

Distributed Resource Allocation Optimization for User-Centric Cell-Free MIMO Networks

Hussein A. Ammar*, *Student Member, IEEE*, Raviraj Adve*, *Fellow, IEEE*, Shahram Shahbazpanahi†*, *Senior Member, IEEE*, Gary Boudreau‡, *Senior Member, IEEE*, and Kothapalli Venkata Srinivas‡, *Member, IEEE*

Abstract—We develop two distributed downlink resource allocation algorithms for user-centric, cell-free, spatially-distributed, multiple-input multiple-output (MIMO) networks. In such networks, each user is served by a subset of nearby transmitters that we call distributed units or DUs. The operation of the DUs in a region is controlled by a central unit (CU). Our first scheme is implemented at the DUs, while the second is implemented at the CUs controlling these DUs. We define a hybrid quality of service metric that enables distributed optimization of system resources in a proportional fair manner. Specifically, each of our algorithms performs user scheduling, beamforming, and power control while accounting for channel estimation errors. Importantly, our algorithm does not require information exchange amongst DUs (CUs) for the DU-distributed (CU-distributed) system, while also smoothly converging. Our results show that our CU-distributed system provides 1.3- to 1.8-fold network throughput compared to the DU-distributed system, with minor increases in complexity and front-haul load - and substantial gains over benchmark schemes like local zero-forcing. We also analyze the trade-offs provided by the CU-distributed system, hence highlighting the significance of deploying multiple CUs in user-centric cell-free networks.

Index Terms—Distributed resource allocation, user scheduling, user-centric clustering, cell-free MIMO, cooperative cellular networks, distributed antenna system, scalable resource allocation, fairness.

I. INTRODUCTION

Scalability requires a system to accommodate growing demands gracefully [1]. Scalability is critical motivation for user-centric, cell-free, spatially-distributed, multiple-input multiple-output (MIMO) networks [2], [3], where users can be served by many transmitters [4], [5], denoted herein as distributed units (DUs). In these networks, a serving cluster of DUs is defined for each user based on a metric, e.g., channel power [5] or serving distance [6]; each user is effectively located at the center of its serving cluster, thereby eliminating conventional cell edges.

Designing flexible distributed resource allocation schemes for cell-free networks is still an open issue [7] due to the

difficulties arising from the lack of a regular cell structure and the overlapping serving clusters for the users. Theoretically, centralized resource allocation, wherein all transmissions are jointly optimized, provides the performance upper bound. However, such a scheme implies an enormous overhead in terms of real-time exchange of channel state information (CSI) and computation load. Furthermore, since the optimization problems involved are usually non-convex, even effective solutions may not lead to the global optimum.

A distributed system, on the other hand, provides a trade-off between performance and scalability. Compared to a centralized scheme, it provides an advantage due to lower front-haul load as well as lower computational and storage complexity per network node. Importantly, it allows for lower communication overhead with limited exchange of CSI. Hence, a distributed system is more practical to deploy than a centralized one [8]. On the other hand, such an approach *may* provide worse performance than a centralized scheme [9], because the resource allocation is performed without a global view of the network and with limited coordination.

A major challenge in designing a distributed resource allocation scheme is that the crucial signal-to-interference-plus-noise ratio (SINR) metric is coupled between all the transmitters. Hence, SINR-based approaches, like in [10], [11], are not suitable for distributed resource allocation. One alternative is to use metrics like the signal-to-leakage-plus-noise ratio (SLNR) [12] which decouples the allocation problem between the different transmitters. However, SLNR does not allow for effective power allocation because the beam power scales equally in both the signal and leakage terms. Here, we modify the approach proposed for cellular networks [13] based on a mix of inter-cell leakage and intra-cell interference. Specifically, for the cell-free MIMO network at hand, the combinatorial approach to user scheduling in [13] is not possible. Additionally, unlike most works in distributed processing, we include fairness as a key criterion for user-scheduling, thereby avoiding repeatedly serving the same users with strongest channels.

The available efforts to implement a distributed resource allocation scheme in cell-free networks have been based on algorithms that require iterative exchange of signals between the base stations [14]. The investigation in [15] studies a distributed framework for cooperative precoding in cell-free MIMO requiring over-the-air signaling between the base stations instead of front-haul/back-haul signaling. Similarly, the work in [16] maximizes the weighted sum rate (WSR) within joint transmission clusters without centralized beamformer

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and in part by Ericsson Canada.

*H. A. Ammar and R. Adve are with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: ammarhus@ece.utoronto.ca; rsadve@comm.utoronto.ca).

†S. Shahbazpanahi is with the Department of Electrical, Computer, and Software Engineering, University of Ontario Institute of Technology, Oshawa, ON L1H 7K4, Canada. He also holds a Status-Only position with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto.

‡G. Boudreau and K. V. Srinivas are with Ericsson Canada, Ottawa, ON K2K 2V6, Canada.

processing. However, the problem is solved through an equivalent minimum square error-based problem that requires a feedback channel to update weights during algorithm iterations.

The authors in [17] maximize the uplink minimum SINR by decoupling the problem into two sub-problems, one that designs the receiver filter and another that optimizes the power allocation, using an alternating optimization. The study in [18] propose two distributed variants of zero-forcing (ZF); however, this work does not consider user scheduling. The authors of [19] optimize the energy efficiency in the uplink of cell-free massive MIMO networks under different scenarios of signal quantization. The investigations in [2] and [20] study a suboptimal but scalable power control policy that uses large-scale fading decoding. Moreover, the study in [21] uses dual decomposition and the gradient method for resource allocation in a relaying system by relaxing the binary scheduling variables. Furthermore, the work in [8] analyzes the uplink spectral efficiencies under four different levels of cooperation in cell-free implementations. However, this work does not optimize resource allocation, but rather numerically studies the performance under minimum mean-square error (MMSE) combining.

Distributed resource allocation can also be tackled through a game-theoretic framework. The studies in [22], [23] use a distributed form of auction theory for user scheduling in conventional networks. Briefly, the users compete for the resources through bidding and assignment phases [24]. One disadvantage of such approaches is the considerable communications required between the network nodes before an allocation occurs, hence leading to a substantial overhead. Furthermore, these have been developed for simple schemes like carrier sense multiple access.

The DUs that connect to users are, themselves, controlled by central units (CUs). While some models use a single CU to control all DUs [9], [20], this would limit the scalability of the network. As in [18], we consider a system of multiple central units (CUs), each controlling the DUs in its region. Under the user-centric cell-free MIMO scheme at hand, a serving cluster of DUs is defined specifically for each user. In contrast to the schemes available in the literature, we develop two *totally* distributed resource allocation schemes that perform user scheduling, beamforming, and implicit power control in a user-centric MIMO network. The first solution is implemented on the DUs, while the second one is deployed on the CUs¹. We acquire *local* CSI between each DU and its users, an approach shown to be scalable [26]. We then define a weighted pseudo-rate function that depends on a hybrid leakage and intra-(DU\CU) interference; the weights implement user fairness. Our objective decouples the problem between the different network nodes, with each node making decisions independent of the other nodes in the network.

To solve our formulated problem, we employ tools such as block coordinate descent, fractional programming, and

compressive sensing to develop an algorithm that converges smoothly in a non-decreasing manner. In addition to standard CSI estimation, we further propose using either statistical CSI or the spatial traffic distribution to compute the leakage needed to construct our objective function. These alternative approaches further reduce any required information exchange amongst processing nodes. Notably, our results show that the proposed methods to compute the leakage are very effective with a dense distribution of users.

To the best of the authors' knowledge, this work is one of the first studies to consider fully distributed resource allocation, and is the first study which investigates fully distributed user-scheduling and beamforming (not only power control) for user-centric cell-free MIMO networks. In most of the literature, user scheduling is neglected and the users are assumed pre-selected. Hence, our work fills two gaps by focusing on distributed schemes and on user scheduling. Our scheme is different from the literature, e.g., [15], in the sense that it conducts resource allocation without using feedback channels between the network entities. Additionally, to maintain scalability, we use a network with distributed CUs [9], [20].

A few works in the literature focus on scalability. The study in [9] proposes a simple sub-optimal power allocation using conjugate beamforming (to be able to perform a distributed beamforming), but this study does not optimize user scheduling or beamforming. In [2] the authors study the uplink of a single-CU network, using large-scale fading decoding and power control; however, the scheme is not distributed. The authors of [27] employ a combinatorial search for user-association based on the position of the access points. The system uses conjugate beamforming and maximizes the sum rate, however, this scheme is also not distributed. The scheme sacrifices performance to enable a low fronthaul load. Thus, our developed schemes are fundamentally different from those found in literature.

The theory developed here sets the foundation for deployment of an optimized resource allocation scheme. Specifically, the contributions of our paper are:

- Developing two distributed resource allocation schemes that optimize resources, including user scheduling and beamforming (not only power allocation) in a cell-free, user-centric, MIMO network. These schemes are based on a hybrid leakage and interference metric that eliminates the need for information exchange and implements fairness amongst users.
- Proposing and testing three approaches to calculate the leakage term in the metric. These approaches require different levels of computational complexity and required real-time information while producing comparable performance, especially at high user densities.
- Analyzing the computation complexity of our two proposed approaches, illustrating the substantial reduction in complexity compared to a centralized resource allocation scheme, with, in some cases, improved performance.
- Highlighting the importance of a multiple-CU user-centric cell-free network by illustrating trade-offs in performance and front-haul load provided by the CU-

¹Although this study is not concerned with the core-network protocols that should be used to deploy a distributed CU system under the user-centric cell-free network architecture, the authors note that distributed software defined network (SDN) [25] is a promising avenue to implement the flexible network architecture suggested in this paper.

distributed system compared to DU-distributed and centralized systems.

The rest of the paper is organized as follows. Section II presents the system model. Section III formulates the resource allocation problem in a DU-distributed system and develops the steps required to solve this problem. Section IV casts the problem as a CU-distributed system and develops an approach to perform the resource allocation. Section V proposes further methods to enhance the scalability of calculating the leakage term. Section VI reports on our numerical results and findings. Finally, Section VII concludes our discussion.

Notation: a lower or upper case letter, e.g., a or A , represents a scalar, a bold lower case letter, e.g., \mathbf{a} , represents a vector, while a bold upper case letter, e.g., \mathbf{A} , represents a matrix. The term $[\mathbf{A}]_{ij}$ is the (i, j) th entry of matrix \mathbf{A} . The operators $(\cdot)^{-1}$, $(\cdot)^T$ and $(\cdot)^H$, used as superscripts, denote the inverse, transpose, and conjugate transpose, respectively. $\|\cdot\|_2$ and $|\cdot|$ are the vector and scalar Euclidean norms (ℓ_2 norm), $\|\cdot\|_p$ is the ℓ_p norm, and \mathbf{I}_m is the m -dimensional identity matrix. Calligraphic letters, e.g., \mathcal{A} , are used to indicate sets with its cardinality represented by $|\mathcal{A}|$. We use $\mathbf{A} = [\mathcal{A}]$ to construct a matrix or vector using the elements of set \mathcal{A} . The spaces \mathbb{B} , \mathbb{C}^m , and $\mathbb{C}^{m \times n}$ represent the set of binary numbers, complex $m \times 1$ column vectors, and complex $m \times n$ matrices, respectively. $\mathbb{E}\{\cdot\}$ is the expectation operator, and $\mathbf{x} \sim \mathcal{CN}(\mathbf{m}, \mathbf{R})$ indicates that \mathbf{x} is a complex Gaussian random vector with mean \mathbf{m} and covariance \mathbf{R} .

II. SYSTEM MODEL

A. Network and Signal Model

We consider a user-centric cell-free MIMO network, which operates in time-division duplex (TDD) mode. Our network, illustrated in Fig. 1, comprises Q CUs, each controlling N DUs represented by the set \mathcal{B}_q . We denote the region containing the DUs in \mathcal{B}_q as a *virtual cell*. Accordingly, we have a total of NQ DUs represented by the set of sets $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_Q\}$. Each DU is equipped with M antennas, while each user is equipped with a single antenna. We use \mathcal{U} to represent the set of users that need to be served, where $|\mathcal{U}|$ is a random number, but is much larger than the number of available resources. The DUs serve users coherently².

Based on this model, for each user u , we define a serving cluster \mathcal{C}_u comprising the DUs that *can potentially* serve the user. Specifically, here we define \mathcal{C}_u based on the criterion $\{\psi_{ru}\beta(d_{ru}) \geq \rho : r \in \mathcal{B}\}$, where the term ψ_{ru} accounts for the (lognormal) shadowing, $\beta(d_{ru})$ accounts for the path loss, which depends on the distance d_{ru} between DU r and user u . If no DU can meet this connection criterion, \mathcal{C}_u comprises the DU providing the highest average received power, i.e., largest $\psi_{ru}\beta(d_{ru})$. Based on this scheme, we can formally

²Coherent transmission requires phase synchronization across the DUs, which can be achieved through synchronization protocols like the IEEE 1588v2 (Precision Time Protocol PTP) [28]. The system can achieve this synchronization by properly choosing the serving clusters and the use of a cyclic prefix [6]. We note that synchronization issues are out of the scope of this paper; they need to be studied in a dedicated work using different tools from those used in this paper.

