User-centric C-RAN Architecture for Ultra-dense 5G Networks: Challenges and Methodologies

Cunhua Pan, *Member, IEEE*, Maged Elkashlan, *Member, IEEE*, Jiangzhou Wang, *Fellow, IEEE*, Jinhong Yuan, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

Abstract

Ultra-dense networks (UDN) constitute one of the most promising techniques of supporting the fifth generation (5G) mobile system. By deploying more small cells in a fixed area, the average distance between users and access points can be significantly reduced, hence a dense spatial frequency reuse can be exploited. However, severe interference is the major obstacle in UDNs. Most of the contributions investigate the interference by designing distributed algorithms based on cooperative game theory. This paper advocates the application of dense user-centric cloud radio access network (C-RAN) philosophy to UDNs, thanks to the recent development of cloud computing techniques. Under dense C-RAN architectures, centralized signal processing can be invoked for supporting Coordinated Multiple Points Transmission/Reception (CoMP). We summarize the main challenges in dense user-centric C-RANs. One of the most challenging issues is the requirement of the global CSI for the sake of cooperative transmission. We investigate this requirement by only relying on partial channel state information (CSI), namely, on inter-cluster large-scale CSI. Furthermore, the estimation of the intra-cluster CSI is considered, including the pilot allocation and robust transmission. Finally, we highlight several promising research directions to make the dense user-centric C-RAN become a reality, with special emphasis on the application of the 'big data' techniques.

Index Terms

Ultra-dense networks (UDN), user-centric C-RAN, virtual cells, DAS, imperfect CSI, pilot allocation.

C. Pan and M. Elkashlan are with the Queen Mary University of London, London E1 4NS, U.K. (Email:{c.pan, maged.elkashlan}@qmul.ac.uk). J. Wang is with the School of Engineering and Digital Arts, University of Kent, Canterbury, Kent, CT2 7NZ, U.K. (e-mail: j.z.wang@kent.ac.uk). J. Yuan is with the University of New South Wales, Sydney, NSW 2052, Australia. (e-mail:j.yuang@unsw.edu.au). L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K. (e-mail:lh@ecs.soton.ac.uk).

I. Introduction

The capacity demand for new emerging mobile applications, such as the Internet of things (IoTs) and three dimensional (3D) wireless video streaming has been growing explosively. The fifth generation (5G) wireless system of the near future is anticipated to offer a substantially increased data throughput compared to the fourth generation (4G) system. To achieve this ambitious goal, ultra dense networks (UDNs) have been widely regarded as one of the most promising solutions relying on small-cell base stations (BSs) [1]. In UDNs, the average distance between users and small cell BSs is significantly reduced, hence the link quality is dramatically improved, which further increases the network capacity.

However, the drastic interference generated by the neighboring small cells is a limiting factor in UDNs. The attainable network performance may even be decreased when the BS density is extremely high [2]. Hence, the interference should be carefully managed in order to reap the potential benefits of UDNs. Most of the existing contributions deal with the interference by designing partially distributed algorithms based on powerful game-theoretical approaches [3]. By adopting cooperative game theory, multiple small cell BSs exchange the necessary information for their coordination through the wired backhaul (BH) links, which works well for small-scale networks. However, for UDNs, the heavy overhead of coordination and the increasing cost of deploying the wired BH links will preclude the application of this approach. Apart from the interference issues, employing more small cell BSs will also increase the maintenance and operational costs. Furthermore, small cells relying on unity frequency reuse will incur more frequent handoffs, which leads to a high latency and/or frequent outages. Hence, a new network architecture should be adopted to support reliable communications in UDNs.

Due to the recent developments in cloud computing and the maturity of multi-core processors, the ultra-dense cloud radio access network (C-RAN) concept has been widely regarded as a promising network architecture that can efficiently address the issues arising in UDNs.

The ultra-dense C-RAN architecture is shown in Fig. 1, which consists of three key components:

1) Baseband unit (BBU) pool hosted in a cloud data center, supported by the techniques of cloud computing, network function virtualization (NFV), software-defined networks (SDN), etc;
2) Low-cost, low power radio remote heads (RRHs) geographically distributed over the coverage area;
3) Wireless fronthaul links that connect the RRHs to the BBU pool.

