proposed for better implementation of the DLAS-aided detection. Moreover, the DER architecture supports traditional distributed iterative algorithms applicable in DBP and ring architectures with reduced bandwidth and parallel computation.

A. Proposed DER Architecture

As illustrated in Algorithm 2, the execution of proposed DLAS-aided detection requires the summation of the local residual vector $\mathbf{e}_c$ along with the diagonal vector $\mathbf{D}_c$ in the CPU to update the global estimate, which results in considerable bandwidth requirements for the CPU interface. To this end, we consider a novel data exchange approach to obtain $\mathbf{e} = \sum_{c=1}^{C} \mathbf{e}_c$ and $\mathbf{D} = \sum_{c=1}^{C} \mathbf{D}_c$ in all DUs.

For clarity, we employ a data exchange scenario of the residual vector $\mathbf{e}$ in DLAS, as shown in Fig. 3, as an example to explain this data exchange approach in detail. To begin with, consider BS with a total of $C = 3$ DUs, each of which directly connects to its two adjacent DUs through unidirectional links and holds the local information $\mathbf{e}_c$ to be aggregated. Assuming that K is an integer multiple of C, we divide $\mathbf{e}_c$ into C equal-sized data blocks a, b and c, with each block containing K/C elements. Subsequently, all DUs obtain the global information $\mathbf{e}$ through the following two stages.

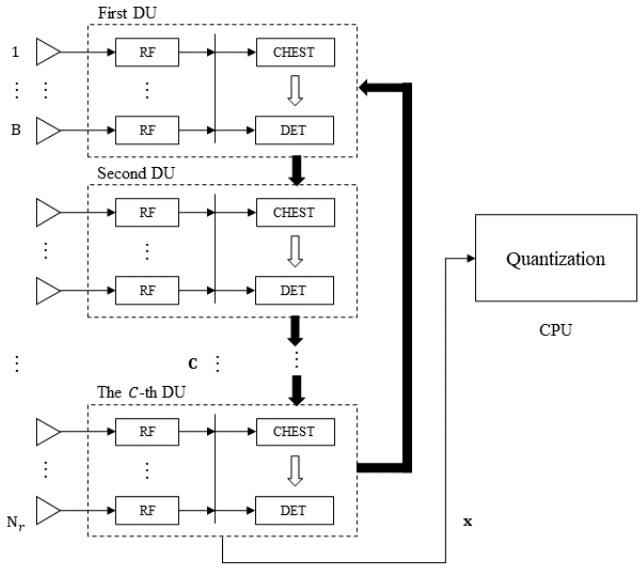


Fig. 4. Illustration of the decentralized effective ring (DER) architecture with C DUs. Each DU is equipped with $B = N_r/C$ antennas and an independent computing fabric for channel estimation (CHEST) and detection (DET) in parallel, while the CPU is only responsible for quantization ($\mathcal{Q}$).

1. **Data Aggregation**: Each DU adds the received data block to its own corresponding one and then transmits the summation to the next DU in parallel for $C-1$ times.

Specifically, as illustrated in Fig. 3, during the first pass, DU1 transmits the first data block a_1 to DU2, resulting in the accumulation of $a_1 + a_2$ in DU2's first data block. At the same time, DU1 and DU3 obtain the accumulation of $c_1 + c_3$ and $b_2 + b_3$ in their third and second data blocks, respectively. Then, in the second pass, DU1, DU2, and DU3 forward the third, first, and second data blocks to the following DUs, respectively. Therefore, the accumulations on DU1, DU2, and DU3 are $b_1 + b_2 + b_3$, $c_1 + c_2 + c_3$, and $a_1 + a_2 + a_3$ in that order.

2. **Data Replacement**: Each DU replaces the received data block with its own corresponding one, and then transmits it to the next DU in parallel for $C-1$ times.

Specifically, during the third and fourth passes, the aggregated data blocks $a_1 + a_2 + a_3$, $b_1 + b_2 + b_3$, and $c_1 + c_2 + c_3$ are sequentially transmitted twice until they are obtained by all DUs.

Following that, the local diagonal vector $\mathbf{D}_c$ can also be summed through these two stages, which completes the DLAS data exchange process. In addition, traditional distributed iterative detection schemes applicable in DBP and ring architectures such as CG [6], ADMM [7], DN [8], and EPA [17] can also aggregate the local information in the same way. Therefore, we summarize it as the decentralized effective ring (DER) architecture in Fig. 4, which shares the same topological structure as the ring architecture [8], but with different data exchange approach.

