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


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
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 is designed in a grant-free fashion. For users with light-load traffic demands, for formulating the user activity detection (UAD) and channel estimation (CE) into a compressed sensing problem, we provide a low-complexity algorithm 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.

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

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.
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 traffic 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 heterogenous 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 \( \{A\}_i \) and \( a_{ij} \) to denote the \( i \)th row and \((i,j)\)th entry in \( A \), respectively. The operation \( \|A\|_p \) denotes the \( l_p \)-norm of the matrix \( A \), and \( \text{diag}(\cdot) \) denotes the diagonal matrix of the vector \( \cdot \). The superscripts \( (\cdot)^H \), \( \text{tr} (\cdot) \) and \( \text{E} [\cdot] \) denote the matrix Hermitian transpose, trace and the expectation of the input random entity, respectively. Finally, \( (A \cdot B) \) denotes the Hadamard product (element-wise product between matrices \( A \) and \( B \)), and \( A^{\circ(a)} \) denotes an element-wise exponentiation operation on \( A \), i.e., \( (A^{\circ(a)})_{i,j} = (A_{i,j})^{a} \).

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...
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 $K_A \subseteq \{1, \ldots, K\}$ are active ($|K_A| = K_A < M$).

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
  \[ g_{ij}^{(t)} = h_{ij}^{(t)} \sqrt{\beta_{ij}}, \]
  where $\beta_{ij}$ 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,
  \[ h_{ij}^{(t)} = \sqrt{\frac{\kappa}{\kappa + 1}} \tilde{h}_{ij} + \sqrt{\frac{1}{\kappa + 1}} \tilde{h}_{ij}^{(t)}, \]
  where $\tilde{h}_{ij}$ is a deterministic value, $\tilde{h}_{ij}^{(t)}$ is a random complex component with zero mean and unit variance, while the parameter $\kappa$ 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 $i$th F-AP and from the $k$th user to the $m$th L-AP as $g_{i,m,k}^{(t)}$ and $g_{m,k}^{(t)}$, 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 scenarios in IoT, once APs receive the pilot sequences, users are simply associated to the nearest AP to improve quality of service, shown in Fig. 2.

Assume that users send the pilot sequences simultaneously, and the $k$th user is assigned the pilot sequence $p_k \in \mathbb{C}^{1 \times \tau}$ of length $\tau$ with $\|p_k\|^2 = \tau p_p$, where $p_p$ represents the transmit power of pilot symbol. The received signal vector $y_{1,m} \in \mathbb{C}^{1 \times \tau}$ at the $m$th F-AP is given by
  \[ y_{1,m}^{(t)} = \sum_{k \in K_A} g_{i,m,k}^{(t)} p_k + n_{m}^{(t)}, \]
  where $n_{m}^{(t)} \sim \mathcal{CN}(0, \sigma^2 I)$ denotes the additive Gaussian noise. Similarly, the received signal vector $y_{1,m'}^{(t)} \in \mathbb{C}^{1 \times \tau}$ at the $m'$th L-AP can be is expressed as follows:
  \[ y_{1,m'}^{(t)} = \alpha \sum_{k \in K_A} g_{i,m',k}^{(t)} p_k + \alpha n_{m'}^{(t)}, \]
  where $n_{m'}^{(t)}$ is the noise vector at the $m'$th L-AP, and $\alpha$ represents the quantization noise coefficient, which is related to the quantization bit $\kappa$, as listed in Table I [35]. Along with the fact $\|p_k\|^2 = \tau p_p$, the variance of $n_{q,m'}$' can be evaluated as [36]
  \[ \text{Var}(n_{q,m'}) = \tau \alpha (1 - \alpha) \left( \sigma^2 + p_p \sum_{k \in K_A} \beta_{m'm'} \right). \]

We now introduce the active indicator function
  \[ \delta_A(a) = \begin{cases} 
  1, & \text{if } a \in A, \\
  0, & \text{if } a \notin A. 
\end{cases} \]