The main characteristic of dense C-RAN is that all the baseband signal processing units of conventional small cell BSs have been incorporated in the BBU pool, where the computing resources can be shared among the BBUs. Then the conventional full-functionality small cell BSs can be replaced by the low-cost,

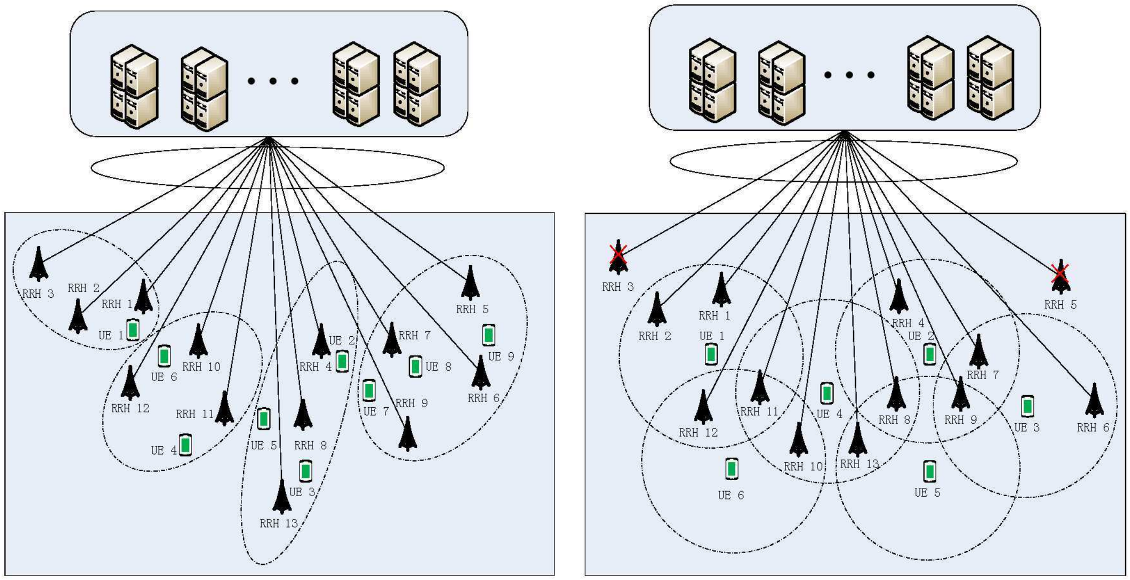

Fig. 1. Illustration of a C-RAN architecture: (a) Disjoint cluster, where the whole network is divided into several non-overlapped clusters, and the cluster-edge users still suffer from high inter-cluster interference; (b) User-centric cluster, where cluster is formed from the user side and the cluster-edge issues are eliminated.

low-complexity RRHs, which are only responsible for the transmission/reception of radio frequency (RF) signals. Due to their simple functionalities, RRHs can be densely deployed at a low hardware cost to provide ultra-high throughput and seamless coverage for a large number of users in tele-traffic hot spots, such as airports, metro/train stations and shopping malls. Thanks to the centralized architecture of C-RAN, the global network information can be shared in the BBU pool and cooperative communication techniques can be realized with the aid of powerful cloud computing, such as large-scale network coordination, global resource management, coordinated multi-point (CoMP) processing, joint flow scheduling and advanced mobility management. These techniques cannot be implemented in conventional ultra-dense small cell networks due to the strict rate/latency requirements. Hence, the cochannel interference that is a limiting factor in UDNs can be efficiently mitigated under the ultra-dense C-RAN architecture. Additionally, the operating status of RRHs and the computing resources of the BBU pool can be dynamically controlled in order to adapt to the capacity demand fluctuations of the users over time, which leads to significant energy and operational cost reductions.

Table I contrasts the differences between the conventional small cell network and the ultra-dense C-RAN.

TABLE I

DIFFERENT TYPES OF UDN DEPLOYMENT

Type of UDN	**Functionality**	**Interference Management**	**Connectivity**	**Deployment Cost**
Small Cells	Fully Functioning	Distributed Algorithm	Wired Backhaul	High Maintenance Cost
Dense C-RAN	PHY Layer Functioning	Centralized Algorithm	Wireless Fronthaul	Low Hardware Cost

II. RESEARCH CHALLENGES AND EXISTING SOLUTIONS

In this section, we first summarize the challenges arising in dense C-RAN and then provide an overview of the existing solutions.

A. Research Challenges

Despite its appealing advantages, there are many technical and deployment challenges associated with dense C-RAN, which include, but are not limited to, the following aspects:

High computational complexity: In dense C-RAN, the BBU pool usually supports a large number of RRHs and the number of variables to optimize, such as beamforming-vectors or transmit powers will become excessive, even in the context of cloud computing.

