Massive Access in Cell-Free Massive MIMO-Based Internet of Things: Cloud Computing and Edge Computing Paradigms

Malong Ke, *Student Member, IEEE,* Zhen Gao, *Member, IEEE,* Yongpeng Wu, *Senior Member, IEEE,* Xiqi Gao, *Fellow, IEEE,* and Kat-Kit Wong, *Fellow, IEEE*

***Abstract*—This paper studies massive access in cell-free massive multi-input multi-output (MIMO)-based Internet of Things and solves the challenging active user detection (AUD) and channel estimation (CE) problems. For the uplink transmission, we propose an advanced frame structure design to reduce the access latency. Moreover, by considering the cooperation of all access points (APs), we investigate two processing paradigms at the receiver for massive access: cloud computing and edge computing. For cloud computing, all APs are connected to a centralized processing unit (CPU), and the signals received at all APs are centrally processed at the CPU. While for edge computing, the central processing is offloaded to part of APs equipped with distributed processing units, so that the AUD and CE can be performed in a distributed processing strategy. Furthermore, by leveraging the structured sparsity of the channel matrix, we develop a structured sparsity-based generalized approximated message passing (SS-GAMP) algorithm for reliable joint AUD and CE, where the quantization accuracy of the processed signals is taken into account. Based on the SS-GAMP algorithm, a successive interference cancellation-based AUD and CE scheme is further developed under two paradigms for reduced access latency. Simulation results validate the superiority of the proposed approach over the state-of-the-art baseline schemes. Besides, the results reveal that the edge computing can achieve the similar massive access performance as the cloud computing, and the edge computing is capable of alleviating the burden on CPU, having a faster access response, and supporting more flexible AP cooperation.**

Manuscript received February 1, 2020; revised June 9, 2020; accepted July 17, 2020. The work of Z. Gao was supported in part by the National Natural Science Foundation of China under Grant 61701027, in part by Beijing Natural Science Foundation under Grants 4182055 and L182024, in part by Young Elite Scientists Sponsorship Program by CAST, and in part by Talent Innovation Project of BIT. The work of Y. Wu was supported in part by the National Key R&D Program of China under Grant 2018YFB1801102, in part by JiangXi Key R&D Program under Grant 20181ACE50028, in part by National Science Foundation under Grant 61701301, in part by Young Elite Scientist Sponsorship Program by CAST, and in part by the open research project of State Key Laboratory of Integrated Services Networks (Xidian University) under Grant ISN20-03. The work of X. Gao was supported by the National Key R&D Program of China under Grant 2018YFB1801103. *(Corresponding author: Zhen Gao.)*

M. Ke and Z. Gao are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China, and also with the Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: kemalong@bit.edu.cn; gaozhen16@bit.edu.cn).

Y. Wu is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xian 710071, China (e-mail: yongpeng.wu@sjtu.edu.cn).

X. Gao is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: xqgao@seu.edu.cn).

K.-K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: kai-kit.wong@ucl.ac.uk).

***Index Terms*—Massive access, cell-free massive MIMO, cloud computing, edge computing, active user detection, structured sparsity.**

I. Introduction

With the advent of the Internet-of-Things (IoT) era, massive machine-type communications (mMTC) have been identified as the indispensable services in future wireless networks [1], [2]. Against this background, the future base stations (BSs) are expected to enable massive connectivity with billions of user equipments (UEs). However, the reliable support of low-latency massive access for mMTC is still challenging in current wireless networks [3]. On the one hand, assigning orthogonal pilot sequences to all potential UEs would be impractical for massive access. On the other hand, for traditional grant-based random access protocols, the complex signaling information interaction would lead to the extremely high access latency when the number of UEs becomes large [4]. Fortunately, a key characteristic of mMTC is the sporadic traffic of UEs, i.e., among a large pool of UEs, only a small fraction are active in any given time interval [5]. Hence, the grant-free random access protocol is recently proposed as a promising alternative, where each active UE transmits its pilots and data to the BS simultaneously without scheduling in advance [6]. In grant-free random access, the BS has to utilize the received pilot signals to detect the active UEs and estimate their channels, which are vital for the subsequent data detection [7]. However, due to the large number of UEs but the limited radio resources for massive access, the active user detection (AUD) has been emerging as a challenging problem [5]–[7].

Moreover, since the power limited IoT UEs are usually distributed in a vast area, multiple BSs should cooperate to offer a better coverage and to save the transmit power of UEs. Different from the massive access for single-BS scenarios, the multiple BS scenarios pose a new massive access problem known as "multi-cell massive access" [8] or "random access for crowded massive MIMO systems" [9], [10]. For traditional network architecture, each BS operates independently to perform AUD and channel estimation (CE) for the UEs distributed in its own cell while treating the inter-cell interference as noise [8]. Consequently, the inter-cell interference is a severely limiting factor for reliable massive access. Fortunately, the promising cell-free massive MIMO network brings new opportunities to facilitate the massive access, where the massive

MIMO BSs are regarded as access points (APs) and deployed in a vast area to serve massive IoT UEs, and these APs are connected to one or multiple processing units for joint signal processing [1], [11]. Since there are no "cells" or "cell boundaries", the inter-cell interference can be avoided. However, the design of an efficient AUD and CE scheme for grant-free massive access in cell-free massive MIMO systems is still an open issue.

A. Related Work

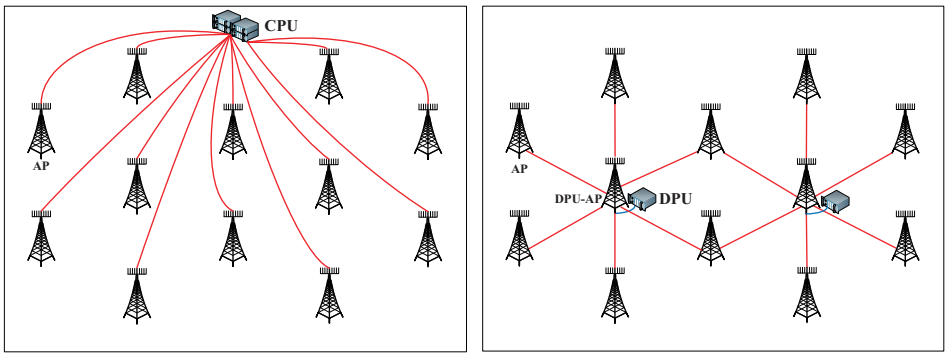
Exploiting the sparse UE activity, several compressive sensing (CS)-based approaches have been proposed to detect active UEs for grant-free massive access. In [12], a CS-based multi-user detection method was suggested, where the concerned AUD was formulated as a sparse signal recovery problem. But this method only considered the detection in one symbol period. In typical massive access scenarios, the active UEs generally transmit uplink signals in several successive time slots [13]. By assuming the UE activity remains unchanged in several adjacent time slots, the authors in [14] proposed a structured iterative support detection algorithm to jointly detect the active UEs and the transmitted data, where the structured sparsity pattern observed in multiple time slots was leveraged for improved detection performance. However, for practical IoT applications, the UEs can randomly access or leave the system, which yields a time-varying UE activity. On the other hand, although the active UE set (AUS) can be changed over time, the variation would be gradual [15]. This leads to the temporal correlation of UE activity within several successive time slots. Hence, a dynamic CS-based multi-user detection approach was proposed in [16], where the AUS obtained in current time slots was used as the a priori information to estimate AUS in the next time slot. However, the solution [16] assumes the availability of the sparsity level, i.e., the number of active UEs, which can be unrealistic. To overcome this shortcoming, in [17], the authors developed an efficient prior-information-aided adaptive subspace pursuit algorithm to detect active UEs without the knowledge of the sparsity level. Furthermore, by leveraging the a priori information of the transmitted signals, the authors of [18] proposed an approximate message passing (AMP)-based joint AUD and data detection scheme for further improved performance.

The solutions [12]–[18] focus on joint AUD and data detection, which assume the availability of perfect channel state information (CSI). In practice, the channels between the active UEs and the BS should be estimated before the following coherent data detection. Based on the idea of the orthogonal matching pursuit, the authors of [19] proposed an efficient greedy algorithm to realize joint AUD and CE, where only single-antenna is considered at the BS. The analysis and numerical results in [20] reveal that the detection error probability of AUD can always be driven to zero by equipping a large-scale antenna array at the BS. Against this background, an advanced grant-free massive access scheme was developed for multi-antenna systems [21], where both sparse UE activity and the sparsity of the delay-domain channel impulse response (CIR) were leveraged for facilitating AUD and CE. To reduce the computational complexity in the case of a large number of UEs and antennas, a dimension reduction-based joint AUD and CE approach was further proposed in [22]. The solutions [19], [21], [22] are developed from the CS greedy (non-Bayesian) algorithms to achieve the sparse signal recovery, where the a priori distribution of the channels is not taken into account. By exploiting the statistical information of the massive access channels based on the Bayesian inference framework, the authors in [23] developed an AMP-based access scheme, which could significantly improve the AUD and CE performance compared to the greedy approaches. Besides, an expectation propagation-based scheme was proposed in [24] for further enhanced performance. However, the work [23], [24] assumes that the noise variance and the parameters of the a priori distribution of channels are known in advance. In [25], an expectation maximization (EM) algorithm was incorporated into the AMP-based scheme to learn the unknown hyper-parameters. Meanwhile, the structured sparsity of the massive access channel matrix observed at multiple BS antennas was leveraged to improve AUD performance. Furthermore, the joint AUD and CE for massive access was further extended to the cloud radio network architecture [26] and multi-cell massive access scenarios [8].

B. Main Contributions

In this paper, we investigate grant-free massive access in cell-free massive MIMO-based IoT, where orthogonal frequency division multiplexing (OFDM) technique is employed for uplink transmission. Specifically, we first propose a frame structure design for massive access and then compare two processing paradigms consisting of cloud computing and edge computing for the practical processing of AUD and CE. Due to the limited capacity of the backhaul links between APs and processing units, we further consider the quantization of the APs' received signals. For both paradigms, by exploiting the sporadic traffic of UEs and the angular-domain sparsity of massive MIMO channels, the AUD and CE problems are formulated as two CS problems based on the spatial-domain and angular-domain channel models, respectively. Subsequently, a structured sparsity-based generalized AMP (SS-GAMP) algorithm is developed for CS recovery, where the quantization of the processed signals is considered. On this basis, a successive interference cancellation (SIC)-based AUD and CE algorithm is developed for alternately detecting active UEs and estimating their channels. Our main contributions can be summarized as follows:

- **Massive access in cell-free massive MIMO-based IoT:** We propose to employ the promising cell-free massive MIMO to support massive connectivity service in future IoT applications, and the related massive access problem is investigated. Different from the well studied single-cell massive access [19]–[25], we consider a more general mMTC scenario, where UEs are distributed in a large area and multiple APs cooperate to offer a wide coverage range. Furthermore, compared to the traditional network architecture [8], which extends the aforementioned problem to the multi-cell massive access, the cell-free massive MIMO shows its superiority in combating inter-cell inter-



(a) (b)

Fig. 1. Two processing paradigms for cell-free massive MIMO-based IoT: (a) Cloud computing; (b) Edge computing.

ference, the better massive access performance, and more flexible AP cooperation.

- **A frame structure design for low-latency massive access:** In grant-free massive access, for a specific frame of the uplink signals, the time-frequency radio resource is divided into multiple resource elements to transmit pilots and payload data. We propose a frame structure tailored for massive access with OFDM transmission, where an advanced resource division strategy is considered. Compared to the conventional frame structure in [25], the proposed frame structure reaps a significant access latency reduction.
- **Cloud computing and edge computing processing paradigms for the proposed scheme:** We introduce two network architectures for cell-free massive MIMO systems to support the cloud computing-based and edge computing-based signal processing, respectively. For the proposed massive access scheme, the AUD and CE performance of edge computing can approach that of cloud computing. Moreover, edge computing has the potential to offload the computational burden from the central processing unit (CPU) in cloud computing to multiple distributed processing units (DPUs) and reduce the cooperation cost (e.g., backhaul cost and response time), but increases the price that part of APs should employ DPUs.
- **SS-GAMP algorithm:** Existing CS-based massive access schemes [8], [12]–[25] only consider the ideal processed signals with infinite-resolution quantization. By contrast, the proposed SS-GAMP algorithm provides a general framework to achieve joint AUD and CE, where the quantization of the processed signals is considered. Hence, for processed signals after low-resolution quantization due to the limited capacity of the wireless backhaul, the proposed algorithm has a better massive access performance than conventional algorithms in [8], [25]. Moreover, we propose a weighted message refining strategy to leverage the sparsity properties of the channel matrix, which can further improve the performance in contrast to the strategy in [25].
- **SIC-based AUD and CE algorithm:** This algorithm consists of three modules: spatial-domain AUD, angular-domain CE, and the identified UE cancellation. These three modules are executed alternately in an iterative manner. In contrast to the spatial-domain joint AUD and CE solutions without SIC [19]–[24], this algorithm can dramatically reduce the access latency by further leveraging the angular domain sparsity of massive MIMO channels and the idea of SIC.

Notations: We use normal-face letters to denote scalars, lowercase (uppercase) boldface letters to denote column vectors (matrices). The (k,m)-th element, the k-th row vector, and the m-th column vector of the matrix $\mathbf{H} \in \mathbb{C}^{K\times M}$ are denoted as $[\mathbf{H}]_{k,m}$, $[\mathbf{H}]_{k,:}$, and $[\mathbf{H}]_{:,m}$, respectively. $\{\mathbf{H}_n\}_{n=1}^{N}$ denotes a matrix set with the cardinality of N and $\mathbf{0}_{K\times M}$ is the zero matrix of size $K \times M$. The superscripts $(\cdot)^{\mathrm{T}}$, $(\cdot)^*$, and $(\cdot)^{\mathrm{H}}$ represent the transpose, complex conjugate, and conjugate transpose operators, respectively. $[K]$ denotes the set of integers $\{1, 2, \cdots, K\}$, $|\mathcal{A}|_c$ is the cardinal number of set $\mathcal{A}$, $\emptyset$ is the empty set, and $\mathrm{supp}\{\cdot\}$ denotes the support set of a sparse vector or matrix. $\lceil b \rfloor$ rounds b to the nearest integer greater than or equal to b. $\mathcal{U}(x; a, b)$ denotes the variable x follows the uniform distribution between a and b. Finally, $\mathcal{CN}(x; \mu, v)$ denotes the complex Gaussian distribution of a random variable x with mean μ and variance v, and $\mathbb{E}[\cdot]$ denotes statistical expectation operator.