TABLE III
SIMULATION SCENARIOS FOR EACH FIGURE

Figure	Channel Model	Modulation Schemes	Number of BS Antennas	Number of Users	Number of DUs
Fig. 5	Rayleigh	16-QAM	128	32	4
Fig. 6	Rayleigh	16-QAM	256	32	8
Fig. 7	Rayleigh with imperfect CSI	16-QAM	256	32	8
Fig. 8	Rician	16-QAM	128	32	4
Fig. 9	Rayleigh	4-QAM, 16-QAM	64	16	4
Fig. 10	Rayleigh	4-QAM, 16-QAM	128	16	8
Fig. 11, 12	Rayleigh	16-QAM	128	64	2, 4, 8

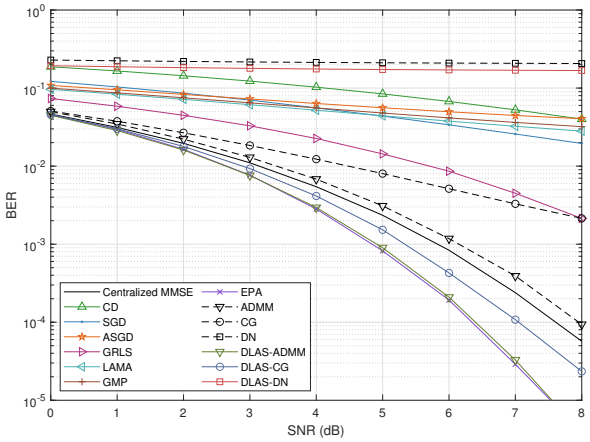


Fig. 5. Bit error rate performance comparison of different methods for the uncoded 128×32 large-scale MIMO system with $C = 4$ and 16-QAM.

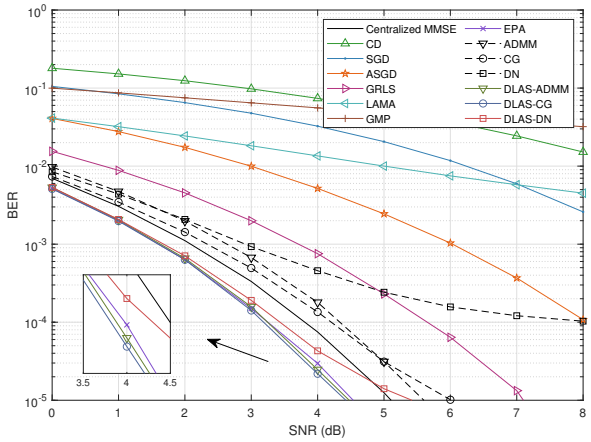


Fig. 6. Bit error rate performance comparison of different methods for the uncoded 256×32 large-scale MIMO system with $C = 8$ and 16-QAM.

B. Reduced Bandwidth and Parallel Computation of DER

For clarity, the bandwidth cost of the proposed DER architecture is also illustrated in Table II. Intuitively, compared to the same bandwidth achieved in both DBP and ring architectures, DER reduces the data bandwidth by 25% and 12.5% for the cases of $C = 4$ and $C = 8$, respectively. In addition, since each DU is directly connected to its two adjacent DUs through unidirectional links in DER, the entire bandwidth concentrated on CPU is distributed to each DU, significantly relieving the pressure on the interface of CPU in the DBP architecture. On the other hand, compared to the sequential manner in the ring architecture, the unique approach to data exchange in DER enables parallel computation among DUs, which greatly reduces latency and thus improves computation efficiency.

In summary, the proposed DER architecture involves parallel operations of all DUs and completely bypasses CPU involvement, which is more in line with the original principle of distributed systems design.

VI. SIMULATION

In this section, we perform a comprehensive simulation study to evaluate the performance of the proposed DLAS-aided detection in uplink large-scale MIMO systems. We begin by comparing the BER performance of DLAS-aided detection with other distributed detection schemes across Rayleigh and Rician channels. Then, we contrast DLAS-aided detection schemes with the ML detection to demonstrate the full diversity gain as established in Theorem 3. Finally, we examine the impact of different C settings and compare the bandwidth cost between DBP and proposed DER architecture. A summary of the simulation scenarios for each figure is presented in Table III. All simulations are performed on uncoded systems with 10,000 Monte Carlo trials.

