Joint Beamforming Design and Stream Allocation for Non-Coherent Joint Transmission in Cell-Free MIMO Networks

Xi Wang, Xiaotong Zhao, Juncheng Wang, You Li, and Qingjiang Shi

Abstract—We consider joint beamforming and stream allocation to maximize the weighted sum rate (WSR) for non-coherent joint transmission (NCJT) in user-centric cell-free MIMO networks, where distributed access points (APs) are organized in clusters to transmit different signals to serve each user equipment (UE). We for the first time consider the common limits of maximum number of receive streams at UEs in practical networks, and formulate a joint beamforming and transmit stream allocation problem for WSR maximization under per-AP transmit power constraints. Since the integer number of transmit streams determines the dimension of the beamformer, the joint optimization problem is mixed-integer and nonconvex with coupled decision variables that is inherently NP-hard. In this paper, we first propose a distributed low-interaction reduced weighted minimum mean square error (RWMMSE) beamforming algorithm for WSR maximization with fixed streams. Our proposed RWMMSE algorithm requires significantly less interaction across the network and has the current lowest computational complexity that scales linearly with the number of transmit antennas, without any compromise on WSR. We draw insights on the joint beamforming and stream allocation problem to decouple the decision variables and relax the mixed-integer constraints. We then propose a joint beamforming and linear stream allocation algorithm, termed as RWMMSE-LSA, which yields closed-form updates with linear stream allocation complexity and is guaranteed to converge to the stationary points of the original joint optimization problem. Simulation results demonstrate substantial performance gain of our proposed algorithms over the current best alternatives in both WSR performance and convergence time.

Index Terms—Cell-free networks, non-coherent joint transmission, beamforming, low-interaction, stream allocation, mixed-integer programming.

I. INTRODUCTION

THE exponential increase in the number of dense and heterogeneous terminals and their diverse requirements present major challenges to modern wireless communication networks, such as increased inter-cell interference and attenuated cell-edge rate. Due to the co-located antennas architecture of traditional cellular paradigm, cell-edge user equipments (UEs) are inevitably far away from the base stations (BSs), and

Xi Wang and Xiaotong Zhao are with the School of Software Engineering, Tongji University, Shanghai 201804, China (e-mail: {wangxi_w,xiaotongzhao}@tongji.edu.cn). Juncheng Wang is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China (e-mail: jcwang@comp.hkbu.edu.hk). You Li is with Huawei Technologies Co. Ltd, Chengdu 611730, China (e-mail: liyou1992@163.com). Qingjiang Shi is with the School of Software Engineering, Tongji University, Shanghai 201804, China, and also with the Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: shiqj@tongji.edu.cn).

thus suffer from edge rate attenuation. The paradigm of cell-free multiple-input multiple-output (MIMO) was pioneered in [1] to provide uniform service to all the UEs. In cell-free MIMO networks, geographically distributed access points (APs) are coordinated by one or multiple central units (CUs) to jointly serve the UEs. In such networks, all the UEs are ensured to be located at the effective center of their serving APs, and cell edges no longer exist. Therefore, the inherent large data rate variation and inter-cell interference are mitigated [2]. Moreover, the proximity of densely deployed APs to their serving UEs in cell-free MIMO networks brings enhanced energy efficiency [3], low communication latency, and augmented service reliability [4], [5]. These exceptional benefits of cell-free MIMO networks position it as a promising paradigm for future wireless communication systems.

To provide high quality service and effectively manage interference, APs in cell-free MIMO networks need to perform joint transmission (JT) to enhance constructive signals and suppress destructive signals at the UEs [6], [7]. There are two JT approaches for cell-free MIMO networks: coherent joint transmission (CJT) and non-coherent joint transmission (NCJT). In CJT, all the APs cooperate as a virtual MIMO system to transmit the same signals to their serving UEs. Therefore, CJT requires strict synchronization across the network [8]. However, in practical cell-free networks, APs are coordinated by multiple CUs instead of a single CU to avoid single-point failure [9], [10]. It is difficult to deploy multiple CUs in CJT due to its strict synchronization requirement across the network, especially when the serving APs of a UE are controlled by different CUs [11]. NCJT differs from CJT in that the signals transmitted to a UE from its serving APs are different. Since each UE can decode the received signals independently, strict synchronization between CUs and APs is no longer required for NCJT [8], [12].