Stringent fronthaul capacity requirement: In conventional C-RAN, the fronthaul links are typically fixed links, such as optical fibers or high-speed Ethernet. However, in ultra-dense C-RAN a large number of fronthaul links are required. Laying wired links requires high operational and maintenance costs. Additionally, some RRHs may be located at inaccessible places. An attractive alternative is to use wireless fronthaul links, such as millimeter wave (mmWave) Communication links, which are much more scalable and cost-effective than fixed links. However, the bandwidth of wireless links is much lower than that of wired links, which means that the number of users supported by each wireless link is lower.

Huge training overhead for channel state information (CSI) estimation: To facilitate CoMP transmission, the global CSI should be available at the BBU pool, which constitutes an excessive overhead for dense C-RAN. The estimation of the global CSI will also impose a high training overhead that scales with the size of the network. Caire *et al.* [4] showed that the increasing amount of training overhead may in fact outweigh the cooperative gains provided by CoMP transmission.

B. State of the Art Solutions

The most common technique of reducing the signal processing complexity is to adopt the related cluster technique. In general, there are two types of cluster techniques, as shown in Fig. 1: Disjoint clustering and user-centric clustering. In the disjoint cluster, all the RRHs in the network are partitioned into several non-overlapped clusters, and the RRHs in each cluster employ the CoMP technique to serve the users within the coverage area of this cluster. Although the intra-cluster interference can be mitigated, the inter-cluster interference still persists. As a result, the cluster-edge users still suffer from inter-cluster interference. For example, in Fig. 1-(a), user equipment (UE) 1 of cluster 1 suffers from a high interference imposed by the nearby cluster 2. By contrast, in the user-centric cluster, each user is individually served by its nearby RRHs and the cluster is formed from the user's perspective. The scheduled user is the center of the corresponding cluster. Different clusters may overlap with each other, which will eliminate the potential cluster-edge effect. Hence, both the intra- and inter-cluster interference can be efficiently mitigated. Kim *et al.* [5] demonstrated the benefits of employing user-centric clusters over the cell-centric clusters. Hence, the user-centric cluster constitutes the focus of this paper.

The fronthaul capacity constraint has been extensively studied, which can be divided into two categories: the compression strategy and the data sharing strategy. In the compression strategy, the BBU pool first computes the beamforming-vectors for each RRH. Then, the beamformed signals are generated at the BBU pool, which are compressed and sent to the corresponding RRHs. The fronthaul capacity is related to how fine is the resolution of the compressed signals: higher resolutions require a higher fronthaul capacity. Hence, the compression resolution should be optimized under the fronthaul capacity constraints [6]. In the second strategy, the beamforming-vectors computed at the BBU pool are directly sent to the corresponding RRHs. Then, the BBU pool shares each user's data directly with its serving RRH cluster. The beamformed signals are generated at each RRH. In this strategy, the fronthaul capacity depends on the number of users served by each link: a higher number of users requires a higher fronthaul capacity. Hence, the user-RRH associations should be optimized under the fronthaul capacity constraints [7].

To reduce the channel estimation overhead, a promising technique is to rely on partial CSI case under the user-centric cluster, where each user only estimates the CSI of the links from the RRHs within its own cluster (termed as intra-cluster CSI) and only tracks the path loss and shadowing outside its own cluster (termed as inter-cluster CSI), which are sent back to the corresponding RRHs and then collected at the BBU pool. These parameters are necessary for the joint CoMP transmission design at the BBU pool. Indeed, the large-scale fading may be readily tracked since it changes slowly compared to the instantaneous CSI. The design of CoMP transmission weight vectors under this partial CSI case is a challenging task. Hence, there is a paucity of contributions [8]–[10] based on partial CSI for dense C-RAN. To elaborate, compressive CSI acquisition was proposed in [8] for determining the set of instantaneous CSIs and large scale fading gains. However, its complexity is high, hence cannot be readily implemented in dense C-RAN. Lakshmana *et al.* [9] proposed an innovative beamforming design to maximize the weighted sum-rate of UEs under user-centric cluster formation for partial CSI, where the path loss was also incorporated into the optimization problem. However, neither the rate requirements of each user nor the fronthaul capacity constraints were considered in this paper. Recently, we provided a design framework in [10] to deal with most of the emerging challenges arising in dense C-RAN, including the network's power consumption, the limited fronthaul capacity constraints, the computational complexity and the channel estimation overhead. However, the intra-cluster CSI was assumed to be perfect in [8]–[10], which is difficult to achieve in practice.