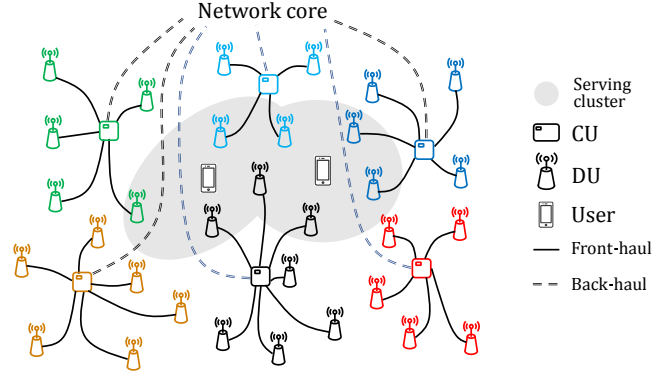


Fig. 1: User-centric cell-free network with *either* DU-distributed or CU-distributed processing.

define $\mathcal{C}_u = \{r : \psi_{ru}\beta(d_{ru}) \geq \rho\} \cup \{\arg \max_r \psi_{ru}\beta(d_{ru})\}$. The set \mathcal{C}_u , therefore excludes DUs that cannot contribute significantly to the user's useful signal while also limiting the number of users each DU serves. We assume a low-mobility profile for the users and hence we use the popular block fading model for the channel. Hence, the channel (small-scale and large-scale fading) is assumed static within each channel coherence interval. The large-scale fading also changes much more slowly than the small-scale fading, and hence stays constant for many coherence intervals. The block fading model is widely accepted and used in the literature when high-mobility scenarios are not being studied [4] as the case of this paper.

We develop two different distributed algorithms for resource allocation. The first variant runs on the DUs, while the second variant runs on the CUs. The developed algorithms provide user scheduling and perform resource allocation. These procedures include channel estimation, formation of serving clusters, user scheduling, beamforming, and fairness control for the users. Finally, it is worth noting that we will be using the *global* indices q , r , and u to, respectively, refer to the CUs, DUs and users found in the network.

In the downlink, the signal received by user $u \in \mathcal{U}$ can be written as

$$y_u = \sum_{r \in \mathcal{C}_u} \sqrt{s_{ru}} \mathbf{h}_{ru}^H \mathbf{w}_{ru} x_u + \sum_{u' \in \mathcal{U}, u' \neq u} \sum_{r' \in \mathcal{C}_{u'}} \sqrt{s_{r'u'}} \mathbf{h}_{r'u'}^H \mathbf{w}_{r'u'} x_{u'} + z_u. \quad (1)$$

The first term in (1) is the useful signal, the second term is the interference, and the third is the additive white Gaussian noise (AWGN) $z_u \sim \mathcal{CN}(0, \sigma_z^2)$. The scalar $s_{ru} \in \mathbb{B}$ is the scheduling decision at DU r for user u (that is, when $s_{ru} = 1$, DU r schedules user u , otherwise $s_{ru} = 0$), $\mathbf{h}_{ru} \in \mathbb{C}^{M \times 1}$ is the channel vector between the two peers, \mathbf{w}_{ru} is the beamforming weight or the transmit precoder used by DU r to serve the user with an overall power budget of p for the DU, and x_u is the zero-mean data symbol intended for user u with $\mathbb{E}\{|x_u|^2\} = 1$.

We model the channel between DU r and user u as $\mathbf{h}_{ru} = \sqrt{\psi_{ru}\beta(d_{ru})} \mathbf{g}_{ru} \in \mathbb{C}^{M \times 1}$, where the term $\mathbf{g}_{ru} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_M)$ accounts for the small-scale fading, and as noted earlier, ψ_{ru}

and $\beta(d_{ru})$ account for lognormal shadowing and path loss, respectively. We define the set \mathcal{E}_r representing the users that need to be served by DU r . These sets $\{\mathcal{E}_r : r \in \mathcal{B}\}$ can be obtained directly from the sets $\{\mathcal{C}_u : u \in \mathcal{U}\}$. Crucially, because of user-centric clustering, each \mathcal{E}_r can partially or totally overlap with other $\mathcal{E}_{r'}$, for $r' \neq r$; similarly, the same user may “belong” to multiple CUs.

B. Channel Estimation

For a TDD system, channel estimation is performed in the uplink and the estimated channel is used in downlink assuming channel reciprocity. In the pilot-training phase of length τ_p , the signal $\mathbf{Y}_r \in \mathbb{C}^{M \times \tau_p}$ observed at DU r can be written as

$$\mathbf{Y}_r = \sum_{u \in \mathcal{U}} \sqrt{p_u} \mathbf{h}_{ru} \phi_u + \mathbf{Z}_r, \quad (2)$$

where $\phi_u \in \mathbb{C}^{1 \times \tau_p}$ is the unit norm ($\phi_u \phi_u^H = 1$) pilot sequence used by user u , p_u is the transmit power, and \mathbf{Z}_r is the noise with entries distributed as $\mathcal{CN}(0, \sigma_Z^2)$. Following the assumptions in [4], [5], we assume the knowledge of the the large-scale fading and the transmit power used.

Unfortunately, using pilots that are orthogonal among the users, i.e., $\phi_u \phi_{u'}^H = 0, \forall u' \neq u$, requires having $\tau_p \geq |\mathcal{U}|$ which is unlikely to be feasible. We assume that the users are grouped such that the pilot sequences inside each group are orthogonal, while the same pilot sequences are used across groups. Specifically, as in [29], we use the hierarchical agglomerative clustering (HAC) algorithm to cluster the users into subsets each containing a number of users less than or equal to τ_p , the length of the pilot sequence. The available orthogonal pilot sequences are randomly assigned to the users inside each subset. The intuition is to keep users sharing the same pilot sequence as far as possible from each other, thus minimizing pilot contamination in our user-centric cell-free network [30]. Additionally, the choice of HAC is based on its consistency and its lack of sensitivity to the choice of the distance-metric used to construct the subsets [31].

We can extract the channel of user u at DU r by first defining $\check{\mathbf{y}}_{ru} = \frac{1}{\sqrt{p_u}} \mathbf{Y}_r \phi_u^H$, hence eliminating all users' contributions other than the ones using the same pilot sequence ϕ_u . The channel $\{\mathbf{h}_{ru} : u \in \mathcal{U}\}$ can then be estimated using linear MMSE as [32]

$$\hat{\mathbf{h}}_{ru} = \mathbf{D}_{ru} \left(\sum_{u' \in \mathcal{U}_u} \mathbf{D}_{ru'} + \frac{\sigma_Z^2}{p_u} \mathbf{I}_M \right)^{-1} \check{\mathbf{y}}_{ru}, \quad (3)$$

where $\mathbf{D}_{ru} \in \mathbb{C}^{M \times M}$ is a diagonal matrix with entries $[\mathbf{D}_{ru}]_{mm} \triangleq \psi_{ru} \beta(d_{ru})$, and \mathcal{U}_u are the users using the same pilot sequence as user u (including user u).

Remark 1. Due to path loss, each DU r only needs to estimate the channel vectors of nearby users. In Section V, we show that each DU need only estimate the channels to users $u \in \mathcal{E}_r$, i.e., its own users and it can use large-scale fading statistics or a traffic distribution model for the other users, thus enhancing the scalability of the required channel estimation.

The estimated channel $\hat{\mathbf{h}}_{ru}$ is distributed according to $\mathcal{CN}(\mathbf{0}, \mathbf{\Psi}_{ru})$, with $\mathbf{\Psi}_{ru}$ defined as

$$\mathbf{\Psi}_{ru} \triangleq \mathbf{D}_{ru} \left(\sum_{u' \in \mathcal{U}_u} \mathbf{D}_{ru'} + \frac{\sigma_Z^2}{p_u} \mathbf{I}_M \right)^{-1} \mathbf{D}_{ru}. \quad (4)$$

It is known from the theory of MMSE estimation that the channel estimation error $\mathbf{e}_{ru} = \mathbf{h}_{ru} - \hat{\mathbf{h}}_{ru}$ is distributed as $\mathcal{CN}(\mathbf{0}, \mathbf{\Theta}_{ru})$, where $\mathbf{\Theta}_{ru} \triangleq \mathbf{D}_{ru} - \mathbf{\Psi}_{ru}$ [32].

Based on this, the model for the signal received at user u can be written as

$$\begin{aligned} y_u &= \sum_{r \in \mathcal{C}_u} \sqrt{s_{ru}} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru} x_u + \sum_{r \in \mathcal{C}_u} \sqrt{s_{ru}} \mathbf{e}_{ru}^H \mathbf{w}_{ru} x_u \\ &+ \sum_{u' \in \mathcal{U}, u' \neq u} \sum_{r' \in \mathcal{C}_{u'}} \sqrt{s_{r'u'}} \hat{\mathbf{h}}_{r'u'}^H \mathbf{w}_{r'u'} x_{u'} \\ &+ \sum_{u' \in \mathcal{U}, u' \neq u} \sum_{r' \in \mathcal{C}_{u'}} \sqrt{s_{r'u'}} \mathbf{e}_{r'u'}^H \mathbf{w}_{r'u'} x_{u'} + z_u. \end{aligned} \quad (5)$$

Including the random estimation error, \mathbf{e}_{ru} , in (5) allows us to implement robust beamforming that exploits the error covariance matrix $\mathbf{\Theta}_{ru}$ and hence compensate for some of the error.

We now define hybrid expressions that account for both the leakage and intra-(DU\CU) interference. For distributed implementation, these expressions use *locally constructed* beamformers.

III. DU-DISTRIBUTED SYSTEM

A. Hybrid Leakage-Interference

Let $\mathcal{U}_{-r} = \mathcal{U} \setminus \mathcal{E}_r$, represent the set of users that do not belong to \mathcal{E}_r , i.e., $\mathcal{U}_{-r} = \{u \mid u \notin \mathcal{E}_r\}$. We define an expression for the average power (leakage) experienced by the users \mathcal{U}_{-r} from DU r by serving user u . Averaged over the random channel estimation error this power is given by

$$\begin{aligned} L_{ru}(\mathbf{w}_{ru}) &= \mathbb{E}_{\mathbf{e}} \left\{ \sum_{u' \in \mathcal{U}_{-r}} t_{r,u'} \left| \hat{\mathbf{h}}_{r'u'}^H \mathbf{w}_{ru} \right|^2 + \sum_{u' \in \mathcal{U}_{-r}} t_{r,u'} \left| \mathbf{e}_{r'u'}^H \mathbf{w}_{ru} \right|^2 \right\} \\ &= \sum_{u' \in \mathcal{U}_{-r}} t_{r,u'} \left| \hat{\mathbf{h}}_{r'u'}^H \mathbf{w}_{ru} \right|^2 + \sum_{u' \in \mathcal{U}_{-r}} t_{r,u'} \mathbf{w}_{ru}^H \mathbf{\Theta}_{r'u'} \mathbf{w}_{ru}. \end{aligned} \quad (6)$$

The vector $\hat{\mathbf{h}}_{r'u'}$ for $u' \in \mathcal{U}_{-r}$ is the estimated *leakage* channel between DU r and user $u' \in \mathcal{U}_{-r}$. The term $t_{r,u'} \in \mathbb{B}$ is defined at DU r and user $u' \in \mathcal{U}_{-r}$. The term $t_{r,u'}$ represents the *assumption* about user $u' \in \mathcal{U}_{-r}$ being scheduled by at least one of its serving DUs $r' \in \mathcal{C}_{u'}$.

We also define our optimization variables as the matrix $\mathbf{W}_r = [\{\mathbf{w}_{ru} : u \in \mathcal{E}_r\}] \in \mathbb{C}^{M \times |\mathcal{E}_r|}$, and $\mathbf{s}_r = [\{s_{ru} : u \in \mathcal{E}_r\}]^T \in \mathbb{B}^{|\mathcal{E}_r| \times 1}$ the beamformers and scheduling variables used by DU r . For each pair of DU r and user u , we now define a hybrid signal-to-leakage-and-intra-DU-interference-and-noise ratio (SLINR-D) as

$$\xi_{ru} \triangleq \frac{s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru}}{A_{ru}(\mathbf{s}_r, \mathbf{W}_r)}, \quad (7)$$

where $A_{ru}(\mathbf{s}_r, \mathbf{W}_r)$, used to simplify the notation, is the

leakage plus intra-DU interference and noise averaged over the unknown random channel estimation error, and it is defined in (8).

In (7) we treat channel estimation error \mathbf{e}_{ru} as additional noise with covariance Θ_{ru} as defined earlier. The first term in (8) is the power leakage to the users not served by DU r , the second is the self-interference resulting from imperfect CSI of the serving channel, the third and fourth terms are the intra-DU interference experienced by user u , and the final term is the noise power.

Remark 2. It is worth commenting on the role of self-interference in (8). Previous works based on leakage, do not mix in interference. However, from (7) and (8), using purely leakage means that the beam power ($\|\mathbf{w}_{ru}\|^2$) would effectively cancel out in all but the noise term. This effectively takes away the role of power allocations, to the detriment of performance [13].

Remark 3. Even for a user served by a *cluster of DUs*, we can still define an SLINR-D for each DU-user pair. Also, (7) is still indirectly coupled across the scheduling decisions of the other DUs through the terms $\{t_{r,u'} : u' \in \mathcal{U}_{-r}\}$. In centralized resource allocation, the scheduling of users is performed by a single entity, hence we can simply set $t_{r,u} = \min(1, \sum_{r' \in \mathcal{C}_u} s_{r'u})$. However, in a DU-distributed system, a DU cannot know which users are scheduled by the other DUs, complicating the development of a distributed scheduling algorithm. In Section V, we address this issue by introducing different methods to calculate the leakage.