II. System Model

In this section, we first introduce two processing paradigms for cell-free massive MIMO-based IoT. Subsequently, we detail the procedure, the proposed frame structure, and the related signal model for massive access in cell-free massive MIMO systems. Finally, the sparsity properties of the massive access channel matrix represented in the spatial and angular domains are illustrated.

A. Proposed Cell-Free Massive MIMO-Based IoT

Consider a typical cell-free massive MIMO system to serve massive IoT UEs, where quantities of APs equipped with massive antennas cooperate in the network to serve a vast area, as illustrated in Fig. 1. The APs are connected to the processing unit (i.e., CPU or DPU) via backhaul links, thus the received signals and information obtained at multiple APs

can be jointly processed at the processing units to realize AP cooperation. In this context, the concepts of cell and cell boundary do not exist. Here, we consider two different processing paradigms to enable the AP cooperation for massive access: *(1) Cloud computing paradigm* for centralized cooperation, where all APs will collect the signals from UEs and then transfer them to the CPU far away from the UEs, see Fig. 1(a). The CPU will perform high computational complexity signal processing for the whole network. Since the APs are only designed for receiving and transmitting signals, this architecture can significantly reduce the APs' cost for their large-scale deployment. *(2) Edge computing paradigm* for distributed cooperation, which offloads the signal processing from one CPU to multiple DPUs (also mobile edge computing [MEC] severs) deployed at part of the APs, and these APs are referred to as the DPU-APs, MEC-APs, or fog-APs, see Fig. 1(b). Furthermore, other APs are connected to several adjacent DPU-APs for distributed signal processing. In this case, the processing work is offloaded from the CPU to multiple DPUs at the corresponding DPU-APs. Compared to the cloud computing, this paradigm can alleviate the burden on backhaul links and CPU, and support more flexible signal processing implementation. These advantages make the edge computing-based massive access has a faster access response, while at the cost that the DPU-APs should employ extra DPUs (MEC severs).

Remark 1: Note that most existing cell-free massive MIMO papers consider the cell-free architecture with distributed massive MIMO configuration, i.e., each AP is equipped with one antenna or few antennas [11]. By contrast, this paper considers the co-located massive MIMO configuration, where each AP is equipped with massive antennas, as shown in Fig. 1. Compared with the former one, we believe the cell-free architecture using co-located massive MIMO configuration is more practical, since most commercialized massive MIMO systems are co-located and can be easily upgraded to the cell-free architecture.

B. Massive Access in Cell-Free Massive MIMO Systems

The procedure of the proposed grant-free massive access scheme for cell-free massive MIMO-based IoT can be summarized as follows.

- **Step 1:** During the uplink transmission phase, all active UEs directly transmit their non-orthogonal access pilot sequences and the following payload data to the APs without waiting for the access permission.
- **Step 2:** Each AP collects the received signals over multiple successive time slots, and sends the collected signals to the processing unit, i.e., CPU in cloud computing or the DPUs equipped at the adjacent DPU-APs in edge computing, via backhaul links.
- **Step 3:** By jointly processing the received signals from multiple APs, the processing unit performs AUD and CE for the cell-free massive MIMO-based IoT, and then the obtained AUS and corresponding channel estimates are used for subsequent data detection.

Next, we will detail the proposed technical components.

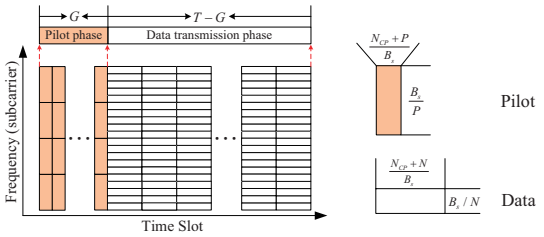


Fig. 2. The proposed frame structure for the uplink transmission in grant-free massive access.

1) The Proposed Frame Structure Design: At the UEs side, we propose an advanced frame structure design to transmit the uplink access pilot sequence and payload data. The proposed frame structure is illustrated in Fig. 2, where the cyclic prefix (CP)-OFDM is employed to combat time dispersive channels and the length of CP is denoted by N_{CP}. By adopting OFDM, the time-frequency radio resource can be divided into multiple resource elements to convey the pilot signals and payload data. Specifically, a frame comprising T time slots is divided into two phases in the time domain, where the first G time slots (i.e., pilot phase) are used to transmit access pilot signals, and the remaining $(T-G)$ time slots (i.e., data phase) are reserved for payload data transmission. In the pilot phase, we consider the OFDM's discrete Fourier transform (DFT) length is $P = N_{CP}$, so that the subcarrier spacing is B_s/P and thus each CP-OFDM symbol's duration is $(N_{CP}+P)/B_s$, where B_s is the two-sided bandwidth. In the data phase, we consider the OFDM symbol's DFT length is $N \gg P$ and thus each CP-OFDM symbol's duration is $(N_{CP}+N)/B_s$. In grant-free massive access, the pilot signals will be used for both AUD and CE, and thus the pilot transmission latency in the proposed scheme is $G(N_{CP}+P)/B_s$. While for the frame structure adopted by existing broadband massive access scheme in [25], the OFDM symbol's DFT lengths in both pilot and data phases are N, thus the corresponding latency required is $G(N_{CP}+N)/B_s$. Compared to the traditional frame structure, the proposed frame structure will significantly reduce the access latency as usually $P \ll N$. For example, we consider $P = N_{CP} = 64$ and $N = 2048$ in the simulations, the proposed frame structure can reap a reduction of approximately 94% in access latency.

2) Received Signal Model at APs: We investigate a massive access problem in cell-free massive MIMO systems, where B APs are employed to serve K UEs, and the UEs are distributed in a vast area. Here, K is usually large (e.g., $K = 10^3$ in [20]). Each AP is equipped with an M_c-antenna uniform linear array (ULA), and each UE has only one antenna without loss of generality. Here we focus on the pilot phase with OFDM's size being P. For the subchannel of the p-th pilot subcarrier $(1 \le p \le P)$, the signal $\mathbf{y}^t_{p,b,k} \in \mathbb{C}^{M_c \times 1}$ received at the b-th AP from the k-th UE in the t-th time slot (i.e., the t-th OFDM symbol) is expressed as

$$\mathbf{y}^t_{p,b,k} = \sqrt{P_k}\mathbf{h}_{p,b,k}s^t_{p,k} + \mathbf{n}^t_{p,b}, \tag{1}$$

where P_k denotes the transmit power of the k-th UE, $\mathbf{h}_{p,b,k} \in \mathbb{C}^{M_c \times 1}$ is the subchannel associated with the k-th UE and the b-th AP, $s^t_{p,k}$ is the uplink access pilot, and $\mathbf{n}^t_{p,b}$ denotes the

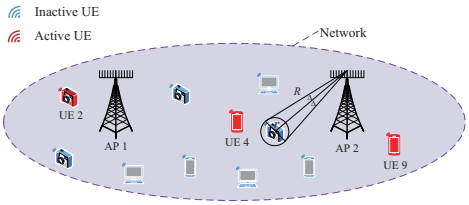


Fig. 3. Sparse UE activity in massive access scenarios. A classical one-ring channel model is considered for the channels between the UEs and the massive MIMO APs.

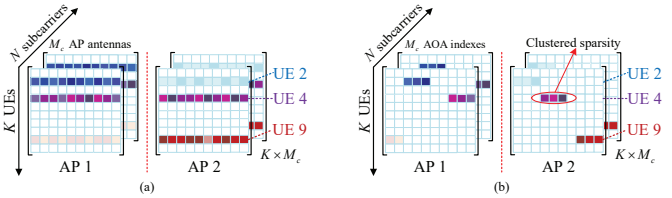


Fig. 4. The massive access channel matrix exhibits two forms of structured sparsity in the spatial and angular domains: (a) Spatial-domain structured sparsity due to sparse UE activity; (b) Angular-domain structured sparsity due to the limited angular spread of the MPCs. A darker color denotes a higher channel gain.

additive white Gaussian noise (AWGN). Due to the sporadic traffic of UEs, within a given time duration, only a small number of UEs are activated and try to transmit uplink signals to the APs, as illustrated in Fig. 3. We define an activity indicator α_k to indicate the UEs' activity, which equals 1 when the k-th UE is active and 0 otherwise. Meanwhile, the set of active UEs is defined as $\mathcal{A} = \{k|\alpha_k = 1, 1 \le k \le K\}$, and the number of active UEs is denoted by $K_a = |\mathcal{A}|_c$. Hence, for the p-th pilot subcarrier and the t-th time slot, the signal received at the b-th AP from all active UEs is given as follows

$$\mathbf{y}_{p,b}^t = \sum\nolimits_{k=1}^{K} \sqrt{P_k}\alpha_k \mathbf{h}_{p,b,k} s_{p,k}^t + \mathbf{n}_{p,b}^t. \tag{2}$$

The channel $\mathbf{h}_{p,b,k}$ can be modeled as $\mathbf{h}_{p,b,k} = \rho_{b,k}\widetilde{\mathbf{h}}_{p,b,k}$, where both the large-scale fading and small-scale fading are taken into account. Here, $\rho_{b,k}$ is the large-scale fading coefficient caused by path loss, and $\widetilde{\mathbf{h}}_{p,b,k}$ is the small-scale fading vector. For the p-th pilot subcarrier, the subchannel between the k-th UE and the b-th AP is modeled as follows [27], [28]

$$\widetilde{\mathbf{h}}_{p,b,k} = \sum\nolimits_{l=1}^{L_{b,k}} \beta_{b,k}^l \mathbf{a}_R\left(\phi_{b,k}^l\right) e^{-j2\pi\tau_{b,k}^l f_p}, \tag{3}$$

where $f_p = -\frac{B_s}{2} + \frac{B_s p}{P}$, $L_{b,k}$ denotes the number of multi-path components (MPCs) between the k-th UE and the b-th AP, $\beta_{b,k}^l$ and $\tau_{b,k}^l$ are the complex path gain and the path delay of the l-th MPC, respectively. The array response vector $\mathbf{a}_R\left(\phi_{b,k}^l\right)$ is given by $\mathbf{a}_R\left(\phi_{b,k}^l\right) = \left[1, e^{-j2\pi\phi_{b,k}^l}, \cdots, e^{-j2\pi(M_c-1)\phi_{b,k}^l}\right]^{\mathrm{T}}$, where $\phi_{b,k}^l = \frac{\widetilde{d}}{\lambda}\sin\left(\varphi_{b,k}^l\right)$. Here, $\varphi_{b,k}^l$ is the angle of arrival (AOA) observed at the AP side, λ denotes the wavelength, and the antenna spacing $\widetilde{d} = \lambda/2$ is considered.

C. Sparsity Properties of the Massive Access Channel Matrix

Define $\mathbf{H}_{p,b} = \left[\sqrt{P_1}\alpha_1\mathbf{h}_{p,b,1}, \cdots, \sqrt{P_K}\alpha_K\mathbf{h}_{p,b,K}\right]^{\mathrm{T}} \in \mathbb{C}^{K\times M_c}$ as the massive access channel matrix between all UEs and the b-th AP at the p-th pilot subcarrier. In this section, we first present the structured sparsity of the spatial-domain channel matrices $\{\mathbf{H}_{p,b}\}_{p=1}^P, \forall b$. Furthermore, by representing the MIMO channels in the virtual angular domain, the structured sparsity of the angular-domain channel matrices $\{\mathbf{W}_{p,b}\}_{p=1}^P, \forall b$, is further illustrated.

1) Spatial-Domain Structured Sparsity: For a typical massive access scenario, only a small number of UEs out of total K UEs are active, i.e., most of $\alpha_k, \forall k$ are equal to 0. Thus, the channel vector $[\mathbf{H}_{p,b}]_{:,m}$ observed at the m-th receive antenna of the b-th AP is sparse, i.e.,

$$\left|\mathrm{supp}\left\{[\mathbf{H}_{p,b}]_{:,m}\right\}\right|_c = K_a \ll K. \tag{4}$$

Moreover, given the UE activity, i.e., the value of α_k, all elements of the k-th row of $\{\mathbf{H}_{p,b}\}_{p=1}^P$ will be zero or non-zero simultaneously. Therefore, the sparsity pattern (4) can be simultaneously observed at different AP antennas and different subcarriers, which can be expressed as

$$\mathrm{supp}\{[\mathbf{H}_{p,b}]_{:,1}\} = \mathrm{supp}\{[\mathbf{H}_{p,b}]_{:,2}\} = \cdots = \mathrm{supp}\{[\mathbf{H}_{p,b}]_{:,M_c}\}, \tag{5}$$

and

$$\mathrm{supp}\left\{\mathbf{H}_{1,b}\right\} = \mathrm{supp}\left\{\mathbf{H}_{2,b}\right\} = \cdots = \mathrm{supp}\left\{\mathbf{H}_{P,b}\right\}, \tag{6}$$

respectively. We refer to the structured sparsity in (4)-(6) as the spatial-domain structured sparsity of $\{\mathbf{H}_{p,b}\}_{p=1}^P$. Particularly, the signals of active UEs can be received by all APs, and this structured sparsity caused by sporadic UEs' traffic would be the same for different APs. On the other hand, due to the large-scale fading caused by path loss, the channel strength from a specific active UE to far away APs can be approximate zero. Hence, the channel matrices between UEs and different APs, i.e., $\{\mathbf{H}_{p,b}\}_{p=1}^P, \forall b$, exhibit approximate common sparsity pattern. To illustrate this structured sparsity more explicitly, we provide an example in Fig. 3 and Fig. 4(a), where we assume that $K_a = 3$ active UEs out of $K = 10$ total UEs access the network and each AP is equipped with $M_c = 10$ antennas. Given the locations of active UEs and APs described in Fig. 3, the 4-th row vectors in both channel matrices $\{\mathbf{H}_{p,1}\}_{p=1}^P$ and $\{\mathbf{H}_{p,2}\}_{p=1}^P$ (corresponding to the 4-th UE in the active state in Fig. 4) have large gain (strong common support). However, due to the large path loss gap, for the 2-th (or 9-th) UE in the active state, only the 2-th (or 9-th) row vectors in $\{\mathbf{H}_{p,1}\}_{p=1}^P$ (or $\{\mathbf{H}_{p,2}\}_{p=1}^P$) have the sufficiently large gain while those in $\{\mathbf{H}_{p,2}\}_{p=1}^P$ (or $\{\mathbf{H}_{p,1}\}_{p=1}^P$) can be negligible, which can be illustrated in Fig. 4(a).