III. Transmission Scheme Designed for Imperfect Intra-Cluster CSI

For time-division duplex (TDD) C-RANs, the training sequences sent from different users to the same serving RRH should be mutually orthogonal so that the RRH can distinguish the channel vectors of the different users. However, the number of pilots required scales linearly with the number of users, which becomes excessive for dense C-RANs. Hence, the number of time slots remaining for data transmission will be significantly reduced. To reduce the number of pilots, they may be reused by a group of users under the condition that none of users in the group is allowed to share the same serving RRH with the other users. For example, in Fig. 1-(b), UE 1 and UE 2 can reuse the same pilot, since they do not share the same RRH. It is widely recognized that the pilot reuse scheme will however impose the pilot contamination issue, which results in sizeable channel estimation errors. Similar problems occur also in frequency division duplexing (FDD) C-RANs. In the following, we propose a two-stage optimization method to optimize the transmissions for dense C-RAN: In Stage I, a novel pilot reuse scheme is proposed; In Stage II, a robust beamforming-vector optimization algorithm is conceived for maximizing the number of users admitted while considering the pilot contamination incurred by Stage I.

A. Stage I: Novel Pilot Reuse Scheme

The pilot reuse issues have been extensively studied in massive multiple-input multiple-output (MIMO) systems. The basic idea is to reuse the same pilot within the specific group of users having different angles of arrival. However, in dense C-RAN, the number of antennas used at each RRH is limited due to the limited space and hardware cost. Hence, the schemes developed for massive MIMOs are not applicable in dense C-RANs.

Recently, Chen *et al.* [11] proposed a novel pilot allocation scheme for dense C-RANs by using the classic Dsatur graph coloring algorithm [11], which minimizes the number of pilots required for

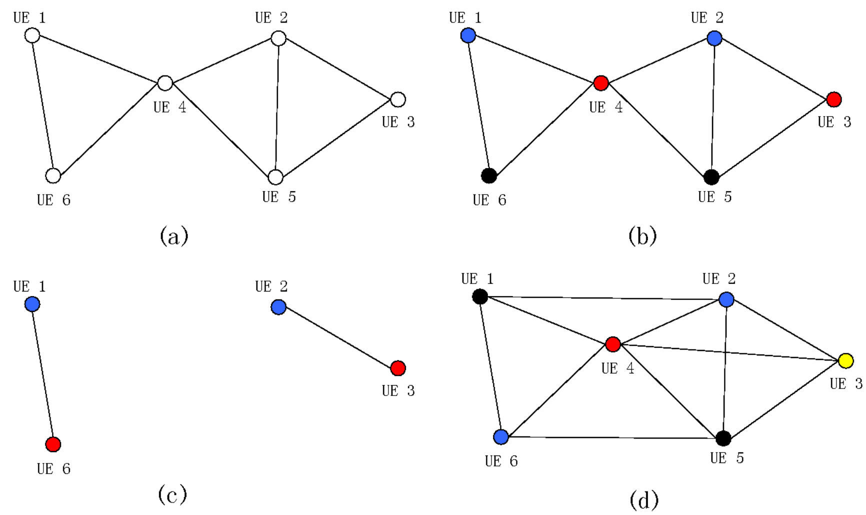

Fig. 2. (a) Construction of the undirected graph for the network in Fig. 1-(b), where the vertexes represent the corresponding users; (b) The colored graph after applying the Dsatur algorithm [11] with $n_{\max} = 2$, the minimum number of required pilots is $n^* = 3$; (c) The user selection and pilot allocation result after using the algorithm for Case I when $\tau = 2$ and $n_{\max} = 2$; (d) The pilot reallocation result after using the algorithm for Case II. In this pilot reallocation result, user 1 and use 2 are allocated with different pilots to reduce the pilot interference, and the same holds for user 5 and user 6, or UE 3 and UE 4.

a given set of users. However, for simplicity of implementation, the 4G long term evolution (LTE) system suggests that the proportion of pilots designated for channel estimation should be fixed within the channel's coherence time, i.e. 1% for a 10 ms training period [12]. Given the fixed number of available pilots, some users may not be allocated any pilots, since again users sharing the same RRH must not reuse the same pilot. Hence, we aim for providing a joint user selection and pilot allocation scheme for maximizing the number of users admitted at a given number of available pilots, while satisfying the following two conditions:

1) The users sharing the same RRH must not reuse the same pilot;
2) There is an upper-bound on how many time each pilot is reused, denoted as $n_{\max}$.

The second constraint is imposed for guaranteeing a fair use of the available pilots in order to avoid the extreme case, where some pilots are reused many times, while there are still some unallocated pilots. The true optimal solution can only be obtained by an exhaustive search method. However, its complexity increases exponentially with the number of users, which is not practical for dense C-RANs. Hence, a low-complexity nearly-optimal algorithm is proposed for solving this problem.