Based on (7), we define a weighted *pseudo-rate* between each DU r and its user $u \in \mathcal{E}_r$ as

$$\text{WPS}_{ru} \triangleq \delta_u \log(1 + \xi_{ru}). \quad (9)$$

As the name suggests, the weighted pseudo-rate employs the SLINR-D metric in a rate-like function while providing for fairness. We will optimize this pseudo-rate. Here, δ_u is a term that accounts for the proportional fairness of user u . At time slot $i+1$, these weights are the inverse of the long-term exponentially averaged achieved data rate, i.e., $\delta_u^{(i+1)} = \frac{1}{\bar{R}_u^{(i)}}$ [33], where $\bar{R}_u^{(i)} = \eta R_u^{(i)} + (1-\eta)\bar{R}_u^{(i-1)}$ is the user's exponentially weighted rate averaged over the previous time slots. Here, $0 < \eta < 1$ acts as a forgetting factor, and $R_u^{(i)}$ is the data rate at time slot i .

The motivation to optimize the weighted pseudo-rate is, in fact, two-fold. First, by only using local variables, it allows for the development of a DU-distributed system. Second, it acknowledges that maximizing the useful signal while minimizing the leakage and intra-DU interference is a prudent

practice. Although the SLINR lacks an operational meaning (from an engineering viewpoint) still it can be interpreted as a proxy to enhance the performance [12]. Specifically, when a DU r minimizes the leaked interference to users \mathcal{U}_{-r} , it minimizes the interference terms experienced by these users, and hence it enhances the SINR for these users. The same applies for the intra-DU interference experienced by users \mathcal{E}_r . The mix between the leakage and the intra-DU interference instead of using the leakage only prevents the power of the beam of the DU from scaling equally in both the signal and leakage terms when it is being optimized. In our numerical results, we compare our approach to an SINR-based approach which, unfortunately, requires centralized deployment, and we show that our approach is effective.

B. Problem Definition

For each DU r , we define the following optimization problem:

$$(P1)(r) \quad \max_{\mathbf{w}_r, \mathbf{s}_r} \quad \sum_{u \in \mathcal{E}_r} \delta_u \log(1 + \xi_{ru}) \quad (10a)$$

$$\text{s.t.} \quad \sum_{u \in \mathcal{E}_r} s_{ru} \leq M \quad (10b)$$

$$\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}\|_2^2 \leq p \quad (10c)$$

$$\xi_{ru} = \frac{s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru}}{A_{ru}(\mathbf{s}_r, \mathbf{W}_r)}, \quad u \in \mathcal{E}_r \quad (10d)$$

$$s_{ru} \in \{0, 1\}, \quad u \in \mathcal{E}_r. \quad (10e)$$

The term $A_{ru}(\mathbf{s}_r, \mathbf{W}_r)$ is defined in (8). We note that, in (10d), $\{t_{r,u'} : u' \in \mathcal{U}_{-r}\}$ are not decision variables, but rather they represent the *assumptions* of DU r about the scheduled users served by other DUs. When this knowledge is not available by any means, $\{t_{r,u'} : u' \in \mathcal{U}_{-r}\}$ are simply set to 1. In Section V, we evaluate alternative methods to handle this important issue.

The aim in (10) is to optimize, at DU r , the decision variables \mathbf{s}_r and \mathbf{W}_r , respectively the user scheduling and beamforming vectors. The constraint in (10b) restricts the number of users served by DU r to the number of antennas M , (10c) specifies the power budget of DU r , (10d) treats the SLINR-D expression as an auxiliary variable, and (10e) shows that a user may be either scheduled or not. It is well established that problems having the form of (10) are NP-hard [34], thus obtaining the global optimum is computationally prohibitive, and only a local optimum can be obtained. The problem is mixed-integer and non-convex due to the binary variables and

$$\begin{aligned} A_{ru}(\mathbf{s}_r, \mathbf{W}_r) &= L_{ru}(\mathbf{w}_{ru}) + \mathbb{E}_{\mathbf{e}} \left\{ s_{ru} |\mathbf{e}_{ru}^H \mathbf{w}_{ru}|^2 + \sum_{u' \in \mathcal{E}_r, u' \neq u} s_{ru'} \left| \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru'} \right|^2 + \sum_{u' \in \mathcal{E}_r, u' \neq u} s_{ru'} |\mathbf{e}_{ru}^H \mathbf{w}_{ru'}|^2 + \sigma_z^2 \right\} \\ &= L_{ru}(\mathbf{w}_{ru}) + s_{ru} \mathbf{w}_{ru}^H \Theta_{ru} \mathbf{w}_{ru} + \sum_{u' \in \mathcal{E}_r, u' \neq u} s_{ru'} \mathbf{w}_{ru'}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru'} + \sum_{u' \in \mathcal{E}_r, u' \neq u} s_{ru'} \mathbf{w}_{ru'}^H \Theta_{ru} \mathbf{w}_{ru'} + \sigma_z^2. \end{aligned} \quad (8)$$

their presence in both the numerator and denominator of the utility function.

Using fractional programming, the problem in (10) can be reformulated and written as

$$(P2)(r) \quad \max_{\mathbf{w}_r, \mathbf{s}_r, \boldsymbol{\xi}_r, \boldsymbol{\zeta}_r} f_2(\mathbf{s}_r, \mathbf{W}_r, \boldsymbol{\xi}_r, \boldsymbol{\zeta}_r) \quad (11a)$$

$$\text{s.t.} \quad \sum_{u \in \mathcal{E}_r} s_{ru} \leq M \quad (11b)$$

$$\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}\|_2^2 \leq p \quad (11c)$$

$$s_{ru} \in \{0, 1\}, \quad u \in \mathcal{E}_r. \quad (11d)$$

with the function $f_2(\mathbf{s}_r, \mathbf{W}_r, \boldsymbol{\xi}_r, \boldsymbol{\zeta}_r)$ found in (11a) is defined as

$$\begin{aligned} f_2(\mathbf{s}_r, \mathbf{W}_r, \boldsymbol{\xi}_r, \boldsymbol{\zeta}_r) &= \sum_{u \in \mathcal{E}_r} \delta_u (\log(1 + \xi_{ru}) - \zeta_{ru}) \\ &+ \sum_{u \in \mathcal{E}_r} \left(2\text{Re} \left\{ \zeta_{ru}^* \sqrt{\delta_u (1 + \xi_{ru})} s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \right\} \right. \\ &\left. - |\zeta_{ru}|^2 \left(s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru} + A_{ru}(\mathbf{s}_r, \mathbf{W}_r) \right) \right). \quad (12) \end{aligned}$$

where $\boldsymbol{\zeta}_r \in \mathbb{C}^{|\mathcal{E}_r| \times 1}$ is a new auxiliary variable vector introduced by fractional programming [10]. Fractional programming refers to optimizing functions composed of ratios. The function can be composed of a single ratio, or, as in our case, a sum of ratios. The numerator and denominator can be nonlinear. Techniques used to obtain local optimum for such optimization problems include the Charnes-Cooper method [35], Dinkelbach's method [36], and quadratic transform [10].

The reformulation in (11) is based on the Lagrangian formulation and fractional programming. Please refer to Appendix A for details.

In what follows, we obtain optimal expressions for the optimization variables, one set of variables at a time, i.e., we use block coordinate descent. As for the scheduling variables \mathbf{s}_r , we optimize them using a combinatorial search as described below.

When the variables other than ζ_{ru} are fixed, the first optimality condition of (12) with respect to ζ_{ru} results in the optimal value as

$$\zeta_{ru} = \frac{s_{ru} \sqrt{\delta_u (1 + \xi_{ru})} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru}}{s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru} + A_{ru}(\mathbf{s}_r, \mathbf{W}_r)}. \quad (13)$$

We write the Lagrangian formulation of (11), and when variables $(\mathbf{s}_r, \boldsymbol{\xi}_r, \boldsymbol{\zeta}_r)$ are fixed, we derive the optimality condition to obtain the optimal expression for the transmit precoder \mathbf{w}_{ru} as shown in (14).

The variable μ_r denotes the Lagrange multiplier for the power budget constraint (11c), and is inversely related to the beamformers' power at DU r ; μ_r can be determined through the complementary slackness condition for the power budget. Specifically, let $\mathbf{w}_{ru}(\mu_r)$ denote the right-hand side of (14). If $\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}(0)\|_2^2 \leq p$, then $\mu_r = 0$, otherwise the value of μ_r can be determined through a bisection search to satisfy $\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}(\mu_r)\|_2^2 = p$.

Algorithm 1: Distributed resource allocation at each DU r

- 1 Initialize \mathbf{W}_r for all users.
 - 2 Initialize \mathbf{s}_r by selecting M users with max (9).
 - 3 **while** *NOT* converged **do**
 - 4 Update $\{\xi_{ru} : u \in \mathcal{E}_r\}$ using (10d).
 - 5 Update $\{\zeta_{ru} : u \in \mathcal{E}_r\}$ using (13).
 - 6 Update $\{\mathbf{w}_{ru} : u \in \mathcal{E}_r\}$ using (14) and μ_r through the complementary slackness condition of power budget.
 - 7 Update \mathbf{V}_r by solving (15) using Hungarian algorithm, then $\{s_{ru} = \sum_{m=1}^M v_{ru,m} : u \in \mathcal{E}_r\}$
 - 8 **end**
-

When the variables $(\mathbf{W}_r, \boldsymbol{\xi}_r, \boldsymbol{\zeta}_r)$ are fixed, the problem of optimizing \mathbf{s}_r at each DU r is a combinatorial problem, where the users are matched to the available non-zero beams $\{\tilde{\mathbf{w}}_{rm} : m \in \{1, \dots, M\}\}$ found at the DU. Thus, the problem of optimizing the scheduling is cast as a combinatorial problem as follows.

Constraint (15b) states that a user can be assigned at most one non-zero beam, and constraint (15c) states that all non-zero beams are assigned. In (15a), $\tilde{A}_{ru,m}$ is the term that contains the leakage, interference and noise terms resulting from assigning the non-zero beam m on DU r to user u , and is defined as (16).

It is important to note that, $\tilde{A}_{ru,m}$ is not affected if the non-zero beams other than m are assigned to different users, because *in the scheduling phase* the beamformers are fixed.

This problem in (15) is an agent-task assignment problem [37], and can be written as finding an $(|\mathcal{E}_r| \times M)$ column-permutating matrix \mathbf{V}_r that maximizes $([1 \dots 1] \mathbf{A}_r \mathbf{V}_r^T [1 \dots 1]^T)$ for a given matrix \mathbf{A}_r . A column-permutating matrix is binary with a single 1 in each column in a unique location. Here, $[\mathbf{V}_r]_{um} = v_{ru,m}$ and the entries of $\mathbf{A}_r \in \mathbb{R}^{|\mathcal{E}_r| \times M}$ are the pseudo-rates in (15a) representing the utility of the users served by DU r . The matrix \mathbf{V}_r , hence, serves to assign each non-zero beam (column of \mathbf{V}_r) to a specific user (row of \mathbf{V}_r).

This assignment problem can be solved efficiently in polynomial time using the Hungarian algorithm [38] (also known as the Kuhn–Munkres algorithm). This algorithm executes a series of iterative manipulations for the rows and columns of the matrix \mathbf{A}_r (or alternatively a cost matrix), which allows us to find the maximum (or minimum for cost matrix) entries for the assignment of the rows of \mathbf{A}_r to the columns (agent-task assignment). Implementations for the Hungarian algorithm can be readily found in many scripting languages.

Finally, user u is scheduled by DU r if it is assigned a non-zero beam, i.e., $s_{ru} = \sum_{m=1}^M v_{ru,m}$. After defining the optimal expression for each variable type when the other variables are fixed, we can construct Algorithm 1 that uses block coordinate descent to optimize the resource allocation in a distributed fashion on each DU. The initialization of the beamforming in Step 1 can be performed using conjugate beamforming, while Steps 3-7 optimize one variable at a time until convergence.

We will analyze convergence in our section on numerical results.

IV. CU-DISTRIBUTED SYSTEM

The previous section designed an algorithm that allows each DU to perform its own allocation decisions. However, if a CU can coordinate the DUs under its control, we can include the intra-CU interference, i.e., inside the virtual cell of each CU. A CU-distributed algorithm would, therefore, provide a balance between some coordination (and information exchange) and the completely centralized case where all CUs are jointly optimized. Additionally, the CU would decide on user scheduling for all the DUs under its control. We emphasize, however, that a user may be associated with DUs under the control of different CUs (user-centric clustering). A user may, therefore, be scheduled by one CU, but not by the other. For each CU q , we define the set of users $\bar{\mathcal{U}}_q \triangleq \bigcup_{r' \in \mathcal{B}_q} \mathcal{E}_{r'}$ that includes all the users connected to at least one DU under the control of CU q . Due to user-centric clustering, the different sets $\{\bar{\mathcal{U}}_q\}_{q=1}^Q$ overlap.

For each user $u \in \bar{\mathcal{U}}_q$ and DUs $\mathcal{D}_{qu} = (\mathcal{C}_u \cap \mathcal{B}_q)$, we define the concatenation of beamformers, estimated channels, channel estimation error and scheduling variables as

$$\bar{\mathbf{w}}_{qu} = [\{\mathbf{w}_{ru}^T : r \in \mathcal{D}_{qu}\}]^T \in \mathbb{C}^{M|\mathcal{D}_{qu}| \times 1} \quad (17)$$

$$\bar{\mathbf{h}}_{qu,u'} = [\{\hat{\mathbf{h}}_{ru'}^T : r \in \mathcal{D}_{qu}\}]^T \in \mathbb{C}^{M|\mathcal{D}_{qu}| \times 1} \quad (18)$$

$$\bar{\mathbf{e}}_{qu,u'} = [\{\mathbf{e}_{ru'}^T : r \in \mathcal{D}_{qu}\}]^T \in \mathbb{C}^{M|\mathcal{D}_{qu}| \times 1} \quad (19)$$

$$\mathbf{S}_{qu} = (\text{diag}(\{s_{ru} : r \in \mathcal{D}_{qu}\}) \otimes \mathbf{I}_M) \in \mathbb{B}^{M|\mathcal{D}_{qu}| \times M|\mathcal{D}_{qu}|} \quad (20)$$

We note that (18) and (19) represent the estimated channel and estimation error, respectively, between user u' and the serving DUs \mathcal{D}_{qu} for user u . As in the previous section, we define an expression for the power leakage experienced by the users in $\bar{\mathcal{U}}_{-q} = \mathcal{U} \setminus \bar{\mathcal{U}}_q$ from all the DUs due to serving user u . Averaged over the random channel estimation error this leakage is given by (21), where the term $\mathbb{E}\{\bar{\mathbf{e}}_{qu,u'} \bar{\mathbf{e}}_{qu,u'}^H\} = \bar{\boldsymbol{\Theta}}_{qu,u'}$ in (21) represents the covariance of the channels' estimation error obtained from treating the unknown random Gaussian error as additional noise.