2) Angular-Domain Structured Sparsity: By representing the massive MIMO channels in the virtual angular domain, we can find some additional sparsity properties of the massive access channel matrix. Specifically, the angular-domain massive MIMO channel between the k-th UE and the b-th AP at the p-th pilot subcarrier can be represented as

$$\widetilde{\mathbf{w}}_{p,b,k} = \mathbf{A}_R^{\mathrm{H}}\widetilde{\mathbf{h}}_{p,b,k}, \tag{7}$$

where the transformation matrix $\mathbf{A}_R \in \mathbb{C}^{M_c\times M_c}$ at the AP side is a unitary matrix. Here, $\mathbf{A}_R$ depends on the geometry of the array, which becomes the DFT matrix for a ULA when $\widetilde{d} = \lambda/2$ [28]. For the practical implementation of the network, the APs are usually deployed at high elevation with few

scatterers around, whereas the UEs are typically distributed at low elevation in a local rich scattering environment far from the APs [29]. We model this typical scenario as the classical one-ring channel model [30], as illustrated in Fig. 3. Here, we assume a UE is located in a rich scattering environment within a radius of r, and the distance between the UE and AP is R, so the angular spread observed at the AP is given as

$$\Delta \approx \arctan\left(r/R\right). \tag{8}$$

Hence, the sparsity level of the angular-domain channels is proportional to Δ, and it is expected to be far less than M_c as usually $R \gg r$. This indicates the virtual angular-domain sparsity of massive MIMO channels, i.e.,

$$\left|\mathrm{supp}\left\{\widetilde{\mathbf{w}}_{p,b,k}\right\}\right|_c \ll M_c, \tag{9}$$

and this sparsity is clustered, as illustrated in Fig. 4(b). Furthermore, as the scattering environment for all subchannels within the bandwidth remains unchanged, the angular spreads of all subchannels are very similar. Hence, all subchannels have a common sparsity pattern as

$$\mathrm{supp}\left\{\widetilde{\mathbf{w}}_{1,b,k}\right\} = \mathrm{supp}\left\{\widetilde{\mathbf{w}}_{2,b,k}\right\} = \cdots = \mathrm{supp}\left\{\widetilde{\mathbf{w}}_{P,b,k}\right\}. \tag{10}$$

We refer to the structured sparsity in (9) and (10) as the angular-domain structured sparsity. Define the virtual angular-domain channel matrix as $\mathbf{W}_{p,b} = \mathbf{H}_{p,b}\mathbf{A}_R^* = \left[\sqrt{P_1}\alpha_1\mathbf{w}_{p,b,1}, \cdots, \sqrt{P_K}\alpha_K\mathbf{w}_{p,b,K}\right]^{\mathrm{T}}$, where $\mathbf{w}_{p,b,k} = \rho_{b,k}\widetilde{\mathbf{w}}_{p,b,k}$. By combining the sparse UE activity and the angular-domain structured sparsity, we further have $\left|\mathrm{supp}\left\{\left[\mathbf{W}_{p,b}\right]_{:,m}\right\}\right|_c \ll K_a$, and

$$\mathrm{supp}\left\{\mathbf{W}_{1,b}\right\} = \mathrm{supp}\left\{\mathbf{W}_{2,b}\right\} = \cdots = \mathrm{supp}\left\{\mathbf{W}_{P,b}\right\}. \tag{11}$$

An illustration of the structured sparsity of $\{\mathbf{W}_{p,b}\}_{p=1}^P, \forall b$ is also provided in Fig. 4(b).

Remark 2: The aforementioned angular-domain structured sparsity of massive MIMO channels is valid even for sub-6 GHz systems [31]. Note that our work considers the cell-free network based on co-located massive MIMO configuration, rather than distributed massive MIMO [11] whose angular-domain sparsity does no exist.

These two forms of structured sparsity will be leveraged to facilitate the design of AUD and CE algorithm in the remainder of this paper. Specifically, the spatial-domain approximate common sparsity can be exploited to enhance the AUD performance, while the angular-domain enhanced sparsity can be utilized to improve the CE performance. Hence, we perform AUD based on the spatial-domain channel model and perform CE for the identified UEs based on the angular-domain channel model.

III. Cloud Computing-Based and Edge Computing-Based Massive Access

This section details the problem formulations at the receiver for massive access based on cloud computing and edge computing paradigms, respectively. In this paper, we adopt a grant-free massive access protocol to avoid complicated access scheduling, where the transmit frame structure proposed in Section II-B is employed. Here, we assume the frame length is far smaller than the channel coherence time, and the activity of the UEs during the channel coherence time remains unchanged. In grant-free massive access, the set of active UEs and the corresponding CSI have to be acquired for the subsequent coherent data detection. In the pilot phase, for the b-th AP and the p-th pilot subcarrier, the pilot signals received in G successive time slots are collected as

$$\begin{aligned}\mathbf{Y}_{p,b} &= \sum\nolimits_{k=1}^{K} \mathbf{s}_{p,k}\sqrt{P_k}\alpha_k\mathbf{h}_{p,b,k}^{\mathrm{T}} + \mathbf{N}_{p,b} \\ &= \mathbf{S}_p\mathbf{H}_{p,b} + \mathbf{N}_{p,b},\ \forall p \in [P] \text{ and } \forall b \in [B],\end{aligned} \tag{12}$$

where $\mathbf{Y}_{p,b} = \left[\mathbf{y}_{p,b}^1, \cdots, \mathbf{y}_{p,b}^G\right]^{\mathrm{T}} \in \mathbb{C}^{G\times M_c}$ and the received signal $\mathbf{y}_{p,b}^t$ is given in (2). Furthermore, $\mathbf{s}_{p,k} = [s_{p,k}^1, \cdots, s_{p,k}^G]^{\mathrm{T}} \in \mathbb{C}^{G\times 1}$ is the access pilot sequence of the k-th UE at the p-th pilot subcarrier, $\mathbf{S}_p = [\mathbf{s}_{p,1}, \cdots, \mathbf{s}_{p,K}] \in \mathbb{C}^{G\times K}$ is the pilot matrix. Here, the pilot of the k-th UE at the p-th pilot subcarrier is given as $\mathcal{CN} \sim \left(s_{p,k}^t; 0, 1\right)$, and the pilots at different pilot subcarriers are different for achieving diversity [29]. Finally, $\mathbf{H}_{p,b},\ \forall p \in [P]$, denotes the massive access channel matrix between all UEs and the b-th AP, and $\mathbf{N}_{p,b} = \left[\mathbf{n}_{p,b}^1, \cdots, \mathbf{n}_{p,b}^G\right]^{\mathrm{T}}$. Based on (12), the AUD problem is to estimate $\alpha_k, \forall k \in [K]$, i.e., find the indices of non-zero rows of $\{\mathbf{H}_{p,b}\}_{p=1}^P, \forall b \in [B]$; on the other hand, the CE problem is to estimate $\mathbf{h}_{p,b,k}$ for $\forall k \in \mathcal{A}$, i.e., the related row coefficients of $\{\mathbf{H}_{p,b}\}_{p=1}^P, \forall b \in [B]$. Therefore, these two problems can be jointly solved by estimating $\{\mathbf{H}_{p,b}\}_{p=1}^P$ based on the known $\mathbf{S}_p$ and $\mathbf{Y}_{p,b}, \forall b \in [B]$.

A. Cloud Computing-Based Massive Access

For cloud computing paradigm, quantities of APs are distributed in a large area and cooperate at the CPU through backhaul links. Here, the APs are only designed for receiving and transmitting signals, thus the corresponding AUD and CE are centrally processed at the CPU. Considering the limited capacity of wireless backhaul links, the signals received at the APs are first quantized and then transmitted via backhaul links[1] to the CPU, i.e., $\forall p \in [P]$ and $\forall b \in [B]$,

$$\overline{\mathbf{Y}}_{p,b} = \psi_b\left(\mathbf{Y}_{p,b}\right) = \psi_b\left(\mathbf{S}_p\mathbf{H}_{p,b} + \mathbf{N}_{p,b}\right), \tag{13}$$

where $\psi_b\left(\cdot\right)$ is the complex-valued quantizer at the b-th AP. The quantizer is applied to the received signal element-wisely, and the real and imaginary parts are quantized separately. Here, we consider a uniform codebook for quantization,

$$\mathcal{C}_b = \left\{-\frac{2^Q-1}{2}\Delta_b, \cdots, \frac{2^Q-1}{2}\Delta_b\right\}, \tag{14}$$

where Q is the number of quantization bits, $\Delta_b = \left(y_b^{\max} - y_b^{\min}\right)/2^Q$, $y_b^{\max}$ and $y_b^{\min}$ are the maximum and the minimum real values of both real and imaginary parts of $\{\mathbf{Y}_{p,b}\}_{p=1}^P$, respectively. At the CPU, the quantized received signals from all APs are concentrated as $\forall p \in [P]$,

$$\overline{\mathbf{Y}}_p = \left[\overline{\mathbf{Y}}_{p,1}, \overline{\mathbf{Y}}_{p,2}, \cdots, \overline{\mathbf{Y}}_{p,B}\right] = \mathbf{S}_p\mathbf{H}_p + \mathbf{N}_p^q + \mathbf{N}_p, \tag{15}$$

[1]Especially for the widely used wireless backhaul with limited capacity, the higher resolution of quantization benefits the better massive access performance but at the cost of larger backhaul latency.

where $\mathbf{N}_p^q$ denotes the quantization error, $\mathbf{H}_p \in \mathbb{C}^{K\times M}$ is expressed as $\mathbf{H}_p = [\mathbf{H}_{p,1}, \mathbf{H}_{p,2}, \cdots, \mathbf{H}_{p,B}]$, $M = BM_c$, and $\mathbf{N}_p = [\mathbf{N}_{p,1}, \mathbf{N}_{p,2}, \cdots, \mathbf{N}_{p,B}]$. In stark contrast to the standard linear model (SLM) with infinite-resolution quantization widely used in [11], [26], the model (15) is a generalized linear model (GLM) due to the nonlinear measurements. By exploiting the sparse UE activity, the AUD problem based on (15) is formulated as a GLM-based CS problem, where we seek to recover the sparse channel matrices $\{\mathbf{H}_p\}_{p=1}^P$ from the quantized measurements $\{\overline{\mathbf{Y}}_p\}_{p=1}^P$. Meanwhile, the spatial-domain structured sparsity of $\{\mathbf{H}_p\}_{p=1}^P$, as described in Section II-C and illustrated in Fig. 4(a), can be exploited to improve the detection performance.

On the other hand, by representing the massive MIMO channels in the virtual angular domain, we can further transform (13) into

$$\mathbf{R}_{p,b} = \overline{\mathbf{Y}}_{p,b}\mathbf{A}_R^* = \mathbf{S}_p\mathbf{W}_{p,b} + \overline{\mathbf{N}}_{p,b}^q + \overline{\mathbf{N}}_{p,b}, \tag{16}$$

where $\overline{\mathbf{N}}_{p,b}^q = \mathbf{N}_{p,b}^q\mathbf{A}_R^*$ and $\overline{\mathbf{N}}_{p,b} = \mathbf{N}_{p,b}\mathbf{A}_R^*$. Thus, the (15) at the CPU can be also expressed as

$$\mathbf{R}_p = [\mathbf{R}_{p,1}, \mathbf{R}_{p,2}, \cdots, \mathbf{R}_{p,B}] = \mathbf{S}_p\mathbf{W}_p + \overline{\mathbf{N}}_p^q + \overline{\mathbf{N}}_p, \tag{17}$$

where $\mathbf{W}_p = [\mathbf{W}_{p,1}, \mathbf{W}_{p,2}, \cdots, \mathbf{W}_{p,B}]$ and $\overline{\mathbf{N}}_p = \left[\overline{\mathbf{N}}_{p,1}, \overline{\mathbf{N}}_{p,2}, \cdots, \overline{\mathbf{N}}_{p,B}\right]$. With the estimate of AUS based on (15), denoted as $\widehat{\mathcal{A}}$, the CE problem based on (17) is equivalent to solving the following CS problem

$$\mathbf{R}_p = [\mathbf{S}_p]_{:,\widehat{\mathcal{A}}}[\mathbf{W}_p]_{\widehat{\mathcal{A}},:} + \widetilde{\mathbf{N}}_p, \tag{18}$$

where $\widetilde{\mathbf{N}}_p$ includes the aggregated AWGN, quantization error, and estimation error of AUD. By leveraging the angular-domain structured sparsity of $\{\mathbf{W}_p\}_{p=1}^P$, as described in Section II-C and illustrated in Fig. 4(b), the CSI estimates of the UEs identified in (15) can be further refined.

Hence, by leveraging the two forms of structured sparsity the channel matrix, the AUD and CE problems based on cloud computing paradigm are equivalent to solving the CS problems in (15) and (18), respectively, i.e., detecting the non-zero rows of $\{\mathbf{H}_p\}_{p=1}^P$ and estimating the corresponding row coefficients of $\{\mathbf{W}_p\}_{p=1}^P$.

B. Edge Computing-Based Massive Access

For edge computing paradigm, the central processing at the CPU is offloaded to the edge of the network, as illustrated in Fig. 1(b). Specifically, a part of the APs, termed as DPU-APs, are equipped with the DPUs or MEC severs having the storage and the computing capabilities. Hence, the signals received at multiple APs are jointly processed at the adjacent DPU-APs. We will further explain this distributed processing strategy from a DPU-AP centric perspective. Specifically, for a specific DPU-AP, its DPU will collect the signals received locally and from the $(N_{co}-1)$ nearest APs for the distributed processing. Here, N_{co} is the number of APs for cooperation, which includes one DPU-AP and $(N_{co}-1)$ conventional APs without DPU. Assume there are I DPU-APs in the network, the signals received at the i-th DPU-AP are organized as

$$\begin{aligned}\overline{\mathbf{Y}}_{p,i} &= \left[\overline{\mathbf{Y}}_{p,\mathcal{B}_i(1)}, \overline{\mathbf{Y}}_{p,\mathcal{B}_i(2)}, \cdots, \overline{\mathbf{Y}}_{p,\mathcal{B}_i(N_{co})}\right] \\ &= \mathbf{S}_p\left[\mathbf{H}_p\right]_{:,\mathcal{M}_i} + \left[\mathbf{N}_p^q\right]_{:,\mathcal{M}_i} + \left[\mathbf{N}_p\right]_{:,\mathcal{M}_i},\end{aligned} \tag{19}$$

where $\mathcal{B}_i$ denotes the set of APs cooperate on the i-th DPU-AP, $\mathcal{B}_i(n)$ is the n-th element of $\mathcal{B}_i$, and the column index set $\mathcal{M}_i$ is defined as $\mathcal{M}_i = \{m | m = (b-1)M_c + 1 : bM_c, \forall b \in \mathcal{B}_i\}$. Meanwhile, the spatial-domain channel model (19) can be further represented in the angular domain as

$$\begin{aligned}\mathbf{R}_{p,i} &= \left[\mathbf{R}_{p,\mathcal{B}_i(1)}, \mathbf{R}_{p,\mathcal{B}_i(2)}, \cdots, \mathbf{R}_{p,\mathcal{B}_i(N_{co})}\right] \\ &= \mathbf{S}_p\left[\mathbf{W}_p\right]_{:,\mathcal{M}_i} + \left[\overline{\mathbf{N}}_p^q\right]_{:,\mathcal{M}_i} + \left[\overline{\mathbf{N}}_p\right]_{:,\mathcal{M}_i}.\end{aligned} \tag{20}$$

For all DPU-APs, i.e., $\forall i \in [I]$, by exploiting the sparsity properties of the channel matrix, the AUD and CE problems based on edge computing paradigm are equivalent to solving the CS problems in (19) and (20), respectively.

