**QUEEN'S
UNIVERSITY
BELFAST**

Towards Massive Connectivity for IoT in Mixed-ADC Distributed Massive MIMO

Yuan, J., He, Q., Matthaiou, M., Quek, T. Q. S., & Jin, S. (2019). Towards Massive Connectivity for IoT in Mixed-ADC Distributed Massive MIMO. *IEEE Internet of Things Journal*. <https://doi.org/10.1109/JIOT.2019.2957281>

Published in:
IEEE Internet of Things Journal

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

General rights

Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Towards Massive Connectivity for IoT in Mixed-ADC Distributed Massive MIMO

Jide Yuan, *Member, IEEE*, Qi He, Michail Matthaiou, *Senior Member, IEEE*,
Tony Q. S. Quek, *Fellow, IEEE*, and Shi Jin, *Senior Member, IEEE*

Abstract—Massive connectivity is a key requirement for the Internet of Things (IoT). In practice, the network should be capable of accessing thousands of devices and meeting their traffic demands. In this paper, we consider the access phase for IoT in a mixed-analog-to-digital converter distributed massive multiple-input multiple-output system, in which users are classified into light-load users and heavy-load users depending on their traffic load requirements. To meet the low-latency and low-cost demands in IoT, the access scheme for both types of users are designed in a grant-free fashion. For users with light-load traffic demands, by formulating the user activity detection (UAD) and channel estimation (CE) into a compressed sensing problem, we provide a low-complexity algorithm solver which requires no prior information. The simulation results verify that the proposed algorithm can effectively detect user activity and estimate channel state information (CSI) between users and access points (APs). To satisfy the throughput requirements of heavy-load users, after UAD and CE, a two-step dynamic clustering is proposed for coordinated multi-point transmission using the large scale fading (LSF) information. The impact of quantization noise on LSF estimation is investigated, as well as, a corresponding compensation method and accuracy bound. By detecting the clustering behavior among users in the first step, the complexity of the joint user and AP clustering is substantially reduced. Numerical results reveal that the proposed algorithm can offer significant performance gains in various scenarios with fast convergence.

Index Terms—Channel estimation, distributed mMIMO, dynamic clustering, Internet of Things (IoT), user activity detection, weighted sum rate maximization.

I. INTRODUCTION

A. Motivation

Internet of Things (IoT) is a typical scenario in fifth generation (5G), in which the communication of the objects around us will be automated and intelligent [2–4].

As the number of devices increases exponentially, a spectrum of novel applications emerge which can be classified based on the type of performance metrics, such as coverage, time latency, reliability, and throughput [5–7]. We categorize the applications into three domains based on their specific requirements: 1) Personal; 2) Public utility; 3) Community. This categorization is depicted in Fig. 1, in which applications in personal IoT occur frequently and have small data requirements; applications in public utility IoT require extremely high reliability and low latency to satisfy the strict security constraints; applications in community IoT need continuous transmission and high throughput. As such, all the communication relevant issues, such as, access, transmission and resource management among different IoT domains should be considered based on their demands. In this paper, our aim is to provide a joint analysis for the access problem in personal and community IoT; for this reason, each type of traffic in each domain will be hereafter classified as **light-load** and **heavy-load** for convenience.

The type of architecture, beyond all doubt, is the cornerstone for the IoT development. Without an appropriate architecture, the discussions on the access scheme for different scenarios and the corresponding techniques are out of the question. Recently, distributed IoT systems have attracted attention for their reliability, robustness, and flexibility, which allows each access point (AP) to share information with the cloud for centralized processing [8–10]. However, these benefits of distributed systems do not come for free. In particular, a massive number of APs will boost the hardware cost and electronic power consumption to prohibitive levels. This motivates us to employ some low-resolution analog-to-digital converters (ADCs) to implement distributed IoT networks for their low power consumption properties [11]. Mixed-ADC distributed systems have been proved to be practical in cellular networks, since they nearly equivalent performance to that of performance of conventional high-cost architectures with much less cost [12–14]. The success of mixed-ADC architectures in cellular networks drives us to thoroughly examine the practicability of the distributed IoT architectures with mixed-ADC APs, and devise possible solutions to address the access problem of **light-load** and **heavy-load** users.

A conference version of this paper has appeared in [1].

Manuscript received September 22, 2018; revised April 14, 2019, and November 12, 2019; accepted November 21, 2019. The work of J. Yuan and M. Matthaiou was supported by the RAEng/The Leverhulme Trust Senior Research Fellowship LTSRF1718\14\2. The work of T. Q. S. Quek was supported in part by the Singapore University of Technology and Design-Zhejiang University (SUTD-ZJU) Research Collaboration under Grant SUTDZJU/RES/01/2016, and the SUTD-ZJU Research Collaboration under Grant SUTD-ZJU/RES/05/2016. The work of S. Jin was supported in part by the National Science Foundation of China (NSFC) under Grant 61531011 and the NSFC for Distinguished Young Scholars of China with Grant 61625106. The associate editor coordinating the review of this paper and approving it for publication was Prof. C. Stefanović. (*Corresponding author: Shi Jin.*)

J. Yuan and S. Jin are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, P. R. China (e-mail: {yuanjide, jinshi}@seu.edu.cn).

Q. He is with the School of Engineering, Westlake University, Hangzhou 310024, P. R. China (e-mail: heqi.tech@hotmail.com).

M. Matthaiou is with the Institute of Electronics, Communications and Information Technology (ECIT), Queen's University Belfast, Belfast, BT3 9DT, U.K. (e-mail: m.matthaiou@qub.ac.uk).

T. Q. S. Quek is with the Singapore University of Technology and Design, Singapore 487382, Singapore (e-mail: tonyquek@sutd.edu.sg).

Fig. 1. Internet of Things schematic illustration showing the applications based on their traffic requirements.

In current Long Term Evolution (LTE), multiple requests and authorizations are required for the connections to be established [15]. Unfortunately, such a procedure may become inappropriate to support the massive connectivity requirements in IoT due to three main reasons. First, such a grant-based access scheme incurs significant overhead which is unaffordable for a massive number of devices. Some contention-based schemes have been studied [16, 17], in which a random access protocol is proposed that allows users who randomly select the unoccupied preambles to build the connections. However, these schemes are not suitable for IoT, since a large amount of devices may cause frequent collisions. Second, the current onefold access scheme is incompetent to treat objects separately based on their substantially different requirements [18]. For instance, **light-load** users who require very limited resources can access to and release from the network in advance, and to satisfy the throughput of **heavy-load** users, the network needs to schedule enough resources for devices during the access phase. Third, a grant-based access scheme cannot be applied to distributed multiple-input-multiple-output (MIMO) systems, one of the emerging architectures, since with a large number of APs spread over a vast area, it would be impractical to force users to communicate with all the APs to grant access.

As a result, for IoT, a novel access scheme that can effectively reduce the access overhead and be also capable of accessing devices while meeting their traffics demands is of pivotal importance.

B. Related work

We consider a massive connectivity scenario in mixed-ADC distributed massive MIMO (mMIMO) systems to support IoT [19, 20], in which a portion of APs are equipped with low-resolution ADCs. However, due to the low-quality of received signals from the antennas equipped with low-resolution ADCs,

it is still unclear whether low-resolution ADCs can satisfy the strict accuracy requirements of control signaling during the access phase.

In distributed mMIMO, for the devices with **light-load** traffic, a reasonable access method is to associate them to their neighboring AP after user active detection (UAD) and channel estimation (CE), and for the devices with **heavy-load** traffic, besides UAD and CE, dynamic clustering is further needed to achieve coordinated multi-point transmission for supporting the high throughput requirements. As such, we investigate the access scheme for two types of traffic; particularly, we focus on accuracy and complexity of UAD and CE for **light-load** scenarios, whereas the average throughput is our main concern for **heavy-load** scenarios.

For **light-load** scenarios in IoT, since reducing access overhead is one of the fundamental challenges during the access phase, grant-free schemes have attracted some research attention, where UAD and CE can be performed in one time slot simultaneously based on received pilots. The grant-free scheme is feasible due to the sparsity in user activity pattern since only a small proportion of devices in the system are active. The UAD and CE problems can then be formulated into a compressive sensing (CS) form [21–25]. In [22], a joint UAD and data estimation scheme was investigated for code division multiple access system with perfect channel state information (CSI) at the base station (BS). The authors in [23] proposed a modified Bayesian compressive sensing algorithm for cloud radio access network (C-RAN), which can achieve the same level of performance by exploiting the large-scale fading (LSF) information. However, the assumption that LSF is perfectly known at the BS is controversial since the prediction of LSF between each user and each AP can be computationally intensive [24]. Another approach has also been studied in [25], in which the proposed algorithm exploited no prior information of the transmitted symbols. Nevertheless, the performance of such scheme is analyzed under the assumption

of ideal hardware.

For **heavy-load** scenarios in IoT, as we focus on improving throughput, coordinated multi-point transmission is an inherently reasonable approach where devices and distributed APs should be jointly clustered according to their throughput demands. Related works have been reported in the literature. For instance, [26, 27] considered a fixed BS cluster scheme, and investigated the network utility with limited backhaul capacity constraints. Further, [28] proposed a heuristic method aiming to maximize the overall achievable rate, while [29] investigated the problem from the perspective of maximizing the cell-edge user rate. Also, [30] designed a dynamic clustering scheme by forming the overall channel information into a block diagonal matrix, in which way, the user clustering and downlink beamforming are performed simultaneously by leveraging its sparsity. However, from an engineering perspective, the above schemes may not be suitable for practical deployment as they require the instantaneous CSI. The fast fading within the CSI brings unstable clustering results and leads to an excessive time delay because of the unaffordable complexity. Therefore, to guarantee the stability of the clustering results, this paper advocates a LSF information based clustering scheme. A similar problem has been investigated in the context of heterogeneous networks [31, 32], in which BS clustering and multipoint joint transmissions are addressed via graph theory. Note that LSF is generally extracted from instantaneous CSI [33], which implies that we need to store the previous CSI as the observations to estimate LSF. Fortunately, a large proportion of devices which require high throughput, such as surveillance and video conference, remain geographically stationary, which allow us to acquire the CSI data over a long time scale.

C. Main Contributions

In this paper, we provide a grant-free access scheme for **light-load** and **heavy-load** scenarios in IoT. More specifically, the LSF information required for accessing users with **heavy-load** requirements is obtained via the proposed UAD and CE algorithm used for **light-load** users. We consider Rician fading channels in mixed-ADC distributed mMIMO, in which APs employ mixed resolution ADCs.

For **light-load** scenarios in IoT, as accuracy and complexity of UAD and CE is our main concern, we formulate them into a CS problem, and adopt the alternating direction method of multipliers (ADMM) algorithm. The impact of the low-resolution ADCs on the performance is evaluated, and a solution that can compensate such impact is proposed. Simulation results reveal that the proposed algorithm can simultaneously detect the user activity and estimate CSI in various propagation environments with much lower computational complexity compared with conventional CS-based methods.