In order to perform JT, the APs and the CUs need to exchange channel state information (CSI) and cooperative beamforming matrices via fronthaul communication links [13]. However, both the CSI and the beamformer is of high dimension, imposing a large burden on the fronthaul. In practical communication networks, wireless fronthaul is widely deployed due to its high deployment flexibility and low installation cost. Even at the millimeter wave frequency, the available fronthaul capacity is limited, resulting in a bottleneck on efficient JT in cell-free MIMO networks [8]. It is therefore of critical importance to take into account the interaction between the APs and CUs in cell-free MIMO networks.

A. Motivations and Challenges

Most existing algorithms on beamforming design for NCJT are centralized [2], [8], [14], [15], [16], [17], which suffer from high interaction (requiring raw CSI exchange across the network) and high computational complexity in general [10]. Furthermore, they did not consider the possibility of maximizing the weighted sum rate (WSR) through data stream allocation among the UEs. Since the number of received data streams at each UE is limited by its receive antennas [18], while there are fluent APs possibly equipped with a large number of transmit antennas around each UE in cell-free MIMO systems, there is a plenty of room to further improve the WSR by properly allocating data streams from the APs to each UE. However, *there is a scarcity of existing literature on how to jointly allocate data streams and design cooperative beamformer for NCJT in cell-free MIMO systems.*

Due to the above discrepancies, in this work, we consider joint beamforming design and stream allocation to maximize the WSR of cell-free MIMO networks with NCJT, under individual transmit power constraints at the APs. We aim at developing low-interaction and low-complexity distributed joint beamforming and stream allocation algorithms. To achieve this goal, we must address several challenges: 1) Since the signal-to-interference-plus-noise ratio (SINR) of each UE is coupled among its serving APs, the CUs intrinsically require global CSI data to effectively mitigate inter-UE interference. However, directly exchanging of raw CSI data (the high-dimensional channel matrices) between the APs and the CUs is practically prohibited due to the limited fronthaul capacity in practical cell-free MIMO systems, while communicating partial CSI generally degrades the system performance [19]. It is therefore challenging to reduce the interaction between the APs and CUs without sacrificing the WSR. 2) Beamforming design and stream allocation are intrinsically coupled and hard to be jointly optimized, since the dimension of the beamformer is determined by the number of data streams. Even with fixed transmit streams, the pure beamforming optimization problem remains NP-hard. 3) Due to the unique sum-of-quadratic form of the signal covariance matrix in the WSR expression for NCJT, whether the efficient weighted minimum mean square error (WMMSE) approach for conventional CJT can be applied to reduce the computational complexity for beamforming design in NCJT remains an open problem [8], [17].

B. Contributions

Different from existing centralized beamforming algorithms for NCJT that are of high interaction (requiring raw CSI exchange between the APs and the CUs) and high complexity, we propose a distributed low-interaction and low-complexity beamforming algorithm using the WMMSE techniques. Furthermore, we *for the first time* study joint beamforming and stream allocation for user-centric cell-free MIMO networks with NCJT. Specifically, the main contributions of this paper are as follows:

1) **An Answer to Whether the WMMSE Approach is Applicable to NCJT:** Prior works have assumed that the WMMSE approach is not applicable to NCJT [8], [17]. Based on unique observations on the structure of the WSR expression for NCJT, we equivalently transform the original WSR maximization problem into a standard WMMSE form, showing the applicability of the WMMSE approach to NCJT. We then propose a centralized WMMSE based beamforming algorithm that has the current lowest computational complexity.