IV. Proposed Active User Detection and Channel Estimation Algorithm

As described in Section III, the AUD and CE problems for grant-free massive access are formulated as the CS problems. In this section, we first develop a SS-GAMP algorithm to realize the related sparse signal recovery with quantized measurements. On this basis, a SIC-based AUD and CE algorithm is further proposed. Here, the explanations of the proposed algorithms are based on the cloud computing as an example, which can be easily extended to the edge computing.

A. SS-GAMP Algorithm

For the CS problem with quantized measurements, we adopt the unified Bayesian inference framework proposed in [32], which can iteratively reduce the GLM problem to a series of SLM problems. Moreover, based on the message passing theory and employing the low-complexity heuristics for approximating the messages, we develop a SS-GAMP algorithm to reap both the better performance than greedy methods [29] and the lower complexity than conventional message passing algorithms [33]. To simplify the derivations, we focus on the spatial-domain channel model (15) and the p-th pilot subcarrier first. The acquired key steps of the proposed algorithm can be easily extended to the angular-domain channel model (17) and multiple pilot subcarriers cases. Furthermore, for notational simplicity, the index p in $\overline{\mathbf{Y}}_p$, $\mathbf{S}_p$, and $\mathbf{H}_p$ is dropped and will be reused when the multiple pilot subcarriers case is considered.

The block diagram of the proposed SS-GAMP algorithm is illustrated in Fig. 5, which comprises two modules: nonlinear module and SLM module. Based on the quantized received signal $\overline{\mathbf{Y}}$ and the noise variance σ, nonlinear module performs minimum mean square error (MMSE) estimate of the linear received signal $\mathbf{Y} = \mathbf{SH} + \mathbf{N}$, and the corresponding posterior mean and variance are denoted by $\mathbf{Y}^{\text{post}}$ and V^{post}, respectively. The extrinsic messages of nonlinear module, i.e., the equivalent linear measurement $\widehat{\mathbf{Y}}$ and noise variance $\widehat{\sigma}$, form the input of SLM module. In SLM module, the concerned GLM problem has been transformed into an equivalent SLM problem as

$$\widehat{\mathbf{Y}} = \mathbf{SH} + \widehat{\mathbf{N}}, \tag{21}$$

where the variance of the equivalent noise $\widehat{\mathbf{N}}$ is given as $\widehat{\sigma}$. Hence, SLM module employs the SLM-based AMP algorithm

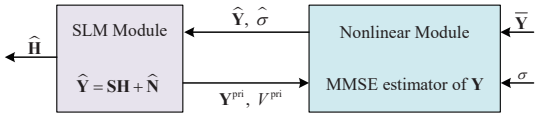


Fig. 5. Block diagram of the proposed SS-GAMP algorithm.

to estimate the channel matrix $\mathbf{H}$, and its extrinsic messages, i.e., $\mathbf{Y}^{\mathrm{pri}}$ and V^{pri}, are passed to nonlinear module as the a priori information of $\mathbf{Y}$. These two modules are executed alternately in a turbo manner until convergence.

1) Nonlinear Module: The posterior probability of $\mathbf{Y}$ is expressed as

$$p\left(\mathbf{Y}|\overline{\mathbf{Y}}\right) \propto p\left(\overline{\mathbf{Y}}|\mathbf{Y}\right)\mathcal{CN}\left(\mathbf{Y};\mathbf{Y}^{\mathrm{pri}},V^{\mathrm{pri}}\right), \tag{22}$$

where $p\left(\overline{\mathbf{Y}}|\mathbf{Y}\right)$ is the likelihood function. Since the processed signals are quantized element-wisely, we can compute $p\left(\mathbf{Y}|\overline{\mathbf{Y}}\right)$ element-wisely, and the real and imaginary parts are calculated separately. Furthermore, as the quantization codebooks of different APs are different, the signals from different APs are also processed separately. According to the derivations in [34], the posterior mean and variance of the real part of $\mathbf{Y}_{p,b}$ are finally given as

$$y_{b,g,m}^{\mathrm{post}} = y_{b,g,m}^{\mathrm{pri}} + \frac{\mathrm{sign}\left(\overline{y}_{b,g,m}\right)V^{\mathrm{pri}}}{\sqrt{2\left(\sigma+V^{\mathrm{pri}}\right)}}\left(\frac{\phi\left(\eta_1\right)-\phi\left(\eta_2\right)}{\Phi\left(\eta_1\right)-\Phi\left(\eta_2\right)}\right), \tag{23}$$

$$V^{\mathrm{post}} = \frac{V^{\mathrm{pri}}}{2} - \frac{\left(V^{\mathrm{pri}}\right)^2}{2\left(\sigma+V^{\mathrm{pri}}\right)} \times\left(\frac{\eta_1\phi\left(\eta_1\right)-\eta_2\phi\left(\eta_2\right)}{\Phi\left(\eta_1\right)-\Phi\left(\eta_2\right)}+\left(\frac{\phi\left(\eta_1\right)-\phi\left(\eta_2\right)}{\Phi\left(\eta_1\right)-\Phi\left(\eta_2\right)}\right)^2\right), \tag{24}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the cumulative distribution function and the probability density function of the standard normal distribution, respectively. In (23) and (24), η_1 and η_2 are defined as

$$\eta_1 = \frac{\mathrm{sign}\left(\overline{y}_{b,g,m}\right)-\min\left\{|\overline{y}_{b,g,m}-\Delta_b/2|,|\overline{y}_{b,g,m}+\Delta_b/2|\right\}}{\sqrt{\frac{\sigma+V^{\mathrm{pri}}}{2}}}, \tag{25}$$

$$\eta_2 = \frac{\mathrm{sign}\left(\overline{y}_{b,g,m}\right)-\max\left\{|\overline{y}_{b,g,m}-\Delta_b/2|,|\overline{y}_{b,g,m}+\Delta_b/2|\right\}}{\sqrt{\frac{\sigma+V^{\mathrm{pri}}}{2}}}, \tag{26}$$

respectively. For ease of notation, we have abused $y_{b,g,m}^{\mathrm{post}}$, $y_{b,g,m}^{\mathrm{pri}}$, and $\overline{y}_{b,g,m}$ to denote the real part of these variables, and the imaginary part can be computed analogously. Furthermore, the extrinsic messages of nonlinear module are computed as

$$\widehat{\sigma} = \frac{V^{\mathrm{post}}V^{\mathrm{pri}}}{V^{\mathrm{pri}}-V^{\mathrm{post}}}, \tag{27}$$

$$\widehat{\mathbf{Y}} = \widehat{\sigma}\left(\mathbf{Y}^{\mathrm{post}}/V^{\mathrm{post}}-\mathbf{Y}^{\mathrm{pri}}/V^{\mathrm{pri}}\right). \tag{28}$$

Note that $\widehat{\mathbf{Y}}$ and $\widehat{\sigma}$ are actually the equivalent measurement of $(\mathbf{SH}+\mathbf{N})$ and the equivalent noise variance, respectively.

2) SLM Module: In this module, the quantized CS problem (15) has been transformed into the conventional SLM problem, as in (21). Thus, the AMP algorithm proposed in [25], which are designed for CS problems with linear measurements, can be directly applied to acquire the estimate of $\mathbf{H}$. Due to the limited paper length, here we only clarify the key steps, and please refer to [25] for more details. Based on the derivations in [25], the AMP algorithm can be explained intuitively. In the large system limit, i.e., as $K\to\infty$, while $\gamma = K_a/K$ and $\kappa = G/K$ are fixed, the AMP algorithm decouples the matrix estimation problem based on (21) into KM scalar estimation problems, as $\forall k\in[K]$ and $\forall m\in[M]$,

$$\widehat{\mathbf{Y}} = \mathbf{SH}+\widehat{\mathbf{N}} \to A_{k,m}^q = h_{k,m}+\widehat{n}_{k,m}^q, \tag{29}$$

where $A_{k,m}^q \sim \mathcal{CN}\left(A_{k,m}^q;h_{k,m},B_{k,m}^q\right)$ is the equivalent measurement of $h_{k,m}$ obtained in the q-th iteration of AMP algorithm, and $\widehat{n}_{k,m}^q \sim \mathcal{CN}\left(\widehat{n}_{k,m}^q;0,B_{k,m}^q\right)$ denotes the effective noise. The effective noise includes AWGN and the estimation error of $h_{k,m}$ in the q-th iteration. In this way, the posterior distribution of $h_{k,m}$, $\forall k,m$, can be approximated as

$$\begin{aligned} p\left(h_{k,m}|\widehat{\mathbf{Y}}\right) &\approx p\left(h_{k,m}|A_{k,m}^q,B_{k,m}^q\right) \\ &\approx \frac{1}{F_1}p_0\left(h_{k,m}\right)\mathcal{CN}\left(h_{k,m};A_{k,m}^q,B_{k,m}^q\right), \end{aligned} \tag{30}$$

where F_1 is a normalization factor and $p_0\left(h_{k,m}\right)$ denotes the a priori distribution of $h_{k,m}$. In (30), $B_{k,m}^q$ and $A_{k,m}^q$ are updated as follows

$$B_{k,m}^q = \left[\sum\nolimits_{g=1}^{G}\frac{|s_{g,k}|^2}{\widehat{\sigma}+C_{g,m}^q}\right]^{-1}, \tag{31}$$

$$A_{k,m}^q = \hat{h}_{k,m}^q + B_{k,m}^q\sum\nolimits_{g=1}^{G}\frac{s_{g,k}^*\left(\widehat{y}_{g,m}-D_{g,m}^q\right)}{\widehat{\sigma}+C_{g,m}^q}, \tag{32}$$

and $C_{g,m}^q$ and $D_{g,m}^q$ are updated as follows

$$C_{g,m}^q = \sum\nolimits_{k=1}^{K}|s_{g,k}|^2 v_{k,m}^q, \tag{33}$$

$$D_{g,m}^q = \sum\nolimits_{k=1}^{K}s_{g,k}\hat{h}_{k,m}^q - \frac{C_{g,m}^q}{\widehat{\sigma}+C_{g,m}^{q-1}}\left(\widehat{y}_{g,m}-D_{g,m}^{q-1}\right), \tag{34}$$

where $v_{k,m}^q$ is the posterior variance of $h_{k,m}$.

To characterize the sparsity of the channel matrix, this paper adopts the spike and slab distribution [31] to model the a priori distribution of $\mathbf{H}$, which can be expressed as

$$\begin{aligned} p_0\left(\mathbf{H}\right) &= \prod_{m=1}^{M}\prod_{k=1}^{K}p_0\left(h_{k,m}\right) \\ &= \prod_{m=1}^{M}\prod_{k=1}^{K}\left[\left(1-\gamma_{k,m}\right)\delta\left(h_{k,m}\right)+\gamma_{k,m}f\left(h_{k,m}\right)\right], \end{aligned} \tag{35}$$

where $0<\gamma_{k,m}<1$ is the sparsity ratio, i.e., the probability of $h_{k,m}$ being non-zero, $\delta(\cdot)$ is the Dirac delta function. The a priori distribution of channel gains $f\left(h_{k,m}\right)$ is related to the channel model $\mathbf{h}_{b,k} = \rho_{b,k}\widetilde{\mathbf{h}}_{b,k}$, where $b = \lceil m/M_c\rceil$ and $\widetilde{\mathbf{h}}_{b,k}$ is given in (3). Furthermore, this paper adopts the one-ring channel model, where the UEs are located in a

local rich scattering environment, i.e., the number of MPCs $L_{b,k}$ can be large but the angular spread can be limited. Hence, given $\beta_{b,k}^{l} \sim \mathcal{CN}\left(\beta_{b,k}^{l}; 0, 1\right)$, we assume $f\left(h_{k,m}\right) = \mathcal{CN}\left(h_{k,m}; \mu_{k,m}, \tau_{k,m}\right)$ according to the central limit theorem, where

$$\begin{aligned}\mu_{k,m} &= \mathbb{E}\left[\sqrt{P_k}\rho_{b,k}\sum_{l=1}^{L_{b,k}}\beta_{b,k}^{l}e^{-j2\pi(m-bM_c)\phi_{b,k}^{l}}e^{-j2\pi\tau_{b,k}^{l}f}\right] \\ &= 0,\end{aligned} \tag{36}$$

and

$$\begin{aligned}\tau_{k,m} =& \mathbb{E}\left[P_k\rho_{b,k}^{2}\left(\sum_{l=1}^{L_{b,k}}\beta_{b,k}^{l}e^{-j2\pi(m-bM_c)\phi_{b,k}^{l}}e^{-j2\pi\tau_{b,k}^{l}f}\right)^{2}\right] \\ &- \mu_{k,m}^{2} = P_k\rho_{b,k}^{2}L_{b,k}.\end{aligned} \tag{37}$$