Then, to approach coordinated multiple-point transmission for **heavy-load** scenarios in IoT, a two-step dynamic clustering algorithm is proposed. At the first step, the clustering features are identified among users; then, a joint clustering

algorithm is performed to solve the sum rate maximization problem. In contrast to prior works [30, 34], the fronthaul capacity constraints and AP operating power are considered. By detecting the clustering features among the devices, the complexity of the algorithm is significantly reduced. Numerical results verify the fast convergence of our proposed algorithm, with which the achievable sum rate and antenna selection matrix converge within only a few iterations. Moreover, remarkable gains on the sum rate are achieved with the proposed algorithm; these gains are more pronounced in the high fronthaul limit or the high AP operating power scenarios.

It is worth mentioning that, to provide stable clustering results for **heavy-load** scenarios, the clustering algorithm is operated based on the LSF which can be predicted by averaging the estimated CSI obtained in the same way as in **light-load** scenarios. For practical reference, the relationship between the amount of observations and the accuracy of LSF estimation is rigorously investigated, and upper bounds on the accuracy are provided. We further investigate the impact of signal quantization on LSF estimation, and propose a compensation solution based on the theoretical results. We point out that our numerical results tend to approach the accuracy bound for high transmit power and converge to the bound in a dense system. The estimated CSI can be stored and is further used as observations to obtain LSF for dynamic clustering in **heavy-load** scenarios.

D. Paper Organization

The paper is organized as follows: Section II presents the system model of IoT in mixed-ADC distributed mMIMO. Section III provides the UAD and CE algorithm for light-load scenarios. Section IV details the dynamic clustering procedure for heavy-load scenarios as well as the corresponding analysis. The numerical results presented in Section V demonstrates the feasibility of our scheme, and Section VI includes the conclusions while proofs are relegated to appendices.

Notation—Throughout this paper, vectors and matrices are denoted in bold lowercase letters and bold uppercase letters, respectively. We use $\{\mathbf{A}\}_i$ and a_{ij} to denote the i th row and (i, j) th entry in $\mathbf{A}$, respectively. The operation $\|\mathbf{A}\|_p$ denotes the ℓ_p -norm of the matrix $\mathbf{A}$, and $\text{diag}(\mathbf{a})$ denotes the diagonal matrix of the vector $\mathbf{a}$. The superscripts $(\cdot)^H$, $\text{tr}(\cdot)$ and $\mathbb{E}[\cdot]$ denote the matrix Hermitian transpose, trace and the expectation of the input random entity, respectively. Finally, $(\mathbf{A} \cdot \mathbf{B})$ denotes the Hadamard product (element-wise product between matrices $\mathbf{A}$ and $\mathbf{B}$), and $\mathbf{A}^{\circ(a)}$ denotes an element-wise exponentiation operation on $\mathbf{A}$, i.e., $(\mathbf{A}^{\circ(a)})_{i,j} = (\mathbf{A}_{i,j})^a$.

II. SYSTEM MODEL

Consider a multi-user distributed mMIMO system consisting of M single-antenna APs, in which the number of APs with full-resolution ADCs (F-APs) is M_f , and the number of APs with low-resolution ADCs (L-APs) is $M_l = M - M_f$. Specifically, F-APs can be regarded to macro base stations (BSs) which are deployed on a regular grid for coverage, whereas L-APs can be regarded to micro BSs which are

Fig. 2. Illustrative system model of the considered mixed-ADC distributed mMIMO, where F-APs and L-APs are connected to the centralized processor via a capacity limited fronthaul.

deployed randomly to improve quality of service, shown in Fig. 2. In this paper, we assume that the fronthaul used for connecting central processor and all the APs is error-free with limited capacity. There are K users randomly distributed in the system, and only a proportion of them $\mathcal{K}_A \subseteq \{1, \dots, K\}$ are active ($|\mathcal{K}_A| = K_A \ll M$).

The following assumptions are made in this paper:

- The complex propagation coefficient at the t th time slot between the i th antenna and j th antenna are modelled as

$$g_{ij}^{(t)} = h_{ij}^{(t)} \sqrt{\beta_{ij}}, \quad (1)$$

where β_{ij} represents the LSF coefficient which remains constant over many coherence, and $h_{ij}^{(t)}$ represents the small-scale fading coefficient which comprise a deterministic component corresponding to the line-of-sight (LoS) signal and a Rayleigh-distributed component corresponding to the scattered signals. Thus,

$$h_{ij}^{(t)} = \sqrt{\frac{\kappa}{\kappa+1}} \hat{h}_{ij} + \sqrt{\frac{1}{\kappa+1}} \tilde{h}_{ij}^{(t)} \quad (2)$$

where $\hat{h}_{ij}$ is a deterministic value, $\tilde{h}_{ij}^{(t)}$ is a random complex component with zero mean and unit variance, while the parameter κ is the Ricean K -factor which denotes the ratio of the power of the deterministic component to the power of the scattered signals.¹ We denote the channel coefficient from the k th user to the m th F-APs and from the k th user to the m' th L-AP as $g_{f,mk}^{(t)}$ and $g_{l,m'k}^{(t)}$, respectively.

- For **light-load** scenarios in IoT, once APs receive the pilot sequences from active users, users are simply associated to their neighboring APs after UAD and CE. For these

users, we focus on the accuracy and complexity of UAD and CE algorithm, while the scheduling will not be addressed in this paper.

- For **heavy-load** scenarios in IoT, after UAD and CE, dynamic clustering is further pursued using the estimated LSF. To ensure accurate decoding of controlling signals, we assume that there is at least one F-AP in each cluster for its high signal quality.

Assume that users send the pilot sequences simultaneously, and the k th user is assigned the pilot sequence $\mathbf{p}_k \in \mathbb{C}^{1 \times \tau}$ of length τ with $\|\mathbf{p}_k\|^2 = \tau p_p$, where p_p represents the transmit power of pilot symbol. The received signal vector $\mathbf{y}_{f,m} \in \mathbb{C}^{1 \times \tau}$ at the m th F-AP is given by

$$\mathbf{y}_{f,m}^{(t)} = \sum_{k \in \mathcal{K}_A} g_{f,mk}^{(t)} \mathbf{p}_k + \mathbf{n}_m^{(t)}, \quad (3)$$

where $\mathbf{n}_m^{(t)} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$ denotes the additive Gaussian noise. Similarly, the received signal vector $\mathbf{y}_{l,m'}^{(t)} \in \mathbb{C}^{1 \times \tau}$ at the m' th L-AP can be expressed as follows:

$$\mathbf{y}_{l,m'}^{(t)} = \alpha \sum_{k \in \mathcal{K}_A} g_{l,m'k}^{(t)} \mathbf{p}_k + \alpha \mathbf{n}_{m'}^{(t)} + \mathbf{n}_{q,m'}^{(t)}, \quad (4)$$

where $\mathbf{n}_{m'}^{(t)}$ is the noise vector at m' th L-AP, $\mathbf{n}_{q,m'}^{(t)}$ denotes the quantization noise vector, α represents the quantization coefficient corresponding to the quantization bits b , as listed in Table I [35]. Along with the fact $\|\mathbf{p}_k\|^2 = \tau p_p$, the variance of $\mathbf{n}_{q,m'}^{(t)}$ can be evaluated as [36]

$$\text{Var}(\mathbf{n}_{q,m'}) = \tau \alpha (1 - \alpha) \left(\sigma^2 + p_p \sum_{k' \in \mathcal{K}_A} \beta_{m'k'} \right). \quad (5)$$

We now introduce the active indicator function

$$\delta_{\mathcal{A}}(a) = \begin{cases} 1, & \text{if } a \in \mathcal{A}, \\ 0, & \text{if } a \notin \mathcal{A}. \end{cases} \quad (6)$$

The central processor is able to stack the received pilot data from M APs into a matrix, given as

$$\mathbf{Y}^{(t)} = \mathbf{G}^{(t)} \mathbf{\Delta} \mathbf{P} + \mathbf{N}^{(t)}, \quad (7)$$

where

$$\mathbf{Y}^{(t)} = \left[\left(\mathbf{y}_{f,1}^{(t)} \right)^T, \dots, \left(\mathbf{y}_{f,M_f}^{(t)} \right)^T, \left(\mathbf{y}_{l,1}^{(t)} \right)^T, \dots, \left(\mathbf{y}_{l,M_l}^{(t)} \right)^T \right]^T,$$

while $\mathbf{\Delta} = \text{diag}(\delta_{\mathcal{K}_A}(k))_{k=1, \dots, K}$, and $\mathbf{P} = [\mathbf{p}_1^T, \dots, \mathbf{p}_K^T]^T \in \mathbb{C}^{K \times \tau}$. The noise term $\mathbf{N}^{(t)}$ consists of both AWGN and quantization noise, and is described as $\mathbf{N}^{(t)} = \left[\left(\mathbf{N}_f^{(t)} \right)^T, \left(\mathbf{N}_l^{(t)} \right)^T \right]^T \in \mathbb{C}^{M \times K}$ with

$$\mathbf{N}_f^{(t)} = \left[\left(\mathbf{n}_{f,1}^{(t)} \right)^T, \dots, \left(\mathbf{n}_{f,M_f}^{(t)} \right)^T \right]^T$$

and

$$\mathbf{N}_l^{(t)} = \left[\left(\mathbf{n}_{q,1}^{(t)} + \alpha \mathbf{n}_{l,1}^{(t)} \right)^T, \dots, \left(\mathbf{n}_{q,M_l}^{(t)} + \alpha \mathbf{n}_{l,M_l}^{(t)} \right)^T \right]^T.$$

Moreover, $\mathbf{G}^{(t)} = [\mathbf{g}_1^{(t)}, \dots, \mathbf{g}_K^{(t)}] \in \mathbb{C}^{M \times K}$, where the k th column represents the channel propagation coefficient vector

¹In a practical scenario, the κ -factor is normally different across users. We here assume uniform κ across users in order to investigate its impact on the LSF estimation algorithm shown in Section IV.

TABLE I
THE QUANTIZATION COEFFICIENT α FOR DIFFERENT ADCS
QUANTIZATION BITS b

b	1	2	3	4	5
α	0.6336	0.8825	0.95645	0.990503	0.997501

of the k th user, and is given by

$$\mathbf{g}_k^{(t)} = \left[\left(\mathbf{g}_{f,k}^{(t)} \right)^T, \alpha \left(\mathbf{g}_{l,k}^{(t)} \right)^T \right]^T.$$

III. UAD AND CE FOR LIGHT-LOAD SCENARIOS

In this section, the UAD and CE for users with light-load traffic is formulated into a CS problem without any prior information.

A. Problem Formulation

The sparsity of the user activity pattern motivates us to adopt a CS technique to solve the UAD and CE problem. We, thus, transform the expression in (7) into a standard CS form:

$$\left(\mathbf{Y}^{(t)} \right)^H = \mathbf{P}^H \Delta \left(\mathbf{G}^{(t)} \right)^H + \left(\mathbf{N}^{(t)} \right)^H. \quad (8)$$

As shown in (8), the expression perfectly matches the standard CS formulation, where $\left(\mathbf{Y}^{(t)} \right)^H$ and $\mathbf{P}^H$ are the compressed observation and the measurement matrix, respectively.