2) **A Distributed Low-interaction and Low-Complexity Beamforming Algorithm:** We draw some unique observations on the beamformer structure for WSR maximization in cell-free networks MIMO with NCJT. Based on these observations, we propose a low-interaction reduced WMMSE (RWMMSE) algorithm, which converges smoothly and achieves the same WSR as the centralized WMMSE method. Without sacrificing any WSR performance, our proposed RWMMSE algorithm yields much lower interaction (independent of the number of transmit antennas) than the centralized WMMSE algorithm. Furthermore, our proposed RWWMSE algorithm achieves a computational complexity that scales linearly with M, which is much lower than the current lowest $\mathcal{O}\left(M^3\right)$ complexity in [2], where M is the number of transmit antennas.

3) **A Joint Beamforming and Linear Stream Allocation Algorithm:** We for the first time consider the individual maximum number of receive streams constraints in NCJT, and formulate a joint beamforming and stream allocation problem to maximize the WSR of cell-free MIMO networks. Note that the beamforming and stream variables are coupled, in the sense that the beamforming dimension is determined by the number of streams. Moreover, the joint optimization problem is mixed-integer and nonconvex. We first introduce auxiliary stream indicator matrices to decouple the beamforming and stream variables. We then utilize the unique quadratic-linear property of the stream indicator matrices to transform the decoupled quadratic 0-1 integer stream allocation problem into an equivalent linear form, and relax it to a continuous optimization problem. We propose a joint beamforming and linear stream allocation algorithm, termed as RWMMSE-LSA, which consists of closed-form updates with linear stream allocation complexity and is shown to converge to the stationary points of the original joint optimization problem.

4) Our simulation results demonstrate that the proposed algorithms converge fast without any WSR sacrifice, while substantially reducing the computation time compared to the current best alternatives. We further compare the WSR between CJT and NCJT approaches, showing a balance between synchronization cost and WSR performance provided by NCJT. In addition, our proposed joint beamforming and stream allocation algorithm outperforms the pure beamforming algorithms.

Notations: The notation $\mathbf{A} \succ \mathbf{0}$ indicates positive definite. The notation $\mathrm{blkdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ denotes a block diagonal matrix with matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$. The column space of $\mathbf{A}$ is

the span of its column vectors. The null space of $\mathbf{A}$ is the linear subspace of the domain of the mapping to the zero vector. Orthogonal complement of the column space of $\mathbf{A}$ is defined by $\prod_{\mathbf{A}}^{\perp} \triangleq \mathbf{I} - \mathbf{A} \left(\mathbf{A}^H \mathbf{A} \right)^{-1} \mathbf{A}^H$. The binary and complex space are denoted as $\mathbb{B}$ and $\mathbb{C}$.

II. RELATED WORK

A. Beamforming Algorithms

1) Centralized Beamforming Algorithms: Most existing beamforming algorithms for NCJT are centralized, which demand massive interaction between the CUs and the APs and suffer from high computational cost [2], [8], [14], [16], [17]. For dense small cell networks, a semi-definite relaxation (SDR) based algorithm was proposed to minimize power consumption under transmit rate constraints [16]. Although the SDR based algorithm converges in polynomial time, it is not applicable to the WSR maximization problem. The SCA method was used in [8] to relax the nonconvex WSR maximization problem to a second order cone programming (SOCP) problem. Similar scheme was adopted in [14] for non-orthogonal multiple access systems. The SOCP problem in [8] is solved via CVX [20], which involves the interior point method that causes high computational complexity.

The authors in [2] utilized the fractional programming (FP) approach and the block coordinate descent (BCD) method to optimize the beamformer for WSR maximization under the special case that each UE is equipped with a single receive antenna. The FP based algorithm in [2] has much lower complexity than the SCA based algorithms, but still suffers from a $\mathcal{O}\left(M^3\right)$ computational complexity that is still high especially when massive antennas are deployed. Moreover, all of the aforementioned beamforming algorithms require the interaction of channel and beamforming matrices, both of which are related to the number of transmit antennas, which puts a significant burden on the fronthaul links.

2) Low-Interaction Beamforming Algorithms: Great efforts have been made to develop low-interaction beamforming algorithms especially for CJT [19]. Beamforming approaches such as local minimum mean square error (MMSE) combining [10], weighted MMSE (WMMSE) [2], [21], local partial zero forcing, and local protective partial zero forcing [22], reduce the interaction by limiting the CSI sharing.