By exploiting this a priori model in (30), the posterior distribution of $h_{k,m}$ is obtained as follows

$$\begin{aligned}p\left(h_{k,m}|A_{k,m}^{q}, B_{k,m}^{q}\right) =& \left(1-\theta_{k,m}^{q}\right)\delta\left(h_{k,m}\right) \\ &+ \theta_{k,m}^{q}\mathcal{CN}\left(h_{k,m}; Z_{k,m}^{q}, V_{k,m}^{q}\right),\end{aligned} \tag{38}$$

where

$$Z_{k,m}^{q} = \frac{\tau_{k,m}A_{k,m}^{q} + \mu_{k,m}B_{k,m}^{q}}{B_{k,m}^{q} + \tau_{k,m}}, \tag{39}$$

$$V_{k,m}^{q} = \frac{\tau_{k,m}B_{k,m}^{q}}{\tau_{k,m} + B_{k,m}^{q}}, \tag{40}$$

$$\mathcal{J}_{k,m}^{q} = \ln\frac{B_{k,m}^{q}}{B_{k,m}^{q}+\tau_{k,m}} + \frac{\left|A_{k,m}^{q}\right|^{2}}{B_{k,m}^{q}} - \frac{\left|A_{k,m}^{q}-\mu_{k,m}\right|^{2}}{\left(B_{k,m}^{q}+\tau_{k,m}\right)}, \tag{41}$$

$$\theta_{k,m}^{q} = \frac{\gamma_{k,m}}{\gamma_{k,m} + \left(1-\gamma_{k,m}\right)\exp\left(-\mathcal{J}_{k,m}^{q}\right)}, \tag{42}$$

and $\theta_{k,m}^{q}$ is referred to as the *belief indicator*. The posterior mean and variance of $h_{k,m}$ can now be explicitly calculated as

$$\hat{h}_{k,m} = \theta_{k,m}^{q}Z_{k,m}^{q}, \tag{43}$$

$$v_{k,m}^{q} = \theta_{k,m}^{q}\left(\left|Z_{k,m}^{q}\right|^{2} + V_{k,m}^{q}\right) - \left|\hat{h}_{k,m}\right|^{2}, \tag{44}$$

respectively. The equations (31)-(34) and (39)-(44) make up the key steps of the basic AMP algorithm, which provides a simplified approach to calculate the MMSE estimate of $\mathbf{H}$. Here, we assume the CPU can acquire the full knowledge of the sparsity ratio $\gamma_{k,m}$ and the noise variance $\widehat{\sigma}$, which is an impractical assumption. The reason is that, for practical cell-free massive MIMO systems, the varying numbers of active UEs leads to the varying channel sparsity level $\gamma_{k,m}$. Moreover, when performing angular-domain CE, the variance of the effective noise $\widetilde{\mathbf{N}}_p$ is hard to compute as the estimation error of AUD would be unknown. For facilitating the practical implementation of the algorithm, the EM is employed to learn the unknown hyper-parameters,

$$\widehat{\sigma}^{q+1} = \frac{1}{GM}\sum\nolimits_{g=1}^{G}\sum\nolimits_{m=1}^{M}\left[\frac{\left|\widehat{y}_{g,m}-D_{g,m}^{q}\right|^{2}}{\left|1+C_{g,m}^{q}/\widehat{\sigma}^{q}\right|^{2}} + \frac{\widehat{\sigma}^{q}C_{g,m}^{q}}{\widehat{\sigma}^{q}+C_{g,m}^{q}}\right], \tag{45}$$

$$\gamma_{k,m}^{q+1} = \theta_{k,m}^{q+1} = \frac{\gamma_{k,m}^{q}}{\gamma_{k,m}^{q} + \left(1-\gamma_{k,m}^{q}\right)\exp\left(-\mathcal{J}_{k,m}^{q}\right)}. \tag{46}$$

Finally, given the posterior mean and variance of the channel matrix, the extrinsic messages of SLM module are given as [32]

$$\mathbf{Y}^{\mathrm{pri}} = \mathbf{S}\widehat{\mathbf{H}} + \frac{\mathbf{C}^{q}}{\widehat{\sigma}+\mathbf{C}^{q-1}}\circ\left(\widehat{\mathbf{Y}}-\mathbf{D}^{q-1}\right), \tag{47}$$

$$V^{\mathrm{pri}} = \frac{1}{GM}\|\mathbf{C}\|_{\mathrm{F}}^{2}, \tag{48}$$

where $\circ$ denotes the Hadamard product.

Next, we extend the key steps of the generalized AMP algorithm derived above, i.e., (23)-(28), (31)-(34), (39)-(48), to the multiple subcarriers case, where the spatial-domain or angular-domain structured sparsity of the channel matrix is exploited to enhance the CS recovery performance. The resulted algorithm is referred to as the SS-GAMP algorithm, which is summarized in *Algorithm 1*. Specifically, in *lines 3-10*, the messages are updated independently for all subcarriers. Moreover, as the matrix estimation problem is decoupled into multiple scalar estimation problems, as shown in (29), the variables for calculating the associated messages are also computed independently for all b, k, g, and m. *Line 7* employs a damping parameter $\rho = 0.3$ to prevent the SS-GAMP algorithm from diverging [35]. Note that except for *line 11*, all variables in the SS-GAMP algorithm are updated independently. In *line 11*, the sparsity ratio $\gamma_{p,b,k,m}^{q+1}$ associated with different p, b, and m, are jointly refined based on the spatial-domain or angular-domain structured sparsity of the channel matrix.

When applying the SS-GAMP algorithm to the spatial-domain channel model (15) for AUD, the spatial-domain structured sparsity of $\{\mathbf{H}_p\}_{p=1}^{P}$ is considered. For the channel matrix between all UEs to a specific AP b, the channel vectors $[\mathbf{H}_{p,b}]_{:,m}$ observed at different pilot subcarriers and different AP antennas have a common sparsity, as described in (4)-(6) and illustrated in Fig. 4(a). Meanwhile, the sparsity ratio $\gamma_{p,b,k,m}$ is the probability that the (k,m)-th element of $[\mathbf{H}_{p,b}]_{k,m}$ is non-zero. Hence, for channel matrices $\{\mathbf{H}_{p,b}\}_{p=1}^{P}$, the elements associated with the same UE share a common sparsity ratio, so we consider

$$\widetilde{\gamma}_{b,k}^{q+1} = \frac{1}{\left|\mathcal{N}_{p,b,k,m}\right|_c}\sum\nolimits_{(o,b,k,u)\in\mathcal{N}_{p,b,k,m}}\theta_{o,b,k,u}^{q+1}, \tag{49}$$

where

$$\mathcal{N}_{p,b,k,m} = \left\{(o,b,k,u)\,|o=1,\cdots,P;\ u=1,\cdots,M_c\right\}. \tag{50}$$

Additionally, by further considering the approximate common sparsity between channel matrices $\{\mathbf{H}_{p,b}\}_{p=1}^{P}$ for different b,

Algorithm 1 SS-GAMP Algorithm

Input: $\forall p, b$: Quantized received signals $\overline{\mathbf{Y}}_{p,b}$, $\mathbf{S}_p$, Δ_b, $y_b^{\max}$, and $y_b^{\min}$; ρ, the maximum numbers of AMP and turbo iterations, $T_{\rm amp}$ and $T_{\rm tur}$, and the termination threshold η.

Output: $\forall p, b, k, m$: Estimated channel matrices $\{\widehat{\mathbf{H}}_{p,b}\}_{p=1}^{P}$ and the related belief indicators $\theta_{p,b,k,m}$. % In the reminder, $\forall m$ denotes $\forall m \in [M_c]$

$\forall p, b, k, m, g$: Set AMP iteration index q to 1, set turbo iteration index i to 1, initialize the $\gamma_{p,b,k,m}$ and $\widehat{\sigma}$ as in [25], and initialize other parameters as $y_{p,b,g,m}^{\rm pri}(1) = 0$, $V^{\rm pri}(1) = 10^6$, $C_{p,b,g,m}^{0} = 1$, $D_{p,b,g,m}^{0} = y_{p,b,g,m}$, $\hat{h}_{p,b,k,m}^{1} = \mu_{p,b,k,m}^{1}$, $v_{p,b,k,m}^{1} = \tau_{p,b,k,m}^{1}$.

for $i \le T_{\rm tur}$ **do**

  $\forall p, b$: Compute the posterior mean $\mathbf{Y}_{p,b}^{\rm post}(i)$ and posterior variance $V_{p,b}^{\rm post}(i)$ of the un-quantized received signals $\mathbf{Y}_{p,b}$, as in (23)-(26).

  $\forall p, b$: Compute the extrinsic messages of nonlinear module, $\widehat{\mathbf{Y}}_{p,b}(i)$ and $\widehat{\sigma}_{p,b}(i)$, as in (27) and (28), respectively, and $\widehat{\sigma}(i) = \frac{1}{PB}\widehat{\sigma}_{p,b}(i)$.

  repeat

    $\forall p, b, g, m$: Update $C_{p,b,g,m}^{q}$ and $D_{p,b,g,m}^{q}$ according to (33) and (34).

    $C_{p,b,g,m}^{q} = \rho C_{p,b,g,m}^{q-1} + (1-\rho)\, C_{p,b,g,m}^{q}$, $D_{p,b,g,m}^{q} = \rho D_{p,b,g,m}^{q-1} + (1-\rho)\, D_{p,b,g,m}^{q}$.

    $\forall p, b, k, m$: Update $B_{p,b,k,m}^{q}$ and $A_{p,b,k,m}^{q}$ according to (31) and (32).

    $\forall p, b, k, m$: Compute posterior mean $\hat{h}_{p,b,k,m}^{q+1}$ and posterior variance $v_{p,b,k,m}^{q+1}$ of the cannel matrix according to (43) and (44), respectively.

    $\forall p, b, k, m$: Update the $\gamma_{p,b,k,m}^{q+1}$ as in (46). Moreover, $\widehat{\sigma}^{q+1} = \frac{\sum_p \sum_b \widehat{\sigma}_{p,b}^{q+1}}{PB}$, where $\widehat{\sigma}_{p,b}^{q+1}$ is given in (45).

    $\forall p, b, k, m$: Refine the update rule of the sparsity ratio $\gamma_{p,b,k,m}^{q+1}$ based on the structured sparsity of the channel matrix, as in (49)-(53).

    $q = q + 1$.

  until $q \ge T_{\rm amp}$ or $\sum_p \left\| \widehat{\mathbf{H}}_p^{q} - \widehat{\mathbf{H}}_p^{q-1} \right\|_{\rm F}^2 / \sum_p \left\| \widehat{\mathbf{H}}_p^{q-1} \right\|_{\rm F}^2 < \eta$.

  $i = i + 1$.

  Compute the extrinsic messages of SLM module as in (47) and (48).

end for

return $\{\widehat{\mathbf{H}}_{p,b}^{q-1}\}_{p=1}^{P}, \forall b$; $\theta_{p,b,k,m} = \gamma_{p,b,k,m}^{q-1}$, $\forall p, b, k, m$.

the update rule for the sparsity ratio can be finally refined as

$$\gamma_{p,b,k,m}^{q+1} = \gamma_{k}^{q+1} = \sum\nolimits_{b=1}^{B} \frac{1}{d_{b,k}\Delta_k} \widetilde{\gamma}_{b,k}^{q+1}, \tag{51}$$

where $d_{b,k}$ denotes the distance between the k-th UE and the b-th AP, and $\Delta_k = \sum_{b=1}^{B} 1/d_{b,k}$. We can explain (51) intuitively from a UE-centric perspective. Specifically, for a specific UE k, its activity observed from the adjacent APs can be more reliable than that observed from the remote APs. Therefore, we consider a weighted method to refine the sparsity ratio, i.e., compared to the remote APs, the adjacent APs contribute more weights to the value of γ_k^{q+1}.

Due to the DFT transformation in (16), the angular domain received signals $\mathbf{R}_{p,b}, \forall p, b$ are not consistent with the quantization codebook of the corresponding GLM, i.e., $r_{p,b,g,m} \notin \mathcal{C}_b, \forall p, b, g, m$. This will lead to an unreliable estimate of $\mathbf{Y}_{p,b}$, as the nonlinear module of SS-GAMP algorithm is developed based on the quantization codebook. Hence, for CE, we directly apply the SLM module of SS-GAMP algorithm to (17), where the quantization error is treated as noise. Here, the virtual-angular domain sparsity of massive MIMO channels is further taken into account. However, this angular-domain sparsity destroys the structured sparsity over different AP antennas. Hence, the clustered sparsity illustrated in Fig. 4(b) is leveraged to refine the sparsity ratio. Specifically, define the neighbors of $w_{p,b,k,m}$ as

$$\begin{aligned} \widetilde{\mathcal{N}}_{p,b,k,m} = \{&(p-1, b, k, m), (p+1, b, k, m), \\ &(p, b, k, m-1), (p, b, k, m+1)\}, \end{aligned} \tag{52}$$

$w_{p,b,k,m}$ and the elements of $\widetilde{\mathcal{N}}_{p,b,k,m}$ tend to be simultaneously either zero or non-zero, and the update rule of $\gamma_{p,b,k,m}$ is given as

$$\gamma_{p,b,k,m}^{q+1} = \frac{1}{\left|\widetilde{\mathcal{N}}_{p,b,k,m}\right|_c} \sum\nolimits_{(o,b,k,u)\in\widetilde{\mathcal{N}}_{p,b,k,m}} \theta_{o,b,k,u}^{q+1}. \tag{53}$$

Remark 3: In contrast to the conventional CS algorithms for SLM, the proposed SS-GAMP mainly shows its superiority in low-resolution quantization cases. When the quantization accuracy is good enough, i.e., the number of quantization bits Q is large, the quantization error can be negligible. In this case, we can directly apply the SLM module of SS-GAMP algorithm to (15) for AUD, which can reduce the computational complexity with negligible performance loss.

B. SIC-Based AUD and CE Algorithm

The AUD and CE can be jointly realized by applying the SS-GAMP algorithm to (15) or (17). However, these solutions can not fully exploit the enhanced sparsity of $\{\mathbf{W}_p\}_{p=1}^{P}$ and the structured sparsity of $\{\mathbf{H}_p\}_{p=1}^{P}$. In this section, based on the SS-GAMP algorithm, we develop a SIC-based AUD and CE algorithm for alternately detecting active UEs based on (15) and estimating their channels based on (17), so that the massive access performance can be further improved.

The procedure of the proposed algorithm is summarized in *Algorithm 2* and is illustrated in Fig. 6, which mainly consists of three modules. Specifically, in each SIC iteration, the spatial-domain active UE detector (module A) acquires a rough AUS estimate and a relatively reliable AUS estimate (i.e., $\widehat{\mathcal{A}}$ and Ξ^j, respectively, in Fig. 6), which are passed to the angular-domain channel estimator (module B); subsequently, module B estimates the channels of the identified active UEs in $\widehat{\mathcal{A}}$; finally, based on the AUS and CSI estimates, module C updates the residual received signals by cancelling the components associated with the active UEs identified in Ξ^j, and the residual received signals are passed to module A. The three modules are executed alternately in an iterative

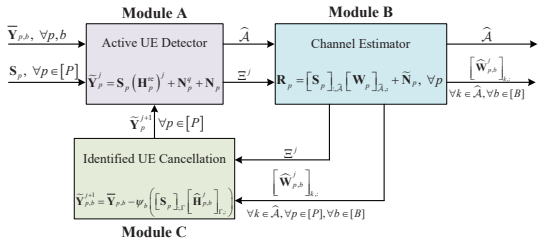


Fig. 6. Block diagram of the proposed SIC-based AUD and CE algorithm.

manner until convergence. Next, we will detail three modules as follow.