It is assumed that the number of users K in dense C-RANs is higher than the number of available pilots τ. For a dense C-RAN, we construct a $(K \times K)$ matrix $\mathbf{B}$, where each element is given by

$$b_{k,k'} = \begin{cases} 1, \text{ if } \mathcal{I}_k \cap \mathcal{I}_{k'} \neq \emptyset \text{ and } k \neq k' \\ 0, \text{ otherwise,} \end{cases} \tag{1}$$

where $\mathcal{I}_k$ denotes the set of RRHs that potentially serve user k. The above definitions mean that if user k and user k' do not have common RRHs, the corresponding matrix element is set to $b_{k,k'} = 1$. Otherwise, the element is set to zero. Based on matrix $\mathbf{B}$, a unidirectional graph can be constructed for the network of Fig. 1-(b) as seen in Fig. 2-(a), where the vertices represent the corresponding users. Fig. 2-(b) illustrates the pilot allocation results after applying the Dsatur algorithm [11], which shows that to adequately serve all the six users, at least three different pilots are required.

To solve our pilot allocation optimization problem, we first adopt the Dsatur algorithm to find the minimum number of pilots required for serving all users in the network. If this number is higher than the number of pilots available, some users should be removed. Otherwise, all users can be admitted. In the latter case, some pilots may be reused by up to $n_{\max}$ users, while some pilots are not allocated, which wastes the pilot resources. To resolve this problem, all pilots should be reallocated so that all available pilots should be used to reduce the pilot contamination. Let us denote the minimum number of pilots required by the Dsatur algorithm as n^*. In the following, we provide a detailed algorithm to deal with each case: 1) $n^* > \tau$; 2) $n^* < \tau$. When $n^* = \tau$, no additional operations are required.

Case I: $n^* > \tau$. In this case, the number of available pilots is insufficient for supporting all users in the network, hence some users should be removed. Let us define $\theta_k \triangleq \sum_{k' \neq k, k' \in \bar{\mathcal{U}}} b_{k,k'}$ as the degree of the vertex associated with user k, which represents the total number of users connected to this user. The user having the largest value of θ_k should be deleted with high priority, since many users should be allocated different pilots compared to this user. However, there may exist several users with the same value of θ_k, and randomly removing one of them will lead to a reduced performance. Hence, the user earmarked for deletion should be carefully selected. Intuitively, the user suffering from the highest pilot interference should be removed. To this end, we define the metric $\eta_{k,k'}$ to quantify the level of pilot interference between any two users, when they are reusing the same pilot:

$$\eta_{k,k'} = \log\left(1 + \frac{\sum_{i \in \mathcal{I}_{k'}} \alpha_{i,k}}{\sum_{i \in \mathcal{I}_k} \alpha_{i,k}}\right) + \log\left(1 + \frac{\sum_{i \in \mathcal{I}_k} \alpha_{i,k'}}{\sum_{i \in \mathcal{I}_{k'}} \alpha_{i,k'}}\right). \tag{2}$$

where $\alpha_{i,k}$ represents the large-scale fading power from RRH i to user k. Obviously, a higher value of $\eta_{k,k'}$ means higher pilot interference between these two users, when they employ the same pilot. Then, we quantify the level of pilot interference incurred by this user by $\xi_k = \sum_{k' \in \mathcal{K}_{\pi_k} \setminus \{k\}} \eta_{k,k'}$ as our figure of merit. The user with the largest value of ξ_k should be removed. Based on this idea, we conceive the user selection and pilot allocation method formulated in Algorithm 1. By invoking this algorithm for the network in Fig. 1-(b) with $\tau = 2$ and $n_{\max} = 2$, the final assignment is shown in Fig. 2-(c), where four users are admitted.

Algorithm 1 User selection and pilot allocation algorithm for Case I

1: Initialize matrix $\mathbf{B}$, user set $\mathcal{U} = \{1, \cdots, K\}$, initial number of required pilots n^* obtained from the Dsatur algorithm;

2: While $n^* > \tau$

3: Find $k^* = \arg\max_{k \in \mathcal{U}} \theta_k$. If there are multiple users with the same θ_k, remove the user with the largest ξ_k;

4: Remove user k^* from $\mathcal{U}$, i.e., $\mathcal{U}$=$\mathcal{U}/k^*$, and update matrix $\mathbf{B}$ with current $\mathcal{U}$;

5: Use the Dsatur algorithm to calculate n^* with $\mathbf{B}$ and $\mathcal{U}$;