The unique WSR structure (a summation term resulting from different data streams in the SINR numerator) of NCJT makes it challenging to directly extend the above-mentioned low-interaction beamforming algorithms to NCJT. Authors in [8] and [17] imply that the widely deployed WMMSE approach is not applicable to NCJT, whereas we suggest otherwise in this paper with a unique but equivalent WMMSE reformulation of the original WSR maximization problem. For NCJT, [17] used the inner approximation (InAp) method to relax the nonconvex WSR maximization problem to a convex approximation subproblem, which is then solved by the alternating direction method of multipliers (ADMM) method. The algorithm presented in [17] only requires scalar interactions across the network. However, the scalar interaction is required for each inner ADMM iteration, and the two-layer iterative
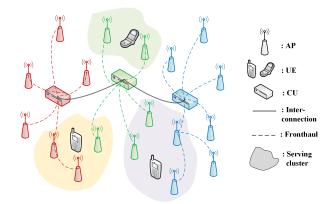

Fig. 1. An illustration of a user-centric cell-free network with multiple CUs.

approach adopted in the algorithm results in unavoidably high complexity. Furthermore, [17] focuses on the single receive antenna case, and the interaction is no longer scalar for the general multiple receive antenna cases.

B. Stream Allocation and User Scheduling

As stated in [18], *the number of usable spatial streams should be less than both of the number of transmit and receive antennas*. However, stream allocation is overlooked in prior works on cell-free networks, and the number of transmit streams is pre-defined. Besides, distant APs occupy power and bandwidth but contribute little receive power due to pathloss in cell-free networks [13], [23]. Therefore, UEs are not necessary served by all the APs but only nearby APs, which is generally referred to as user-centric cell-free networks [10], [24].

Existing works on user-centric cell-free networks define the serving cluster by serving distance [25] or follow the dynamic cooperation clustering framework [8], [17], [26]. The authors in [26] treated the mixed-integer clustering problem as an agent-task assignment problem, and used the Hungarian algorithm to solve it in polynomial time. In [17], a branch reduce-and-bound (BRnB) framework was developed to find the global optimal serving cluster for WSR maximization but with exponential complexity. In [8], a joint beamforming and user scheduling problem was formulated under limits on the maximum number of UEs served by each AP. The mixed-integer problem was transformed to ℓ_0-norm and then approximated using weighted ℓ_1-norm. Since the algorithm in [8] involves in two-layer iterations, the authors merge the two-layer iteration into a one-layer iteration to reduce the computational complexity. However, the resulting one-layer iteration algorithm cannot converge strictly.

III. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a downlink user-centric cell-free MIMO network that comprises I APs and K UEs, denoted by indices $\mathcal{I} = \{1, \ldots, I\}$ and $\mathcal{U} = \{1, \ldots, K\}$, respectively. Each AP i is equipped with M_i transmit antennas, serving a set of UEs $\mathcal{U}_i \subseteq \mathcal{U}$. Each UE k is equipped with N_k receive antennas and is jointly served by APs $\mathcal{I}_k \subseteq \mathcal{I}$. As illustrated in Fig. 1, geographically distributed APs collaborate with each other either through CUs or nearby APs.

A. Non-Coherent Joint Transmission Model

We consider that the APs cooperate in a non-coherent joint transmission mode to relieve the burden of strict synchronization among the CUs and APs. Specifically, each UE k receives different signals $\mathbf{s}_{i,k} \in \mathbb{C}^{D_{i,k} \times 1}$ from all the AP i with $D_{i,k}$ being the number of transmit streams. Note that only AP $i \in \mathcal{I}_k$ transmits useful signals to UE k. Denote the channel matrix between AP i and UE k as $\mathbf{H}_{i,k} \in \mathbb{C}^{N_k \times M_i}$. The received signal at UE k is given by