The central processor is able to stack the received pilot data from $M$ APs into a matrix, given as
  \[ Y^{(t)} = G^{(t)} \Delta P + N^{(t)}, \]
  where
  \[ Y^{(t)} = \begin{bmatrix} \left( y_{1,1}^{(t)} \right)^T, \ldots, \left( y_{1,M}^{(t)} \right)^T, \left( y_{1,1}^{(t)} \right)^T, \ldots, \left( y_{1,M}^{(t)} \right)^T \end{bmatrix}^T, \]
  \[ \Delta = \text{diag} (\delta_{K_A} (k))_{k=1,\ldots,K}, \quad \text{and} \quad P = [P_1^T, \ldots, P_K^T]^T \in \mathbb{C}^{K \times \tau}. \]
  The noise term $N^{(t)}$ consists of both AWGN and quantization noise, and is described as
  \[ N^{(t)} = \begin{bmatrix} N_f^{(t)}^T, N_i^{(t)} \end{bmatrix}^T \in \mathbb{C}^{M \times \tau} \]
  with
  \[ N_f^{(t)} = \begin{bmatrix} \left( n_{q,1}^{(t)} \right)^T, \ldots, \left( n_{q,M}^{(t)} \right)^T \end{bmatrix}^T \]
  and
  \[ N_i^{(t)} = \begin{bmatrix} \left( n_{m,1}^{(t)} + \alpha n_{m,1}^{(t)} \right)^T, \ldots, \left( n_{m,M}^{(t)} + \alpha n_{m,M}^{(t)} \right)^T \end{bmatrix}^T. \]

Moreover, $G^{(t)} = \left[ g_1^{(t)}, \ldots, g_K^{(t)} \right] \in \mathbb{C}^{M \times K}$, where the $k$th column represents the channel propagation coefficient vector.

\(^1\)In a practical scenario, the $K$-factor is normally different across users. We here assume uniform $K$ across users in order to investigate its impact on the LSF estimation algorithm shown in Section IV.
of the $k$th user, and is given by

$$g_k^{(t)} = \left[ \left( g_{1,k}^{(t)} \right)^T, \alpha \left( g_{2,k}^{(t)} \right)^T \right]^T.$$

### 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:

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

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

By denoting $A^{(t)} = \Delta \left( G^{(t)} \right)^H \in \mathbb{C}^{K \times M}$, it is straightforward to reveal the “row sparsity” of $A^{(t)}$ due to the sparse active users. Moreover, “element sparsity” can be observed from $A^{(t)}$ 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 $f(A^{(t)})$:

$$f(A^{(t)}) \triangleq \min_{A^{(t)}} \frac{1}{2} \left\| (Y^{(t)})^H - P^H A^{(t)} \right\|^2_2 + \gamma_1 \sum_{k=1}^{K} \sum_{m=1}^{M} \delta \left( a^{(t)}_{km} \right) + \gamma_2 \sum_{k=1}^{K} \left\| \left\{ A^{(t)} \right\}_k \right\|_1, \tag{9}$$

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

$$\delta(A) = \left\{ \begin{array}{ll} 0, & \text{if } \|A\|_0 = 0, \\ 1, & \text{otherwise}. \end{array} \right.$$  

The function of (9) considers jointly the element sparsity and row sparsity, and the parameters $\gamma_1$ and $\gamma_2$ 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 $A^{(t)}$, which consists of the indication of active users $\Delta$ and channel gain $G^{(t)}$.

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 $A^{(t)}$, 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:

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

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

$$\tilde{Y}^{(t)} \triangleq \left( Y^{(t)} \right)^T, \left( \sqrt{1/\alpha} Y_1^{(t)} \right)^T.$$

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 $\hat{A}^{(t)}$ and a proportion of APs whose corresponding column is zero can be set in idle mode to save power. Note that the choice of $\gamma_1$ and $\gamma_2$ are crucial for the performance of (10). A small $\gamma_1$ and $\gamma_2$ 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 $\gamma_1$ and smaller $\gamma_2$ 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 $\gamma_2$ becomes more effective for processing. The upper bounds of choosing $\gamma_1$ and $\gamma_2$ are given in the following proposition.

**Proposition 1:** Denote by $A^{(t)}$ the solution of (10). We have $\hat{A}^{(t)} = 0$ if $\gamma_1 \geq \| p_k \left( \tilde{Y}^{(t)} \right)^H \|_2$ or $\gamma_2 \geq \frac{1}{\alpha} \| p_k \left( \tilde{Y}^{(t)} \right)^H \|_2$.