As before, we define the hybrid signal-to-leakage-and-intra-CU-interference-and-noise-ratio (SLINR-C) between each CU q and user $u \in \bar{\mathcal{U}}_q$ as

$$\bar{\xi}_{qu} \triangleq \frac{\bar{\mathbf{w}}_{qu}^H \mathbf{S}_{qu}^{1/2} \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \mathbf{S}_{qu}^{1/2} \bar{\mathbf{w}}_{qu}}{B_{qu}(\mathcal{S}_q, \mathcal{W}_q)}, \quad (22)$$

where the leakage plus intra-CU interference and noise, averaged over the random estimation error, is defined as (23).

Here in (23), \mathbf{S}_{qu} is defined in (20) with diagonal entries $s_{ru} \in \mathbb{B}$ denoting the scheduling variable of user u at DU r , and $\bar{t}_{q,u} \in \mathbb{B}$ is defined at CU q to represent the *assumption* for user u being scheduled by at least one of its serving DUs $r' \notin \mathcal{B}_q$ not under the control of CU q .

Similar to the previous section, for each CU q we can optimize the pseudo-rates as (24). The set $\mathcal{W}_q = \{\mathbf{W}_r : r \in \mathcal{B}_q\}$ in (24) represents the matrix of beamformers used by the DUs \mathcal{B}_q , where, as noted earlier, $\mathbf{W}_r = [\{\mathbf{w}_{ru} : u \in \mathcal{E}_r\}] \in \mathbb{C}^{M \times |\mathcal{E}_r|}$ are the beamformers used by each DU r to serve its users. Similarly, the scheduling variables $\mathcal{S}_q = \{\mathbf{S}_{qu} : u \in$

$$\begin{aligned} \mathbf{w}_{ru} &= s_{ru} \zeta_{ru}^* \sqrt{\delta_u (1 + \xi_{ru})} \left(|\zeta_{ru}|^2 (s_{ru} (\hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H + \boldsymbol{\Theta}_{ru}) + \sum_{u' \in \bar{\mathcal{U}}_{-r}} t_{r,u'} \hat{\mathbf{h}}_{ru'} \hat{\mathbf{h}}_{ru'}^H + \sum_{u' \in \bar{\mathcal{U}}_{-r}} t_{r,u'} \boldsymbol{\Theta}_{ru'}) \right. \\ &\quad \left. + s_{ru} \sum_{u' \in \mathcal{E}_r, u' \neq u} |\zeta_{ru'}|^2 (\hat{\mathbf{h}}_{ru'} \hat{\mathbf{h}}_{ru'}^H + \boldsymbol{\Theta}_{ru'}) + \mu_r \mathbf{I}_M \right)^{-1} \hat{\mathbf{h}}_{ru} \\ &= s_{ru} \zeta_{ru}^* \sqrt{\delta_u (1 + \xi_{ru})} \left(|\zeta_{ru}|^2 \left(\sum_{u' \in \bar{\mathcal{U}}_{-r}} t_{r,u'} \hat{\mathbf{h}}_{ru'} \hat{\mathbf{h}}_{ru'}^H + \sum_{u' \in \bar{\mathcal{U}}_{-r}} t_{r,u'} \boldsymbol{\Theta}_{ru'} \right) + s_{ru} \sum_{u' \in \mathcal{E}_r} |\zeta_{ru'}|^2 (\hat{\mathbf{h}}_{ru'} \hat{\mathbf{h}}_{ru'}^H + \boldsymbol{\Theta}_{ru'}) + \mu_r \mathbf{I}_M \right)^{-1} \hat{\mathbf{h}}_{ru}, \quad (14) \end{aligned}$$

$$(P3)(r) \quad \max_{\{v_{ru,m} : u \in \mathcal{E}_r, 1 \leq m \leq M\}} \sum_{m=1}^M v_{ru,m} \sum_{u \in \mathcal{E}_r} \delta_u \log \left(1 + \frac{\tilde{\mathbf{w}}_{rm}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \tilde{\mathbf{w}}_{rm}}{\tilde{A}_{ru,m}} \right) \quad (15a)$$

$$\text{s.t.} \quad \sum_{m=1}^M v_{ru,m} \leq 1, \quad u \in \mathcal{E}_r \quad (15b)$$

$$\sum_{u \in \mathcal{E}_r} v_{ru,m} = 1, \quad m = 1, \dots, M \quad (15c)$$

$$v_{ru,m} \in \{0, 1\}, \quad u \in \mathcal{E}_r, m = 1, \dots, M \quad (15d)$$

$$\tilde{A}_{ru,m} = \sum_{u' \in \bar{\mathcal{U}}, u' \notin \mathcal{E}_r} t_{r,u'} \tilde{\mathbf{w}}_{rm}^H (\hat{\mathbf{h}}_{ru'} \hat{\mathbf{h}}_{ru'}^H + \boldsymbol{\Theta}_{ru'}) \tilde{\mathbf{w}}_{rm} + \tilde{\mathbf{w}}_{rm}^H \boldsymbol{\Theta}_{ru} \tilde{\mathbf{w}}_{rm} + \sum_{m'=1, m' \neq m}^M \tilde{\mathbf{w}}_{rm'}^H (\hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H + \boldsymbol{\Theta}_{ru}) \tilde{\mathbf{w}}_{rm'} + \sigma_z^2. \quad (16)$$

$\bar{\mathcal{U}}_q\}$, where \mathbf{S}_{qu} contains the scheduling variables s_{ru} by each DU $r \in \mathcal{D}_{qu}$. \mathcal{W}_q and \mathcal{S}_q are the optimization variables.

Because of the coupling across DUs under the control of CU q , we require a new approach to the scheduling problem. The scheduling variables can be related to the indicator function of the beamformer which, in turn, can be written as an ℓ_0 -norm, i.e., $s_{ru} = \mathbb{1}\{\|\mathbf{w}_{ru}\|_2^2\} = \|\|\mathbf{w}_{ru}\|_2^2\|_0$. From compressive sensing [39], we can approximate the ℓ_0 -norm as a weighted ℓ_1 -norm that is easier to work with as $\|\mathbf{x}\|_0 \simeq \sum_m \alpha_m |x_m| = \|\boldsymbol{\alpha}^T \mathbf{x}\|_1$, [40]. The variables α_m are weights that can be updated iteratively. For our problem, we can define the weights α_{ru} as

$$\alpha_{ru} = \frac{1}{\|\mathbf{w}_{ru}\|_2^2 + \epsilon}, \quad (25)$$

where $\epsilon > 0$ provides stability and ensures that a zero-valued component in $\|\mathbf{w}_{ru}\|_2^2$ does not strictly prohibit a nonzero estimate in the next iteration; importantly, the results are not very sensitive to ϵ and it can be chosen to be slightly smaller than the expected value of $\|\mathbf{w}_{ru}\|_2^2$ for the scheduled users [40].

Based on this, we can define the following optimization problem at each CU q

$$(P5)(q) \quad \max_{\mathcal{W}_q} \sum_{u \in \bar{\mathcal{U}}_q} \delta_u \log(1 + \bar{\xi}_{qu}) \quad (26a)$$

$$\text{s.t.} \quad \sum_{u \in \mathcal{E}_r} \alpha_{ru} \|\mathbf{w}_{ru}\|_2^2 \leq M, \quad r \in \mathcal{B}_q \quad (26b)$$

$$\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}\|_2^2 \leq p, \quad r \in \mathcal{B}_q \quad (26c)$$

$$\bar{\xi}_{qu} = \frac{\bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu}}{C_{qu}(\mathcal{W}_q)}, \quad u \in \bar{\mathcal{U}}_q \quad (26d)$$

where the auxiliary variable (26d) does not contain the scheduling variables in the set \mathcal{S}_q because the expression includes the beamformers \mathcal{W}_q which will be zero for the users not scheduled, thus, the expression of C_{qu} is the same as B_{qu} without including \mathcal{S}_q .

Using fractional programming, our optimization problem in (26) can be reformulated as

$$(P6)(q) \quad \max_{\mathcal{W}_q, \bar{\xi}_q, \bar{\zeta}_q} f_2(\mathcal{W}_q, \bar{\xi}_q, \bar{\zeta}_q) \quad (27a)$$

$$\text{s.t.} \quad \sum_{u \in \mathcal{E}_r} \alpha_{ru} \|\mathbf{w}_{ru}\|_2^2 \leq M, \quad r \in \mathcal{B}_q \quad (27b)$$

$$\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}\|_2^2 \leq p, \quad r \in \mathcal{B}_q. \quad (27c)$$

with the function $f_2(\mathcal{W}_q, \bar{\xi}_q, \bar{\zeta}_q)$ in (27a) defined as

$$\begin{aligned} f_2(\mathcal{W}_q, \bar{\xi}_q, \bar{\zeta}_q) &= \sum_{u \in \bar{\mathcal{U}}_q} \delta_u (\log(1 + \bar{\xi}_{qu}) - \bar{\xi}_{qu}) \\ &+ \sum_{u \in \bar{\mathcal{U}}_q} \left(2\text{Re} \left\{ \bar{\zeta}_{qu}^* \sqrt{\delta_u (1 + \bar{\xi}_{qu})} \bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \right\} \right. \\ &\left. - |\bar{\zeta}_{qu}|^2 (\bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu} + C_{qu}(\mathcal{W}_q)) \right), \quad (28) \end{aligned}$$

where $\bar{\zeta}_q = [\bar{\zeta}_{qu_1} \dots \bar{\zeta}_{qu_{|\bar{\mathcal{U}}_q|}}]^T$ are required auxiliary variables. In Appendix B we provide the details of the reformulation leading to (27).

Using the first optimality condition of (28) with respect to $\bar{\zeta}_{qu}$. The optimal value of $\bar{\zeta}_{qu}$ is

$$\bar{\zeta}_{qu} = \frac{\sqrt{\delta_u (1 + \bar{\xi}_{qu})} \bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u}}{\bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu} + C_{qu}(\mathcal{W}_q)}. \quad (29)$$

Now, using the constraints (27b) and (27c), we define the

$$\begin{aligned} \bar{L}_{qu}(\bar{\mathbf{w}}_{qu}) &= \mathbb{E}_{\mathbf{e}} \left\{ \sum_{u' \in \bar{\mathcal{U}}_{-q}} \bar{t}_{q,u'} |\bar{\mathbf{h}}_{qu,u'}^H \bar{\mathbf{w}}_{qu}|^2 + \sum_{u' \in \bar{\mathcal{U}}_{-q}} \bar{t}_{q,u'} |\bar{\mathbf{e}}_{qu,u'}^H \bar{\mathbf{w}}_{qu}|^2 \right\} \\ &= \sum_{u' \in \bar{\mathcal{U}}_{-q}} \bar{t}_{q,u'} |\bar{\mathbf{h}}_{qu,u'}^H \bar{\mathbf{w}}_{qu}|^2 + \sum_{u' \in \bar{\mathcal{U}}_{-q}} \bar{t}_{q,u'} \bar{\mathbf{w}}_{qu}^H \bar{\Theta}_{qu,u'} \bar{\mathbf{w}}_{qu}, \quad (21) \end{aligned}$$

$$\begin{aligned} B_{qu}(\mathcal{S}_q, \mathcal{W}_q) &= \bar{L}_{qu}(\bar{\mathbf{w}}_{qu}) + \mathbb{E}_{\mathbf{e}} \left\{ \left| \bar{\mathbf{e}}_{qu,u}^H \mathbf{S}_{qu}^{1/2} \bar{\mathbf{w}}_{qu} \right|^2 + \sum_{u' \in \bar{\mathcal{U}}_q, u' \neq u} \bar{\mathbf{w}}_{qu'}^H \mathbf{S}_{qu'}^{1/2} \bar{\mathbf{h}}_{qu',u} \bar{\mathbf{h}}_{qu',u}^H \mathbf{S}_{qu'}^{1/2} \bar{\mathbf{w}}_{qu'} \right. \\ &+ \left. \sum_{u' \in \bar{\mathcal{U}}_q, u' \neq u} \bar{\mathbf{w}}_{qu'}^H \mathbf{S}_{qu'}^{1/2} \bar{\mathbf{e}}_{qu',u} \bar{\mathbf{e}}_{qu',u}^H \mathbf{S}_{qu'}^{1/2} \bar{\mathbf{w}}_{qu'} + \sigma_z^2 \right\}, \\ &= \bar{L}_{qu}(\bar{\mathbf{w}}_{qu}) + \bar{\mathbf{w}}_{qu}^H \mathbf{S}_{qu}^{1/2} \bar{\Theta}_{qu,u} \mathbf{S}_{qu}^{1/2} \bar{\mathbf{w}}_{qu} + \sum_{u' \in \bar{\mathcal{U}}_q, u' \neq u} \bar{\mathbf{w}}_{qu'}^H \mathbf{S}_{qu'}^{1/2} \bar{\mathbf{h}}_{qu',u} \bar{\mathbf{h}}_{qu',u}^H \mathbf{S}_{qu'}^{1/2} \bar{\mathbf{w}}_{qu'} \\ &+ \sum_{u' \in \bar{\mathcal{U}}_q, u' \neq u} \bar{\mathbf{w}}_{qu'}^H \mathbf{S}_{qu'}^{1/2} \bar{\Theta}_{qu',u} \mathbf{S}_{qu'}^{1/2} \bar{\mathbf{w}}_{qu'} + \sigma_z^2 \quad (23) \end{aligned}$$