$$\mathbf{y}_k = \underbrace{\sum_{i \in \mathcal{I}_k} \mathbf{H}_{i,k} \mathbf{P}_{i,k} \mathbf{s}_{i,k}}_{\text{useful signals}} + \underbrace{\sum_{l \in \mathcal{U}_{-k}} \sum_{j \in \mathcal{I}_l} \mathbf{H}_{j,k} \mathbf{P}_{j,l} \mathbf{s}_{j,l}}_{\text{inter-user interference}} + \mathbf{z}_k \quad (1)$$

where $\mathbf{P}_{i,k} \in \mathbb{C}^{M_i \times D_{i,k}}$ is the beamforming matrix between AP i and UE k, $\mathbf{z}_k \sim \mathcal{CN}\left(\mathbf{0}, \sigma_k{}^2 \mathbf{I}\right)$ is the additive Gaussian noise at UE k. Note that the first term in (1) represents the useful signals transmitted from APs $\mathcal{I}_k$ to UE k, while the second term is the inter-user interference.

For NCJT, successive interference cancellation (SIC) is usually adopted at the UEs to detect useful signals from their serving APs [8], [17]. With SIC, the achievable data rate of UE k is

$$R_k = \log \det \left(\mathbf{I} + \left(\sum_{i \in \mathcal{I}_k} \mathbf{H}_{i,k} \mathbf{P}_{i,k} \mathbf{P}_{i,k}^H \mathbf{H}_{i,k}^H \right) \mathbf{N}_k^{-1} \right) \quad (2)$$

where $\mathbf{N}_k \in \mathbb{C}^{N_k \times N_k}$ is the interference-plus-noise term

$$\mathbf{N}_k = \sum_{l \in \mathcal{U}_{-k}} \sum_{j \in \mathcal{I}_l} \mathbf{H}_{j,k} \mathbf{P}_{j,l} \mathbf{P}_{j,l}^H \mathbf{H}_{j,k}^H + \sigma_k^2 \mathbf{I}. \quad (3)$$

B. Problem Formulation

We aim to jointly optimize downlink beamforming and stream allocation to maximize the WSR of the user-centric cell-free MIMO networks with NCJT, subject to both individual power budgets at the APs and data stream limits at the UEs. The optimization problem is formulated as

$$\max_{\substack{\{\mathbf{P}_{i,k}\}, \\ \{D_{i,k}\}}} \sum_{k=1}^K \alpha_k \log \det \left(\mathbf{I} + \left(\sum_{i \in \mathcal{I}_k} \mathbf{H}_{i,k} \mathbf{P}_{i,k} \mathbf{P}_{i,k}^H \mathbf{H}_{i,k}^H \right) \mathbf{N}_k^{-1} \right) \quad (4a)$$

$$\text{s.t.} \quad \sum_{k \in \mathcal{U}_i} \text{Tr}\left(\mathbf{P}_{i,k} \mathbf{P}_{i,k}^H \right) \leq P_{\max,i}, \ \forall i, \quad (4b)$$

$$D^k \leq N_k, \ \forall k \quad (4c)$$

where $\alpha_k > 0$ is the weight of the data rate on UE k, $P_{\max,i}$ is the power budget of AP i, and $D^k = \sum_{i \in \mathcal{I}_k} D_{i,k}$ denotes the total number of receive streams at UE k.

The joint beamforming design and stream allocation problem (4) is non-convex and mixed-integer in nature, which is known to be NP-hard. Furthermore, each beamforming matrix $\mathbf{P}_{i,k}$ is tightly coupled with the number of data streams $D_{i,k}$, since it determines the dimension of $\mathbf{P}_{i,k}$. These bring new challenges to the algorithm design to decouple the optimization variables $\{\mathbf{P}_{i,k}\}$, $\{D_{i,k}\}$, and solve (4) with low computational complexity.

We note here that existing works [2], [8], [14], [17] on WSR maximization for NCJT treat the pure beamforming optimization problem with fixed data streams $\{D_{i,k}\}$ as an NP-hard

problem, and their beamforming algorithms for NCJT in cell-free networks are centralized approaches. This necessitates that all beamforming calculations be performed by the CUs, creating unavoidable computational complexity and communication overhead between APs and their connected CUs. In contrast, we propose the RWMMSE beamforming algorithm in Section IV that achieves lower computational complexity and less interaction, without any WSR compromise.