Module A: Spatial-domain AUD. In the first SIC iteration ($j = 1$), module A detects active UEs based on the spatial-domain channel model. Specifically, the SS-GAMP algorithm is applied to model (15) to acquire the belief indicators $\theta^j_{p,b,k,m}$, $\forall p, b, k, m$, based on which the AUS estimate, denoted as $\widehat{\mathcal{A}}$, is determined. It has been proved in [25] that if a reliable estimate of $\{\mathbf{H}_b\}_{p=1}^P$ is acquired after the convergence of the SLM module of SS-GAMP algorithm, belief indicator $\theta^j_{p,b,k,m}$ tends to 1 for $h_{p,b,k,m} \neq 0$ and 0 for $h_{p,b,k,m} = 0$. Hence, we design a belief indicator-based active UE (BI-AUE) detector as follows

$$\widehat{\alpha}_k = \begin{cases} 1, & \frac{1}{PM_c}\sum_{p=1}^{P}\sum_{m=1}^{M_c}\theta^j_{p,b^*,k,m} \geq p_{\rm th}, \\ 0, & \frac{1}{PM_c}\sum_{p=1}^{P}\sum_{m=1}^{M_c}\theta^j_{p,b^*,k,m} < p_{\rm th}. \end{cases} \tag{54}$$

Here, $k \in [K]$, $b^* = \{b|\min(d_{b,k}), \forall b\}$, i.e., the k-th UE's activity is mainly dependent on the belief indicators $\theta^j_{p,b^*,k,m}$ inferred from the pilot signals received at the b^*-th AP, which has the shortest spatial distance (also has the smallest path loss) with the k-th UE. For facilitating the subsequent CE and SIC processing, we utilize the BI-AUE detector to acquire two AUS estimates having different reliability: a rough AUS estimate $\widehat{\mathcal{A}}$ based on a lower threshold $p_{\rm th} = p_{\rm det}$ and a relatively reliable AUS estimate Ξ^j based on a higher threshold $p_{\rm th} = p_{\rm rel}$, as shown in *lines 5-13* in *Algorithm 2*, so that $\Xi^j \subseteq \widehat{\mathcal{A}}$. Here, we set $p_{\rm det}$ to 0.1 to reduce the missed detection probability, and $p_{\rm rel}$ is set to 0.9 to guarantee the UEs in Ξ^j are active with high probability. These two AUS estimates, $\widehat{\mathcal{A}}$ and Ξ^j, are passed to module B.

2) Module B: Angular-domain CE for identified active UEs. In module B, given the rough AUS estimate $\widehat{\mathcal{A}}$, the angular-domain channel vectors of the UEs in $\widehat{\mathcal{A}}$, i.e., $[\mathbf{W}_p]_{\widehat{\mathcal{A}},:}$, are estimated based on the model in (17) as follows

$$\mathbf{R}_p = [\mathbf{S}_p]_{:,\widehat{\mathcal{A}}}\,[\mathbf{W}_p]_{\widehat{\mathcal{A}},:} + \widetilde{\mathbf{N}}_p,\ \forall p \in [P]\,, \tag{55}$$

where $[\mathbf{S}_p]_{:,\widehat{\mathcal{A}}} \in \mathbb{C}^{G\times|\widehat{\mathcal{A}}|_c}$ and $[\mathbf{W}_p]_{\widehat{\mathcal{A}},:} \in \mathbb{C}^{|\widehat{\mathcal{A}}|_c\times M}$ are sub-matrices of $\mathbf{S}_p$ and $\mathbf{W}_p$, respectively, $\widetilde{\mathbf{N}}_p = [\mathbf{S}_p]_{:,\mathcal{K}-\widehat{\mathcal{A}}}\,[\mathbf{W}_p]_{\mathcal{K}-\widehat{\mathcal{A}},:} + \overline{\mathbf{N}}^q_p + \overline{\mathbf{N}}_p$, $\mathcal{K}$ is the set of all potential UEs, and $\mathcal{K} - \widehat{\mathcal{A}}$ denotes the difference set of sets $\mathcal{K}$ and $\widehat{\mathcal{A}}$. Note that $\widetilde{\mathbf{N}}_p$ is the effective noise including AWGN, quantization error, and the estimation error of AUD. Furthermore, if $\mathcal{A} \subseteq \widehat{\mathcal{A}}$, we have $\widetilde{\mathbf{N}}_p = \overline{\mathbf{N}}^q_p + \overline{\mathbf{N}}_p$. Hence, to reduce the power of $\widetilde{\mathbf{N}}_p$, a low missed detection probability is desirable. According to the angular-domain structured sparsity of $\{\mathbf{W}_p\}_{p=1}^P$, as described in (9)-(11) and illustrated in Fig. 4(b), the low-dimensional channel matrix $[\mathbf{W}_p]_{\widehat{\mathcal{A}},:}$ is still sparse. Hence, we can estimate $[\mathbf{W}_p]_{\widehat{\mathcal{A}},:}, \forall p$, by applying the SLM module of SS-GAMP algorithm to (55), see *line 15* of *Algorithm 2*. Finally, the reliable AUS estimate Ξ^j and the related CSI estimate are passed to module C.

3) Module C: Identified UE cancellation. Since the active UEs in Ξ^j are reliably detected in module A and their CSI is estimated in module B, the signals received from the UEs in Γ, a subset of Ξ^j, are removed from $\overline{\mathbf{Y}}_p$ to enhance the sparsity of the channel matrix for AUD in the next SIC iteration. The residual received signals $\widetilde{\mathbf{Y}}^j_p, \forall p$ are computed in *lines 16-18*, and are passed to module A. In the following SIC iterations ($j > 1$), the AUD problem in module A is to recover $\left(\mathbf{H}^{\rm re}_p\right)^j$ based on the following model

$$\widetilde{\mathbf{Y}}^j_p = \mathbf{S}_p\left(\mathbf{H}^{\rm re}_p\right)^j + \mathbf{N}^q_p + \mathbf{N}_p, \forall p \in [P]\,, \tag{56}$$

where $\widetilde{\mathbf{Y}}^j_p$ denotes the residual received signals in the j-th SIC iteration, $\left(\mathbf{H}^{\rm re}_p\right)^j = \mathbf{H}_p - \widetilde{\mathbf{H}}^j_p$, and $\widetilde{\mathbf{H}}^j_p \in \mathbb{C}^{K\times M}$ is defined as $[\widetilde{\mathbf{H}}^j_p]_{\Gamma,:} = [\widehat{\mathbf{H}}^{j-1}_p]_{\Gamma,:}$, while $[\widetilde{\mathbf{H}}^j_p]_{\mathcal{K}-\Gamma,:} = \mathbf{0}_{|\mathcal{K}-\Gamma|_c\times M}$. To

Algorithm 2 SIC-Based AUD and CE Algorithm

Input: $\forall p, b$: Quantized received signals $\overline{\mathbf{Y}}_{p,b}$, $\mathbf{S}_p$; the number of SIC iterations $T_{\rm sic}$.

Output: $\forall p, b$: The AUS estimate $\widehat{\mathcal{A}}$ and the related CSI estimates $\{\widehat{\mathbf{h}}_{p,b,k}\}_{p=1}^P, \forall k \in \widehat{\mathcal{A}}$.

1: Initialization: $j = 1$, $\Xi^0 = \emptyset$, $\widetilde{\mathbf{Y}}^1_p = \overline{\mathbf{Y}}_p$.
2: **repeat**
3: $k = 0$, $\widehat{\mathcal{A}} = \Gamma = \emptyset$.
4: $\forall p, b, k, m$: Acquire $\theta^j_{p,b,k,m}$ by applying the SS-GAMP algorithm to model (56).
5: **for** $k \leq K$ **do**
6: **if** $\frac{1}{PM_c}\sum_{p=1}^{P}\sum_{m=1}^{M_c}\theta^j_{p,b^*,k,m} \geq p_{\rm det}$ **then**
7: $\widehat{\mathcal{A}} = \widehat{\mathcal{A}} \cup \Xi^{j-1} \cup \{k\}$.
8: **end if**
9: **if** $\frac{1}{PM_c}\sum_{p=1}^{P}\sum_{m=1}^{M_c}\theta^j_{p,b^*,k,m} \geq p_{\rm rel}$ **then**
10: $\Xi^j = \Xi^{j-1} \cup \{k\}$.
11: **end if**
12: $k = k + 1$.
13: **end for** % Here, $b^* = \{b|\min(d_{b,k}), \forall b \in [B]\}$.
14: $\forall p$: $\mathbf{R}_p = \left[\overline{\mathbf{Y}}_{p,1}\mathbf{A}^*_R, \overline{\mathbf{Y}}_{p,2}\mathbf{A}^*_R, \cdots, \overline{\mathbf{Y}}_{p,B}\mathbf{A}^*_R\right]$, $\widehat{\mathbf{W}}^j_p = \mathbf{0}_{K\times M}$.
15: $\forall p, b$: Acquire the channel vectors $\left[\widehat{\mathbf{W}}^j_{p,b}\right]_{k,:}, \forall k \in \widehat{\mathcal{A}}$, by applying the SLM module of SS-GAMP algorithm to model (55).
16: Acquire set Γ, $\Gamma \subseteq \Xi^j$, and $|\Gamma|_c \,/\, \left|\Xi^j\right|_c = \lambda_{\rm aus}$. % The elements in Γ are randomly selected from Ξ^j.
17: $\forall p$: $\widehat{\mathbf{H}}^j_p = \left[\widehat{\mathbf{W}}^j_{p,1}\mathbf{A}^{\rm T}_R, \widehat{\mathbf{W}}^j_{p,2}\mathbf{A}^{\rm T}_R, \cdots, \widehat{\mathbf{W}}^j_{p,B}\mathbf{A}^{\rm T}_R\right]$.
18: $\forall p, b$: $\widetilde{\mathbf{Y}}^{j+1}_{p,b} = \overline{\mathbf{Y}}_{p,b} - \psi_b\left([\mathbf{S}_p]_{:,\Gamma}\left[\widehat{\mathbf{H}}^j_{p,b}\right]_{\Gamma,:}\right)$. % $\psi_b(\cdot)$ is applied only when the low-resolution quantization is considered.
19: $j = j + 1$.
20: **until** $j > T_{\rm tur}$.
21: **return** $\forall p, b$: $\widehat{\mathcal{A}}$; $\widehat{\mathbf{h}}_{p,b,k} = \left[\widehat{\mathbf{H}}^{j-1}_{p,b}\right]_{k,:}, \forall k \in \widehat{\mathcal{A}}$.

guarantee the robustness of the SS-GAMP-based AUD, we only remove the signals received from a part of the UEs in Ξ^j, i.e., $\lambda_{\rm aus} < 1$ (e.g., we consider $\lambda_{\rm aus} = 0.8$).

Modules A, B, and C will be executed alternately in an iterative manner. Since the $\left(\mathbf{H}_p^{\rm re}\right)^j$ becomes sparser and the CSI estimates of the UEs in $\widehat{\mathcal{A}}$ are iteratively re-estimated as the SIC iterations proceed, the $\widehat{\mathcal{A}}$ and the corresponding CSI estimates are constantly refined. Therefore, compared to the joint AUD and CE solutions without SIC, the proposed SIC-based scheme facilitates more reliable AUD and CE with a significant reduction in access latency. However, as the SS-GAMP algorithm is called twice in each SIC iteration and the identified UE cancellation requires additional matrix multiplication, the performance improvement is at the cost of a higher computational complexity.

V. Differences Between Cloud Computing and Edge Computing Paradigms

This section compares cloud computing and edge computing in terms of their algorithm implementation, computational complexity, access latency, and the cost of AP deployment. Compared with cloud computing, edge computing has the advantages of alleviating the burden on backhaul links and CPU, a faster access response, and supporting more flexible AP cooperation, while increases the cost of large-scale AP deployment.

A. Algorithm Implementation

For cloud computing paradigm, the detailed procedure of the proposed AUD and CE approach is summarized in *Algorithms 1* and *2*. It is clear that the signals collected from B APs are processed in parallel in *lines 3-10* of *Algorithm 1*, and are centrally processed in *line 11* only. Intuitively, *line 11* leverages the structured sparsity described in Section II-C to refine the update rule of the sparsity ratio $\gamma_{p,b,k,m}$, as in (49)-(53). Note that this paper considers a large-scale network to serve a vast area, so the channel strength from a specific active UE to far away APs approximates zero due to the large-scale fading caused by severe path loss. Hence, for this specific UE, the signals received at the remote APs have a negligible effect on refining the sparsity ratio. This reveals that centrally processing all the APs' received signals at the CPU for jointly refining sparsity ratio maybe not an efficient way.

While for edge computing paradigm, the SIC-based AUD and CE algorithm summarized in *Algorithm 2* can be directly applied based on (19) and (20) for detecting active UEs and estimating their channels, respectively. Here, the AUD and CE problems for the whole network are locally processed at multiple DPU-APs in close proximity to the UEs. Clearly, the APs in the edge computing paradigm are divided into several groups, and each group seeks to detect only part of the total UEs. We can also explain the AP and UE association from a UE-centric perspective, that is to say, for a specific UE, its activity and CSI can be estimated by jointly processing the signals received at its nearest one DPU-AP and $(N_{co}-1)$ APs. Compared to cloud computing, the edge computing enables more flexible AP cooperation by considering different numbers of cooperative APs and reduces the transmission burden on backhaul links.