Case II: $n^* < \tau$. In this case, we aim for reallocating all the available pilots to all users for additionally reducing the pilot contamination. Note that in Fig. 2-(b) having three allocated pilots, there may be measurable pilot interference between user 1 and user 2. If four pilots are available, these two users can be allocated different pilots as seen in Fig. 2-(d), which enhances the system performance. To this end, we reconstruct the undirected graph by introducing a threshold $\eta_{\rm th}$. If $\eta_{k,k'} > \eta_{\rm th}$, user k and user k' have measurable pilot interference if reusing the same pilot, hence they should be connected. Otherwise, they can reuse the same pilot. Based on this idea, matrix $\mathbf{B}$ can be reconstructed as

$$b_{k,k'} = \begin{cases} 1, \text{ if } \mathcal{I}_k \cap \mathcal{I}_{k'} \neq \emptyset \text{ and } k \neq k', \\ 1, \text{ if } \eta_{k,k'} > \eta_{\rm th},\ \mathcal{I}_k \cap \mathcal{I}_{k'} = \emptyset \text{ and } k \neq k', \\ 0, \text{ otherwise.} \end{cases} \tag{3}$$

As expected, a smaller value of $\eta_{\rm th}$ will result in more users becoming connected with each other and hence more pilots are required. In the extreme case of $\eta_{\rm th} < \min\{b_{k,k'}\}$, all users become connected and the number of pilots required is equal to K. On the other hand, if $\eta_{\rm th} \geq \max\{b_{k,k'}\}$, the reconstructed matrix $\mathbf{B}$ reduces to the initial $\mathbf{B}$ defined in (1), where the number of pilots required is n^*. Since $\tau < K$, there must exist a $\eta_{\rm th}$ value so that the number of pilots required is equal to τ. The bisection search algorithm can be adopted to find this threshold $\eta_{\rm th}$, but the details of this must be omitted. Again, Fig. 2-(d) shows the pilot allocation results after using this algorithm.

B. Stage II: Robust Beamforming-vector Design

In Stage II, we aim for designing the beamforming-vectors by considering the pilot contamination due to the pilot reuse scheme in Stage I. Specifically, we formulate a transmit power minimization problem, while considering the following three constraints:

1) Each user's data rate should be higher than its minimum requirement;
2) Each fronthaul link capacity constraint is imposed;
3) Each RRH has its individual power constraint.

There are four challenges to solve this optimization problem:

1) Since we consider the partial CSI case where only the inter-cluster large-scale fading parameters are available, the exact data rate of each user is difficult to obtain.
2) Due to the fronthaul capacity constraint and RRH power constraint, this system may not be able to support all users at their specific rate requirements.
3) The fronthaul capacity constraint contains the non-smooth and non-differentiable indicator function, which results in a mixed-integer non-linear programming (MINLP) problem that is NP-hard to solve.
4) Due to the channel estimation error, each user suffers from residual self-interference. The conventional weighted minimum mean square error (WMMSE) method of [13] that has been successfully applied in the perfect intra-cluster CSI scenario [7], [10], cannot be used for solving the problem considered here.

We provide brief descriptions of the associated methods to address the above four challenges.

First, Jensen's inequality is used for finding the lower-bound of the exact data rate, which is more amenable to the design of our algorithm. In [10], we have shown that for the specific scenario of non-overlapped cluster, the gap between the lower-bound and the exact data rate is within 3% for both sparse and dense C-RAN scenarios. Hence, a simple correction factor may be used for practical applications.

Second, we construct an alternative optimization problem to deal with the infeasibility of the original problem by introducing a series of auxiliary non-negative variables. This idea is inspired by [14], which can maximize the number of users admitted, while simultaneously minimizing the transmit power.

Third, the non-smooth indicator function is replaced by a concave function $f_\theta(x) = \frac{x}{x+\theta}$, where θ is a small positive value. The transformed problem is the well-known difference of convex (d.c.) program, which can be efficiently solved by the successive convex approximation (SCA) method under convergence guarantee.

Finally, to deal with the fourth challenge, we adopt the semi-definite relaxation approach and formally prove that semidefinite relaxation is tight with a probability of 1.

C. Simulation Results

We now present our simulation results for evaluating the performance of the proposed algorithms. The dense C-RAN is assumed to cover a square shaped area of 2 km $\times$ 2 km. The numbers of RRHs and users are set to $I = 30$ and $K = 18$, respectively. Both the users and RRHs are uniformly and independently distributed in this area. It is assumed that each user is potentially served by its nearest L RRHs. Each fronthaul link is assumed to only support three users, since mmWave communication is employed as the wireless fronthaul link. The other systems parameters can be found in [15].