By denoting $\mathbf{A}^{(t)} = \Delta \left(\mathbf{G}^{(t)} \right)^H \in \mathbb{C}^{K \times M}$, it is straightforward to reveal the “row sparsity” of $\mathbf{A}^{(t)}$ due to the sparse active users. Moreover, “element sparsity” can be observed from $\mathbf{A}^{(t)}$ since most of APs are far away from a particular user. To leverage these two properties, we formulate our problem into a $\ell_{2,0}$ -regularization function denoted as $f(\mathbf{A}^{(t)})$:

$$f(\mathbf{A}^{(t)}) \triangleq \min_{\mathbf{A}^{(t)}} \frac{1}{2} \left\| \left(\mathbf{Y}^{(t)} \right)^H - \mathbf{P}^H \mathbf{A}^{(t)} \right\|_2^2 + \gamma_1 \sum_{k=1}^K \sum_{m=1}^M \delta(a_{km}^{(t)}) + \gamma_2 \sum_{k=1}^K \delta \left(\left\{ \mathbf{A}^{(t)} \right\}_k \right), \quad (9)$$

where $\delta(\cdot)$ is an indicator function defined as following:

$$\delta(\mathbf{A}) = \begin{cases} 0, & \text{if } \|\mathbf{A}\|_0 = 0, \\ 1, & \text{otherwise.} \end{cases}$$

The function of (9) considers jointly the element sparsity and row sparsity, and the parameters γ_1 and γ_2 are used to fit the sparsity pattern, and balance the tradeoff between the sparsity solutions. Note that $\gamma_1 = 0$ gives the group-LASSO fit, in which the sparsity of user activities is detected [37], while $\gamma_2 = 0$ gives the LASSO fit, in which the sparse relationship between APs and each active user is detected. Both forms belong to the regression analysis framework which aim to enhance the prediction accuracy by performing both variable selection and regularization. Our aim is to estimate $\mathbf{A}^{(t)}$, which consists of the indication of active users Δ and channel gain $\mathbf{G}^{(t)}$.

Unfortunately, solving (9) is very difficult since we need to sweep across all possible subsets of variables, which would be

impractical for a distributed mMIMO system. In the following, we provide two algorithms that can effectively solve the problem.

B. Sparse Group Lasso Approach

We note that both “row sparsity” and “element sparsity” exist in $\mathbf{A}^{(t)}$, which allow us to relax our problem to comply with the “Sparse Group Lasso (SGLasso)” remit [37]. We first extend the “SGLasso” fit into a matrix pattern, shown as:

$$\min_{\mathbf{A}^{(t)}} \frac{1}{2} \left\| \left(\bar{\mathbf{Y}}^{(t)} \right)^H - \sum_{k=1}^K \mathbf{p}_k^H \left\{ \mathbf{A}^{(t)} \right\}_k \right\|_2^2 + \gamma_1 \sum_{k=1}^K \left\| \left\{ \mathbf{A}^{(t)} \right\}_k \right\|_2 + \gamma_2 \sum_{k=1}^K \left\| \left\{ \mathbf{A}^{(t)} \right\}_k \right\|_1, \quad (10)$$

where $\bar{\mathbf{Y}}^{(t)}$ is defined as

$$\bar{\mathbf{Y}}^{(t)} \stackrel{(a)}{=} \left[\left(\mathbf{Y}_f^{(t)} \right)^T, \left(\sqrt{1/\alpha} \mathbf{Y}_l^{(t)} \right)^T \right]^T. \quad (11)$$

The equation (a) is due to the signal strength attenuation because of the introduced quantization noise. Since this attenuation affects the second-order statistics of the estimated CSI, i.e., LSF, a rigorous technique to compensate the distortion is given in *Proposition 2*.

In equation (10), we simply set each row as a group of interests. The solution of (10) is rather useful because active users can be detected by the nonzero rows of $\mathbf{A}^{(t)}$, and a proportion of APs whose corresponding column is zero can be set in idle mode to save power. Note that the choice of γ_1 and γ_2 are crucial for the performance of (10). A small γ_1 and γ_2 may be inefficient for processing, while large values may lead to biased estimation. In practice, “row sparsity” dominates in a dense and small distributed system, where larger γ_1 and smaller γ_2 will exploit the two types of sparsities more efficiently. When the system becomes large, the effect of “element sparsity” gradually dominates, in which case a larger γ_2 becomes more effective for processing. The upper bounds of choosing γ_1 and γ_2 are given in the following proposition.

Proposition 1: Denote by $\hat{\mathbf{A}}^{(t)}$ the solution of (10). We have $\hat{\mathbf{A}}^{(t)} = 0$ if $\gamma_1 \geq \|\mathbf{p}_k(\bar{\mathbf{Y}}^{(t)})^H\|_2$ or $\gamma_2 \geq \frac{1}{M} \|\mathbf{p}_k(\bar{\mathbf{Y}}^{(t)})^H\|_2$.

Proof: To prove the proposition, we separate the problem into two parts, namely

$$\min_{\mathbf{A}^{(t)}} \frac{1}{2} \left\| \left(\bar{\mathbf{Y}}^{(t)} \right)^H - \sum_{k=1}^K \mathbf{p}_k^H \left\{ \mathbf{A}^{(t)} \right\}_k \right\|_2^2 + \gamma_1 \sum_{k=1}^K \left\| \left\{ \mathbf{A}^{(t)} \right\}_k \right\|_2 + \min_{(\mathbf{A}')^{(t)}} \gamma_2 \sum_{k=1}^K \left\| \left\{ (\mathbf{A}')^{(t)} \right\}_k \right\|_1 \leq (10), \quad (12)$$

where equality holds when solutions $\hat{\mathbf{A}}^{(t)} = (\hat{\mathbf{A}}')^{(t)}$. Since the objective function is convex, the k th group of the minimum solution $\hat{\mathbf{A}}$ must satisfy

$$\mathbf{p}_k \left(\mathbf{p}_k^H \left\{ \mathbf{A}^{(t)} \right\}_k - \left(\bar{\mathbf{Y}}^{(t)} \right)^H \right) - \gamma_1 u_k = 0, \quad (13)$$

where u_k are the subgradient of $\left\| \left\{ \mathbf{A}^{(t)} \right\}_k \right\|_2$, which have the

following property [41, Eq. (129)]

$$u_k = \begin{cases} \frac{\{\mathbf{A}^{(t)}\}_k}{\|\{\mathbf{A}^{(t)}\}_k\|_2}, & \text{if } \{\mathbf{A}^{(t)}\}_k \neq \mathbf{0}, \\ \in \{u_k : \|u_k\|_2 \leq 1\}, & \text{if } \{\mathbf{A}^{(t)}\}_k = \mathbf{0}. \end{cases} \quad (14)$$

By substituting (14) into (13), we demonstrate that the solution is zero when

$$\gamma_1 \geq \left\| \mathbf{P}_k \left(\bar{\mathbf{Y}}^{(t)} \right)^H \right\|_2. \quad (15)$$

Note that the solution of $\min_{(\mathbf{A}')^{(t)}} \gamma_2 \sum_{k=1}^K \|\{\mathbf{A}'^{(t)}\}_k\|_1$ is zero. The solution of (13) is zero when γ_1 satisfies (15). For γ_2 , by noting that $\|\{\mathbf{A}^{(t)}\}_k\|_1 = \sum_{m=1}^M \|a_{km}^{(t)}\|_2$, we can obtain the result in the same way. ■

In fact, the minimization problem in (10) can be transformed into a standard second-order cone program (SOCP) problem [25], which can be solved directly by the interior point method. According to the complexity analysis in [38], the algorithm needs $O(\sqrt{MK})$ iterations while the cost per iteration is $O((MK)^3)$ to solve SOCP. Therefore, the overall computational complexity of (10) is $O((MK)^{3.5})$. It can be seen that, with a large channel matrix in size, the algorithm may fail due to the extremely high complexity. We hence provide an efficient algorithm to accelerate the estimation procedure.

C. ADMM Approach

The ADMM is a variant of the augmented Lagrangian algorithm that splits the variables and uses partial updates for the dual variables [39], which is perfectly suitable for the considered large system in IoT. In our problem, we split $\mathbf{A}^{(t)}$, and introduce two auxiliary variables $\Theta^{(t)}$ and $\Phi^{(t)}$ to fit the ADMM criterion:

$$\begin{aligned} \min_{\mathbf{A}^{(t)}, \Theta^{(t)}, \Phi^{(t)}} & \frac{1}{2} \left\| \left(\bar{\mathbf{Y}}^{(t)} \right)^H - \sum_{k=1}^K \mathbf{P}_k^H \{\mathbf{A}^{(t)}\}_k \right\|_2^2 \\ & + \gamma_1 \sum_{k=1}^K \left\| \{\Theta^{(t)}\}_k \right\|_2 + \gamma_2 \sum_{k=1}^K \left\| \{\Phi^{(t)}\}_k \right\|_1, \\ \text{s.t.} \quad & \Theta^{(t)} = \mathbf{A}^{(t)}, \Phi^{(t)} = \mathbf{A}^{(t)}. \end{aligned} \quad (16)$$

The problem (16) is equivalent to (10) due to the extra constraints, and is separable in $\mathbf{A}^{(t)}$, $\Theta^{(t)}$, and $\Phi^{(t)}$, and, thus, ADMM is applicable. To solve (16), we first transform the problem into an augmented Lagrangian problem as [40]

$$\begin{aligned} \mathcal{L}(\mathbf{A}^{(t)}, \Theta^{(t)}, \Phi^{(t)}, \lambda_1, \lambda_2) &= \frac{1}{2} \left\| \left(\bar{\mathbf{Y}}^{(t)} \right)^H - \sum_{k=1}^K \mathbf{P}_k^H \{\mathbf{A}^{(t)}\}_k \right\|_2^2 + \gamma_1 \sum_{k=1}^K \left\| \{\Theta^{(t)}\}_k \right\|_2 \\ &+ \gamma_2 \sum_{k=1}^K \left\| \{\Phi^{(t)}\}_k \right\|_1 + \frac{\chi}{2} \left(\left\| \Theta^{(t)} - \mathbf{A}^{(t)} \right\|_2^2 + \left\| \Phi^{(t)} - \mathbf{A}^{(t)} \right\|_2^2 \right) \\ &+ \Re \left\{ \text{tr} \left(\lambda_1^H \left(\Theta^{(t)} - \mathbf{A}^{(t)} \right) \right) + \text{tr} \left(\lambda_2^H \left(\Phi^{(t)} - \mathbf{A}^{(t)} \right) \right) \right\} \end{aligned} \quad (17)$$

where λ_1 and λ_2 are the Lagrangian multipliers which have the same dimension as $\mathbf{A}^{(t)}$, $\Re\{\cdot\}$ stands for real part of a complex number, and χ is the regularization parameter. The algorithm is operated by minimizing $\mathbf{A}^{(t)}$, $\Theta^{(t)}$ and $\Phi^{(t)}$ in an iterative manner. Since each of the subproblems is a convex problem, we now provide the updating algorithm as below:

- For $\mathbf{A}^{(t)}$, by taking the derivative with respect to $\mathbf{A}^{(t)}$ in (17) and setting it to zero, the minimum $\dot{\mathbf{A}}^{(t)}$ can be evaluated as

$$\begin{aligned} \dot{\mathbf{A}}^{(t)} &= (\mathbf{P}\mathbf{P}^H + 2\chi\mathbf{I})^{-1} \left(\mathbf{P} \left(\mathbf{Y}^{(t)} \right)^H \right. \\ &\quad \left. + \chi \left(\Theta^{(t)} + \Phi^{(t)} \right) - (\lambda_1 + \lambda_2) \right). \end{aligned} \quad (18)$$

The expression can be obtained by utilizing [41, Eq. (100-104), (119)].