following Lagrangian function

$$\begin{aligned}
f_3(\mathcal{W}_q, \boldsymbol{\mu}_q, \boldsymbol{\lambda}_q) &= - \sum_{r \in \mathcal{B}_q} \lambda_r \left(\sum_{u \in \mathcal{E}_r} \alpha_{ru} \|\mathbf{w}_{ru}\|_2^2 - M \right) \\
&\quad - \sum_{r \in \mathcal{B}_q} \mu_r \left(\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}\|_2^2 - p \right) \\
&\stackrel{(a)}{=} - \sum_{u \in \mathcal{U}_q} \sum_{r \in \mathcal{D}_{qu}} \lambda_r \alpha_{ru} \|\mathbf{w}_{ru}\|_2^2 + M \sum_{r \in \mathcal{B}_q} \lambda_r \\
&\quad - \sum_{u \in \mathcal{U}_q} \sum_{r \in \mathcal{D}_{qu}} \mu_r \|\mathbf{w}_{ru}\|_2^2 + p \sum_{r \in \mathcal{B}_q} \mu_r \\
&= - \sum_{u \in \mathcal{U}_q} \sum_{r \in \mathcal{D}_{qu}} \omega_{ru} \|\mathbf{w}_{ru}\|_2^2 + \sum_{r \in \mathcal{B}_q} (p\mu_r + M\lambda_r), \quad (30)
\end{aligned}$$

where $\omega_{ru} = \mu_r + \lambda_r \alpha_{ru}$, and (a) follows from $\sum_{r \in \mathcal{B}_q} \sum_{u \in \mathcal{E}_r} (\cdot) = \sum_{u \in \mathcal{U}_q} \sum_{r \in \mathcal{D}_{qu}} (\cdot)$.

When the variables other than \mathcal{W}_q are fixed, using the Lagrangian formulation of (27), we can write the corresponding expression for the optimal beamformer \mathbf{w}_{ru} as (31), where $\boldsymbol{\Omega}_{qu} = (\text{diag}(\{\omega_{ru} : r \in \mathcal{C}_u\}) \otimes \mathbf{I}_M)$. Note that the beamformer \mathbf{w}_{ru} is obtained from $\bar{\mathbf{w}}_{qu}$ using (17).

The Lagrangian multipliers $\mu_r \geq 0$ and $\lambda_r \geq 0$ found in $\boldsymbol{\Omega}_{qu}$ and (30) can be determined using their power budget in (27c) and capacity (27b) constraints, respectively. Importantly, both constraints are related to the power used at DU r , where constraint (27b) can be seen as a weighted power constraint. Hence, both cannot be tight simultaneously; from complementary slackness, one of these Lagrangian multipliers needs to be zero. Unfortunately, we do not know, a priori, which constraint will remain tight. Therefore, we propose a heuristic that, at each iteration, the algorithm checks for whether the capacity constraint is satisfied (allowing $\lambda_r = 0$); if it is not satisfied, we update λ_r to a small value. As for μ_r , we update it using the bisection search method as was discussed in the previous section for a DU-distributed system.

It is worth noting that the off-diagonal entries of the terms of the inverse term in (31) are very small compared to the diagonal entries; this directly follows from summing up a large number of multiplied independent zero-mean random elements. Hence, in practice, changing μ_r has little effect on the other beamformers $\{\mathbf{w}_{r'u} : r' \in \mathcal{D}_{qu} \setminus r\}$ in (31) (see (17)).

Finally, we implement Algorithm 2 on each CU to perform the resource allocation and the user scheduling. The discussion

Algorithm 2: Distributed resource allocation at each CU q

- 1 Initialize \mathcal{W} and α_{ru} for all users.
 - 2 **while** *NOT* converged **do**
 - 3 Update $\{\bar{\xi}_{qu} : u \in \mathcal{U}_q\}$ using (26d).
 - 4 Update $\{\bar{\zeta}_{qu} : u \in \mathcal{U}_q\}$ using (29).
 - 5 Update \mathcal{W} using (31).
 - 6 Update $\{\lambda_r, \mu_r : r \in \mathcal{B}_q\}$ as described using complementary slackness.
 - 7 Update weights α using (25).
 - 8 **end**
-

of Algorithm 2 is similar to that in the DU-distributed system and is omitted for brevity.

V. SCALABILITY OF CALCULATING THE LEAKAGE

The main issue in calculating the leakage expressions in (6) or (21), is that each DU needs to estimate the channels to all the users in the network. This is, likely, impractical and not scalable. Another issue is that in a distributed algorithm, a DU in the DU-distributed system (or CU in the CU-distributed system) cannot know which users will be scheduled by other DUs (or CUs), i.e., does not know the value of the terms $s_{r,u'}$ (or $\bar{s}_{q,u'}$) to accurately choose $t_{r,u'}$ (or $\bar{t}_{q,u'}$). As such, for a practical implementation, alternative methods to calculate the leakage are needed.

In this section, we exploit the fact that *leakage is a convenience* we use to estimate the impact a DU (CU) has on other DUs (CUs). We propose here three methods to calculate this impact:

- Standard CSI estimation: assume all the users \mathcal{U}_{-r} in DU-distributed system (or $\bar{\mathcal{U}}_{-q}$ in CU-distributed system) are scheduled and estimate the CSI to all of them. This means that for each DU r , we set $\{t_{r,u} = 1 : u \in \mathcal{U}_{-r}\}$; similarly for $\bar{t}_{q,u}$ for the CU-distributed system. This technique is already used above in the development of our algorithms.
- Statistical CSI: similar to the first method, except using only the *large-scale fading* statistics to calculate the leakage, eliminating the need for continual training of the leakage channels.

$$(P4)(q) \quad \max_{\mathcal{W}_q, \mathcal{S}_q} \sum_{u \in \mathcal{U}_q} \delta_u \log(1 + \bar{\xi}_{qu}) \quad (24a)$$

$$\text{s.t.} \quad \sum_{u \in \mathcal{E}_r} s_{ru} \leq M, \quad r \in \mathcal{B}_q \quad (24b)$$

$$\sum_{u \in \mathcal{E}_r} \|\mathbf{w}_{ru}\|_2^2 \leq p, \quad r \in \mathcal{B}_q \quad (24c)$$

$$\bar{\xi}_{qu} = \frac{\bar{\mathbf{w}}_{qu}^H \mathbf{S}_{qu}^{1/2} \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \mathbf{S}_{qu}^{1/2} \bar{\mathbf{w}}_{qu}}{B_{qu}(\mathcal{S}_q, \mathcal{W}_q)}, \quad r \in \mathcal{B}_q, u \in \mathcal{E}_r \quad (24d)$$

$$s_{ru} \in \{0, 1\}, \quad r \in \mathcal{B}_q, u \in \mathcal{E}_r \quad (24e)$$

- Traffic distribution: use traffic distribution to calculate the leakage. Based on the statistical distribution of traffic, this *completely eliminates* the need for real-time leakage information.

Statistical CSI: Each DU r uses the large-scale fading statistics to calculate the leakage to *all* the users $u' \in \mathcal{U}_{-r}$ other than the ones it serves. This means that the leakage term in the DU-distributed system can be alternatively defined as

$$L_{ru}(\mathbf{w}_{ru}) = \sum_{u' \in \mathcal{U}_{-r}} \mathbf{w}_{ru}^H \mathbf{\Lambda}_{ru'} \mathbf{w}_{ru},$$

where $\mathbf{\Lambda}_{ru'} = \psi_{ru'} \beta(d_{ru'}) \mathbf{I}_M$. (32)

Similarly, for the CU-distributed system, we define the leakage in the CU-distributed system as

$$\bar{L}_{qu}(\bar{\mathbf{w}}_{qu}) = \sum_{u' \in \mathcal{U}_{-q}} \bar{\mathbf{w}}_{qu}^H \bar{\mathbf{\Lambda}}_{qu,u'} \bar{\mathbf{w}}_{qu},$$

where $\bar{\mathbf{\Lambda}}_{qu,u'} = (\text{diag}(\{\psi_{ru'} \beta(d_{ru'}) : r \in \mathcal{D}_{qu}\}) \otimes \mathbf{I}_M)$. (33)

Traffic distribution: Instead of calculating the leakage to all the users that may or may not be scheduled, each node in the network can use a spatial traffic distribution for the network region around it. Such an approach can be easily based on a simple traffic survey, resulting in a traffic probability density function (PDF), $\Upsilon_n(\tilde{x}_u, \tilde{y}_u)$ at location $(\tilde{x}_u, \tilde{y}_u)$. we can redefine the leakage terms found in (8) and (23) to use the PDF of the traffic distribution in the region around the node instead of calculating the leakage to the users in these regions. Hence, for the DU-distributed system, the leakage term can be written as

$$L_{ru}(\mathbf{w}_{ru}) = \mathbb{E}_{\tilde{x}_u, \tilde{y}_u} \left\{ \mathbf{w}_{ru}^H \tilde{\mathbf{\Lambda}}_{ru} \mathbf{w}_{ru} \right\}$$

$$= \iint_{\tilde{x}_u, \tilde{y}_u \in \iota_r} \mathbf{w}_{ru}^H \tilde{\mathbf{\Lambda}}_{ru} \mathbf{w}_{ru} \Upsilon(\tilde{x}_u, \tilde{y}_u) d\tilde{x}_u d\tilde{y}_u, \quad (34)$$

where ι_r is the boundary of the region considered to calculate the leakage. Essentially, (34) calculates the average leakage power weighted by the traffic PDF, i.e., it emphasizes the regions with hotspots. The term $\tilde{\mathbf{\Lambda}}_{ru} = \left(\beta(\tilde{d}_{ru} + d_{\text{excl}}) \right) \mathbf{I}_M$, with $\tilde{d}_{ru} = \sqrt{(\tilde{x}_r - \tilde{x}_u)^2 + (\tilde{y}_r - \tilde{y}_u)^2}$ written as a function of the location of the DU $(\tilde{x}_r, \tilde{y}_r)$ and the point $(\tilde{x}_u, \tilde{y}_u)$ used to calculate the leakage, and d_{excl} is an exclusion region around the DU, which is chosen to be larger than the reference distance of the path loss.

	<i>Parameter</i>	<i>Value</i>
Cell con-fig.	Q, N, M , density of PPP λ_{users} when no hotspots exist	7, 10, 8, 200 users/km ²
Power, pilots	p, p_u, τ_d, τ_p	30 dBm, 20 dBm, 200, (32 or 64)
Noise	noise spectral efficiency S_z , noise figure F_z , Bandwidth	-174 dBm/Hz, 8 dBm, 180 KHz
Fading	$\sigma_{\text{shadowing}}, \rho$	4 dB, $\beta(0.4)$

TABLE I: Simulation parameters.

Similarly, in the CU-distributed system, the leakage term can be defined as

$$\bar{L}_{qu}(\bar{\mathbf{w}}_{qu}) = \mathbb{E}_{\tilde{x}_u, \tilde{y}_u} \left\{ \bar{\mathbf{w}}_{qu}^H \check{\mathbf{\Lambda}}_{qu} \bar{\mathbf{w}}_{qu} \right\}$$

$$= \iint_{\tilde{x}_u, \tilde{y}_u \in \bar{\iota}_q} \bar{\mathbf{w}}_{qu}^H \check{\mathbf{\Lambda}}_{qu} \bar{\mathbf{w}}_{qu} \Upsilon(\tilde{x}_u, \tilde{y}_u) d\tilde{x}_u d\tilde{y}_u, \quad (35)$$

where $\check{\mathbf{\Lambda}}_{qu} = \left(\text{diag} \left(\left\{ \beta(\tilde{d}_{ru} + d_{\text{excl}}) : r \in \mathcal{D}_{qu} \right\} \right) \otimes \mathbf{I}_M \right)$ is the concatenated path loss between the serving cluster of user u found in cell q , i.e., \mathcal{D}_{qu} , and locations $(\tilde{x}_u, \tilde{y}_u) \in \bar{\iota}_q$. We emphasize that the traffic distribution *eliminates the need for any inter-node information exchange*.

VI. NUMERICAL RESULTS

In this section, we illustrate the efficacy of the proposed algorithms. We simulate a network of 7 hexagonal *virtual* cells, i.e., 7 CUs, each has a radius of 500 meters. We assume that the N DUs in each virtual cell are uniformly distributed inside the cell boundaries. We use wraparound to eliminate the network border effect. As stated in the system model, the user-centric clustering is applied, thus the concept of cells is only applicable on the front-haul, hence the term “virtual”. We use an exclusion region of radius 20 m around each DU and model the user locations as a Poisson point process (PPP) with local density determined by the traffic model.