Fig. 5 presents a comparison of detection performance between the proposed DLAS-aided detection and other distributed detection schemes in a 128×32 uncoded large-scale MIMO system operating over a Rayleigh channel with $C = 4$ and 16-QAM. The centralized MMSE detection and distributed schemes such as CG [6], ADMM [7], DN [8], CD [9], SGD [10], ASGD [11], GRLS [12], LAMA [15], GMP [16], and EPA [17] are employed for comparison. Throughout the context, the number of DUs of the SGD, and ASGD algorithms is set to $C = N_r$ due to the constraint $B = 1$, and the CG, ADMM, DN, CD, LAMA, GMP, and EPA algorithms

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3541209

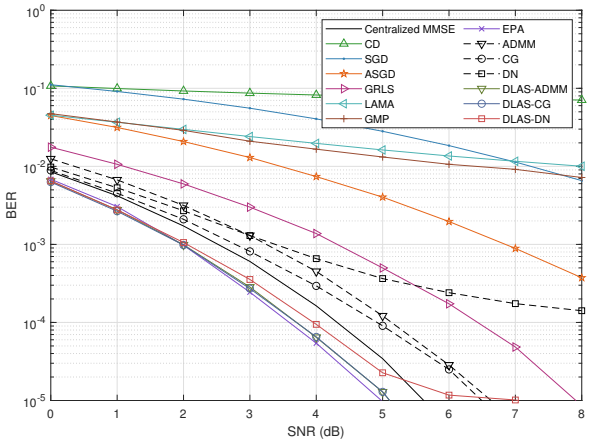


Fig. 7. Bit error rate performance comparison of different methods for the uncoded 256×32 large-scale MIMO system with imperfect CSI using $C = 8$ and 16-QAM.

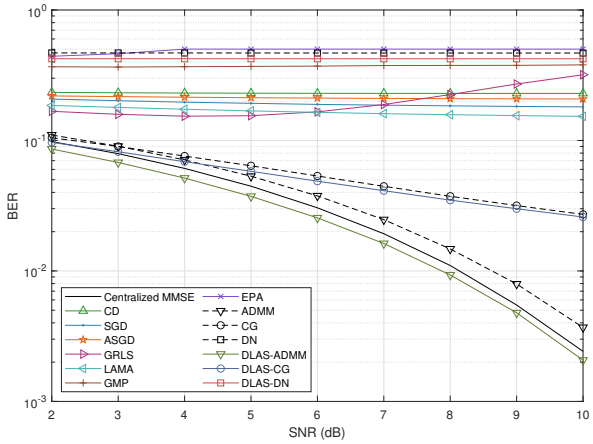


Fig. 8. Bit error rate performance comparison of different methods for the uncoded 128×32 large-scale MIMO system over Rician channels using $C = 4$ and 16-QAM.

are applied with $C = 4$ as the comparison, corresponding to $B = N_r/4$ in each DU. As in [10], the step-size of SGD is set to 0.025 and the averaging procedure in ASGD starts at $k_0 = N_r/2$. ADMM is applied with $\rho = 3$ and $\gamma = 2$. Moreover, GRLS employs configurations of $C = 64$ and $C_0 = 12$ as recommended in [12]. The number of inner iterations for CD, LAMA, GMP, and EPA schemes is set to $I_{\max} = 4$. Meanwhile, the numbers of outer iterations $T_{\max}$ for ADMM, CG, and EPA schemes are set to 4 and for DN to 8. Specifically, the DLAS mechanism is integrated into distributed iterative algorithms such as ADMM, CG, and DN, initialized with $\mathbf{x}^0 = \mathbf{x}^{T_{\max}}$, leading to the DLAS-ADMM, DLAS-CG, and DLAS-DN algorithm, respectively. Due to the high scattering at BS antennas, the local estimation results obtained by CD, LAMA, GMP, and DN algorithms are inaccurate, leading to significant performance loss. In addition, although SGD, ASGD, and GRLS algorithms require lower complexity and bandwidth, they all suffer significant performance degradation compared to MMSE detection. Despite its high complexity and bandwidth requirements, the state-of-the-art EPA scheme achieves a 1.2dB gain over the MMSE detection at BER = 1×10^{-4}. It is evident that the traditional CG algorithm suffers from an SNR loss of nearly 3 dB at a BER of 2×10^{-3} when compared to MMSE detection. However, with the assistance of DLAS, DLAS-CG not only outperforms MMSE detection but also achieves a gain exceeding 3 dB at the same BER level. Moreover, the BER performance of DLAS-ADMM gradually approaches that of EPA, but with considerably lower complexity overhead.

Fig. 6 extends the BER performance comparison to a 256×32 uncoded large-scale MIMO system with $C = 8$ and 16-QAM, employed with the same parameters as in Fig. 5. While the convergence performance of CD, SGD, ASGD, GRLS, LAMA, and GMP algorithms improves, their BER performance in uplink large-scale MIMO systems still remains unsatisfactory. In addition, the ADMM, CG, and DN detection schemes narrow the gap to MMSE detection over iterations, incurring minor performance degradation. Besides, EPA outperforms other decentralized algorithms with higher complexity cost. Clearly, with the help of DLAS, all of the DLAS-aided detection schemes exceeds the linear MMSE performance threshold. For example, the proposed DLAS-ADMM and DLAS-DN detection algorithms achieve gains of nearly 0.7 dB and 0.4 dB, respectively, over MMSE detection at the BER of 10^{-4}. Moreover, the detection performance of the DLAS-CG algorithm gradually surpasses the EPA scheme, but with significantly reduced complexity and bandwidth requirements.