- For $\Theta^{(t)}$, we reform the minimization problem by completing the square, given as

$$\begin{aligned} \sum_{k=1}^K \min_{\{\Theta^{(t)}\}_k} & \left\{ \gamma_1 \left\| \{\Theta^{(t)}\}_k \right\|_2 \right. \\ & \left. + \frac{\chi}{2} \left\| \{\Theta^{(t)}\}_k - \dot{\mathbf{A}}^{(t)} - \frac{1}{\chi} \lambda_1 \right\|_2^2 \right\}. \end{aligned} \quad (19)$$

By taking the derivative with respect to $\{\Theta^{(t)}\}_k$, and utilizing the subgradients property as in (14), the optimal solution of each row in (19) can be evaluated as

$$\begin{aligned} \{\dot{\Theta}^{(t)}\}_k &= \max \left\{ 1 - \frac{2\gamma_1}{\chi} \left\| \left\{ \dot{\mathbf{A}}^{(t)} + \frac{1}{\chi} \lambda_1 \right\}_k \right\|_2^{-1}, 0 \right\} \\ &\quad \times \left\{ \dot{\mathbf{A}}^{(t)} + \frac{1}{\chi} \lambda_1 \right\}_k. \end{aligned} \quad (20)$$

- Noting that $\|\{\Phi^{(t)}\}_k\|_1 = \sum_{m=1}^M \|\phi_{mk}^{(t)}\|_2$, we are able to separate $\Phi^{(t)}$ and reform the minimization problem in element-wise as following:

$$\sum_{k=1}^K \sum_{m=1}^M \min_{\phi_{mk}^{(t)}} \left\{ \gamma_1 \|\phi_{mk}^{(t)}\|_2 + \frac{\chi}{2} \left\| \phi_{mk}^{(t)} - \dot{a}_{mk}^{(t)} - \frac{1}{\chi} \lambda_{2,mk} \right\|_2^2 \right\}. \quad (21)$$

Similar to (20), the minimum value of each element can be evaluated as

$$\begin{aligned} \dot{\phi}_{mk}^{(t)} &= \max \left\{ 1 - \frac{2\gamma_1}{\chi} \left| \dot{a}_{mk}^{(t)} + \frac{\lambda_{2,mk}}{\chi} \right|^{-1}, 0 \right\} \\ &\quad \times \left(\dot{a}_{mk}^{(t)} + \frac{\lambda_{2,mk}}{\chi} \right). \end{aligned} \quad (22)$$

- The multipliers λ_1 and λ_2 are updated as following:

$$\begin{cases} \dot{\lambda}_1 \leftarrow \lambda_1 + \chi \left(\dot{\Theta}^{(t)} - \dot{\mathbf{A}}^{(t)} \right), \\ \dot{\lambda}_2 \leftarrow \lambda_2 + \chi \left(\dot{\Phi}^{(t)} - \dot{\mathbf{A}}^{(t)} \right). \end{cases} \quad (23)$$

The ADMM is guaranteed to converge to the optimal solution of (16) from any initial point [39]. Moreover, it will approach the optimal point within a few tens of iterations according to our simulation experiment, but is very slow to achieve high accuracy. Specifically, for a convergence

threshold of ϵ , each variable in this algorithm requires $O(1/\epsilon)$ iterations (see [37]) to converge with the complexity of (18), (19) and (22) per iteration being $O(MK^2)$, $O(KM)$ and $O(KM)$, respectively, and the overall computational complexity is therefore $O(MK^2/\epsilon)$. The algorithm is summarized in Algorithm 1.

IV. DYNAMIC CLUSTERING FOR HEAVY-LOAD TRAFFIC SCENARIOS

In this section, a stable access scheme for the users with heavy-load traffic is provided. In contrast to the users with light-load traffic, which are associated to their neighboring APs, a two-step dynamic clustering algorithm is further devised to achieve coordinated multi-point transmission for the high throughput demands.

Note that to ensure the stability of clustering results, the joint clustering algorithm is performed using the channel statistics information, which is averaged over the instantaneous CSI estimated via the same UAD and CE algorithm proposed in **light-load** scenarios.

A. LSF Estimation Analysis

Since the impact of the low-resolution ADC on the LSF estimation can be predetermined, we first present some preliminary results for compensating the impact of the quantization distortion, which complement the manipulations in (11).

Proposition 2: The received signal strength will deteriorate when replacing a full-resolution ADC with a low-resolution ADC at any AP, and such distortion can be compensated after multiplying with $\sqrt{1/\alpha}$ at the receiver.

Proof: Consider a typical AP i , equipped with full-resolution ADC, the power of the received signal can be evaluated according to (3), and is given by

$$\text{Var}(\mathbf{y}_{i,i}) = \tau \left(p_p \sum_{k \in \mathcal{K}_A} \beta_{ik} + \sigma^2 \right). \quad (24)$$

Similarly, when replacing the full-resolution ADC by a low-resolution one, the power of the received signal is then calculated as

$$\begin{aligned} \text{Var}(\mathbf{y}_{1,i}) &= \tau \alpha^2 \left(p_p \sum_{k \in \mathcal{K}_A} \beta_{ik} + \sigma^2 \right) + \text{Var}(\mathbf{n}_{q,i}) \\ &= \tau \alpha \left(p_p \sum_{k \in \mathcal{K}_A} \beta_{ik} + \sigma^2 \right). \end{aligned} \quad (25)$$

Comparing (24) and (25), we obtain the result. $\blacksquare$

Algorithm 1 ADMM

Initialization: variables $\Theta^{(t)}$ and $\Phi^{(t)}$, multiplier λ_1 and λ_2 , parameters γ_1 , γ_2 , χ and the counter Loop = 1.
while not converge or Loop $\leq$ MaxLoop **do**
 1) Update $\mathbf{A}^{(t)}$, $\Theta^{(t)}$ and $\Phi^{(t)}$ according to (18), (20) and (22), respectively.
 2) Update λ_1 and λ_2 with (23).
 3) Loop $\leftarrow$ Loop + 1.
end while

As the accuracy of LSF estimation is mainly dominated by the power of the received signal, the manipulation in *Proposition 2* can indeed improve the LSF estimation performance. Having established *Proposition 2*, we now elaborate on our LSF estimation scheme.

The proposed LSF estimation scheme is consistent with LTE, in which LSF is obtained by averaging the instantaneous CSI over several time slots, i.e.,

$$\hat{\beta}_{mk}(T) = \frac{1}{T} \sum_{t=1}^T \left| \hat{g}_{mk}^{(t)} \right|^2, \quad k \in \mathcal{K}_A, \quad (26)$$

where $\hat{g}_{mk}^{(t)}$ is the estimated channel coefficient in the t th slot, and T represents the estimation time. With the fact that $\hat{\beta}_{mk}(T)$ is the arithmetic mean, we conclude that the limit of the accuracy is given by

$$\bar{\beta}_{mk} = \lim_{T \rightarrow \infty} \hat{\beta}_{mk}(T). \quad (27)$$

In fact, it is most sensible to collect instantaneous CSI over several coherence slots to predict the LSF. However, it is impractical to spend too many resources on estimating β_{mk} , and hence we provide the principle of choosing T according to the mean square error (MSE) requirement in the next proposition.

Proposition 3: When the channel is perfectly estimated (i.e., $\hat{g}_{mk} = g_{mk}$), $\hat{\beta}_{mk}(T)$ is an unbiased estimator of β_{mk} whose normalized mean and variance are $\mathbb{E}[\hat{\beta}_{mk}(T)/\beta_{mk}] = 1$ and $\text{Var}[\hat{\beta}_{mk}(T)/\beta_{mk}] = \frac{1+2\kappa}{T(\kappa+1)^2}$, respectively. Moreover, to achieve η ($\eta \in (0, 1)$) confidence interval with μ probability, T should be the nearest integer greater than the solution of the following function:

$$Q_{\frac{V}{2}} \left(\sqrt{V\kappa}, \sqrt{1-\mu} \right) - Q_{\frac{V}{2}} \left(\sqrt{V\kappa}, \sqrt{1+\mu} \right) = \eta, \quad V \in \mathbb{N}^+. \quad (28)$$

where $Q_q(\cdot, \cdot)$ is the Marcum Q -function of order q [42].

Proof: Due to the perfect estimation of the channel gain, we hence have $\hat{g}_{mk}(t) \sim \mathcal{CN}(\sqrt{\frac{\kappa\beta_{mk}}{\kappa+1}}, \frac{\beta_{mk}}{\kappa+1})$. Since $\hat{g}_{mk}$ consists of a real part and imaginary part, let $b_{mk}^V = \sum_{t=1}^V \left| \sqrt{\frac{2(\kappa+1)}{\beta_{mk}}} \hat{g}_{mk}(t) \right|^2$, and $b_{mk}^V \sim \chi_{NC}^2(2V, 2V\kappa)$, where $\chi_{NC}^2(2V, 2V\kappa)$ represents the *Noncentral Chi-square distribution* whose probability density function is given by

$$f_{\chi_{NC}^2(2V, 2V\kappa)}(x) = \frac{e^{-V\kappa}}{2^V \Gamma(V)} {}_0F_1 \left(; V; \frac{V\kappa x}{2} \right) e^{-\frac{x}{2}} x^{V-1},$$

where ${}_0F_1(; \omega; \theta)$ is the *Confluent Hypergeometric Limit Function*. According to the properties of *Noncentral Chi-square distribution*, we can straightforwardly obtain $\mathbb{E}[b_{mk}^V] = 2V + 2V\kappa$ and $\text{Var}[b_{mk}^V] = 4V(1 + 2\kappa)$. Recalling the definition of $\hat{\beta}_{mk}(V)$, we have that

$$\mathbb{E}[\hat{\beta}_{mk}] = \frac{\beta_{mk}}{2V(\kappa+1)} \mathbb{E}[b_{mk}^V] = \beta_{mk}, \quad (29)$$

$$\text{Var} \left[\frac{\hat{\beta}_{mk}}{\beta_{mk}} \right] = \frac{\text{Var}[b_{mk}^V]}{4V^2(\kappa+1)^2} = \frac{1+2\kappa}{V(\kappa+1)^2}. \quad (30)$$

To achieve η confidence interval within μ deviation, V should satisfy

$$\int_{1-\mu}^{1+\mu} f_{\chi_{NC}^2(V, V\kappa)}(x) dx = \eta. \quad (31)$$

The left hand of (31) can further be evaluated as

$$\begin{aligned} & \int_{1-\mu}^{1+\mu} f_{\chi_{NC}^2(V, V\kappa)}(x) dx \\ &= \Pr_{\chi_{NC}^2(V, V\kappa)}(x \leq 1 + \mu) - \Pr_{\chi_{NC}^2(V, V\kappa)}(x \leq 1 - \mu) \\ &\stackrel{(b)}{=} Q_{\frac{V}{2}}(\sqrt{V\kappa}, \sqrt{1 - \mu}) - Q_{\frac{V}{2}}(\sqrt{V\kappa}, \sqrt{1 + \mu}), \end{aligned} \quad (32)$$

where equation (b) is obtained using the result in [42]. Substituting (32) into (31), we complete the proof. ■

Note that $\text{Var}[\hat{\beta}_{mk}(T)/\beta_{mk}]$ is a monotonically decreasing function with respect to T and κ ($\kappa \geq 0$), which implies that the estimation of LSF can be more precisely with more ensemble data in LoS-dominated environments. Here, we further study *Proposition 2* in the following two cases.