IV. BEAMFORMING DESIGN WITH FIXED STREAMS

It is challenging to directly solve the joint optimization problem (4) due to the coupled decision variables $\mathbf{P}_{i,k}$ and $D_{i,k}$. In this section, we first consider the fixed $D_{i,k}$ case to draw theoretical insights on the low-dimension beamforming structures.

With fixed $\{D_{i,k}\}$ in (4), the problem of beamforming design for WSR maximization subject to per-AP power constraints can be reduced to

$$\max_{\{\mathbf{P}_{i,k}\}} \sum_{k=1}^K \alpha_k \log \det \left(\mathbf{I} + \left(\sum_{i \in \mathcal{I}_k} \mathbf{H}_{i,k} \mathbf{P}_{i,k} \mathbf{P}_{i,k}^H \mathbf{H}_{i,k}^H \right) \mathbf{N}_k^{-1} \right) \quad (5)$$

$$\text{s.t.} \quad (4b).$$

In the following, we propose three beamforming algorithms for solving problem (5): 1) centralized WMMSE algorithm, by firstly demonstrating the applicability of the classic WMMSE approach in NCJT; 2) distributed low-interaction RWMMSE algorithm, via studying the beamformer structure for the WSR maximization problem (5); [1] and 3) fully distributed Local EZF method, which can be used to initialize the RWMMSE algorithm.

A. Centralized WMMSE Algorithm

Prior works [8] and [17] assume that the efficient WMMSE approach that has been widely adopted in CJT is not applicable to NCJT. Their proposed SCA based algorithms require solving transformed SOCP problems via the interior-point method in CVX, which is of high computational cost. We show here that, based on our unique observations of the WSR maximization problem structure, the classic WMMSE approach is still applicable for NCJT.

In essence, the classic WMMSE algorithm is proposed to reformulate a non-convex Shannon capacity to a convex weighted mean square error (MSE) minimization problem by introducing auxiliary variables [21]. Complicated WSR optimization problems can then be solved by the BCD method. We reveal the following lemma in [27] to illustrate the equivalent transformation in the WMMSE approach.

Lemma 1. (WMMSE Transformation [27, Lemma 4.1]) For any $\mathbf{A} \in \mathbb{C}^{n \times p}$, $\mathbf{B} \in \mathbb{C}^{p \times m}$ and $\mathbf{N} \succ \mathbf{0} \in \mathbb{C}^{n \times n}$, the following transformation holds:

$$\log \det(\mathbf{I} + \mathbf{A} \mathbf{B} \mathbf{B}^H \mathbf{A}^H \mathbf{N}^{-1})$$
$$= \max_{\mathbf{W} \succ \mathbf{0}, \mathbf{U}} \log \det(\mathbf{W}) - \text{Tr}(\mathbf{W} \mathbf{E}(\mathbf{U}, \mathbf{B})) + m \quad (6)$$

---

[1]Our proposed RWMMSE algorithm is distributed, in the sense that all the beamformers at the APs are *locally* obtained, instead of directly obtaining from the CUs. Furthermore, the RWMMSE algorithm does not require raw CSI and beamformer exchange between the APs and CUs.

where $\mathbf{W} \in \mathbb{C}^{m \times m}$ and $\mathbf{U} \in \mathbb{C}^{n \times m}$ are auxiliary variables, the MSE matrix $\mathbf{E} \in \mathbb{C}^{m \times m}$ is defined as

$$\mathbf{E}(\mathbf{U}, \mathbf{B}) \triangleq \left(\mathbf{I} - \mathbf{U}^H \mathbf{A}\mathbf{B}\right)\left(\mathbf{I} - \mathbf{U}^H \mathbf{A}\mathbf{B}\right)^H + \mathbf{U}^H \mathbf{N}\mathbf{U}. \quad (7)$$