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

$$\min_{A^{(t)}} \frac{1}{2} \left\| (\tilde{Y}^{(t)})^H - \sum_{k=1}^{K} p_k H \left\{ A^{(t)} \right\}_k \right\|^2_2 + \gamma_1 \sum_{k=1}^{K} \left\| \left\{ A^{(t)} \right\}_k \right\|_2 + \gamma_2 \sum_{k=1}^{K} \left\| \left\{ A^{(t)} \right\}_k \right\|_1 \leq (10), \tag{12}$$

where equality holds when solutions $\hat{A}^{(t)} = \left( \hat{A}^{(t)} \right)^T$. Since the objective function is convex, the $k$th group of the minimum solution $\hat{A}$ must satisfy

$$p_k \left( p_k H \left\{ A^{(t)} \right\}_k - \left( \tilde{Y}^{(t)} \right)^H \right) - \gamma_1 u_k = 0, \tag{13}$$

where $u_k$ are the subgradient of $\left\| \left\{ A^{(t)} \right\}_k \right\|_2^2$, which have the
following property [41, Eq. (129)]

$$u_k = \begin{cases} \frac{\{A(t)\}}{\|\{A(t)\}\|}, & \text{if } \{A(t)\} \neq 0, \\ \in \{u_k : \|u_k\| \leq 1\}, & \text{if } \{A(t)\} = 0. \end{cases}$$  \hspace{1cm} (14)

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

$$\gamma_1 > \left\| P_k \left( Y(t) \right)^H \right\|_2.$$  \hspace{1cm} (15)

Note that the solution of

$$\min_{\{A(t)\}} \gamma_2 \sum_{k=1}^K \|\{A(t)\}_k\|_1$$

is zero. The solution of (13) is zero when \(\gamma_1\) satisfies (15). For \(\gamma_2\), by noting that \(\|\{A(t)\}_k\|_1 = \sum_{m=1}^M \|a_{km}\|_2\), 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 \(O((MK)^{3.5})\) 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 \(A(t)\), and introduce two auxiliary variables \(\Theta(t)\) and \(\Phi(t)\) to fit the ADMM criterion:

$$\min_{A(t), \Theta(t), \Phi(t)} \frac{1}{2} \left\| (Y(t))^H - X(t) \right\|^2 \left\| P_k \left( \{A(t)\}_k \right)^H \right\|^2 + \gamma_1 \sum_{k=1}^K \|\{\Theta(t)\}_k\|_1 + \gamma_2 \sum_{k=1}^K \|\{\Phi(t)\}_k\|_1,$$

s.t. \(\Theta(t) = A(t), \Phi(t) = A(t).\)  \hspace{1cm} (16)

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

$$L \left( A(t), \Theta(t), \Phi(t), \lambda_1, \lambda_2 \right)$$

$$= \frac{1}{2} \left\| (Y(t))^H - X(t) \right\|^2 \left\| P_k \left( \{A(t)\}_k \right)^H \right\|^2 + \gamma_1 \sum_{k=1}^K \|\{\Theta(t)\}_k\|_1 + \gamma_2 \sum_{k=1}^K \|\{\Phi(t)\}_k\|_1$$

$$+ \frac{\chi}{2} \left\| \{\Theta(t)\}_k - \{A(t)\}_k \right\|^2 + \|\{A(t)\}_k - \{\Phi(t)\}_k\|^2_2$$

$$+ R \left( \lambda_1^H \left( \{\Theta(t)\}_k - \{A(t)\}_k \right) + \lambda_2^H \left( \{\Phi(t)\}_k - \{A(t)\}_k \right) \right).$$  \hspace{1cm} (17)

where \(\lambda_1\) and \(\lambda_2\) are the Lagrangian multipliers which have the same dimension as \(A(t), \) \(\mathbb{R} \{\} \) stands for real part of a complex number, and \(\chi\) is the regularization parameter. The algorithm is operated by minimizing \(A(t), \Theta(t),\) 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 \(A(t)\), by taking the derivative with respect to \(A(t)\) in (17) and setting it to zero, the minimum \(\hat{A}(t)\) can be evaluated as