- In a scattering-dominated propagation environment (Rayleigh channel), i.e., $\kappa = 0$, the LSF estimation experiences the worst performance, and the normalized MSE (NMSE) is simplified to

$$\text{Var}[\hat{\beta}_{mk}(T)/\beta_{mk}] = \frac{1}{T}. \quad (33)$$

To achieve the same η confidence interval with μ probability, T can be determined by solving the following simplified equation:

$$P\left(\frac{V}{2}, \frac{1+\mu}{2}\right) - P\left(\frac{V}{2}, \frac{1-\mu}{2}\right) = \eta, \quad V \in \mathbb{N}^+, \quad (34)$$

where $P(s, x) = \frac{\gamma(s, x)}{\Gamma(s)}$ with *Gamma function* $\Gamma(\cdot)$ and $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$.

- In a LoS-dominated propagation environment, i.e., $\kappa \rightarrow \infty$, the LSF estimation is extremely precise even with only one set of data, and the NMSE of the estimation is reduced to

$$\text{Var}[\hat{\beta}_{mk}(T)/\beta_{mk}] \rightarrow 0, \quad \forall T. \quad (35)$$

Recall that in practical IoT scenarios, channel state is likely to be LoS-dominated, especially in a dense system, which enables us to estimate LSF with less amount of data exchange [43].

Remark 1: Having established *Proposition 2*, the estimation of LSF matrix can thus be extracted from (26), and is given by

$$\hat{\mathbf{B}} = \frac{1}{T} \sum_{t=1}^T \left(\hat{\mathbf{A}}^{(t)} \right)^{\circ(2)}, \quad (36)$$

where $\hat{\mathbf{B}} \in \mathbb{C}^{K \times M}$ denotes the estimated LSF matrix of all users, whose i th column is described as $\hat{\mathbf{b}}_i = [\hat{\beta}_{1i}(T), \dots, \hat{\beta}_{Mi}(T)]^T$, and $\hat{\mathbf{A}}^{(t)}$ is the solution of $f(\mathbf{A}^{(t)})$ in (9).

B. Problem Formulation For Dynamic Clustering

After estimating LSF, we are ready to process with the dynamic clustering for users with heavy-load traffics.

In a distributed mMIMO network, the number of APs and users is extremely huge, which renders the computational complexity of joint processing at the central processor unmanageable. To reduce the complexity, we propose a two steps dynamic clustering scheme, in which the clustering features are identified at the first step, and the joint clustering algorithm is implemented at the second step for sum-rate maximization. By exploring the number of clusters among users, the complexity of clustering algorithm can be significantly reduced due to the reduced dimensions of the clustering matrix.

Regarding dynamic clustering, the main idea is to group users that are geographically adjacent into one cluster, so that the inter-cell interference can be reduced when a cluster is established. We now assume that K_A active users and M APs are grouped into N clusters, and denote the set of users and APs in n th cluster as $\mathcal{K}_{A_n}$ and $\mathcal{M}_n$, respectively. We aim to group users such that

$$\begin{cases} \bigcup_{\forall n} \mathcal{K}_{A_n} = \mathcal{K}_A, \\ \mathcal{K}_{A_n} \cap \mathcal{K}_{A_{n'}} = \emptyset. \end{cases} \quad (37)$$

We emphasize that users in each $\mathcal{K}_{A_n}$ should be geographically adjacent, such that the margin between $\mathcal{K}_{A_n}$ and $\mathcal{K}_{A_{n'}}$ is maximized. The antenna selection vector and power control weight are jointly considered as an $M \times N$ matrix $\mathbf{S}$, which is given by

$$\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_N], \quad (38)$$

where the n th column represents the selection vector of n th cluster with $s_{mn} \in [0, 1]$. Specifically, $s_{mn} > 0$ represents that the m th AP belongs to the n th cluster, and $s_{mn} = 0$ otherwise. Moreover, $s_{mn} = 1$ means that the m th AP transmits signal with full power. Assume that users and APs in each cluster comprise a virtual MIMO network, and the achievable rate for the n cluster can thus be evaluated as

$$R_n = W \log_2 \left(1 + \text{SINR}(\hat{\mathbf{B}}_n, \mathbf{s}_n) \right), \quad (39)$$

where

$$\text{SINR}(\hat{\mathbf{B}}_n, \mathbf{s}_n) \approx \frac{\|(\hat{\mathbf{B}}_n^{\frac{1}{2}})^T \mathbf{s}_n\|_2^2}{\sum_{n' \in N, n' \neq n} \|(\hat{\mathbf{B}}_n^{\frac{1}{2}})^T \mathbf{s}_{n'}\|_2^2 + \frac{\sigma^2}{P_{\max}}}, \quad (40)$$

where W is the operating bandwidth, $P_{\max}$ represents the per-AP power constraint, and $\hat{\mathbf{B}}_n$ represents the LSF matrix of users in the n th cluster, which is the subset of $\hat{\mathbf{B}}$. Therefore, $\hat{\mathbf{B}}_n$ consists of the columns $\hat{\mathbf{b}}_{i \in \mathcal{K}_{A_n}}$. We have to clarify that (40) is not the “real SINR” that the system will achieve after clustering, but it is used as an indicator to detect the channel quality for clustering. The replacement of instantaneous CSI by long-term information, i.e., LSF in (40) can effectively reduce the complexity of the clustering algorithm, and provide more stable scheduling results to prevent the frequent update of the cluster formations. We omit the intra-cell interference in (40) which conforms to a ZF-type of beamforming which

can largely reduce the interference within a cluster.² We herein quote (40) as the “SINR” for convenience.

Note that the power consumption of fronthaul network is proportional to the achievable sum rate in the cluster, the per-AP fronthaul constraint should be considered as a major restriction in our weight sum rate (WSR) maximization problem:

$$\text{maximize}_{\{\mathbf{s}_n | n \in N\}} \sum_{n=1}^N a_n R_n \quad (41a)$$

$$\text{s.t.} \quad s_{mn} \in [0, 1], \quad \forall m, n, \quad (41b)$$

$$\|(s_{1n}, \dots, s_{Mfn})\|_0 \geq 1, \quad \forall n, \quad (41c)$$

$$\sum_{n \in N} (\|\mathbf{s}_n\|_1 + \mu \|\mathbf{s}_n\|_0) \leq \frac{P_T}{P_{\max}}, \quad (41d)$$

$$R_n \leq C_t, \quad \forall n. \quad (41e)$$

where P_T represents the total power budget consisting of the power consumption of transmission and circuit components, $\mu = \frac{P_c}{P_{\max}}$ is a constant value with P_c denoting the operation power of a AP; C_t denotes the per-AP fronthaul capacity constraint, and a_n denotes the weight associated with the n th cluster which is designed for satisfying specific purposes, such as, fairness [30]. More importantly, (41c) is used to ensure that at least one F-AP is included in each cluster, while (41e) gives the fronthaul capacity constraints. Finally, (41d) is adopted to inactivate the APs which contributes less to the achievable sum rate to save energy.

The conventional WSR maximization problem in (41a) is a well-known nonconvex problem, whose optimal solution is very challenging to find even without discrete constraints in (41b), (41c) and (41e) [30]. Although the WSR maximization problem has been proved that can be approached by adopting a WMMSE method [49], required as prior information, the knowledge of the cluster number N is still missing. As a matter of fact, we are inclined to gather as many users and APs in one cluster as possible to reduce the inter-cell interference. In extreme cases, all the users gather in one group, and the interference can be canceled. However, in consideration of the fronthaul constraint and the complexity of beamforming, we need to balance the number of clusters and the number of users and APs in each cluster.

C. K-Means based Cluster Detection

We now provide a K-means based algorithm which can intelligently detect “the right number of clusters” without information of the users’ locations.

The conventional K-means clustering is a standard heuristic method which aims at partitioning observations into clusters, and each observation belongs to the cluster with the nearest mean [44]. We hence adopt the estimated LSF vector of each user, i.e., $\hat{\mathbf{b}}_k$ to form the observation space. To make LSF information more easy to acquire, we first scale the fading vectors into the logarithmic field, and then evaluate

the *distance* between LSF vectors from the k th user and the k' user, which is defined by

$$d_c(\hat{\mathbf{b}}_k, \hat{\mathbf{b}}_{k'}) \triangleq \|\mathbf{L}(\hat{\mathbf{b}}_k) - \mathbf{L}(\hat{\mathbf{b}}_{k'})\|_2^2, \quad (42)$$

where $\mathbf{L}(\cdot) = \log_{10}(\cdot)$ in a element-wise manner. In a similar fashion, the mean of vectors $\bar{\mathbf{b}}_{\mathcal{K}_{A_n}}$ is defined as

$$\mathbf{L}(\bar{\mathbf{b}}_{\mathcal{K}_{A_n}}) = \frac{1}{\|\mathcal{K}_{A_n}\|_0} \sum_{k \in \mathcal{K}_{A_n}} \mathbf{L}(\hat{\mathbf{b}}_k). \quad (43)$$

Hence, the cluster detection can be transformed into a within-cluster sum of squares minimization, mathematically speaking,

$$SQ_N = \arg \min_{\{\mathcal{K}_{A_n} | n \in N\}} \sum_{n=1}^N \sum_{k \in \mathcal{K}_{A_n}} d_c(\hat{\mathbf{b}}_k, \bar{\mathbf{b}}_{\mathcal{K}_{A_n}}). \quad (44)$$

By solving (44) via K-means algorithm, the members of each cluster can be observed. As the K-means is rather widespread, we omit further details.

An important fact about K-means algorithm is that with different number of clusters N as prior information, the algorithm may output completely different clustering results. Therefore, a criterion that can effectively ascertain the “optimal” number of clusters should be established before K-means operates. However, there appears to be no rigorous theoretical rule that can explicitly determine the “optimal” number of clusters. Fortunately, there have been a number of different principles in the literature for choosing N by running K-means for multiple times [44]. According to the experimental study in [44], each approach may output a different “optimal” number of clusters with different cluster recovery performance. Among all the principles, a experimental rule suggested by Hartigan [45] is superior to other principles in detecting N , and is given in following lemma.