The optimal $\mathbf{U}^*$ and $\mathbf{W}^*$ for (6) are given by

$$\mathbf{U}^* = \left(\mathbf{N} + \mathbf{A}\mathbf{B}\mathbf{B}^H \mathbf{A}^H\right)^{-1} \mathbf{A}\mathbf{B}, \quad (8)$$

and

$$\mathbf{W}^* = \left(\mathbf{I} - \mathbf{U}^H \mathbf{A}\mathbf{B}\right)^{-1}. \quad (9)$$

Remark 1. Prior works assume that the WMMSE approach is not applicable to the WSR maximization problem in NCJT with the unique data rate structure of $\log\left(1 + \mathbf{b}^H \mathbf{A}\mathbf{b}n^{-1}\right)$, *since the rank of* $\mathbf{A} = \bar{\mathbf{a}}\bar{\mathbf{a}}^H$ *is generally higher than one* [8], [17]. However, following the fact that $1 + \mathbf{v}^H \mathbf{u} = \det\left(\mathbf{I} + \mathbf{u}\mathbf{v}^H\right)$, the above data rate structure can be subtly rewritten as $\log \det\left(\mathbf{I} + \bar{\mathbf{a}}^H \mathbf{b}\mathbf{b}^H \bar{\mathbf{a}}n^{-1}\right)$. The WMMSE transformation in (6) of **Lemma 1** can then be applied as follows.

For ease of illustration, we first define the concatenation of the channel matrices and beamformers from APs $\mathcal{I}_k$ to UE k as

$$\mathbf{H}_k \triangleq \left[\mathbf{H}_{i_1,k}, \mathbf{H}_{i_2,k}, \dots, \mathbf{H}_{i_{|\mathcal{I}_k|},k}\right] \in \mathbb{C}^{N_k \times M^k}, \quad (10)$$

and

$$\mathbf{P}_k \triangleq \mathrm{blkdiag}\left(\mathbf{P}_{i_1,k}, \mathbf{P}_{i_2,k}, \dots, \mathbf{P}_{i_{|\mathcal{I}_k|},k}\right) \in \mathbb{C}^{M^k \times D^k} \quad (11)$$

where $M^k = \sum_{i \in \mathcal{I}_k} M_i$, and $|\mathcal{I}_k|$ is the number of serving APs $\mathcal{I}_k$ to UE k. Then problem (5) can be rewritten to yield the structure of the WMMSE transformation as

$$\max_{\{\mathbf{P}_{i,k}\}} \sum_{k=1}^{K} \alpha_k \log \det\left(\mathbf{I} + \mathbf{H}_k \mathbf{P}_k \mathbf{P}_k^H \mathbf{H}_k^H \mathbf{N}_k^{-1}\right) \quad (12)$$
$$\text{s. t.} \quad (4b).$$

Following the WMMSE transformation in (6) in **Lemma 1**, problem (12) can be equivalently transformed to

$$\max_{\substack{\{\mathbf{P}_{i,k}\}, \{\mathbf{U}_k\} \\ \{\mathbf{W}_k \succ \mathbf{0}\}}} \sum_{k=1}^{K} \alpha_k (\log \det(\mathbf{W}_k) - \mathrm{Tr}\left(\mathbf{W}_k \mathbf{E}_k(\mathbf{U}_k, \mathbf{P})\right)) \quad (13)$$
$$\text{s. t.} \quad (4b)$$

where $\mathbf{P} = \{\mathbf{P}_{i,k}\}$, $\mathbf{U}_k \in \mathbb{C}^{N_k \times D^k}$ and $\mathbf{W}_k \succ \mathbf{0} \in \mathbb{C}^{D^k \times D^k}$ are auxiliary variables, the MSE matrix $\mathbf{E}_k \in \mathbb{C}^{D^k \times D^k}$ is defined as

$$\mathbf{E}_k\left(\mathbf{U}_k, \mathbf{P}\right) \triangleq \left(\mathbf{I} - \mathbf{U}_k^H \mathbf{H}_k \mathbf{P}_k\right)\left(\mathbf{I} - \mathbf{U}_k^H \mathbf{H}_k \mathbf{P}_k\right)^H \quad (14)$$
$$+ \mathbf{U}_k^H \mathbf{N}_k \mathbf{U}_k.$$