We compare our proposed algorithm to the following algorithms:

1) Orthogonal pilot allocation (with legend "Ortho"): As the terminology implies, all users are allocated orthogonal pilot sequences, hence the maximum number of users that can be admitted in Stage I is equal to the number of available pilots τ. These τ users are randomly selected from K users.
2) No reallocation operations for Case II in Stage I (with legend "NoCaseII"): This algorithm is similar to our proposed algorithm, except when Case II would occur, no additional operations are performed.
3) Conventional pilot allocation method (with legend "Con"): This algorithm is similar to the "NoCaseII" algorithm, except when Case I would occur, the users are randomly removed until the number of required pilots is equal to τ.
4) Perfect CSI estimation (with legend "Perfect"): This is the baseline algorithm, where the intra-cluster CSI is assumed to be perfectly known. This algorithm is provided to quantify the effect of channel estimation errors.

Figs. 3 and 4 illustrate the number of users admitted versus the candidate RRH set size L in Stage I and Stage II of Section III-A and III-B, respectively. It is seen from Fig. 3 that the number of admitted users monotonically decreases upon increasing the candidate set size L. This is due to the fact that with a larger candidate RRH set size for each user, more users will become connected with each other, which requires more pilots. This figure also illustrates the superior performance of our proposed algorithm over the "Con" algorithm, highlighting the necessity of carefully considering the pilot interference, when removing users.

It is interesting to observe from Fig. 4 that the number of users admitted by all algorithms (except the "Ortho" algorithm) initially increases with the candidate set size and then decreases. The reason is

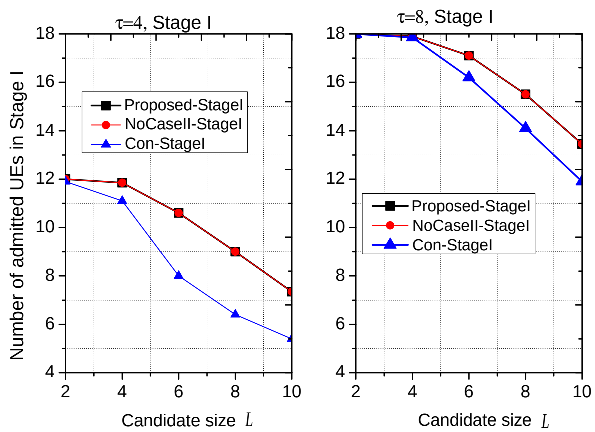

Fig. 3. Number of admitted UEs in Stage I versus the candidate size L. The left subplot corresponds to the case when the number of available pilots is $\tau = 4$ while the right one is $\tau = 8$.

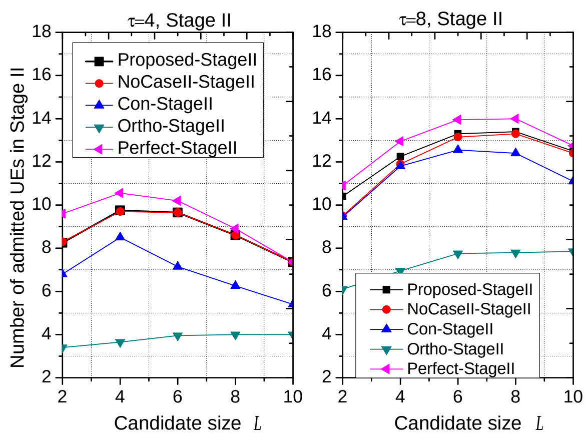

Fig. 4. Number of admitted UEs in Stage II versus the candidate size L. The left subplot corresponds to the case when the number of available pilots is $\tau = 4$ while the right one is $\tau = 8$.

that when L starts to increase, a higher spatial degree of freedom is available to support more users. However, when L continues to increase, many users are rejected in Stage I, as seen in Fig. 3. This trend is different from the widely accepted concept that increasing the candidate set size will always lead to better performance. Hence, the channel estimation process should be taken into account, when designing the cluster. This figure also shows the performance advantage of our proposed algorithms over the other algorithms.

IV. Conclusions and Future Research Challenges

We advocated the application of a user-centric dense C-RAN architecture for UDNs due to its appealing features such as the facilitation of centralized signal processing, low hardware cost, low Capital Expenditure (CAPEX) and Operational Expenditure (OPEX), etc. However, we also identified the challenge of requiring a heavy training overhead for estimating all the CSIs for cooperative transmission. As a remedy, we adopted the partial CSI model, where only the large-scale inter-cluster CSI is available. The channel estimation required for intra-cluster CSI was also considered, where a novel pilot allocation scheme was proposed. Then, we developed a robust transmission design by considering the effect of channel estimation errors. Our simulation results verified the performance advantages of the proposed algorithm over the existing ones.