Conversely, Fig. 7 complements Fig. 6 by evaluating the BER performance of the proposed DLAS-aided detection under imperfect channel state information (CSI) in a 256×32 uncoded large-scale MIMO system using 16-QAM. Specifically, we model the imperfect CSI at the receiver as

$$\widehat{\mathbf{H}} = \mathbf{H} + \triangle\mathbf{H}, \tag{53}$$

where $\triangle\mathbf{H} \in \mathcal{CN}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{N_r})$ represents the channel estimation errors, with $\sigma_e^2 = \frac{K}{n_p E_p}$ [37]. Here, n_p and E_p denote the number and power of pilot symbols, respectively, and we set $n_p E_p = 160$ (i.e., $\sigma_e^2 = 0.1$) for the simulation. Compared to the perfect CSI results shown in Fig. 6, the performance of all detection schemes in Fig. 7 degrades under imperfect CSI conditions. Nonetheless, the performance gains achieved by the proposed DLAS mechanism remain evident. With the assistance of DLAS, all DLAS-aided detection schemes not only surpass the centralized MMSE detection, but also demonstrate substantial performance improvements over their respective counterparts.

In addition to Rayleigh channels, we also examine the impact of Rician channels on large-scale MIMO systems to assess the convergence performance of the proposed DLAS-aided detection schemes in Fig. 8. Following the Rician channel setup in [38], the channel matrix is configured as

$$\widehat{\mathbf{H}} = \mathbf{H}_{\mathrm{LOS}}[\Omega(\Omega+\mathbf{I}_K)^{-1}]^{1/2} + \mathbf{H}_{\mathrm{NLOS}}[(\Omega+\mathbf{I}_K)^{-1}]^{1/2}, \tag{54}$$

where Ω is a $K \times K$ diagonal matrix with $\Omega_{k,k} = \mathcal{K}_k$,

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3541209

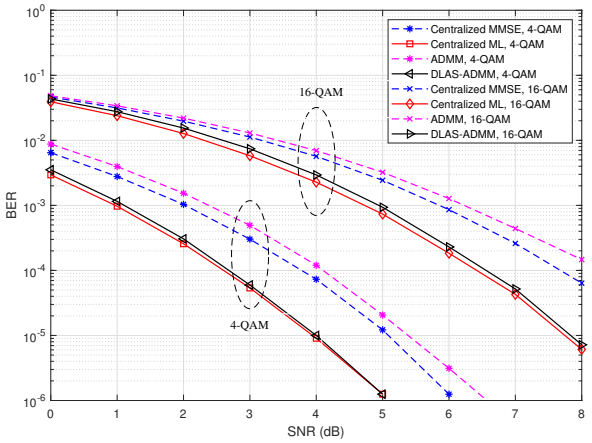


Fig. 9. Bit error rate versus average SNR per bit for the uncoded 64×16 large-scale MIMO system with $C = 4$.

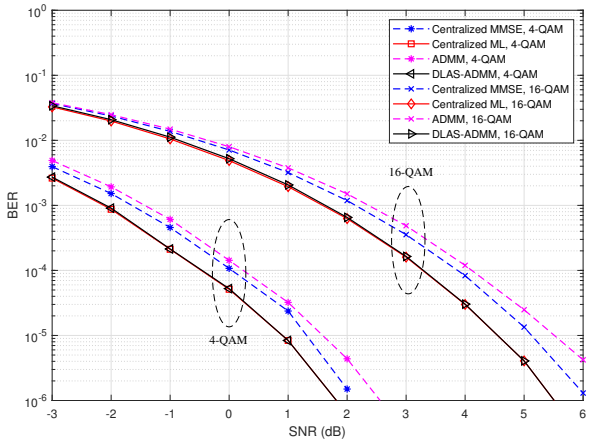


Fig. 10. Bit error rate versus average SNR per bit for the uncoded 128×16 large-scale MIMO system with $C = 8$.