Lemma 1: If N is the result of the K-means algorithm with N groups, then it is justifiable to add an extra group when

$$\left(\frac{SQ_N}{SQ_{N+1}} - 1 \right) (K_A - N - 1) > 10. \quad (45)$$

As clarified in [46], the K-means algorithm is not guaranteed to converge to the global optimum, and it is an NP-hard problem to find the optimal solution with undetermined number of clusters. The general practice is to run the algorithm for multiple times with different starting conditions and select the result with smallest SQ_N . As it requires I (upper bounded by $2^{\Omega(\sqrt{K_A})}$) iterations for K-means to converge, the computational complexity of running K-means algorithm once is $O(MN I K_A)$ [47]. Assume finding SQ_N requires L times running of K-means, the overall complexity for detecting the number of clusters is upper bounded by $O(\Gamma(N+2) M I L K_A)$. The procedure is summarized in Algorithm 2.

D. WMMSE based Clustering

Even after completing the user grouping, the WSR maximization problem in (41a) is still not solvable, since the antenna selection and power constraints in (41a) are discrete,

²Note that, once clusters are established, the “real SINR” accounts for the intra-cell interference caused by quantization noise and channel estimation error as a part of noise-plus-interference. Since the approximation is used to indicate channel quality, these terms are omitted at the clustering phrase.

Algorithm 2 K-means based Cluster Detection

Initialization: $N = 1, L$.

- 1) Compute SQ_N for L times and select the optimal one.
- 2) Compute SQ_{N+1} for L times and select the optimal one.

if (45) is satisfied **then**

- Go to 2).

else

- Output N .

end if

which makes our maximization be a non-convex problem. To solve this issue, we approximate the ℓ_0 -norm with the reweighted ℓ_1 -norm [48]; mathematically speaking,

$$\|\mathbf{s}_n\|_0 \approx \|(\boldsymbol{\omega}_n \cdot \mathbf{s}_n)\|_1, \quad (46)$$

where $\boldsymbol{\omega}_n \in \mathbb{C}^{M \times 1}$ represents the weight vector associated with $\mathbf{s}_n$ whose the m th element is $\omega_{mn} = \frac{1}{s_{mn} + \varpi}$ with ϖ being a very small positive value to provide stability. Generally, with a proper value of ϖ , the ℓ_0 -norm can be effectively approximated by minimizing the ℓ_1 -norm.

By employing the approximation above, we transform the selection constraints in (41c) and (41d):

$$\|\phi(\boldsymbol{\omega}_n \cdot \mathbf{s}_n)^T\|_1 \geq 1, \quad (47)$$

and

$$\sum_{n \in N} \|(\mathbf{1} + \mu \boldsymbol{\omega}_n) \cdot \mathbf{s}_n\|_1 \leq \frac{P_T}{P_{\max}}, \quad (48)$$

where $\phi = [\mathbf{1}^{1 \times M_f}, \mathbf{0}^{1 \times M_f}]$ is a constant vector to ensure that each cluster is associated with no less than one F-AP. However, as R_n appears in both the objective function and constraint, the problem (41) is still unsolvable after replacing (41c) and (41d) by (47) and (48), respectively. Inspired by the iterative method in [30], we address this issue by solving the problem (41a) iteratively with fixed $\tilde{R}_n$ which is obtained from the previous iteration, i.e.,

$$\tilde{R}_n \leq C_t. \quad (49)$$

Regarding the objective function in (41a), as WMMSE has been proved to be equivalent to the WSR problem, the problem (41) can be effectively approached by WMMSE algorithm [49, 50]. An important observation is that the WMMSE proposed in [50] is operated in complex field. We modify the algorithm to an equivalent real interference channel to fit our problem, that is given as

$$\hat{x}_n = \mathbf{u}_n^T \left(\rho_p \hat{\mathbf{B}}_n^T \mathbf{s}_n x_n + \rho_p \sum_{n' \neq n} \hat{\mathbf{B}}_n^T \mathbf{s}_{n'} x_{n'} + \mathbf{n}_n \right), \quad \forall n \quad (50)$$

where $x_n \in \mathbb{C}$ represents the transmitted signal with unit variance, and $\mathbf{u}_n \in \mathbb{R}^{M \times 1}$ denotes the amplifier gain at the received side. Noting that x_n is independent of the noise vector $\mathbf{n}_n$, the MSE between x_n and $\hat{x}_n$ can be calculated as

$$\begin{aligned} e_n &= \mathbb{E} \left[\|\hat{x}_n - x_n\|_2^2 \right] \\ &= \mathbf{u}_n^T (\mathbf{F}_{\hat{\mathbf{B}}_n} + \mathbf{I}_n) \mathbf{u}_n - \rho_p \mathbf{u}_n^T \hat{\mathbf{B}}_n^T \bar{\mathbf{s}}_n + 1 \end{aligned} \quad (51)$$

Algorithm 3 WMMSE based Clustering

Initialization: $\boldsymbol{\omega}_n$, power control matrix $\mathbf{S}$, the iteration counter $\text{Loop} = 1$, $\tilde{R}_n$, and N from Algorithm 2.

while not converge or $\text{Loop} \leq L_{\max}$ **do**

- 1) Fix $\mathbf{S}$, evaluate $\mathbf{u}_n$ and e_n , $\forall n$ via (53) and (51), respectively.
- 2) Update ε_n with (54).
- 3) Evaluate $\mathbf{S}$ with fixed $\mathbf{u}_n$ and ε_n by solving (55).
- 4) Update $\tilde{R}_n$ and $\boldsymbol{\omega}_n$, $\forall n$.
- 5) $\text{Loop} \leftarrow \text{Loop} + 1$.

end while

where

$$\mathbf{F}_{\hat{\mathbf{B}}_n} = \rho_p^2 \sum_{n' \in N} \hat{\mathbf{B}}_n^T \mathbf{s}_{n'} \mathbf{s}_{n'}^T \hat{\mathbf{B}}_n \in \mathbb{C}^{K \times \tau}$$

with fixed $\mathbf{u}_n$. Thus, the equivalence between the WSR problem and WMMSE problem can be obtained via minimizing the coefficient e_n .

Lemma 2: The WSR maximization problem in (41a) can be accurately evaluated by solving the following WMMSE problem:

$$\begin{aligned} &\underset{\{\varepsilon_n, \mathbf{u}_n, \mathbf{s}_n | n \in N\}}{\text{maximize}} \quad \sum_{n=1}^N a_n (\varepsilon_n e_n - \log_2 \varepsilon_n) \quad (52a) \\ &\text{s.t.} \quad (41b), (47), (48), \text{ and } (49), \quad (52b) \end{aligned}$$

where ε_n denotes the MSE weight for cluster n .

Proof: The proposition can be proved directly following the same methodology as in [50, Th. 1] by simplifying from matrix field to vector field. ■

A key observation of problem (52a) the objective function is convex with respect to each of the parameter ε_n , $\mathbf{u}_n$ and $\bar{\mathbf{s}}_n$, respectively. Therefore, this critical fact allows problem to be solved using the block coordinate decent method where each parameter is updated iteratively:

- The optimal $\mathbf{u}_n$ with fixed ε_n and $\mathbf{s}_n$ is given by

$$\mathbf{u}_n^{\text{opt}} = \rho_p (\mathbf{F}_{\hat{\mathbf{B}}_n} + \mathbf{I}_n)^{-1} \hat{\mathbf{B}}_n^T \mathbf{s}_n \quad (53)$$

- The optimal ε_n with fixed $\mathbf{u}_n$ and $\mathbf{s}_n$ is given by

$$\varepsilon_n^{\text{opt}} = e_n^{-1}. \quad (54)$$

- The optimal $\mathbf{s}_n$ with fixed ε_n and $\mathbf{u}_n$ can be obtained by solving following standard quadratically constrained quadratic programming (QCQP) problem:

$$\begin{aligned} &\underset{\{\mathbf{s}_n | n \in N\}}{\text{minimize}} \quad \rho_p^2 \sum_{n=1}^N \mathbf{s}_n^T \left(\sum_{n'=1}^N a_{n'} \varepsilon_{n'} \hat{\mathbf{B}}_{n'} \mathbf{u}_{n'} \mathbf{u}_{n'}^T \hat{\mathbf{B}}_{n'}^T \right) \mathbf{s}_n \\ &\quad - 2 \rho_p \sum_{n=1}^N a_n \varepsilon_n \mathbf{u}_n^T \hat{\mathbf{B}}_n^T \mathbf{s}_n \quad (55a) \\ &\text{s.t.} \quad (52b). \quad (55b) \end{aligned}$$

The problem (55) can be solved using a standard convex optimization solver.

As a summary, the procedure to solve (41) can be separated into two loops, where the inner loop evaluates the antenna selection matrix by solving the WMMSE problem with fixed

Fig. 3. NMSE estimation of $\mathbf{A}$ for “ADMM”, “SGLasso”, and “Lasso”, with $K = 80$, $K_A = 30$, $M = 100$, $b = 2$ bits. (a) NMSE with respect to pilot length with power $p_p = 0$ dBm, $\xi = 0.3$. (b) NMSE with respect to pilot power with $\tau = 32$, $\kappa = 3$ dB.

ω_n and $\tilde{R}_n$, $\forall n$, while the outer loop updates ω_n and $\tilde{R}_n$ according to the inner loop results. The scheme is detailed in Algorithm 3. Regarding the complexity of WMMSE-based clustering scheme, the computational complexity per iteration is dominated by Step 3, whose complexity is at the same level as SOCP, i.e., $O((MN)^{3.5})$. If we set the limit of the iterations as $L_{\max}$, the overall computational complexity is then upper bounded by $O((MN)^{3.5}L_{\max})$. Although the convergence is guaranteed for the WMMSE problem in each inner loop, the whole optimization process cannot guarantee to converge to the global optimal value due to the iterative process of ω_n and $\tilde{R}_n$ in the outer loop. To ensure the robustness of the proposed algorithm, we provide the relevant numerical simulation to demonstrate the convergence. Note that similar conclusion has been provided in [48].

V. NUMERICAL RESULTS

In our simulations, users and all the APs are deployed in a 1km^2 area. We consider the LSF coefficient as a function of distance, which is given by (in dB) [51]

$$\beta(d) = 32.9 + 36.3\log_{10}(d), \quad (56)$$

where d (in m) is the distance between two antennas. We set the guard distance as 20 m, in which β is set to a constant value of 32.9 dB. We set the transmit power budget $P_{\max}$ and the power required for running AP P_c as 23 dBm, and 0.2 W, respectively. The background noise power is -174 dBm/Hz, and the bandwidth $W = 10$ MHz. We choose the NMSE to evaluate the prediction performance, which is mathematically defined as

$$\text{NMSE} = \mathbb{E} \left\{ \frac{\|\hat{\mathbf{X}} - \mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2} \right\}. \quad (57)$$

A. Simulations for UAD and CE

We first examine the performance of the UAD and CE algorithms for **light-load** scenarios, in which we focus on the accuracy and complexity of our algorithms.

In the simulation, we use random pilot assignment scheme, in which each user will be randomly assigned one pilot sequence $\mathbf{p}_k$ which is the k th row of a discrete Fourier transform matrix. The values of γ_1 and γ_2 are empirically set as 5% of the upper bounds in Proposition 1. The regularization parameter χ in ADMM is set as same level as γ_1 from experience. Moreover, by setting $\gamma_1 = 0$, the method “Lasso” is included for comparison. The convergence threshold ϵ and maximum number of iterations MaxLoop are set to 10^{-8} and 10, respectively.