In this paper, we model the traffic distribution as a mix of a uniform distribution, chosen with a probability P_h , and a number (N_h) of hotspots. The hotspot traffic is modeled as bivariate normal distributions. Based on this, we define the PDF of the traffic used to calculate the leakage on a node n (DU or CU) as [41]

$$\Upsilon_n(\tilde{x}_u, \tilde{y}_u) \triangleq f_h \left(P_h \left(\frac{1}{a_n} \right) + (1 - P_h) \right)$$

$$\times \frac{1}{N_h 2\pi\sigma_h^2} \sum_{i=1}^{N_h} \left(\exp \left(-\frac{(\tilde{x}_k - \tilde{x}_{h_i})^2 + (\tilde{y}_k - \tilde{y}_{h_i})^2}{2\sigma_h^2} \right) \right) \quad (36)$$

$$\bar{\mathbf{w}}_{qu} = \bar{\zeta}_{qu}^* \sqrt{\delta_u (1 + \bar{\xi}_{qu})} \left(|\bar{\zeta}_{qu}|^2 \left(\bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H + \bar{\Theta}_{qu,u} + \sum_{u' \in \mathcal{U}_{-q}} \bar{t}_{q,u'} \left(\bar{\mathbf{h}}_{qu,u'} \bar{\mathbf{h}}_{qu,u'}^H + \bar{\Theta}_{qu,u'} \right) \right) \right.$$

$$\left. + \sum_{u' \in \bar{\mathcal{U}}_q, u' \neq u} |\bar{\zeta}_{qu'}|^2 \left(\bar{\mathbf{h}}_{qu,u'} \bar{\mathbf{h}}_{qu,u'}^H + \bar{\Theta}_{qu,u'} \right) + \Omega_{qu} \right)^{-1} \bar{\mathbf{h}}_{qu,u}$$

$$= \bar{\zeta}_{qu}^* \sqrt{\delta_u (1 + \bar{\xi}_{qu})} \left(|\bar{\zeta}_{qu}|^2 \sum_{u' \in \mathcal{U}_{-q}} \bar{t}_{q,u'} \left(\bar{\mathbf{h}}_{qu,u'} \bar{\mathbf{h}}_{qu,u'}^H + \bar{\Theta}_{qu,u'} \right) + \sum_{u' \in \bar{\mathcal{U}}_q} |\bar{\zeta}_{qu'}|^2 \left(\bar{\mathbf{h}}_{qu,u'} \bar{\mathbf{h}}_{qu,u'}^H + \bar{\Theta}_{qu,u'} \right) + \Omega_{qu} \right)^{-1} \bar{\mathbf{h}}_{qu,u}, \quad (31)$$

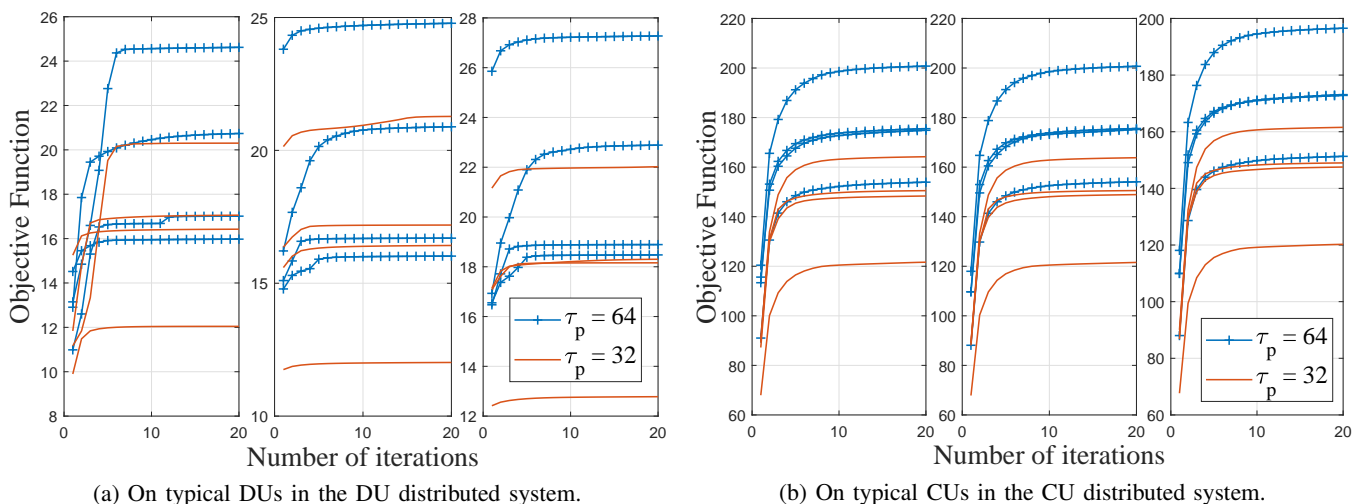


Fig. 2: Evolution of objection function. Subplots in each system: (Left) standard CSI est. case, (center) statistical CSI for out-of-cluster users case, (right) traffic distribution case.

The value of P_h can be used to control the density of the hotspots that are centered at locations $[\tilde{\mathbf{x}}_h, \tilde{\mathbf{y}}_h] = \left[[\tilde{x}_{h_1}, \dots, \tilde{x}_{h_{N_h}}]^T, [\tilde{y}_{h_1}, \dots, \tilde{y}_{h_{N_h}}]^T \right] \in \mathbb{R}^{N_h \times 2}$. These hotspots have equal variances σ_h^2 in both x and y dimensions. The term a_n is the area of the considered region around each node, and f_h is a normalizing factor that can be calculated numerically to normalize the PDF.

We simulate path loss using the COST231 Walfisch-Ikegami model [42], at carrier frequency $f = 1800$ MHz, in a typical urban environment, where we define $\beta(d_{ru})$ (dB) = $-112.4271 - 38 \log_{10}(d_{ru})$ with d_{ru} in km. We use $\tau_d = 200$ symbols as the length of the downlink transmission phase, and we simulate the cases of pilot lengths $\tau_p = 32$ and 64. We average our results using Monte Carlo simulations over both network realizations and time slots.

Simulating multiple time slots captures the effect of the fairness on user scheduling. In this regard, we simulate 30 time slots and average the results over the last 20 time slots after the transient in the fairness weights. These results represent the network steady state performance and is denoted as the long-term performance. In Table I, we summarize the parameters used in all simulations unless specified otherwise.

In Fig. 2, we plot the evolution of the objective function in the DU-distributed system (Fig. 2(a)) and the CU-distributed one (Fig. 2(b)) for different realizations of channels and networks. For each system, we show as subplots the three different methods proposed to calculate the leakage. Additionally, we plot many runs for each case different choices of the pilot length, τ_p . As is clear, the results show that the algorithms converge in a smooth, non-decreasing, pattern.

To test the efficacy of our developed approaches and to benchmark them, we compare our results with three schemes defined as follows.

- Centralized [29]: we use a similar approach to the CU-distributed system but with some differences that include the use of the weighted sum rate objective function, i.e., SINR-based. The implication of this is the need for a

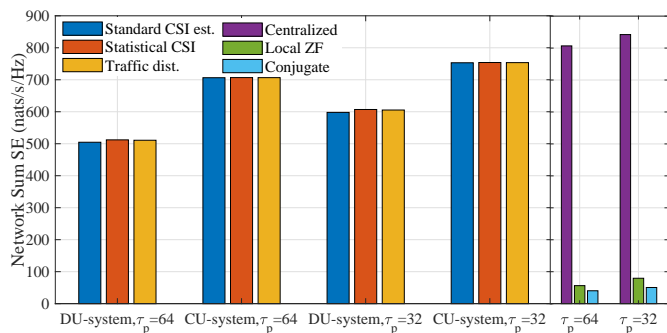
centralized unit to gather all the information about the network and run the algorithm, which means that the scheme is not as scalable as the distributed schemes.

In theory, a centralized solution should outperform a distributed solution. However, since the optimization problem is non-convex, any solution is a local optimum. It is, therefore, not guaranteed that the centralized solution outperforms the distributed solution *in practice*.

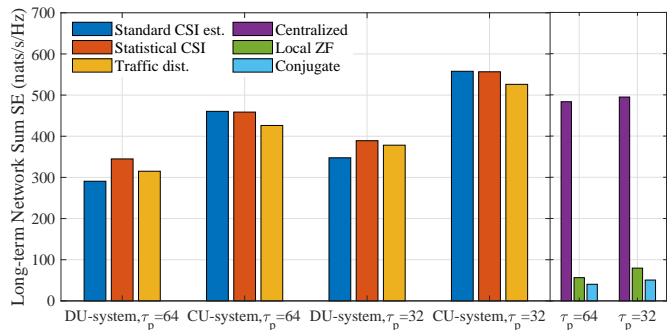
- Local zero-forcing (ZF): we use ZF constructed locally on each DU with round-robin scheduling for the users. Note that the ZF matrix needs to be constructed locally at the DUs, because the sets of users to be served by the DUs overlap with each other. Hence, no disjoint regions can be defined to construct a ZF matrix across multiple DUs.
- Conjugate: we use conjugate beamforming with round-robin scheduling.

Sum Rate: In Fig. 3(a), we plot the network sum of spectral efficiency (SE) averaged over network realizations for a single time slot when the fairness weights are equal for all the users (max sum rate). As expected, the CU-distributed system provides better performance than the DU-distributed system because it implements more coordination. Compared to a centralized approach, the CU-system provides comparable performance, with only a slight loss in network sum SE. This is a very important result because it illustrates the importance of deploying multiple CUs in the cell-free MIMO scheme. Note that unlike the DU- and CU-distributed systems, the centralized scheme is not scalable because it requires gathering all the CSI, including inter-CU CSI, at a central node in the network to perform the optimization.

Clearly, round-robin scheduling and the benchmark beamforming systems provide poor performance due to scheduling the users irrespective of their channel conditions and without any consideration of interference cancellation. Compared to the benchmark schemes, our proposed approaches provide a huge gain. Moreover, the results show the efficacy of the different methods of calculating the leakage, denoted as



(a) At equal fairness weights (max sum rate).



(b) Long-term with evolving fairness weights.

Fig. 3: Network sum SE for the two distributed systems under different techniques to calculate leakage.

standard CSI, statistical CSI and traffic distribution; these three approaches provide approximately equal performance. One note though is that using the traffic distribution to calculate the leakage is more reasonable when the number of users requesting access is very large, which is the main theme for this study. When the number of users requesting access is small, using the traffic distribution will lead to some loss in performance.

In Table II, we compare the gains of our proposed schemes to the benchmark schemes, where our two distributed systems are able to provide different degrees of performance. Note that the CU-distributed system provides a 1.28- and 1.77-fold performance gain compared to the DU-distributed system using $\tau_p = 32$ and $\tau_p = 64$, respectively.

Including Fairness: In Fig. 3(b), we plot the long-term network sum SE, averaged over both network realizations and time slots. The importance of this figure is in capturing the effect of the evolution of fairness weights and steady state algorithm performance. The results show that on the long-term, using the statistical CSI provides the best performance among the other methods used to calculate the leakage. Also, the CU-distributed system provides a 1.3-fold performance gain over the long-term compared to the DU-distributed system. Furthermore, the results show that over the long-term the CU-distributed system can outperform the centralized scheme.

As mentioned, both the centralized scheme and the ones proposed here result in local optima; given the complexity of the objective function and constraints, it is impossible to bound the distance from the optimal solution. However, as in the

	DU-distributed system		CU-distributed system	
	$\tau_p = 64$	$\tau_p = 32$	$\tau_p = 64$	$\tau_p = 32$
Centralized	0.64-fold	0.72-fold	0.88-fold	0.9-fold
Local ZF	9-fold	7.6-fold	12.5-fold	9.5-fold
Conjugate	12.7-fold	12-fold	17.5-fold	15-fold

TABLE II: Performance of our proposed systems compared to different schemes at equal fairness weights.

proposed distributed schemes, fewer variables are optimized at each DU/CU. We speculate that the fewer variables required for the CU-distributed system lead to a solution closer to the global optimum. This is consistent with results for a traditional cellular network [43]. Another possible reason for this is that the fairness weights of the users are different in each system, after the allocation of time slots, which may affect the set of users being scheduled concurrently in the latter time slots. For the DU-distributed system, the centralized scheme still provides a better performance in any scenario. This is also reasonable because the DU-distributed system restricts its performance to the knowledge found at the DU, so it is less cooperative than the CU-distributed system.

All in all, the results show that both schemes are very efficient in performing the resource allocation, though, as expected, the CU-distributed system provides substantially better performance. Our results highlight the importance of deploying multiple CUs in the user-centric cell-free network while the DU-distributed system can still be used if the fronthaul is overloaded.

Using Traffic Distribution: We use the traffic distribution defined in (36) with $\sigma_h = 50$ m, $P_h = 0.5$, and the number of hotspots N_h randomly generated as uniformly random variables between four and six hotspots within the studied network area. User locations are then generated from this traffic distribution. An example of a typical network is shown in Fig. 4 which also plots a realization of DU locations. In Fig. 5, we plot the long-term network sum SE under the scenario of hotspots, where $N = 5$ and density of users is $\lambda_{\text{users}} = 400$ users/km². As before, the results are averaged over network realizations and time slots. Most importantly, the results show that the case using the traffic distribution to calculate the leakage provides comparable performance to the other two methods. We conclude that using the traffic distribution to calculate the leakage can be a promising approach to completely eliminating any information exchange between DUs/CUs.

CDF: In Figs. 6 and 7, we plot the cumulative distribution function (CDF) of the long-term network sum SE and the long-term SE per user, respectively. Note that for the SE per user, we average the SE over all the simulated time slots even if the user is not scheduled; this provides a true measure of the user's overall throughput. The results show the superiority of the CU-distributed system over the DU-distributed system in both network sum SE and SE per user. Notably, for the scenario considered, a pilot length of $\tau_p = 32$ provides better a trade-off between pilot contamination and pilot training overhead compared to $\tau_p = 64$.