B. Computational Complexity

For each SIC iteration in cloud computing, the complexity[2] of SS-GAMP algorithm is in order of $\mathcal{O}\left(T_{\rm amp}\left(4GKMP+3GKP+16GMP+20KMP\right)+T_{\rm tur}\times\left(GKMP+GMP\right)\right)$, the complexity of DFT is $\mathcal{O}\left(2BM_c^2P\right)$, and the complexity of computing residual received signal for SIC is $\mathcal{O}\left(GK_{\rm sic}MP\right)$, $K_{\rm sic}$ is the number of UEs for cancellation. Hence, the overall complexity of the processing tasks at CPU is given as

$$\begin{aligned} C_{\rm cloud} = \mathcal{O}(T_{\rm sic}[&2T_{\rm amp}\left(4GKMP+3GKP+16GMP\right.\\ &\left.+20KMP\right)+T_{\rm tur}\left(GKMP+GMP\right)\\ &+2BM_c^2P+GK_aMP]). \end{aligned} \tag{57}$$

While for edge computing paradigm, the complexity of SIC-based algorithm applied in the i-th DPU-AP is

$$\begin{aligned} C_{\rm edge}^i = \mathcal{O}(T_{\rm sic}[&2T_{\rm amp}\left(4GK_iM_iP+3GK_iP+16GM_iP\right.\\ &\left.+20K_iM_iP\right)+T_{\rm tur}\left(GKM_iP+GM_iP\right)\\ &+2N_{co}M_c^2P+GK_a^iMP]), \end{aligned} \tag{58}$$

where K_i is the number of UEs detected by the i-th DPU-AP, $K_a^i = \gamma K_i$, and $M_i = N_{co}M_c$. Since each DPU-AP seeks to detect only part of the total UEs (i.e., $K_i < K$) and $N_{co} < B$, we have $C_{\rm cloud} > C_{\rm edge}^i$. Hence, by splitting the signal processing task of the whole network and executing related computations at the edge of the network, edge computing can alleviate the computing burden on CPU.

C. Access Latency

The access latency of grant-free massive access consists of three components: pilot transmission time, propagation latency, and computation latency. First, the pilot transmission time depends on the adopted frame structure and the pilot length, which are the same for both cloud computing and edge computing. Second, the DPU-APs in edge computing are deployed at the edge of the network, while the CPU in cloud computing is usually very far away from the UEs. This results in a much smaller propagation delay for edge computing than that for cloud computing. Furthermore, cloud computing requires the information to pass through several networks including the radio access network, backhaul network, and core network, where traffic control, routing, and other network-management operations can contribute to excessive delays. Last, the CPU can have a massive computation capacity than that of DPU. However, the CPU has to be shared by a large number of other services, and the computational complexity of processing tasks at CPU is much larger than that at DPU, as described in Section V-B. Moreover, with the rapid development of the processors, the DPU is powerful enough for running highly sophisticated computing programs. Therefore, the cloud computing and edge computing can have similar computation latencies. According to the analysis above, edge computing can have a faster access response than cloud computing.

[2]Here, we mainly focus on the maximum number of required complex multiplications.

TABLE I: Simulation Parameters

Parameter	Value
Radius of the network coverage	2.65 km
AP-to-AP distance	$\sqrt{3}$ km
Number of active UEs K_a	140
Transmit power P_k	23 dBm
Background noise power	-174 dBm/Hz
OFDM's DFT size P in pilot phase	64
OFDM's DFT size N in data phase	2048
Cyclic prefix length N_{CP}	64
System bandwidth	10 MHz
Number of MPCs $L_{b,k}$	$\mathcal{U}(L_{b,k};40,100)$
Path delay of the l-th MPC $\tau_{b,k}^{l}$	$\mathcal{U}(\tau_{b,k}^{l};0,N_{CP}/B_s)$
Angular spread in degree	10°
Number of SS-GAMP iteration T_{amp}	20
Number of turbo iteration T_{tur}	10
Termination threshold η	10^{-5}
Path loss $\rho_{b,k}$ at distance $d_{b,k}$ in km	$128.1+37.6\log_{10}(d_{b,k})$

D. Cost of AP Deployment

For cloud computing, all APs are only designed for transmitting and receiving signals, where only antennas and radio frequency chains are needed. While for edge computing, part of APs should employ extra DPUs so that these APs can be upgraded to DPU-APs. In cell-free massive MIMO with quantities of APs, this will increase the cost of AP deployment. Furthermore, edge computing also requires some extra links between APs and DPU-APs.

VI. Simulation Results

This section conducts simulations to validate the superiority of the proposed massive access schemes. The simulation parameters are provided in Table I. We consider a typical massive access scenario in cell-free massive MIMO systems, where $K=2800$ UEs are uniformly distributed in the network and $B=7$ APs are geographically distributed to serve these UEs. To reduce the computational complexity, in pilot phase, we only use the signals received at $\widetilde{P}$ out of total P pilot subcarriers for AUD, where $\widetilde{P}\leq P$. With the obtained AUS estimate $\widehat{\mathcal{A}}$, all P pilot subchannels can be estimated by applying the SLM module of SS-GAMP algorithm to (55).

For performance evaluation, we consider the detection error probability of AUD P_e and the normalized mean squire error (NMSE) of CE, which are respectively defined as follows

$$P_e=\frac{\sum_k|\widehat{\alpha}_k-\alpha_k|}{K}, \tag{59}$$

$$\text{NMSE}=10\log_{10}\frac{\sum_p\sum_k\left\|\widehat{\mathbf{h}}_{p,b^*,k}-\mathbf{h}_{p,b^*,k}\right\|_2^2}{\sum_p\sum_k\|\mathbf{h}_{p,b^*,k}\|_2^2}. \tag{60}$$

For AUD in the edge computing paradigm, we obtain the activity estimate of the k-th UE $\widehat{\alpha}_k$ based on the signals received at its nearest one DPU-AP and $(N_{co}-1)$ APs. Moreover, due to the smallest path loss, the UE is expected to be served by the nearest AP in data transmission phase. Thus, for CE, we mainly focus on the estimation reliability of the channel between the k-th UE and the b^*-th AP, which has the shortest spatial distance with the k-th UE. We compare the proposed schemes with the following benchmarks:

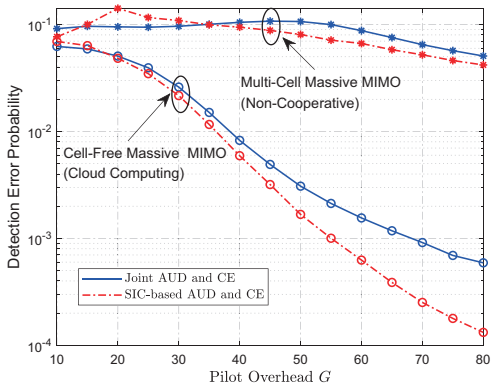


Fig. 7. AUD performance comparison of the proposed cloud computing-based scheme and *Baselines 1 and 2*, where $M_c=16$, $\widetilde{P}=1$, and $T_{\text{sic}}=3$.

- **Baseline 1** (Multi-cell non-cooperative massive MIMO-based IoT): To verify the superiority of the proposed cell-free massive MIMO-based IoT architecture, a conventional multi-cell non-cooperative massive MIMO-based IoT architecture is compared as the baseline 1, where each AP (i.e., massive MIMO BS) only serves its own cell's UEs without multi-cell cooperation and treats the inter-cell interference as noise [8].
- **Baseline 2** (SS-GAMP-based joint AUD and CE): To show the effectiveness of the proposed SIC-based AUD and CE algorithm, the conventional spatial domain-based massive access scheme is compared as the baseline 2, where the proposed SS-GAMP algorithm is applied to (15) for joint AUD and CE.
- **Baseline 3** (SS-GAMP algorithm using SLM to process quantized signals): To demonstrate the advantage of the proposed SS-GAMP-based joint AUD and CE scheme as well as the SIC-based scheme in the case of processed signals with low-resolution quantization, we compare those two schemes based on SS-GAMP algorithm only using SLM as baseline 3.

A. Superiority of Cell-Free Massive MIMO

This section validates the superiority of the proposed cell-free massive MIMO-based IoT architecture, where $Q=10$ is considered. Fig. 7 compares the AUD performance of the proposed cloud computing-based scheme and *Baselines 1 and 2*. It can be observed that the cloud computing-based processing paradigm proposed in cell-free massive MIMO-based IoT can achieve a much better AUD performance than multi-cell non-cooperative massive MIMO-based IoT. The reason is that there are no cell boundaries in cell-free massive MIMO systems, and the inter-cell interference can be avoided via the APs' cooperation. Moreover, for $G\geq 20$, by further leveraging the angular-domain sparsity and the idea of SIC, the proposed SIC-based AUD and CE algorithm outperforms the conventional joint AUD and CE scheme, which is only based on the spatial-domain channel model. However, for the very low pilot overhead region (e.g., $G<20$), the joint AUD and CE scheme performs better than the proposed SIC-based method. This is because the AUS and CSI estimates are extremely inaccurate in this case, which leads to the error

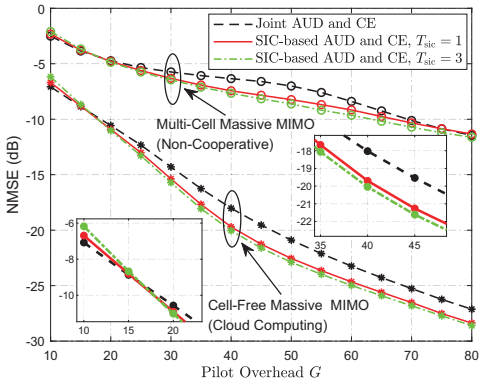


Fig. 8. CE performance comparison of the proposed cloud computing-based scheme and *Baselines 1 and 2*, where $M_c = 16$ and $\widetilde{P} = 1$.

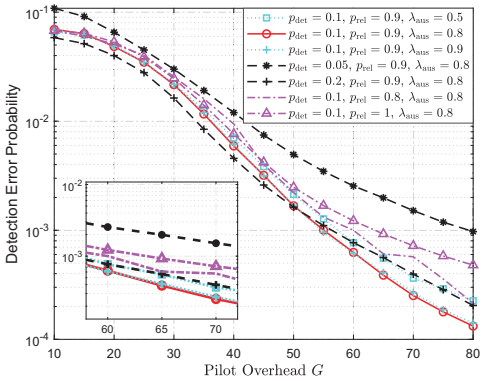


Fig. 9. AUD performance of the proposed cloud computing-based scheme under different $\lambda_{\rm aus}$, $p_{\rm det}$, and $p_{\rm rel}$, where $M_c = 16$, $\widetilde{P} = 1$, and $T_{\rm sic} = 3$.

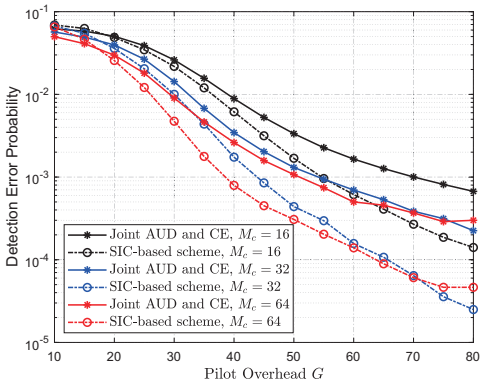


Fig. 10. AUD performance of the proposed SIC-based scheme and *Baseline 2* for different AP antennas M_c, where the cloud computing is considered.

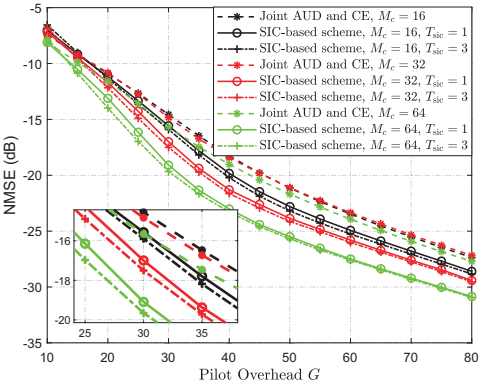


Fig. 11. CE performance of the proposed SIC-based scheme and *Baseline 2* for different AP antennas M_c, where the cloud computing is considered.

propagation of SIC. Fig. 8 depicts the CE performance of the considered schemes, which further validates the superiority of the proposed cell-free massive MIMO architecture and the SIC-based AUD and CE scheme for massive access. Here, the partially enlarged views show the NMSE performance for the pilot overhead regions $G \in [10, 20]$ and $G \in [30, 45]$, respectively. Fig. 9 further studies the influence of parameters $\lambda_{\rm aus}$, $p_{\rm det}$, and $p_{\rm rel}$ on massive access performance. As can be observed, when $\lambda_{\rm aus} = 0.8$, $p_{\rm det} = 0.1$, and $p_{\rm rel} = 0.9$, the proposed approach achieves the best AUD performance.

Fig. 10 and Fig. 11 verify the superiority of massive MIMO-based APs for grant-free massive access, where $T_{\rm sic} = 3$ and $\widetilde{P} = 1$ are considered. It is clear that the proposed cloud computing-based scheme can achieve a better performance by equipping more antennas at the APs. For AUD based on the spatial-domain channel model, a larger number of AP antennas enhances the spatial-domain structured sparsity of the channel matrix $\{\mathbf{H}_p\}_{p=1}^{P}$, which improves the accuracy of AUS estimate. On the other hand, a massive number of antennas can promote the angular-domain sparsity of massive MIMO channels, which can be leveraged to improve the CE performance. If the APs have a relatively small number of antennas (e.g., $M_c = 16$), the angular-domain sparsity of the massive MIMO channels would be weaken, and the performance of the proposed scheme would be degraded. Hence, the proposed scheme shows its superiority for massive MIMO cases. Fig. 12 and Fig. 13 show that the increased $\widetilde{P}$ also improves the AUD and CE performance of the proposed cloud computing-based scheme, where $T_{\rm sic} = 3$ is considered. This is because a larger $\widetilde{P}$ can also enhance the spatial-domain structured sparsity of the channel matrix $\{\mathbf{H}_p\}_{p=1}^{P}$.

B. Comparison of Cloud Computing and Edge Computing Paradigms

For the proposed SIC-based massive access scheme designed for cell-free massive MIMO-based IoT, we further compare two computing paradigms for the processing of AUD and CE, as shown in Fig. 14. By increasing the number of APs for cooperation, i.e., N_{co}, the AUD and CE performance of edge computing approaches that of cloud computing. Furthermore, we observe that only $N_{co} = 4$ APs are required for edge computing to obtain almost the same performance of cloud computing. This is because the channel gains from a specific active UE to the far away APs are approximate zero, the signals received at the remote APs can not further improve the AUD and CE performance. When all the N APs, consisting of a DPU-AP and $(N-1)$ conventional APs, cooperate, i.e., $N_{co} = N$, the edge computing paradigm is equivalent to the cloud one. Meanwhile, note that there is a tradeoff between the performance and the cost of practical DPU-AP deployment. Compared to the cloud computing, the edge computing can reap a more cost-effective cooperation (i.e., CPU burden, backhaul cost, and response time), while

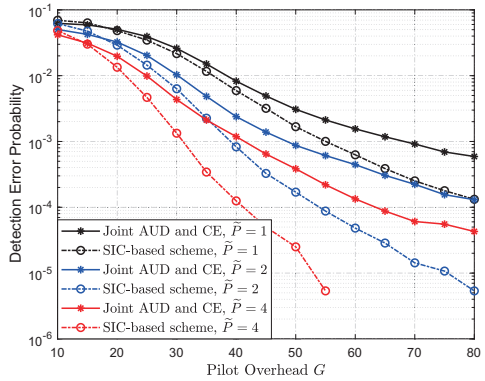


Fig. 12. AUD performance of the proposed SIC-based scheme and *Baseline 2* for different $\widetilde{P}$, where the cloud computing and $M_c = 16$ are considered.