$$\hat{A}(t) = (PP^H + 2\chi I)^{-1} \left( P \left( Y(t) \right)^H + \chi \left( \Theta(t) + \Phi(t) \right) - (\lambda_1 + \lambda_2) \right).$$  \hspace{1cm} (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

$$\min_{\Theta(t)} \frac{1}{2} \left\| \Theta(t) - \hat{A}(t) - \frac{1}{\lambda_1} \right\|^2_2.$$  \hspace{1cm} (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

$$\hat{\Theta}(t)_k = \max \left\{ 1 - 2\gamma_1 \frac{1}{\lambda_1} \|\hat{A}(t) + \frac{1}{\lambda_1}\|_2^{-1}, 0 \right\} \times \left( \hat{A}(t) + \frac{1}{\lambda_1} \right)_k.$$  \hspace{1cm} (20)

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

$$\min_{\Phi_{mk}} \frac{1}{2} \left\| \phi_{mk} \right\|^2_2 + \chi \frac{1}{2} \left\| \phi_{mk} - \hat{a}_{mk} - \frac{1}{\chi} \lambda_{2, mk} \right\|^2_2.$$  \hspace{1cm} (21)

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

$$\hat{\phi}_{mk} = \max \left\{ 1 - 2\gamma_1 \frac{1}{\chi} \hat{a}_{mk} + \frac{\lambda_2}{\chi} \lambda_{2, mk}, 0 \right\} \times \left( \hat{a}_{mk} + \frac{\lambda_2}{\chi} \lambda_{2, mk} \right).$$  \hspace{1cm} (22)

- The multipliers \(\lambda_1\) and \(\lambda_2\) are updated as following:

$$\lambda_1 \leftarrow \lambda_1 + \chi \left( \hat{\Theta}(t) - \hat{A}(t) \right),$$

$$\lambda_2 \leftarrow \lambda_2 + \chi \left( \hat{\Phi}(t) - \hat{A}(t) \right).$$  \hspace{1cm} (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 $\epsilon$, 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 $\sqrt{T/\alpha}$ 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

$$\text{Var}(y_{i,t}) = 7 (p_b \sum_{k \in \mathcal{K}_A} \beta_{ik} + \sigma^2).$$

(24)

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

$$\text{Var}(y_{i,t}) = 7 \alpha^2 (p_b \sum_{k \in \mathcal{K}_A} \beta_{ik} + \sigma^2) + \text{Var}(n_{q_i})$$

$$= 7 \alpha (p_b \sum_{k \in \mathcal{K}_A} \beta_{ik} + \sigma^2).$$

(25)

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

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.,

$$\hat{\beta}_{mk}(T) = \frac{1}{T} \sum_{t=1}^{T} g_{mk}^{(t)}, \quad k \in \mathcal{K}_A,$$

(26)

where $g_{mk}^{(t)}$ 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

$$\hat{\beta}_{mk} = \lim_{T \to \infty} \hat{\beta}_{mk}(T).$$

(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 $\hat{\beta}_{mk}$, 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., $g_{mk} = \hat{g}_{mk}$), $\hat{\beta}_{mk}(T)$ is an unbiased estimator of $\beta_{mk}$ whose normalized mean and variance are $E[\hat{\beta}_{mk}(T)/\beta_{mk}] = 1$ and $\text{Var}[\hat{\beta}_{mk}(T)/\beta_{mk}] = \frac{1}{T+1}$, respectively. Moreover, to achieve $\eta (\eta \in (0, 1))$ confidence interval with $\mu$ probability, $T$ should be the nearest integer greater than the solution of the following function:

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

(28)

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

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

$$f_{\chi_{NC}^2(2V, 2V \kappa)}(x) = \frac{e^{-V \kappa}}{V} F_1 \left( \frac{V \kappa}{2}, \frac{V \kappa}{2} ; \frac{1}{2}x \right),$$

where $a F_1 (\cdot; \cdot; \cdot)$ is the Confluent Hypergeometric Limit Function. According to the properties of Noncentral Chi-square distribution, we can straightforwardly obtain $E[b^{V}_{mk}] = 2V + 2V \kappa$ and $\text{Var}[b^{V}_{mk}] = 4V(1 + 2\kappa)$. Recalling the definition of $\beta_{mk}(V)$, we have that

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

Algorithm 1 ADMM

Initialization: variables $\Theta^{(t)}$ and $\Phi^{(t)}$, multiplier $\lambda_1$ and $\lambda_2$, parameters $\gamma_1$, $\gamma_2$, $\chi$ and the counter Loop = 1.

while not converge or Loop $\leq$ MaxLoop do

1) Update $A^{(t)}$, $\Theta^{(t)}$ and $\Phi^{(t)}$ according to (18), (20) and (22), respectively.