Algorithm Complexity Analysis: We now discuss the complexity of the algorithms developed. For Algorithm 1 ex-

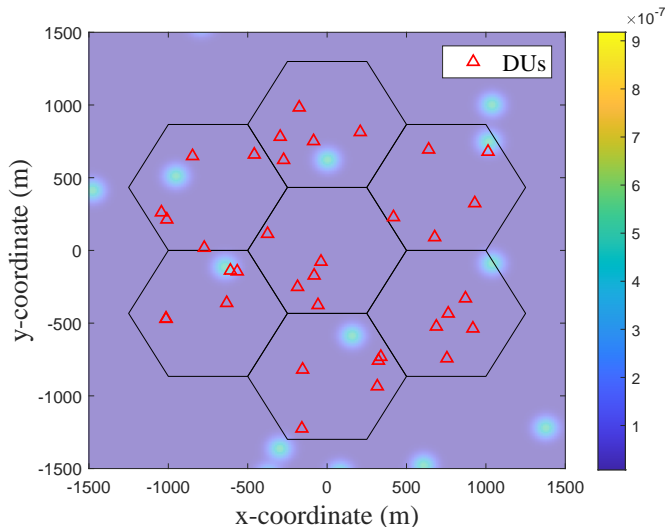


Fig. 4: A typical network with hotspot areas, $N = 5$, $\lambda_{\text{users}} = 400$ users/km².

executed on each DU, per iteration, we have complexity of $\mathcal{O}(|\mathcal{E}_r|)$ and $\mathcal{O}(|\mathcal{E}_r|)$ to update ξ_r and ζ_r , respectively. The combinatorial search step using the Hungarian algorithm has a polynomial time of complexity $\mathcal{O}(|\mathcal{E}_r|^3)$. To finish, the complexity of the beamforming step is equal to the complexity of the weighted MMSE which can be derived to be $\mathcal{O}(|\mathcal{U}_s|^2 M^2 + |\mathcal{U}_s| M^3)$ [44], where $|\mathcal{U}_s| \leq M$ is the number of scheduled users. This means that the complexity per iteration of the algorithm is at most $\mathcal{O}(M^4 + |\mathcal{E}_r|^3)$ at each DU.

For Algorithm 2 executed on each CU, we have a complexity of $\mathcal{O}(|\mathcal{U}_q|)$, $\mathcal{O}(|\mathcal{U}_q|)$, $\mathcal{O}(N|\mathcal{E}_{\text{avg}}|)$, $\mathcal{O}(N)$ to update ξ_q , ζ_q , α , and $\{\lambda_r : r \in \mathcal{B}_q\}$ respectively, where $|\mathcal{E}_{\text{avg}}|$ is the average number of users to be served by a DU. To compare both systems, we can set $\mathcal{E}_r = \mathcal{E}_{\text{avg}}$ in the DU-distributed system. Note that $|\mathcal{U}_q| \ll N|\mathcal{E}_r|$ due to the users overlapped between the clusters $\{\mathcal{E}_r : r \in \mathcal{B}_q\}$. The complexity of the beamforming step is at most $\mathcal{O}(N^2 M^4 + N M^4)$ assuming $|\mathcal{U}_s| \leq N M$. Hence, we obtain a total complexity of $\mathcal{O}(N^2 M^4 + N|\mathcal{E}_{\text{avg}}| + |\mathcal{U}_q|)$ per CU.

As a point of comparison, if we are to construct a centralized resource allocation scheme, we end up with complexity of $\mathcal{O}(|\mathcal{U}|)$, $\mathcal{O}(|\mathcal{U}|)$, $\mathcal{O}(Q N |\mathcal{E}_{\text{avg}}|)$, $\mathcal{O}(Q N)$ to update ξ , ζ , α , and $\{\lambda_r : r \in \mathcal{B}\}$, respectively. Hence, we end up with a total complexity per iteration as $\mathcal{O}(Q^2 N^2 M^4 + Q N |\mathcal{E}_{\text{avg}}| + |\mathcal{U}|)$ on the centralized node that will run the algorithm. Importantly, such a centralized system requires rapidly increasing computation capabilities in the sense that when the network area grows, the number of virtual cells Q grows. The distributed cases do not suffer from this issue.

From a communication point of view, in a centralized system, all the DUs and the CUs in the network need to communicate the algorithm data with a centralized node at the network core to be able to perform the allocation. All in all, the DU-distributed system is the most distributed scheme, while the CU-distributed system provides a balance between the DU-distributed system and the centralized one in both complexity,

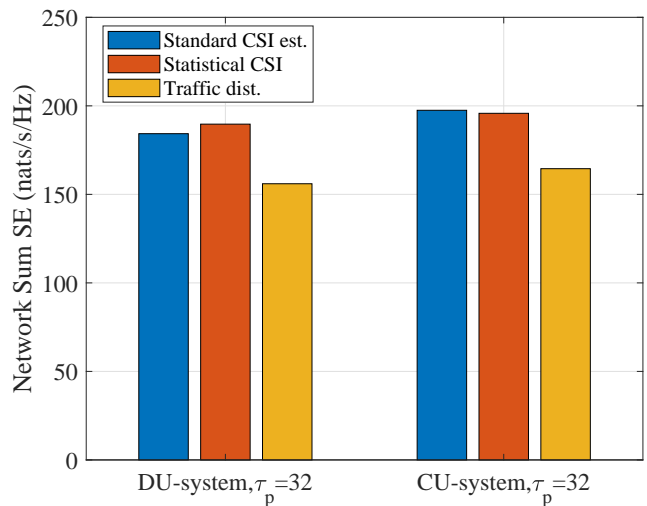


Fig. 5: Long-term network sum SE, hotspots scenario, $N = 5$, $\lambda_{\text{users}} = 400$ users/km².

scalability and performance.

Front-haul Footprint: To analyze the front-haul communications imposed by the algorithms, the DU-distributed system does not require exchanging information with the CU during the execution of the algorithm to perform the resource allocation. For the CU-distributed system, the N DUs need to exchange, with their CU, the CSI for their users, and then receive the scheduling decisions for $|\mathcal{E}_{\text{avg}}|$ users, and the constructed beamformers for at most M users that will be scheduled.

Regarding the CSI exchange, for the standard CSI estimation method used to calculate the leakage, each DU needs to exchange the CSI for $|\mathcal{U}_{\text{ng}}|$ users with the CU, where \mathcal{U}_{ng} represents the users within an area around the DU that have a non-negligible reachable signal, which can simply be based on distance or selected regions. For the statistical CSI approach used to calculate the leakage, we need to exchange $|\mathcal{E}_{\text{avg}}| \leq |\mathcal{U}_{\text{ng}}|$ CSI vectors for the users. For the approach that uses the traffic distribution, we do not need to exchange the leakage CSI, because we use a surveyed spatial traffic distribution to calculate the leakage; however the DUs still need to exchange $|\mathcal{E}_{\text{avg}}| \leq |\mathcal{U}_{\text{ng}}|$ CSI vectors for their users. Note that the size of these data to be exchanged can be reduced using signal quantization techniques [45], [46] proposed for the front-haul. In Table III, we compare the complexity and front-haul footprint of our distributed systems with a centralized one.

Finally, whether to implement the DU-distributed or the CU-distributed system is a system preference, and it depends on the tradeoff between performance and both computational complexity and fronthaul load. However, we believe that the CU-distributed system (in a multi-CU network) is better because it provides a tradeoff between complexity per node and performance compared to the centralized scheme and the DU-distributed system. Moreover, it allows simple design for the DUs.

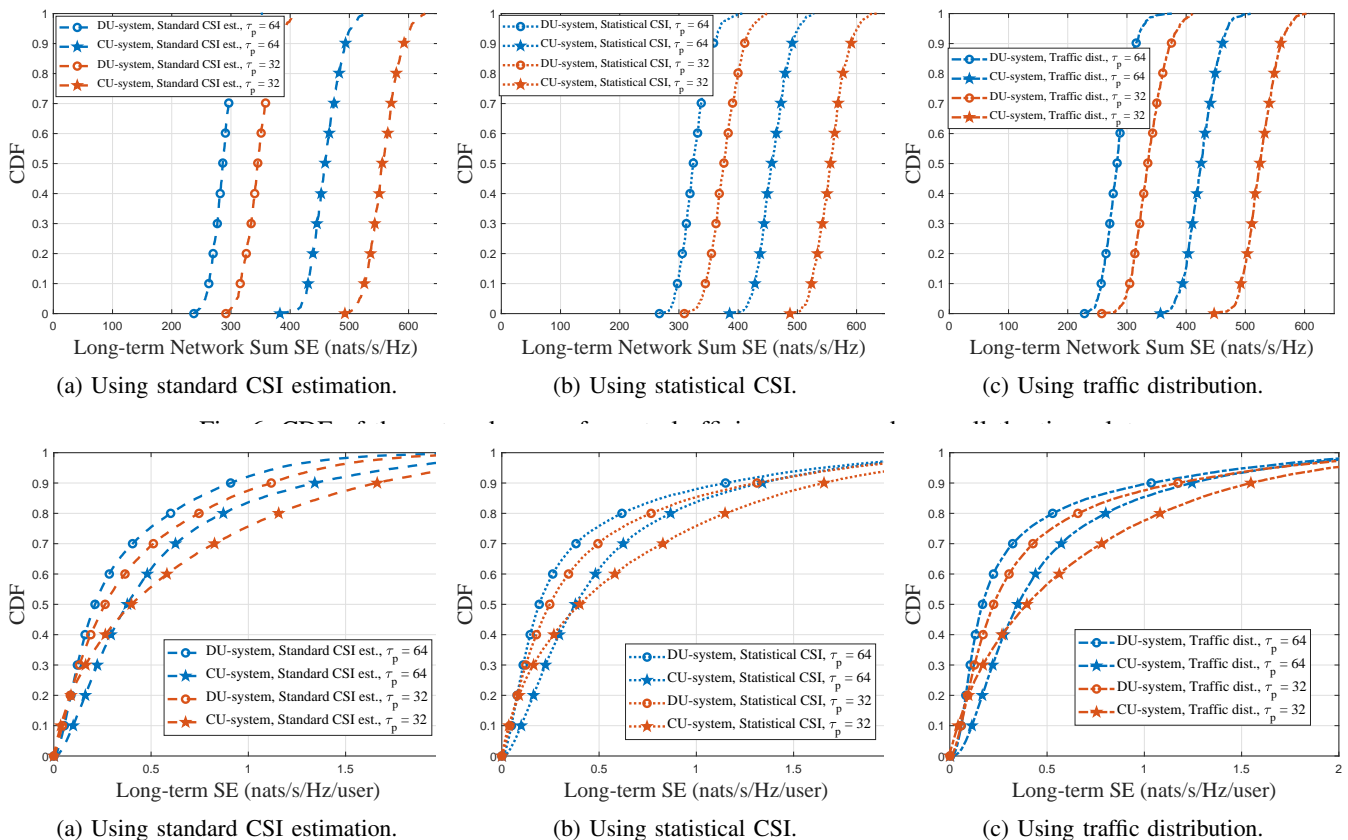


Fig. 7: CDF of the user spectral efficiency averaged over all the time slots.

System	Complexity per node and iteration	Front-haul overhead for algorithm		
		Standard CSI est.	Statistical CSI	Traffic dist.
DU-distributed system	$\mathcal{O}(M^4 + \mathcal{E}_{\text{avg}} ^3)$	0	0	0
CU-distributed system	$\mathcal{O}(N^2 M^4 + N \mathcal{E}_{\text{avg}} + \mathcal{U}_q)$	$\mathcal{O}(\mathcal{U}_{\text{ng}} + N \mathcal{E}_{\text{avg}})$	$\mathcal{O}(N \mathcal{E}_{\text{avg}})$	$\mathcal{O}(N \mathcal{E}_{\text{avg}})$
Centralized	$\mathcal{O}(Q^2 N^2 M^4 + QN \mathcal{E}_{\text{avg}} + \mathcal{U})$	$\mathcal{O}(Q \mathcal{U}_{\text{ng}} + QN \mathcal{E}_{\text{avg}})$		

TABLE III: Comparison of *algorithm complexity and front-haul overhead* for different systems.

VII. CONCLUSION

In this paper, we developed two distributed algorithms that perform user scheduling, beamforming, and implicit power control for user-centric cell-free MIMO networks. The first system is implemented at the DUs, while the second one is implemented at the CUs. Both use a hybrid metric that depends on the leakage and intra-node interference allowing a distributed allocation. Additionally, exploiting the fact that the notion of leakage is a tool to measure a transmission has on neighboring “cells”, we proposed three approaches to calculate the leakage, each requiring a different level of CSI exchange. We present possible deployment scenarios for user-centric cell-free network with manageable complexity with respect to the network size.

In terms of performance, the CU-distributed system provides 1.3- to 1.8-fold network throughput gain compared to the DU-distributed system, with a slight increase in complexity and front-haul load - and can outperform centralized solutions. Further, our results measure the performance gain, computational complexity, and front-haul overhead compared to benchmark schemes and a centralized resource allocation.

Finally, by analyzing the trade-offs provided by the CU-distributed system compared to DU-distributed and centralized ones, we highlight the importance of deploying multiple CUs in user-centric cell-free networks.