representing the $\mathcal{K}$-factor for k-th user, which determines the ratio of power gains between the line-of-sight (LoS) and non-line-of-sight (NLoS) components. The LoS component is defined as $\mathbf{H}_{\mathrm{LOS}_{n,k}} = e^{-j(n-1)\pi \sin(\theta_k)}$, where θ_k denotes the angle of arrival (AoA) for k-th user, uniformly distributed over $[0, 2\pi)$. The NLoS component, $\mathbf{H}_{\mathrm{NLOS}_{n,k}}$, follows an i.i.d $\mathcal{CN}(0, 1)$ distribution. Consistent with [39], we assume the same $\mathcal{K}$-factor for all users, setting $\mathcal{K} = 5$ and $T_{\max} = 6$ to simulate a highly LoS-dominant environment. Compared to the Rayleigh channel results in Fig. 5, all detection schemes exhibit significant performance degradation under the Rician channel conditions, as shown in Fig. 8. This is expected, as the detection schemes were originally designed for large-scale MIMO systems with Rayleigh fading channels. Notably, DN, DLAS-DN and EPA fail to converge due to poor local estimation in this scenario. However, the DLAS mechanism effectively enhances both DLAS-ADMM and DLAS-CG, with significantly low complexity and bandwidth costs. Specifically, the proposed DLAS-ADMM achieves superior BER performance compared to state-of-the-art decentralized schemes under Rician channel conditions.

Fig. 9 evaluates the performance of DLAS-ADMM across 4-QAM and 16-QAM modulation schemes in a 64×16 uncoded large-scale MIMO system with $C = 4$. For a comprehensive comparison, we also present the BER performance of centralized MMSE, ML detection (implemented through classic sphere decoding [40]), and decentralized ADMM with $T_{\max} = 5$. It is evident that both ADMM and MMSE detectors suffer from significant performance degradation relative to ML detection due to their inability to fully exploit antenna diversity. In this case, the proposed DLAS mechanism is recommended for better performance. Clearly, DLAS-ADMM achieves asymptotic ML performance in both 4-QAM and 16-QAM modulation schemes, since it exploits full received diversity as the ML detector, which is in line with the result derived in Theorem 3.

Fig. 10 extends the BER performance comparison to a 128×16 uncoded large-scale MIMO system under 4-QAM and 16-QAM modulation with $C = 8$. Benefiting from full received diversity, the performance curve of the proposed DLAS-ADMM detection closely matches that of ML detector in both 4-QAM and 16-QAM scenarios, indicating its ability to achieve the quasi-ML performance.

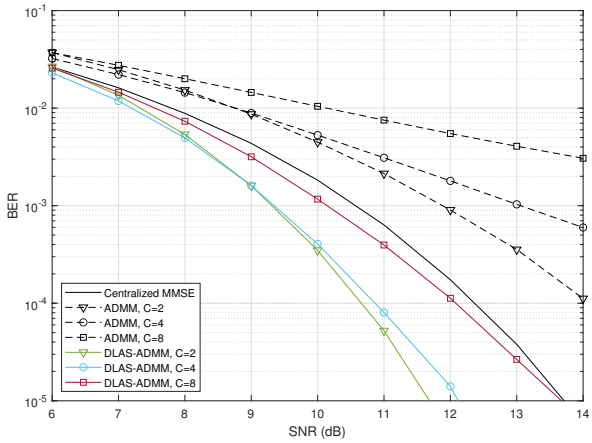


Fig. 11. Bit error rate versus average SNR per bit for the uncoded 128×64 large-scale MIMO system with 16-QAM.

Fig. 11 illustrates the impact of varying C on the BER performance in an uncoded 128×64 large-scale MIMO system. The selection of C depends on factors like BS coverage, user density, and channel characteristics, typically set to $C = 2$, 4, or 8 for DBP and FD [7]–[9], [14]. While a higher C reduces raw data for signal processing and lowers computation and storage demands per unit, it may also lead to overall performance degradation and increased system complexity and bandwidth costs. Specifically, the number of outer iteration $T_{\max}$ is set to 3, 5, and 7 for $C = 2$, 4, and 8, respectively. Notably, despite the additional iterations, decentralized ADMM detection schemes experience greater performance losses compared to MMSE as C increases. This is primarily due to the challenges of managing local information with a larger number of DUs. However, the performance gains

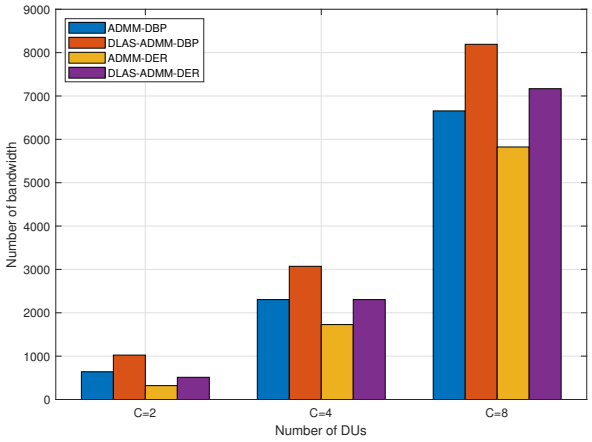


Fig. 12. Bandwidth comparison of DBP and DER for the uncoded 128×64 large-scale MIMO system.

offered by the DLAS mechanism are significant. Regardless of the value of C, the proposed DLAS-ADMM detection schemes consistently outperform centralized MMSE detection while also reducing complexity and bandwidth costs.