Finally, we now highlight several promising research directions to make the user-centric dense C-RAN more amenable for practical implementations.

Dynamic cluster formations: In this work, the cluster for each user was assumed to be fixed, i.e. each user is only connected to its nearest L RRHs. However, in practical systems, the cluster sizes for users should be adapted to the network state, such as the users' rate requirements or traffic load, each RRH's power limit, each fronthaul capacity constraint, etc. Additionally, the channel estimation stage should be taken into account as seen in Fig. 4, where the network performance may even degrade with the cluster size. How to optimize the cluster size individually for each user by jointly considering the above elements remains an inspiring research direction.

User mobility management: User mobility is a very challenging issue in user-centric ultra-dense C-RANs. When the users move from one place to another, the cluster of RRHs assigned for serving this user should be adaptively changed. Explicitly, an adaptive mobility management method should be developed so that the serving cluster can follow each user's behavior (e.g. mobility and service demands) and provide data transmission without the users' involvement. Fortunately, the 'Big Data' technique relying on machine learning is becoming mature, which can track the users' mobility and then predict their future locations. By applying this technology, the users' serving cluster can be formed beforehand in anticipation that significantly reduces the processing time and meets the targeted quality of experience (QoE) levels.

Robust Transmission Design for Dense FDD C-RANs: The FDD mode is another alternative transmission mode to the TDD mode considered in this paper. In dense FDD C-RANs, each user will estimate its CSI wrt the RRHs within its cluster. Then, the CSI will be quantized and fed back to the corresponding serving RRHs. In addition to the channel estimation error considered in TDD C-RANs, other errors will also be introduced, such as the quantization error and that of outdated CSIs. If directly applying the quantized CSI for joint transmission design, not all the users' quality of service (QoS) can be guaranteed. Hence, robust transmission designs have to be developed to guarantee the users' QoS.

Channel prediction is another useful method of providing more accurate CSI. Some prediction techniques, such as Kalman and Weiner filtering can make the performance of the C-RAN network more robust to the delay effects. Finally, Big Data techniques can also be adopted to predict the channel parameters. In a nutshell, the future is bright for user-centric design.

References

[1] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] M. Ding, P. Wang, D. Lpez-Prez, G. Mao, and Z. Lin, "Performance impact of LoS and NLoS transmissions in dense cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2365–2380, Mar. 2016.

[3] Z. Han, *Game theory in wireless and communication networks: theory, models, and applications*. Cambridge University Press, 2012.

[4] G. Caire, S. A. Ramprashad, and H. C. Papadopoulos, "Rethinking network MIMO: Cost of CSIT, performance analysis, and architecture comparisons," in *Information Theory and Applications Workshop (ITA), 2010*, 2010, pp. 1–10.

[5] J. Kim, H. W. Lee, and S. Chong, "Virtual cell beamforming in cooperative networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1126–1138, Jun. 2014.

[6] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[7] 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.

[8] Y. Shi, J. Zhang, and K. B. Letaief, "CSI overhead reduction with stochastic beamforming for cloud radio access networks," in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 5154–5159.

[9] T. R. Lakshmana, A. Tolli, R. Devassy, and T. Svensson, "Precoder design with incomplete feedback for joint transmission," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1923–1936, Mon. 2016.

[10] C. Pan, H. Zhu, N. Gomes, and J. Wang, "Joint user selection and energy minimization for ultra-dense multi-channel C-RAN with incomplete CSI," *to appear in IEEE J. Sel. Areas Commun.*, 2017.

[11] Z. Chen, X. Hou, and C. Yang, "Training resource allocation for user-centric base station cooperation networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2729–2735, Apr. 2016.

[12] G. R1-074068, "Summary of reflector discussions on E-UTRA UL RS," *3GPP TSG RAN WG1 Meeting 50bis*, Oct. 2007.

[13] 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.

[14] E. Matskani, N. D. Sidiropoulos, Z. q. Luo, and L. Tassiulas, "Convex approximation techniques for joint multiuser downlink beamforming and admission control," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2682–2693, Jul. 2008.

[15] C. Pan, H. Mehrpouyan, Y. Liu, M. Elkashlan, and A. Nallanathan, "Joint pilot allocation and robust transmission design for ultra-dense user-centric TDD C-RAN with imperfect CSI," *To appear in Arxiv website.*, 2017.