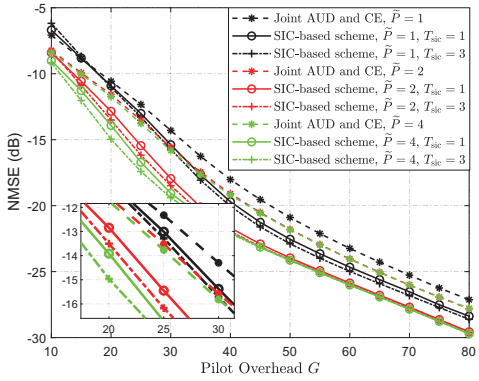
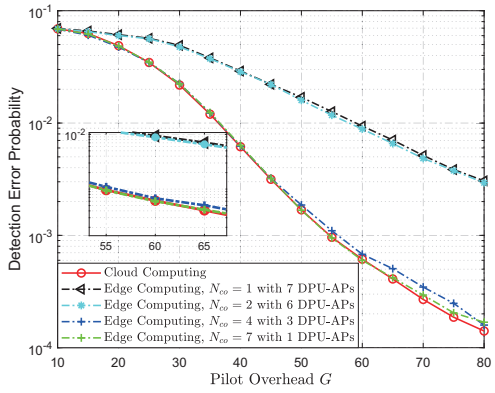
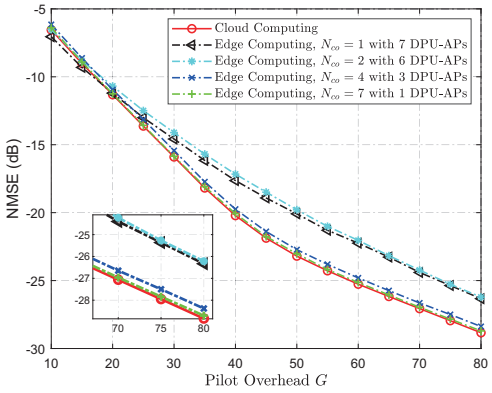


Fig. 13. CE performance of the proposed SIC-based scheme and *Baseline 2* for different $\widetilde{P}$, where the cloud computing and $M_c = 16$ are considered.



(a)



(b)

Fig. 14. Comparison of the proposed cloud computing and edge computing paradigms, where the proposed SIC-based scheme ($T_{\rm sic} = 3$) is considered, $M_c = 16$, $\widetilde{P} = 1$, and $Q = 10$: (a) AUD performance; (b) CE performance. For edge computing paradigm, the number of required DPU-APs is provided for different N_{co}.

may increase the price of network deployment, i.e., the DPU-APs should employ DPUs.

C. Massive Access Under Limited Backhaul Capacity

Fig. 15 and Fig. 16 verify the superiority of the proposed schemes based on SS-GAMP algorithm over those based on *Baseline 3*, where a low-resolution quantization of the processed signals is considered. Here, the number of quantization bits $Q = 3, 4$, and 5 are investigated. As can be observed, the proposed SS-GAMP algorithm can achieve a better performance than *Baseline 3* in both joint AUD and CE scheme and SIC-based scheme. This is because the quantization is taken into account by using the SS-GAMP algorithm with nonlinear module.

VII. Conclusion

This paper studies grant-free massive access in cell-free massive MIMO-based IoT, where multiple APs cooperate in the network to serve massive UEs. By exploiting the structured sparsity of the channel matrix, we develop a SS-GAMP algorithm for the CS recovery, where the quantization accuracy of the processed signals is considered. On this basis, a SIC-based AUD and CE algorithm is further proposed. Compared to the conventional massive access schemes based on the single-cell or multi-cell non-cooperative network architectures, cell-free massive MIMO can offer better coverage and improve the AUD and CE performance via AP cooperation. Furthermore, in contrast to the CS algorithms for SLM and the spatial-domain joint AUD and CE scheme, the proposed SIC-based scheme using SS-GAMP algorithm can significantly reduce the access latency in low-resolution quantization cases. Besides, we consider two computing paradigms, cloud computing and edge computing, to perform AUD and CE. Numerical simulations suggest that the performance of the edge computing can approach that of cloud computing. Meanwhile, edge computing makes the cooperation of APs more flexible and alleviates the burden on CPU and backhaul links, while may increase the cost of network deployment.

References

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," [Online]: arXiv preprint arXiv: 2002.03491, Feb. 2020.

[2] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55-61, Sep. 2017.

[3] C. Bockelmann *et al.*, "Towards massive connectivity support for scalable mMTC communications in 5G networks," *IEEE Access*, vol. 6, pp. 28969-28992, May. 2018.

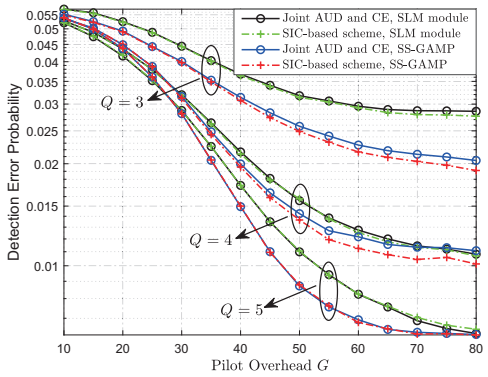


Fig. 15. AUD performance of the proposed SS-GAMP-based schemes and *Baseline 3* in low-accuracy quantization cases, where $M_c = 16$, $\widetilde{P} = 1$, and $T_{\text{sic}} = 3$.

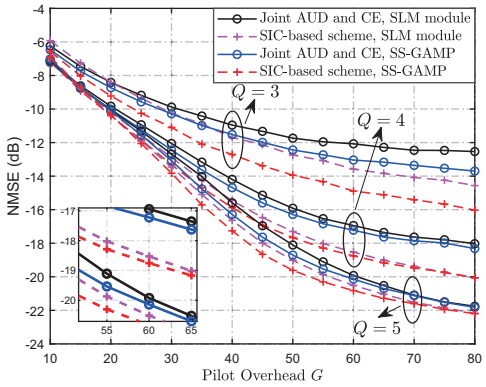


Fig. 16. CE performance of the proposed SS-GAMP-based schemes and *Baseline 3* in low-accuracy quantization cases, where $M_c = 16$, $\widetilde{P} = 1$, and $T_{\text{sic}} = 3$.

[4] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86-93, Jun. 2013.

[5] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88-89, Sep. 2018.

[6] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4-16, 1st Quart. 2018.

[7] X. Shao, X. Chen, C. Zhong, J. Zhao, and Z. Zhang, "A unified design of massive access for cellular internet of things," *IEEE Internet of Things J.*, vol. 6, no. 2, pp. 3934-3947, Apr. 2019.

[8] Z. Chen, F. Sohrabi, and W. Yu, "Multi-cell sparse activity detection for massive random access: Massive MIMO versus cooperative MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 1558-2248, Aug. 2019.

[9] H. Han, Y. Li, and X. Guo, "A graph-based random access protocol for crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7348-7361, Nov. 2017.

[10] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220-2234, Apr. 2017.

[11] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834-1850, Mar. 2017.

[12] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 972-974, Jul. 2012.

[13] A. T. Abebe and C. G. Kang, "Iterative order recursive least square estimation for exploiting frame-wise sparsity in compressive sensing-based MTC," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1018-1021, May. 2016.

[14] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473-1476, Jul. 2016.

[15] N. Vaswani and J. Zhan, "Recursive recovery of sparse signal sequences from compressive measurements: A review," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3523-3549, Jul. 2016.

[16] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320-2323, Nov. 2016.

[17] Y. Du, B. Dong, Z. Chen, X. Wang, Z. Liu, P. Gao, and S. Li, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812-2828, Dec. 2017.

[18] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640-643, Mar. 2017.

[19] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Proc. IEEE Intern. Sympos. Wireless Commun. Systems (ISWCS)*, Ilmenau, Germany, Aug. 2013, pp. 1-5.

[20] L. Liu and W. Yu, "Massive connectivity with massive MIMO-Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933-2946, Jun. 2018.

[21] S. Park, H. Seo, H. Ji, and B. Shim, "Joint active user detection and channel estimation for massive machine-type communications," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1-5.

[22] X. Shao, X. Chen, and R. Jia, "A dimension reduction-based joint activity detection and channel estimation algorithm for massive access," *IEEE Trans. Signal Process.*, vol. 68, pp. 420-435, Jan. 2020.

[23] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890-1904, Apr. 2018.

[24] J. Ahn, B. Shim, and K. B. Lee, "EP-based joint active user detection and channel estimation for massive machine-type communications," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5178-5189, Jul. 2019.

[25] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764-779, Jan. 2020.

[26] Q. He, T. Q. S. Quek, Z. Chen, Q. Zhang, and S. Li, "Compressive channel estimation and multi-user detection in C-RAN with low-complexity methods," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3931-3944, Jun. 2018.

[27] 3GPP TR 38.873 V12.2.0, Study on 3D channel model for LTE, Jun. 2015.

[28] Y. Zhou, M. Herdin, A. M. Sayeed, and E. Bonek, "Experimental study of MIMO channel statistics and capacity via the virtual channel representation," Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep., Feb. 2007.

[29] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169-6183, Dec. 2015.

[30] J. Nam, A. Adhikary, J. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876-890, Oct. 2014.

[31] X. Lin, S. Wu, L. Kuang, Z. Ni, X. Meng, and C. Jiang, "Estimation of sparse massive MIMO-OFDM channels with approximately common support," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1179-1182, May. 2017.

[32] X. Meng, S. Wu, and J. Zhu, "A unified Bayesian inference framework for generalized linear model," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 398-402, Mar. 2018.

[33] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graph and the sum-product algorithm," *IEEE Trans. Inform. Theory.*, vol. 47, no. 2, pp. 498-519, Feb. 2001.

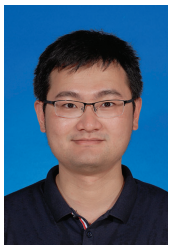
[34] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541-2556, May. 2016.

[35] S. Rangan, P. Schniter, and A. Fletcher, "On the convergence of approximate message passing with arbitrary matrices," in *Proc. Int. Symp. Inform. Theory (ISIT)*, Jun. 2014, pp. 236-240.

Malong Ke (Student Member, IEEE) received the B.S. degree in communication engineering from Shandong University, Jinan, China, in 2017. He is currently working towards the Ph.D. degree with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. His research interests include massive access for mMTC, massive MIMO systems, and sparse signal processing.

Zhen Gao (Member, IEEE) received the B.S. degree in information engineering from 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 Assistant Professor with the Beijing Institute of Technology, Beijing, China. His research interests are in wireless communications, with a focus on multi-carrier modulations, multiple antenna systems, and sparse signal processing. He was the recipient of the IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (Best Paper), the Exemplary Reviewer of IEEE COMMUNICATION LETTERS in 2016, *IET Electronics Letters* Premium Award (Best Paper) 2016, and the Young Elite Scientists Sponsorship Program (2018–2021) from China Association for Science and Technology.



Yongpeng Wu (Senior Member, IEEE) received the B.S. degree in telecommunication engineering from Wuhan University, Wuhan, China, in July 2007, and the Ph.D. degree in communication and signal processing from the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China, in November 2013.

He is currently a Tenure-Track Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. Previously, he was senior research fellow with the Institute for Communications Engineering, Technical University of Munich, Munich, Germany, and the Humboldt research fellow and the senior research fellow with the Institute for Digital Communications, University Erlangen-Nnüberg, Germany. During his doctoral studies, he conducted cooperative research with the Department of Electrical Engineering, Missouri University of Science and Technology, USA. His research interests include massive MIMO/MIMO systems, massive machine type communications, physical layer security, and signal processing for wireless communications.

Dr. Wu was awarded the IEEE Student Travel Grants for IEEE International Conference on Communications (ICC) 2010, the Alexander von Humboldt Fellowship in 2014, the Travel Grants for IEEE Communication Theory Workshop 2016, the Excellent Doctoral Thesis Awards of China Communications Society 2016, the Exemplary Editor Award of IEEE COMMUNICATION LETTERS 2017, and Young Elite Scientist Sponsorship Program by CAST 2017. He was an Exemplary Reviewer of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2015, 2016, and 2018 respectively. He was the lead guest editor for the special issue "Physical Layer Security for 5G Wireless Networks" of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and the guest editor for the special issue "Safeguarding 5G-and-Beyond Networks with Physical Layer Security" of the IEEE WIRELESS COMMUNICATIONS. He is currently an editor of the IEEE WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, and IEEE COMMUNICATIONS LETTERS. He has been a TPC member of various conferences, including Globecom, ICC, VTC, and PIMRC, etc.

Xiqi Gao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Southeast University, Nanjing, China, in 1997.

He joined the Department of Radio Engineering, Southeast University, in April 1992. Since May 2001, he has been a professor of information systems and communications. From September 1999 to August 2000, he was a visiting scholar at Massachusetts Institute of Technology, Cambridge, and Boston University, Boston, MA. From August 2007 to July 2008, he visited the Darmstadt University of Technology, Darmstadt, Germany, as a Humboldt scholar. His current research interests include broadband multicarrier communications, MIMO wireless communications, channel estimation and turbo equalization, and multirate signal processing for wireless communications. From 2007 to 2012, he served as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. From 2009 to 2013, he served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. From 2015 to 2017, he served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Dr. Gao was the recipient of the Science and Technology Awards of the State Education Ministry of China in 1998, 2006, and 2009, the National Technological Invention Award of China in 2011, and the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communications theory.

Kat-Kit Wong (Fellow, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from the Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively.

After graduation, he took up academic and research positions at the University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, the Smart Antennas Research Group of Stanford University, and the University of Hull, U.K. He is currently the Chair of wireless communications with the Department of Electronic and Electrical Engineering, University College London, U.K. His current research centers around 5G and beyond mobile communications. He is fellow of IET and is also on the editorial board of several international journals. He was a co-recipient of the 2013 IEEE Signal Processing Letters Best Paper Award, the 2000 IEEE VTS Japan Chapter Award at the IEEE Vehicular Technology Conference in Japan in 2000, and a few other international best paper awards. He has been the Editor-in-Chief of the IEEE WIRELESS COMMUNICATIONS LETTERS, since 2020.