Fig. 12 compares the bandwidth cost of the DER and DBP architectures under the same configuration as in Fig. 12. The proposed DER architecture effectively eliminates CPU involvement, resulting in a reduction of bandwidth from C to $C-1$, as outlined in Table II. Specifically, the ADMM and DLAS-ADMM schemes based on the DER architecture achieve reductions in bandwidth of 50%, 25%, and 12.5% compared to their DBP-based counterparts for $C = 2$, 4, and 8, respectively. More precisely, when $C = 4$ and $T_{\max} = 4$, DLAS-ADMM in DER attains the same bandwidth cost as ADMM in DBP, yet provides significant performance enhancements, as shown in Fig. 11, achieving a gain of nearly 3.7 dB at BER = 1×10^{-3}.

VII. Conclusion

In this paper, we proposed a novel near-ML scheme, termed DLAS, by modifying and decentralizing the traditional 1-LAS algorithm for distributed large-scale MIMO detection. The DLAS mechanism achieves low computational complexity and significantly reduces data bandwidth requirements. Leveraging the properties of Rayleigh channels, we analyzed the equivalent noise, post-processing SNR, and diversity order for the proposed DLAS-aided detection, which indicates that it achieves the same diversity as the ML detection and matches the optimal performance with the increment of SNR. Moreover, we designed a novel DER architecture to minimize bandwidth requirements and enable parallel processing, making the proposed DLAS-aided detection highly suitable for distributed large-scale MIMO systems.

The high data rate potential of millimeter-wave (mmWave) technology necessitates advanced signal processing and antenna design strategies to mitigate severe path loss and complex channel characteristics, ensuring its benefits are fully realized. Additionally, exploring the error propagation in DUs presents a promising direction for further research.

References

[1] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO Detection Techniques: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 4, pp. 3109-3132, Aug. 2019.

[2] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118-125, Aug. 2020.

[3] I. Tomkos, D. Klonidis, E. Pikasis, and S. Theodoridis, "Toward the 6G networkera: Opportunities and challenges," *IT Professional*, vol.22, no.1, pp. 34-38, Jan./Feb. 2020.

[4] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 742-758, Oct. 2014.

[5] M. A. Albreem, A. Alhabbash, A. M. Abu-Hudrouss, and T. A. Almohamad, "Data detection in decentralized and distributed massive MIMO networks," *Comput. Commun.*, vol. 189, pp. 79-99, Mar. 2022.

[6] K. Li, Y. Chen, R. Sharan, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized data detection for massive MU-MIMO on a Xeon Phi cluster," in *Proc. 2016 50th Conf. Rec. Asilomar Conf. Signals Syst. Comput.*, pp. 468-472, Mar. 2016.

[7] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized Baseband Processing for Massive MU-MIMO Systems," *IEEE J. Emerging Sel. Top. Circuits Syst.*, vol. 7, no. 4, pp. 491-507, Nov. 2017.

[8] A. Kulkarni, M. A. Ouameur, and D. Massicotte, "Hardware Topologies for Decentralized Large-Scale MIMO Detection Using Newton Method," *IEEE Trans. Circuits Syst.*, vol. 68, no. 9, pp. 3732-3745, Jul. 2021.

[9] K. Li, O. Castaeda, C. Jeon, J. R. Cavallaro and C. Studer, "Decentralized Coordinate-Descent Data Detection and Precoding for Massive MU-MIMO," in *Proc. 2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1-5, May. 2019.

[10] J. R. Snchez, F. Rusek, O. Edfors, M. Sarajlic, and L. Liu, "Decentralized Massive MIMO Processing Exploring Daisy-Chain Architecture and Recursive Algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 687-700, Jan. 2020.

[11] J. R. Sanchez, F. Rusek, M. Sarajlic, O. Edfors and L. Liu, "Fully Decentralized Massive MIMO Detection Based on Recursive Methods," in *Proc. 2018 IEEE Workshop Signal. Process. Syst. SiPS Des. Implement.*, pp. 53-58, Oct. 2018.

[12] Q. Chen, Z. Wang, C. Ma, X. Dai and D. W. K. Ng, "General Recursive Least Square Algorithm For Distributed Detection In Massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 73, no. 8, pp. 12137-12142, Aug. 2024.

[13] C. Jeon, K. Li, J. R. Cavallaro and C. Studer, "On the achievable rates of decentralized equalization in massive MU-MIMO systems," in *Proc. 2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1102-1106, Aug. 2017.