Fig. 4. User activity detection results with respect to pilot length, with $K = 80$, $K_A = 20$, $M_f = 20$, $M = 100$ and $p_p = 0$ dBm.

Fig. 5. Execution time of “ADMM”, “SGLasso”, and “Lasso” against the number of APs, with $p_p = 0$ dBm, $K = 80$, $K_A = 20$, $M = 100$, $\kappa = 3$ dB.

In Fig. 3, we verify the accuracy of the channel estimation by comparing the NMSE of $\mathbf{A}$ among three algorithms. From Fig. 3(a), it is clear that the performance of the proposed ADMM and “SGLasso” algorithm is superior to the “Lasso” algorithm with enough pilot length (i.e., $\tau \geq 32$). A key observation is that the algorithms with enough pilot length perform far better than that with insufficient pilot length, which implies that this is a fundamental drawback of the CS-based scheme. In addition, the Rician factor κ also dominates the performance, as the NMSE decreases with increasing κ . This is due to the fact that with higher κ , the channel sparsity is pronounced more significantly, which is more favorable for the CS-based algorithm and results in performance improvement. Denoting by ξ the proportion of the number of F-APs in

the system, i.e., $\xi = \frac{M_f}{M}$, Fig. 3(b) compares the NMSE of channel propagation with respect to ξ and pilot power p_p . As we can see, the performance of the proposed “SGLasso” and “ADMM” are slightly inferior to the “Lasso” algorithm with insufficient pilot power, and such offset is eliminated when the pilot power increases. From Fig. 3(b), it is intuitive that the channel is estimated more accurately with more power and higher ξ . A more important observation is that by increasing the pilot power, the low-cost hardware topology can achieve the same level of accuracy as ideal topologies.

Fig. 4 shows the UAD performance with respect to pilot length τ under different channel conditions. First, we observe that the wrongly detected user number decreases sharply when the pilot length increases for $K_A < \tau$, and gradually converges after $K_A > \tau$. More specifically, τ is required to be approximately as large as $1.5 \times K_A$ to ensure a satisfactory performance due to the random pilot assignment. Also, we find that, all the algorithms can detect active users correctly when pilot length is long enough. Besides, among three algorithms, the proposed “SGLasso” and “ADMM” provide the best performance, and can detect active users correctly in both Rayleigh and Rician channels. In contrast, misdetection can be observed from the results by applying “Lasso”, even with enough pilot length, which is the result of the significant “row sparsity” feature within the CSI matrix in a dense system.

Fig. 5 evaluates the execution time of three algorithms over different system configurations. It is clear that “ADMM” has the lowest computational complexity among the three schemes, while the other two schemes have a rather high complexity, e.g., the computation time of “ADMM” can be reduced into 0.3s with both $\tau = 16$ and $\tau = 32$, which is about in one tenth of that of the other two algorithms. Another important observation from the figure is that the execution time increases with the number of pilot length for “SGLasso” and “Lasso”, while decreases for “ADMM”. This

(a)

(b)

Fig. 6. NMSE of LSF estimation with $K = 80$, $K_A = 30$, $b = 2$ bits, $p_p = 0$ dBm. (a) NMSE of “ADMM” algorithm with respect to κ with $M = 100$. (b) NMSE among “ADMM”, “SGLasso”, and “Lasso” with respect to the number of APs with $\kappa = 3$ dB, $\xi = 0.3$.

is because a larger matrix will increase the problem dimension for any interior point method. In contrast, for our “ADMM” algorithm, such variation will increase the subgradients of each subproblem, which can help the algorithm to converge faster.

B. Simulations for LSF

We now examine the performance of the proposed dynamic clustering for **heavy-load** scenarios.

As the LSF information is necessary to run the clustering algorithm, we demonstrate the performance of LSF estimation in Fig. 6, in which Fig. 6(a) shows the performance of “ADMM” against the Rician factor κ , where “P2” represents the operation in *Proposition 2*. Fig. 6(b) shows the performance among three algorithms with respect to the system size. From Fig. 6(a), a key observation is that the gap of LSF estimation error between two antenna configurations has been significantly reduced by applying the manipulation in *Proposition 2*, which is consistent with our analysis. Moreover, an improvement on the LSF estimation can be observed using such manipulation, especially in high κ scenarios. Also, note that the estimation of LSF is more accurate with higher κ , which allow us to spend less overhead on channel estimation for a LoS-dominated channels. In Fig. 6(b), a small gap between “Lasso” and other two algorithms can be observed when the system becomes dense. This is because the received channel matrix is more likely to have a stronger feature of “row-sparsity” rather “element-sparsity” in a dense system, and this is a feature that “ADMM” and “SGLasso” can leverage. When the system becomes sparse, the performance of “Lasso” is slightly superior to the other two algorithms due to the stronger “element-sparsity”. Also, we include the bounds of NMSE in *Proposition 3* in Fig. 6(b), which are obtained by considering perfect CSI. As we can see, the behavior of LSF estimation improves with growing transmit power and pilot length. More importantly, Fig. 6(b) shows that the algorithms perform poorly in a moderate-sized system, while almost attain the accuracy bound in a dense system, which is also a result of the enhanced signal strength in such scenarios.

C. Simulations for Dynamic Clustering

After verifying the feasibility of the LSF estimation, we now evaluate the performance of the proposed dynamic AP clustering algorithm. We include a heuristic method proposed in [52], namely, “Merge” as a benchmark. In “Merge”, each users select v adjacent APs with the strongest links to establish its own cell. For the cells that overlap with each other, the algorithm merges them, thereby forming a larger cell. System keeps repeating this step until all users are separated into disjoint subsets. With different value of v , the clustering result can be substantially different. With a large v , the size of the cluster is more likely to be larger which will require more operating power, and the transmit power will thus be considerably restricted by the total power constraint. When v is small, the clusters will be denser in which users will suffer more intercell interference.

Fig. 7 shows the cumulative distribution function (CDF) of the average per-user rate with different power consumption budget. In the following simulations, ZF precoding is used after clustering to reduce the intracell interference. To guarantee fairness, the results of “Merge” are also obtained after running the WSR with predetermined clustering sets. First of all, compared with the “Merge” method, we can observe a significant improvement on the per-user rate of “Dynamic clustering”. The reason is that “Merge” cannot decide if it is worth to activate an AP considering its power consumption. With the overall power consumption limit, spending more power on supporting AP operation indicates that the power used for transmission is less, which results in performance degradation. Moreover, by comparing the performance of per-user rate between that with perfect LSF and with estimated LSF, a slight offset can be observed, which indicates that the use of estimated LSF may lead to a different clustering result for “Dynamic clustering”. Nevertheless, the estimation error does not seriously affect the antenna selection, as this distortion occurs only when adjacent APs have similar signal strength.

Fig. 8 shows the behavior of “Dynamic clustering” and “Merge” with different fronthaul capacity limits. We first observe that “Dynamic clustering” remarkably outperforms “Merge” in terms of per-user rate. For example, with fronthaul constraints $C_t = 400$ Mbits/s and 500 Mbits/s, there are about 1 Mbits/s and 2 Mbits/s gains on the per-user rate achieved by “Dynamic clustering” for over 95%-likely users, respectively, and such improvement becomes even more significant when we look at the to the 50th percentile user rate. Besides, by comparing the per-user rate with 400 Mbits/s to 500 Mbits/s fronthaul limits, we can observe a 2 Mbits/s offset for “Dynamic clustering”, while no apparent difference is found for “Merge”. This is because the sum rate achieved with “Merge” is poor in each cell, and hardly reaches the fronthaul capacity constraints.

Fig. 7. Cumulative distribution function (CDF) of per-user rate, with $v = 3$, $K_A = 30$, $M_f = 20$, $M_l/M_f = 4$, $C_t = 50$ Mbits/s and $P_T = 150 \times P_{\max}$.

Fig. 8. CDF of per-user rate w.r.t. fronthaul capacity limits, with $v = 3$, $K_A = 30$, $M_f = 20$, $M_1/M_f = 4$, $\mu = 5$ and $P_T = 150 \times P_{\max}$.

Fig. 9. The convergence of “Dynamic clustering” and the average per-user rate against the number of iterations, with $K_A = 30$, $M_f = 20$, $M_1/M_f = 4$, $C_t = 50$ Mbits and $P_T = 150 \times P_{\max}$.

To illustrate the fast convergence of the proposed dynamic clustering algorithm, we present the ℓ_1 -norm of ω_n and the average per-user rate over each iteration in Fig. 9. The sum of weighted vector norm is drawn in dashed lines while the corresponding average per-user rate are plotted in solid lines. Although a rigorous proof for the convergence is not available when there exist discrete constraints in the WSR problem, we show the average per-user rate and weighted vector evaluation for “Dynamic clustering” to ensure the convergence. From the figure, it is obvious that both the sum of ℓ_1 -norm of weighted vectors and the average per-user rate converge within 5 iterations. Therefore, such rapid convergence demonstrates that our proposed dynamic clustering algorithm is appropriate to meet the time latency requirements in future communications.

VI. CONCLUSION

To meet the massive connectivity requirements in IoT, this paper developed grant-free access schemes for users with different throughput demands. For users with light-load traffic requirements, we provided a low complexity ADMM method that can solve the UAD and CE problems simultaneously. For users with heavy-load traffic requirements, a two-step dynamic clustering was designed based on the LSF information, which is estimated by averaging the instantaneous CSI. The impact of quantization noise on estimation was rigorously investigated, and upper bounds of accuracy were provided as well as a compensation method. By detecting the cluster feature among users, the complexity of clustering algorithm was substantially reduced. Moreover, by leveraging the ℓ_1 -norm reweighted approximation, the proposed algorithm was able to activate APs intelligently to achieve energy-efficiency transmission. The simulation results verified the convergence of the algorithm, and showed a remarkable gain of achievable sum rate of our scheme, especially for the scenarios with high operation power consumptions.