APPENDIX A PROBLEM REFORMULATION FOR DU-DISTRIBUTED SYSTEM

Using the objective function (10a) and the equality constraint in (10d), we define the Lagrangian formulation as

$$f_1(\mathbf{W}_r, \boldsymbol{\xi}_r, \boldsymbol{\nu}_r) = \sum_{u \in \mathcal{E}_r} \delta_u \log(1 + \xi_{ru}) - \sum_{u \in \mathcal{E}_r} \nu_{ru} \left(\xi_{ru} - \frac{s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru}}{A_{ru}(s_r, \mathbf{W}_r)} \right), \quad (37)$$

where $\boldsymbol{\xi}_r = [\{\xi_{ru} : u \in \mathcal{E}_r\}]^T \in \mathbb{R}^{|\mathcal{E}_r| \times 1}$. To satisfy the first-order optimality condition of ξ_{ru} , we set the derivative of (37) with respect to ξ_{ru} to zero, thereby obtaining an expression

for ν_{ru} that satisfies this condition. Then, we substitute this expression for ν_{ru} back in (37) to obtain

$$f_1(\mathbf{s}_r, \mathbf{W}_r, \boldsymbol{\xi}_r) = \sum_{u \in \mathcal{E}_r} \delta_u (\log(1 + \xi_{ru}) - \xi_{ru}) + \sum_{u \in \mathcal{E}_r} \delta_u \left(\frac{(1 + \xi_{ru}) s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru}}{s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru} + A_{ru}(\mathbf{s}_r, \mathbf{W}_r)} \right). \quad (38)$$

The importance of this procedure is that the variables $(\mathbf{s}_r, \mathbf{W}_r)$ are now found outside the logarithmic function, while ξ_{ru} acts as an auxiliary variable or proxy to emulate their effect. When the variables $(\mathbf{s}_r, \mathbf{W}_r)$ are fixed, we can set the derivative of (38) with respect to ξ_{ru} to zero, to obtain the optimal expression for the SLINR auxiliary variable ξ_{ru} , which, as expected, equals (10d). The expression in (38) can be shown to be equivalent to (37) by substituting the optimal expression of ξ_{ru} back into (38), thereby resulting in the same objective function in (10).

We use the fractional programming approach developed in [10] to write (38) as

$$f_2(\mathbf{s}_r, \mathbf{W}_r, \boldsymbol{\xi}_r, \boldsymbol{\zeta}_r) = \sum_{u \in \mathcal{E}_r} \delta_u (\log(1 + \xi_{ru}) - \xi_{ru}) + \sum_{u \in \mathcal{E}_r} \left(2\text{Re} \left\{ \zeta_{ru}^* \sqrt{\delta_u (1 + \xi_{ru})} s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \right\} - |\zeta_{ru}|^2 \left(s_{ru} \mathbf{w}_{ru}^H \hat{\mathbf{h}}_{ru} \hat{\mathbf{h}}_{ru}^H \mathbf{w}_{ru} + A_{ru}(\mathbf{s}_r, \mathbf{W}_r) \right) \right). \quad (39)$$

where $\boldsymbol{\zeta}_r \in \mathbb{C}^{|\mathcal{E}_r| \times 1}$ is a new auxiliary variable vector introduced by fractional programming [10]. The formula in (12) becomes the objective function of the newly formulated problem in (11).

APPENDIX B

PROBLEM REFORMULATION FOR CU-DISTRIBUTED SYSTEM

Using the objective function (26a) and the equality constraint in (26d), we can write the following Lagrangian formulation.

$$f_1(\mathcal{W}_q, \bar{\boldsymbol{\xi}}_q, \boldsymbol{\nu}_q) = \sum_{u \in \bar{\mathcal{U}}_q} \delta_u \log(1 + \bar{\xi}_{qu}) - \sum_{u \in \bar{\mathcal{U}}_q} \nu_{qu} \left(\bar{\xi}_{qu} - \frac{\bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu}}{C_{qu}(\mathcal{W}_q)} \right). \quad (40)$$

To satisfy the first-order optimality condition of $\bar{\xi}_{qu}$, we differentiate (40) with respect to $\bar{\xi}_{qu}$ and set to zero to obtain the corresponding value for ν_{qu} . Substituting ν_{qu} back in (40) we obtain

$$f_1(\mathcal{W}_q, \bar{\boldsymbol{\xi}}_q) = \sum_{u \in \bar{\mathcal{U}}_q} \delta_u (\log(1 + \bar{\xi}_{qu}) - \bar{\xi}_{qu}) + \sum_{u \in \bar{\mathcal{U}}_q} \delta_u \left(\frac{(1 + \bar{\xi}_{qu}) \bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu}}{\bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu} + C_{qu}(\mathcal{W}_q)} \right). \quad (41)$$

When \mathcal{W}_q is fixed, the first optimality condition for $\bar{\xi}_{qu}$ yields the optimal value of $\bar{\xi}_{qu}$ as

$$\bar{\xi}_{qu} = \frac{\bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu}}{C_{qu}(\mathcal{W}_q)}. \quad (42)$$

Equation (41) can be shown to be equivalent to (40) as discussed in the previous section for a DU-distributed system. Again, using fractional programming [10], we write (41) as

$$f_2(\mathcal{W}_q, \bar{\boldsymbol{\xi}}_q, \bar{\boldsymbol{\zeta}}_q) = \sum_{u \in \bar{\mathcal{U}}_q} \delta_u (\log(1 + \bar{\xi}_{qu}) - \bar{\xi}_{qu}) + \sum_{u \in \bar{\mathcal{U}}_q} \left(2\text{Re} \left\{ \bar{\zeta}_{qu}^* \sqrt{\delta_u (1 + \bar{\xi}_{qu})} \bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \right\} - |\bar{\zeta}_{qu}|^2 \left(\bar{\mathbf{w}}_{qu}^H \bar{\mathbf{h}}_{qu,u} \bar{\mathbf{h}}_{qu,u}^H \bar{\mathbf{w}}_{qu} + C_{qu}(\mathcal{W}_q) \right) \right), \quad (43)$$

where $\bar{\boldsymbol{\zeta}}_q = [\bar{\zeta}_{qu_1} \dots \bar{\zeta}_{qu_{|\bar{\mathcal{U}}_q|}}]^T$ are introduced auxiliary variables. The formula in (43) becomes the new objective function of our reformulated problem in (11).

REFERENCES

- [1] A. B. Bondi, "Characteristics of scalability and their impact on performance," in *Proceedings of the 2nd international workshop on Software and performance*, pp. 195–203, 2000.
- [2] S. Chen, J. Zhang, E. Björnson, J. Zhang, and B. Ai, "Structured massive access for scalable cell-free massive MIMO systems," *IEEE Journal on Selected Areas in Comm.*, pp. 1–1, 2020.
- [3] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP Journal on Wireless Comm. and Networking*, vol. 2019, no. 1, p. 197, 2019.
- [4] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. on Wireless Comm.*, vol. 16, no. 3, pp. 1834–1850, 2017.
- [5] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Trans. on Wireless Comm.*, vol. 19, pp. 1250–1264, Feb 2020.
- [6] H. A. Ammar and R. Adve, "Power delay profile in coordinated distributed networks: User-centric v/s disjoint clustering," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, Nov 2019.
- [7] Z. Chen and E. Björnson, "Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry," *IEEE Trans. on Comm.*, vol. 66, no. 11, pp. 5205–5219, 2018.
- [8] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. on Wireless Comm.*, vol. 19, pp. 77–90, Jan 2020.
- [9] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *ICC 2019 - 2019 IEEE International Conference on Comm. (ICC)*, pp. 1–6, 2019.
- [10] K. Shen and W. Yu, "Fractional programming for communication systems—part II: Uplink scheduling via matching," *IEEE Trans. on Signal Processing*, vol. 66, pp. 2631–2644, May 2018.
- [11] A. A. Khan, R. Adve, and W. Yu, "Optimizing downlink resource allocation in multiuser MIMO networks via fractional programming and the Hungarian algorithm," *IEEE Trans. on Wireless Comm.*, pp. 1–1, 2020.
- [12] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Trans. on Wireless Comm.*, vol. 6, pp. 1711–1721, May 2007.
- [13] T. Gamvrelis, A. Khan, and R. Adve, "SLNR- and SLINR-based downlink optimization in MU-MIMO networks," *IEEE Trans. on Wireless Comm.*, 2020. Submitted for review.
- [14] Q.-D. Vu, L.-N. Tran, and M. Juntti, "Noncoherent joint transmission beamforming for dense small cell networks: Global optimality, efficient solution and distributed implementation," *IEEE Trans. on Wireless Comm.*, vol. 19, no. 9, pp. 5891–5907, 2020.

- [15] I. Atzeni, B. Gouda, and A. Tölli, "Distributed precoding design via over-the-air signaling for cell-free massive MIMO," *IEEE Trans. on Wireless Comm.*, vol. 20, no. 2, pp. 1201–1216, 2021.
- [16] J. Kaleva, A. Tölli, M. Juntti, R. A. Berry, and M. L. Honig, "Decentralized joint precoding with pilot-aided beamformer estimation," *IEEE Trans. on Signal Processing*, vol. 66, no. 9, pp. 2330–2341, 2018.
- [17] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max–min SINR of cell-free massive MIMO systems," *IEEE Trans. on Wireless Comm.*, vol. 18, no. 4, pp. 2021–2036, 2019.
- [18] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Trans. on Wireless Comm.*, pp. 1–1, 2020.
- [19] M. Bashar, H. Q. Ngo, K. Cumanan, A. G. Burr, P. Xiao, E. Björnson, and E. G. Larsson, "Uplink spectral and energy efficiency of cell-free massive MIMO with optimal uniform quantization," *IEEE Trans. on Comm.*, pp. 1–1, 2020.
- [20] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. on Comm.*, pp. 1–1, 2020.
- [21] Z. Chen, E. Björnson, and E. G. Larsson, "Dynamic resource allocation in co-located and cell-free massive MIMO," *IEEE Trans. on Green Comm. and Networking*, pp. 1–1, 2019.
- [22] A. Zappone, E. Jorswieck, and A. Leshem, "Distributed resource allocation for energy efficiency in MIMO OFDMA wireless networks," *IEEE Journal on Selected Areas in Comm.*, vol. 34, no. 12, pp. 3451–3465, 2016.
- [23] O. Nappark and A. Leshem, "Fully distributed optimal channel assignment for open spectrum access," *IEEE Trans. on Signal Processing*, vol. 62, no. 2, pp. 283–294, 2014.
- [24] D. P. Bertsekas, "A distributed algorithm for the assignment problem," *Lab. for Inf. Decis. Sys. Working Paper, MIT*, 1979.
- [25] F. Bannour, S. Souihi, and A. Mellouk, "Distributed SDN control: Survey, taxonomy, and challenges," *IEEE Comm. Surveys Tutorials*, vol. 20, no. 1, pp. 333–354, 2018.
- [26] E. Björnson, R. Zakhour, D. Gesbert, and B. Ottersten, "Cooperative multicell precoding: Rate region characterization and distributed strategies with instantaneous and statistical CSI," *IEEE Trans. Sig. Process.*, vol. 58, no. 8, 2010.
- [27] C. D'Andrea and E. G. Larsson, "User association in scalable cell-free massive MIMO systems," *arXiv preprint arXiv:2103.05321*, 2021.
- [28] "IEEE standard for a precision clock synchronization protocol for networked measurement and control systems," *IEEE Std 1588-2019 (Revision of IEEE Std 1588-2008)*, pp. 1–499, 2020.
- [29] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "Downlink resource allocation in multiuser cell-free MIMO networks with user-centric clustering," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2021.
- [30] M. Attarifar, A. Abbasfar, and A. Lozano, "Random vs structured pilot assignment in cell-free massive MIMO wireless networks," in *2018 IEEE International Conference on Comm. Workshops (ICC Workshops)*, pp. 1–6, 2018.
- [31] M. G. Karypis, V. Kumar, and M. Steinbach, "A comparison of document clustering techniques," in *TextMining Workshop at KDD2000 (May 2000)*, 2000.
- [32] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.
- [33] W. Yu, T. Kwon, and C. Shin, "Adaptive resource allocation in cooperative cellular networks," *Cooperative cellular wireless networks*, pp. 233–256, 2011.
- [34] Z. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, pp. 57–73, Feb 2008.
- [35] S. Schaible, "Parameter-free convex equivalent and dual programs of fractional programming problems," *Zeitschrift für Operations Research*, vol. 18, no. 5, pp. 187–196, 1974.
- [36] W. Dinkelbach, "On nonlinear fractional programming," *Management science*, vol. 13, no. 7, pp. 492–498, 1967.
- [37] S. H. Bokhari, *Assignment problems in parallel and distributed computing*. Springer Science & Business Media, 2012.
- [38] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logis. quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [39] D. L. Donoho, "Compressed sensing," *IEEE Trans. on Information Theory*, vol. 52, pp. 1289–1306, April 2006.
- [40] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.
- [41] A. Minasian, R. S. Adve, S. Shahbazpanahi, and G. Boudreau, "On RRH placement for multi-user distributed massive MIMO systems," *IEEE Access*, vol. 6, pp. 70597–70614, 2018.
- [42] J. Walfisch and H. L. Bertoni, "A theoretical model of UHF propagation in urban environments," *IEEE Trans. on Antennas and Propagation*, vol. 36, pp. 1788–1796, Dec 1988.
- [43] A. A. Khan and R. S. Adve, "Centralized and distributed deep reinforcement learning methods for downlink sum-rate optimization," *IEEE Trans. on Wireless Comm.*, vol. 19, no. 12, pp. 8410–8426, 2020.
- [44] Q. Shi, M. Razaviyayn, Z. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. on Signal Processing*, vol. 59, pp. 4331–4340, Sep. 2011.
- [45] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, M. Debbah, and P. Xiao, "Max–min rate of cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. on Comm.*, vol. 67, no. 10, pp. 6796–6815, 2019.
- [46] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, E. G. Larsson, and P. Xiao, "Energy efficiency of the cell-free massive MIMO uplink with optimal uniform quantization," *IEEE Trans. Green Comm. and Net.*, vol. 3, no. 4, pp. 971–987, 2019.