[14] C. Jeon, K. Li, J. R. Cavallaro, and C. Studer, "Decentralized Equalization With Feedforward Architectures for Massive MU-MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4418-4432, Jul. 2019.

[15] K. Li, C. Jeon, J. R. Cavallaro and C. Studer, "Decentralized equalization for massive MU-MIMO on FPGA," in *Proc. 2017 51st IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 1532-1536, Oct. 2017.

[16] Z. Zhang, Y. Dong, K. Long, X. Wang, and X. Dai, "Decentralized Baseband Processing With Gaussian Message Passing Detection for Uplink Massive MU-MIMO Systems," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2152-2157, Feb. 2022.

[17] Z. Zhang, H. Li, Y. Dong, X. Wang, and X. Dai, "Decentralized Signal Detection via Expectation Propagation Algorithm for Uplink Massive MIMO Systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11233-11240, Oct. 2020.

[18] Y. Dong, H. Li, C. Gong, X. Wang and X. Dai, "An Enhanced Fully Decentralized Detector for the Uplink M-MIMO System," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 13030-13042, Dec. 2022.

[19] N. Li and P. Fan, "Distributed Cell-Free Massive MIMO Versus Cellular Massive MIMO Under UE Hardware Impairments," *Chin. J. Electron.*, vol. 33, no. 5, pp. 1274-1285, Sept. 2024.

[20] K. V. Vardhan, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity detector for large MIMO systems and multicarrier CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 473-485, Apr. 2008.

[21] S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low complexity near-ML performance achieving algorithm for large MIMO detection," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2008, pp. 2012-2016.

[22] B. S. Rajan, S. Mohammed, A. Chockalingam, and N. Srinidhi, "Low complexity near-ML decoding of large non-orthogonal STBCs using reactive tabu search," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2009, pp. 1993-1997.

[23] N. Srinidhi, T. Datta, A. Chockalingam, and B. S. Rajan, "Layered tabu search algorithm for large-MIMO detection and a lower bound on ML performance," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 2955-2963, Nov. 2011.

[24] P. Li and R. D. Murch, "Multiple output selection-LAS algorithm in large MIMO systems," *IEEE Commun. Lett.*, vol. 14, no. 5, pp. 399-401, May 2010.

[25] Z. Wang, J. Xu, and X. Tao, "Improved LAS detection based on grouped genetic algorithm for massive MIMO system," in *Proc. 2017 IEEE/CIC International Conference on Communications in China (ICCC)*, Oct. 2017, pp. 1-6.

[26] A. K. Sah and A. K. Chaturvedi, "An Unconstrained Likelihood Ascent Based Detection Algorithm for Large MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2262-2273, April 2017.

[27] M. L. Ammari, C. Issa and J. -Y. Chouinard, "LAS Receiver Exploiting Channel Hardening for Massive MIMO Systems," in *Proc. 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Sept. 2019, pp. 1-5.

[28] E. Aslan and M. E. Celebi, "A Low-Complexity Detector for Very Large MIMO," in *Proc. 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May. 2016, pp. 1-5.

[29] A. K. Sah and A. K. Chaturvedi, "Reduced Neighborhood Search Algorithms for Low Complexity Detection in MIMO Systems," in *Proc. 2015 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2015, pp. 1-6.

[30] A. K. Sah and A. K. Chaturvedi, "Sequential and Global Likelihood Ascent Search-Based Detection in Large MIMO Systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 713-725, Feb. 2018.

[31] Z. Wang, R. M. Gower, Y. Xia, L. He and Y. Huang, "Randomized Iterative Methods for Low-Complexity Large-Scale MIMO Detection," *IEEE Trans. Signal Process.*, vol. 70, pp. 2934-2949, Jun. 2022.

[32] Z. Wang, W. Xu, Y. Xia, Q. Shi and Y. Huang, "A New Randomized Iterative Detection Algorithm for Uplink Large-Scale MIMO Systems," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5093-5107, Sept. 2023.

[33] A. Krishnamoorthy and D. Menon, "Matrix inversion using Cholesky decomposition," in *Proc. 2013 Signal Process. Algorithms, Architect., Arrange., Appl. Conf. Proc., SPA*, pp. 70-72, Sept. 2013.

[34] A. Papoulis, *Probability, random variables, and stochastic processes*, 4th ed. New York, NY, USA: McGraw-Hill, Dec. 2002.

[35] D. A. Gore, R. W. Heath, and A. J. Paulraj, "Transmit selection in spatial multiplexing systems," *IEEE Comm. Letters*, vol. 6, no. 11, pp. 491-493, Nov. 2002.

[36] A. Paulraj, R. U. Nabar, and D. Gore, *Introduction to Space Time Wireless Communications*, Cambridge Univ. Press, Cambridge (UK), 2003.