2) Update $\lambda_1$ and $\lambda_2$ with (23).

3) Loop $\leftarrow$ Loop + 1.

end while
To achieve $\eta$ confidence interval within $\mu$ deviation, $V$ should satisfy
\[
\int_{1-\mu}^{1+\mu} f_{\chi^2_{N,V}}(x) \, dx = \eta. \tag{31}
\]
The left hand of (31) can further be evaluated as
\[
\int_{1-\mu}^{1+\mu} f_{\chi^2_{N,V}}(x) \, dx = \Pr[\chi^2_{N,V}(x \leq 1 + \mu)] - \Pr[\chi^2_{N,V}(x \leq 1 - \mu)] \\
\triangleq Q_\chi(\sqrt{V\kappa}, \sqrt{1 - \mu}) - Q_\chi(\sqrt{V\kappa}, \sqrt{1 + \mu}), \tag{32}
\]
where equation (b) is obtained using the result in [42]. Substituting (32) into (31), we complete the proof.

Remark 1: Having established Proposition 2, the estimation of LSF matrix can thus be extracted from (26), and is given by
\[
\hat{\beta}_{mk}(T)/\beta_{mk} \rightarrow 0, \quad \forall T. \tag{35}
\]
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].

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 $K_{An}$ and $M_n$, respectively. We aim to group users such that
\[
\bigcup_{n=0}^{N} K_{An} = \mathbb{K}, \quad K_{An} \cap K_{An'} = \emptyset. \tag{37}
\]
where $\mathbb{K}$ represents the total number of users, $s_n = \{s_1, \ldots, s_N\}$, (38) where the $n$th column represents the selection vector of $n$th cluster with $s_{mn}$ represents the $n$th user belongs to the $n$th cluster, and $s_{mn} = 0$ otherwise. Moreover, $s_{mn} = 1$ means that the $n$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 $n$th cluster can thus be evaluated as
\[
R_n = W \log_2 \left(1 + \text{SINR} \left(\hat{B}_n, s_n\right)\right), \tag{39}
\]
where
\[
\text{SINR} \left(\hat{B}_n, s_n\right) \approx \frac{\left\|\left(\hat{B}_n^2\right)^T s_n\right\|^2}{\sum_{n'=N,n'\neq n} \left\|\left(\hat{B}_{n'}^2\right)^T s_{n'}\right\|^2 + \frac{\sigma^2}{P_{\text{max}}}}, \tag{40}
\]
where $W$ is the operating bandwidth, $P_{\text{max}}$ represents the per-AP power constraint, and $\hat{B}_n$ represents the LSF matrix of users in the $n$th cluster, which is the subset of $\hat{B}$. Therefore, $\hat{B}_n$ consists of the columns $B_{n} \in K_{An}$. We have to clarify that $\text{SINR}$ is 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.\footnote{Note 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.} 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:

\[
\text{maximize} \quad \sum_{n=1}^{N} a_n R_n \quad \text{s.t.} \quad s_{mn} \in [0, 1], \quad \forall m, n, \quad \| (s_{1n}, \ldots, s_{Mn}) \|_0 \geq 1, \quad \forall n, \quad \sum_{n \in N} (\| s_n \|_1 + \mu \| s_n \|_0) \leq \frac{P_T}{P_{\text{max}}}, \quad R_n \leq C_1, \quad \forall n.
\]

where \( P_T \) represents the total power budget consisting of the power consumption of transmission and circuit components, \( \mu = \frac{P_{\text{max}}}{P_T} \) is a constant value with \( P_T \) denoting the operation power of a AP; \( C_1 \) 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 problem has been proved that can be approached by adopting (41b), (41c) and (41e) [30]. Although the WSR maximization problem in (41a) is still not solvable, since the antenna selection and power constraints in (41a) are discrete, the distance between LSF vectors from the \( k \)-th user and the \( k' \)-th user, which is defined by

\[
d_c (\hat{b}_k, \hat{b}_{k'}) \triangleq \| L (\hat{b}_k) - L (\hat{b}_{k'}) \|_2^2,
\]