REFERENCES

- [1] J. Yuan, Q. He, M. Matthaiou, Y. Wang, T. Q. S. Quek, and S. Jin, “A weighted MMSE approach to amorphous cell for mixed-ADC distributed massive MIMO,” in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Oct. 2018, pp. 969–974.
- [2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (IoT): A vision, architectural elements, and future directions,” *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, Sep. 2013.
- [3] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, “An information framework for creating a smart city through internet of things,” *IEEE Internet Things J.*, vol. 1, no. 2, pp. 112–121, Apr. 2014.
- [4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, “Internet of things for smart cities,” *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.
- [5] D. Minoli, K. Sohraby, and B. Occhiogrosso, “IoT considerations, requirements, and architectures for smart buildings/energy optimization and next-generation building management systems,” *IEEE Internet Things J.*, vol. 4, no. 1, pp. 269–283, Feb. 2017.
- [6] S. Bera, S. Misra, and A. V. Vasilakos, “Software-defined networking for internet of things: A survey,” *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1994–2008, Dec. 2017.
- [7] A. Gluhak, S. Krco, M. Nati, D. Pfisterer, N. Mitton, and T. Razafindralambo, “A survey on facilities for experimental internet of things research,” *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 58–67, Nov. 2011.
- [8] H. Ning and Z. Wang, “Future internet of things architecture: Like mankind neural system or social organization framework?,” *IEEE Commun. Lett.*, vol. 15, no. 4, pp. 461–463, Apr. 2011.
- [9] P. Gope and T. Hwang, “Untraceable sensor movement in distributed IoT infrastructure,” *IEEE Sensors J.*, vol. 15, no. 9, pp. 5340–5348, Sep. 2015.
- [10] C. Encinas, E. Ruiz, J. Cortez, and A. Espinoza, “Design and implementation of a distributed IoT system for the monitoring of water quality in aquaculture,” in *Proc. Wireless Telecommunications Symposium (WTS)*, Apr. 2017, pp. 1–7.
- [11] R. H. Walden, “Analog-to-digital converter survey and analysis,” *IEEE J. Sel. Areas Commun.*, vol. 17, no. 4, pp. 539–550, Apr. 1999.
- [12] N. Liang and W. Zhang, “Mixed-ADC massive MIMO,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 983–997, Apr. 2016.
- [13] J. Yuan, S. Jin, C. K. Wen, and K. K. Wong, “The distributed MIMO scenario: Can ideal ADCs be replaced by low-resolution ADCs?” *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 470–473, Aug. 2017.
- [14] J. Park, S. Park, A. Yazdan, and R. W. Heath Jr., “Optimization of mixed-ADC multi-antenna systems for cloud-RAN deployments,” *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3962–3975, Sep. 2017.
- [15] S. Sesia, I. Toufik, and M. Baker, Eds., *LTE: The UMTS Long Term Evolution: From Theory To Practice*. John Wiley and Sons, 2011.

- [16] 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. 2011.
- [17] 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.
- [18] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.
- [19] Y. Fathy, P. Barnaghi, S. Enshaeifar, and R. Tafazolli, "A distributed in-network indexing mechanism for the internet of things," in *Proc. IEEE World Forum Internet Things*, Dec. 2016, pp. 585–590.
- [20] R. Jain and S. Tata, "Cloud to edge: Distributed deployment of process-aware IoT applications," in *Proc. IEEE Int. Conf. Edge Comput. (EDGE)*, Jun. 2017, pp. 182–189.
- [21] L. Liu and W. Yu, "Massive connectivity with massive MIMO-Part I: Device activity detection and channel estimation," *CoRR*, vol. abs/1706.06438, 2017. [Online]. Available: <http://arxiv.org/abs/1706.06438>
- [22] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454–465, Feb. 2011.
- [23] X. Xu, X. Rao, and V. K. N. Lau, "Active user detection and channel estimation in uplink CRAN systems," in *Proc. IEEE ICC*, Jun. 2015, pp. 2727–2732.
- [24] D. Har, H. H. Xia, and H. L. Bertoni, "Path-loss prediction model for microcells," *IEEE Trans. Veh. Technol.*, vol. 48, no. 5, pp. 1453–1462, Sep. 1999.
- [25] Q. He, T. Q. S. Quek, Z. Chen, and S. Li, "Compressive channel estimation and multi-user detection in C-RAN," in *Proc. IEEE ICC*, May 2017.
- [26] C. T. K. Ng and H. Huang, "Linear precoding in cooperative MIMO cellular networks with limited coordination clusters," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, Dec. 2010.
- [27] S. Kaviani, O. Simeone, W. A. Krzymien, and S. Shamai, "Linear precoding and equalization for network MIMO with partial cooperation," *IEEE Trans. Veh. Technol.*, vol. 61, no. 5, pp. 2083–2096, Jun. 2012.
- [28] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," in *Proc. IEEE ICC*, May 2008, pp. 4033–4037.
- [29] M. Yoon, M. S. Kim, and C. Lee, "A dynamic cell clustering algorithm for maximization of coordination gain in uplink coordinated system," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1752–1760, Mar. 2016.
- [30] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.
- [31] J. Park, N. Lee, and R. W. Heath Jr., "Cooperative base station coloring for pair-wise multi-cell coordination," *IEEE Trans. Commun.*, vol. 64, no. 1, pp. 402–415, Jan. 2016.
- [32] G. Nigam, P. Minero, and M. Haenggi, "Coordinated multipoint joint transmission in heterogeneous networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4134–4146, Nov. 2016.
- [33] X. Li, "RSS-based location estimation with unknown pathloss model," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3626–3633, Dec. 2006.
- [34] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.
- [35] Q. Bai and J. A. Nossek, "Energy efficiency maximization for 5G multi-antenna receivers," *Trans. Emerging Telecommun. Technol.*, vol. 26, no. 1, pp. 3–14, 2015.
- [36] L. Fan, D. Qiao, S. Jin, C. Wen, and M. Matthaiou, "Optimal pilot length for uplink massive MIMO systems with low-resolution ADC," in *Proc. IEEE SAM*, Jul. 2016, pp. 1–5.
- [37] S. Noah, F. Jerome, H. Trevor and T. Robert, "A sparse-group lasso," *J. Comput. and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.
- [38] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebre, "Applications of second-order cone programming," *Linear Algebra and its Applications*, vol. 284, no. 1, pp. 193–228, Jan. 1998.
- [39] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
- [40] Z.-Q. Yu, X.-J. Shi, L. Yan, and W.-J. Li, "Distributed stochastic ADMM for matrix factorization," in *Proc. ACM CIKM*, Nov. 2014, pp. 1259–1268.
- [41] K. Petersen and M. Petersen, "The matrix cookbook," Sep. 2007.
- [42] A. Nuttall, "Some integrals involving the Q_M function," *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 95–96, Mar. 1975.
- [43] Universal Mobile Telecommunications System (UMTS): Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations, Tech. Rep. 125 996, 3GPP ETSI Release 13.1.0, Jan. 2017 [Online]. Available: https://www.etsi.org/deliver/etsi_tr/125900_125999/125996/13.01.00_60/tr_125996v130100p.pdf
- [44] M. M.-T. Chiang, and B. Mirkin, "Intelligent choice of the number of clusters in K -means clustering: An experimental study with different cluster spreads," *Journal of Classification*, vol. 27, no. 1, pp. 3–40, Mar. 2010.
- [45] J. A. Hartigan, *Clustering algorithms*. Wiley New York, 1975, vol. 209.
- [46] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k -means problem is NP-hard," *Theoretical Computer Science*, vol. 442, pp. 13–21, Jul. 2012.
- [47] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k -means clustering algorithm," *Journal of the Royal Statistical Society*, vol. 28, no. 1, pp. 100–108, 1979.
- [48] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.
- [49] S. S. Christensen, R. Agarwal, E. D. Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
- [50] Q. Shi, M. Razaviyayn, Z. Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.
- [51] J. Hoydis, K. Hosseini, S. Ten Brink, and M. Debbah, "Making smart use of excess antennas: Massive MIMO, small cells, and TDD," *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 5–21, Sep. 2013.
- [52] J. Wang and L. Dai, "Downlink rate analysis for virtual-cell based large-scale distributed antenna systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1998–2011, Mar. 2016.

Jide Yuan (S'16–M'19) received the B.S. degree in communication engineering from the Ocean University of China, Qingdao, China in 2011. He then received the M.S. degree from the Harbin Engineering University, Harbin, China in 2014 and Ph.D. degree in Communication and Information Systems from the Southeast University, China in 2018, respectively. From October 2018 through September 2019, he was with the Institute of Electronics, Communications and Information Technology, Queen's University Belfast, U.K. working as a Postdoctoral Research Fellow. He is currently a Research Fellow with the Department of Electronic Systems, Aalborg University, Denmark. His current research interests include random matrix theory, massive MIMO, machine learning and large intelligent surface.

processing, and artificial intelligence.

Qi He received the B.S. and M.S. degrees in Communication Engineering from University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2008 and 2011, respectively. Then he worked in Huawei Technologies Co., Ltd. as a software engineer from 2011 to 2013. From 2013 to 2018, he studied in UESTC again and obtained the Ph.D. degree in communication and information system. He is currently an assistant researcher in Westlake University. His research interests include internet of things, natural language

Michail Matthaiou (S'05–M'08–SM'13) was born in Thessaloniki, Greece in 1981. He obtained the Diploma degree (5 years) in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 2004. He then received the M.Sc. (with distinction) in Communication Systems and Signal Processing from the University of Bristol, U.K. and Ph.D. degrees from the University of Edinburgh, U.K. in 2005 and 2008, respectively. From September 2008 through May 2010, he was with the Institute for Circuit Theory

and Signal Processing, Munich University of Technology (TUM), Germany working as a Postdoctoral Research Associate. He is currently a Reader (equivalent to Associate Professor) in Multiple-Antenna Systems at Queen's University Belfast, U.K. after holding an Assistant Professor position at Chalmers University of Technology, Sweden. His research interests span signal processing for wireless communications, massive MIMO, hardware-constrained communications, and performance analysis of fading channels.

Dr. Matthaiou and his coauthors received the 2017 IEEE Communications Society Leonard G. Abraham Prize. He is currently holding the prestigious 2018/2019 Royal Academy of Engineering/The Leverhulme Trust Senior Research Fellowship and will receive the 2019 EURASIP Early Career Award. His team was the Grand Winner of the 2019 Mobile World Congress Challenge. He was the recipient of the 2011 IEEE ComSoc Best Young Researcher Award for the Europe, Middle East and Africa Region and a co-recipient of the 2006 IEEE Communications Chapter Project Prize for the best M.Sc. dissertation in the area of communications. He has co-authored papers that received best paper awards at the 2018 IEEE WCSP and 2014 IEEE ICC and was an Exemplary Reviewer for IEEE COMMUNICATIONS LETTERS for 2010. In 2014, he received the Research Fund for International Young Scientists from the National Natural Science Foundation of China. In the past, he was an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, Associate Editor/Senior Editor for IEEE COMMUNICATIONS LETTERS and was the Lead Guest Editor of the special issue on "Large-scale multiple antenna wireless systems" of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.

Shi Jin (S06-M07-SM'17) received the B.S. degree in communications engineering from Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from the Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with

the faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory. He serves as an Associate Editor for the IEEE Transactions on Wireless Communications, and IEEE Communications Letters, and IET Communications. Dr. Jin and his co-authors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and a 2010 Young Author Best Paper Award by the IEEE Signal Processing Society.

Tony Q.S. Quek (S'98–M'08–SM'12–F'18) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. Currently, he is the Cheng Tsang Man Chair Professor with Singapore University of Technology and Design (SUTD). He also serves as the Acting Head of ISTD Pillar, Sector Lead of

the SUTD AI Program, and the Deputy Director of the SUTD-ZJU IDEA. His current research topics include wireless communications and networking, network intelligence, internet-of-things, URLLC, and big data processing.

Dr. Quek has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is currently serving as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the Chair of IEEE VTS Technical Committee on Deep Learning for Wireless Communications as well as an elected member of the IEEE Signal Processing Society SPCOM Technical Committee. He was an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, and an Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards – Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, and the 2016-2019 Clarivate Analytics Highly Cited Researcher. He is a Distinguished Lecturer of the IEEE Communications Society and a Fellow of IEEE.