[37] Q. Chen, S. Zhang, S. Xu, and S. Cao, "Efficient MIMO detection with imperfect channel knowledge - A deep learning approach," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2019, pp. 1-6.

[38] T. Liu, J. Tong, Q. Guo, J. Xi, Y. Yu and Z. Xiao, "On the Performance of Massive MIMO Systems With Low-Resolution ADCs and MRC Receivers Over Rician Fading Channels," in *IEEE Syst. J.*, vol. 15, no. 3, pp. 4514-4524, Sept. 2021.

[39] C. Li, J. Yao, H. Wang, U. Ahmed and S. Du, "Effect of Mobile Wireless on Outage and BER Performances Over Rician Fading Channel," *IEEE Access*, vol. 8, pp. 91799-91806, 2020.

[40] M. O. Damen, H. E. Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," in *IEEE Trans. Inform. Theory*, vol. 49, pp. 2389-2401, Oct. 2003.

Qiqiang Chen received the B.S. degree in information engineering from the School of Information Science and Engineering, Southeast University, Nanjing, China, in 2023. He is currently pursuing a Ph.D. degree in signal and information processing at Southeast University. His research interests include massive MIMO systems and decentralized signal processing.

Zheng Wang (Senior Member, IEEE) received the B.S. degree in electronic and information engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, and the M.S. degree in communications from University of Manchester, Manchester, U.K., in 2010. He received the Ph.D degree in communication engineering from Imperial College London, UK, in 2015.

Since 2021, he has been an Associate Professor in the School of Information and Engineering, Southeast University (SEU), Nanjing, China. From 2015 to 2016, he served as a Research Associate at Imperial College London, UK. From 2016 to 2017, he was a senior engineer with the Radio Access Network R&D division, Huawei Technologies Co. From 2017 to 2020, he was an Associate Professor at the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China. His current research interests include massive MIMO systems, machine learning and data analytics over wireless networks, and lattice theory for wireless communications.

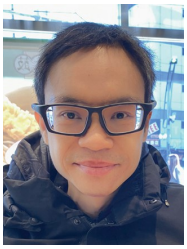
Chenhao Qi (Senior Member, IEEE) received the B.S. degree (Hons.) in information engineering from the Chien-Shiung Wu Honored College, Southeast University, China, in 2004, and the Ph.D. degree in signal and information processing from Southeast University in 2010. From 2008 to 2010, he visited the Department of Electrical Engineering, Columbia University, New York, USA.

Since 2010, he has been a Faculty Member with the School of Information Science and Engineering, Southeast University, where he is currently a Professor and the Head of Jiangsu Multimedia Communication and Sensing Technology Research Center. He received Best Paper Awards from IEEE GLOBECOM in 2019, IEEE/CIC ICCC in 2022, and the 11th International Conference on Wireless Communications and Signal Processing (WCSP) in 2019. He has served as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, and China Communications.

Zhen Gao (Member, IEEE) received the B.S. degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in communication and signal processing from the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China, in 2016. He is currently an Associate Professor with the Beijing Institute of Technology. His research interests include wireless communications, with a focus on multi-carrier modulations, multiple antenna systems, and sparse signal processing. He was a recipient of the IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (Best Paper), an Exemplary Reviewer of IEEE COMMUNICATION LETTERS in 2016, IET Electronics Letters Premium Award (Best Paper) in 2016, and the Young Elite Scientists Sponsorship Program (from 2018 to 2021) from the China Association for Science and Technology.

Yongming Huang (Fellow, IEEE) received the B.S. and M.S. degrees from Nanjing University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 2007.

Since March 2007 he has been a faculty in the School of Information Science and Engineering, Southeast University, China, where he is currently a full professor. He has also been the Director of the Pervasive Communication Research Center, Purple Mountain Laboratories, since 2019. During 2008-2009, Dr. Huang visited the Signal Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden. His current research interests include intelligent 5G/6G mobile communications and millimeter wave wireless communications. He has published over 200 peer-reviewed papers, hold over 80 invention patents. He submitted around 20 technical contributions to IEEE standards, and was awarded a certificate of appreciation for outstanding contribution to the development of IEEE standard 802.11aj. He served as an Associate Editor for the IEEE Transactions on Signal Processing and a Guest Editor for the IEEE Journal Selected Areas in Communications. He is currently an Editor-at-Large for the IEEE Open Journal of the Communications Society and an Associate Editor for the IEEE Wireless Communications Letters.



Dusit Niyato (Fellow, IEEE) received B.Eng. from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand and Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada. He is currently a professor in the College of Computing and Data Science, at Nanyang Technological University, Singapore. His research interests are in the areas of mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.