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

\[
L (\bar{b}_{an}) = \frac{1}{|K_{an}|} \sum_{k \in K_{an}} L (\hat{b}_k).
\]

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

\[
\text{SQ}_N = \arg \min_{\{K_{an} | n \in N\}} \sum_{n=1}^{N} \sum_{k \in K_{an}} d_c (\hat{b}_n, \bar{b}_{K_{an}}).
\]

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

\[
\left( \frac{\text{SQ}_N}{\text{SQ}_{N+1}} - 1 \right) (K_A - N - 1) > 10.
\]

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 \( \text{SQ}_N \). As it requires \( I \) (upper bounded by \( 2^I(\sqrt{\pi \sigma}) \)) iterations for K-means to converge, the computational complexity of running K-means algorithm once is \( O (MNIK_A) \) [47]. Assume finding \( \text{SQ}_N \) requires \( L \) times running of K-means, the overall complexity for detecting the number of clusters is upper bounded by \( O (I (N + 2) MILK_A) \). 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,
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 \( \ell_0 \)-norm with the reweighted \( \ell_1 \)-norm [48]: mathematically speaking,

\[
\|s_n\|_0 \approx \|\omega_n \cdot s_n\|_1, \quad (46)
\]

where \( \omega_n \in \mathbb{C}^{M \times 1} \) represents the weight vector associated with \( s_n \) whose the \( m \)th element is \( \omega_{mn} = \frac{\omega}{x} \) with \( x \) being a very small positive value to provide stability.

Generally, with a proper value of \( x \), the \( \ell_0 \)-norm can be effectively approximated by minimizing the \( \ell_1 \)-norm.

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

\[
\left\| \phi (\omega_n \cdot s_n) \right\|_1 \geq 1,
\]

and

\[
\sum_{n \in N} \left( 1 + \mu \omega_n \right) \cdot s_n \|_1 \leq \frac{P_t}{P_{\text{max}}},
\]

where \( \phi = [1^{1 \times M}, \mathbf{0}^{1 \times M}] \) 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 \( R_n \) which is obtained from the previous iteration, i.e.,

\[
R_n \leq C_1. \quad (49)
\]

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

\[
\hat{x}_n = u_n^T \left( \rho_p \hat{B}^T_n s_n x_n + \rho_p \sum_{n' \neq n} \hat{B}^T_{n'} s_{n'} x_{n'} + u_n \right), \quad \forall n
\]

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

\[
\epsilon_n = E \left[ \| \hat{x}_n - x_n \|^2 \right] = u_n^T (F_{B_n} + I_n) u_n - \rho_p u_n^T \hat{B}_n^T \hat{B}_n s_n + 1
\]

Algorithm 3 WMMSE based Clustering

**Initialization:** \( \omega_n, \) power control matrix \( S, \) the iteration counter \( Loop = 1, R_n, \) and \( N \) from Algorithm 2.

while not converge or \( Loop \leq L_{\text{max}} \) do
   1) Fix \( S, \) evaluate \( u_n \) and \( e_n, \) \( \forall n \) via (53) and (51), respectively.
   2) Update \( e_n \) with (54).
   3) Evaluate \( S \) with fixed \( u_n \) and \( e_n \) by solving (55).
   4) Update \( R_n \) and \( \omega_n, \) \( \forall n. \)
   5) Loop \( \leftarrow \) Loop + 1.
end while

where

\[
F_{B_n} = \rho_p^2 \sum_{n' \in N} \hat{B}_{n'}^T s_{n'} \hat{B}_{n'} \hat{B}_n \in \mathbb{C}^{K \times K}
\]

with fixed \( u_n. \) 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:

\[
\max_{\{ e_n, u_n, s_n \in N \}} \sum_{n=1}^N a_n (e_n e_n - \log_2 e_n) \quad (52a)
\]

s.t. (41b), (47), (48), and (49), (52b)

where \( e_n \) 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 \( e_n, u_n \) and \( s_n, \) respectively. Therefore, this critical fact allows problem to be solved using the block coordinate decent method where each parameter is updated iteratively:

- The optimal \( u_n \) with fixed \( e_n \) and \( s_n \) is given by

\[
u_n^{\text{opt}} = \rho_p (F_{B_n} + I_n)^{-1} \hat{B}_n^T s_n \]

- The optimal \( e_n \) with fixed \( u_n \) and \( s_n \) is given by

\[
e_n = e_n^{\text{opt}}
\]

- The optimal \( s_n \) with fixed \( e_n \) and \( u_n \) can be obtained by solving following standard quadratically constrained quadratic programming (QCQP) problem:

\[
\min_{\{ s_n \in N \}} \rho_p^2 \sum_{n=1}^N s_n^T \left( \sum_{n'=1}^N a_{n'} e_{n'} \hat{B}_{n'}^T u_{n'} u_n^T \hat{B}_{n'}^T \right) s_n - 2 \rho_p \sum_{n=1}^N a_n e_n u_n^T \hat{B}_n s_n
\]

s.t. (52b).

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
\( \omega \) and \( \tilde{R}_n \), \( \forall n \), while the outer loop updates \( \omega_n \) 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., \( \mathcal{O}(MN^3) \). If we set the limit of the iterations as \( \mathcal{L}_{\text{max}} \), the overall computational complexity is then upper bounded by \( \mathcal{O}(MN^3 \mathcal{L}_{\text{max}}) \). 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 \( \omega_n \) and \( \tilde{R}_n \) in the outer loop. To ensure the robustness of the proposed the 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]

\[
\beta(d) = 32.9 + 36.3 \log_{10}(d),
\]

where \( d \) (in m) is the distance between two antennas. We set the guard distance as 20 m, in which \( \beta \) is set to a constant value of 32.9 dB. We set the transmit power budget \( P_{\text{max}} \) and the power required for running AP \( P_r \) 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

\[
\text{NMSE} = \mathbb{E} \left\{ \frac{\| \hat{X} - X \|_2^2}{\| X \|_2^2} \right\}.
\]

#### 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 \( p_k \) which is is the \( k \)th row of a discrete Fourier transform matrix. The values of \( \gamma_1 \) and \( \gamma_2 \) are empirically set as 5% of the upper bounds in Proposition 1. The regularization parameter \( \chi \) in ADMM is set as same level as \( \gamma_1 \) from experience. Moreover, by setting \( \gamma_1 = 0 \), the method “Lasso” is included for comparison. The convergence threshold \( \epsilon \) and maximum number of iterations MaxLoop are set to \( 10^{-8} \) and 10, respectively.
In Fig. 3, we verify the accuracy of the channel estimation by comparing the NMSE of 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 $\kappa$ also dominates the performance, as the NMSE decreases with increasing $\kappa$. This is due to the fact that with higher $\kappa$, the channel sparsity is pronounced more significantly, which is more favorable for the CS-based algorithm and results in performance improvement.

Denoting by $\xi$ 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 $\xi$ 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 $\xi$. 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 $\tau$ 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, $\tau$ 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 Ricean 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 0.3 s 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
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 $\kappa$, 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 $\kappa$ scenarios. Also, note that the estimation of LSF is more accurate with higher $\kappa$, 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 user selects $v$ 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 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.

![Cumulative distribution function (CDF) of per-user rate](image)

**Fig. 7.** Cumulative distribution function (CDF) of per-user rate, with $v = 3$, $K_A = 30$, $M_1 = 20$, $M_2/M_1 = 4$, $C_t = 50$ Mbits/s and $P_f = 150 \times P_{\text{max}}$. 

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 user selects $v$ 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.
To meet the time latency requirements in future communications, our proposed dynamic clustering algorithm is appropriate to iterations. Therefore, such rapid convergence demonstrates that vectors and the average per-user rate converge within 5 iterations. From the figure, it is obvious that both the sum of cluster feature among users, the complexity of clustering algorithm was substantially reduced. Moreover, by leveraging the \( \ell_1 \)-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.

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 \( \ell_1 \)-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


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 2000. 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.

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 1995 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.

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.