Note that problem (13) is convex with respect to (w.r.t.) $\{\mathbf{U}_k\}$, $\{\mathbf{W}_k\}$ and $\{\mathbf{P}_{i,k}\}$, respectively. From (8) and (9) in **Lemma 1**, the optimal $\{\mathbf{U}_k^*\}$ and $\{\mathbf{W}_k^*\}$ with fixed $\{\mathbf{P}_{i,k}\}$ can be updated by

$$\mathbf{U}_k^* = \left(\mathbf{N}_k + \mathbf{H}_k \mathbf{P}_k \mathbf{P}_k^H \mathbf{H}_k^H\right)^{-1} \mathbf{H}_k \mathbf{P}_k, \; \forall k, \quad (15)$$

and

$$\mathbf{W}_k^* = \left(\mathbf{I} - (\mathbf{U}_k^*)^H \mathbf{H}_k \mathbf{P}_k\right)^{-1}, \; \forall k. \quad (16)$$

Algorithm 1 Centralized WMMSE Algorithm

Interact: Each AP i sends its $\{\mathbf{H}_{i,k}\}$ to the connected CUs;
1: Initialize $\{\mathbf{P}_{i,k}\}$ in CUs;
2: **Repeat** ←[CUs]
3:      Update $\{\mathbf{U}_k^*\}$ via (15);
4:      Update $\{\mathbf{W}_k^*\}$ via (16);
5:      Update $\{\mathbf{P}_{i,k}^*\}$ via (17);
6: **Until** Converge
Interact: CUs transmit $\{\mathbf{P}_{i,k}^*, k \in \mathcal{U}_i\}$ to each AP i.

Fixing $\{\mathbf{U}_k^*\}$ and $\{\mathbf{W}_k^*\}$, the WMMSE update of $\{\mathbf{P}_{i,k}^*\}$ is given by

$$\mathbf{P}_{i,k}^* = \left(\sum_{l \in \mathcal{U}} \alpha_l \mathbf{H}_{i,l}^H \mathbf{A}_l \mathbf{H}_{i,l} + \mu_i \mathbf{I}\right)^{-1} \alpha_k \mathbf{H}_{i,k}^H \mathbf{U}_k^* \mathbf{W}_k^* \mathbf{\Xi}_{i,k}^H \quad (17)$$

where $\mu_i \geq 0$ is a Lagrangian constant for the maximum transmit power constraint (4b). In **Algorithm 1**, we summarize our proposed WMMSE algorithm, which requires centralized computation of $\{\mathbf{P}_{i,k}^*\}$ at the CUs.

Remark 2. For WSR maximization in NCJT, the BRnB method in [17] has an $\mathcal{O}\left(\left(\sum_{i \in \mathcal{I}} M_i^2\right)^3\right)$ computational complexity. The SCA based algorithm in [8] achieves a lower $\mathcal{O}\left(\left(\sum_{i \in \mathcal{I}} M_i \sum_{k \in \mathcal{U}_i} D_{i,k}\right)^3\right)$ computational complexity. The InAP and ADMM combined method in [17] has a even lower $\mathcal{O}\left(\left(\sum_{i \in \mathcal{I}} M_i\right)^3\right)$ complexity. The FP based algorithm in [2] achieves the current lowest $\mathcal{O}(\sum_{i \in \mathcal{I}} M_i^3)$ computational complexity for the single UE receive antenna case. Our proposed WMMSE algorithm achieves the same $\mathcal{O}(\sum_{i \in \mathcal{I}} M_i^3)$ computational complexity, for the more general case with multiple UE receive antennas.

B. Distributed Low-Interaction RWMMSE Algorithm

The current lowest $\mathcal{O}(\sum_{i \in \mathcal{I}} M_i^3)$ computational complexity of our proposed WMMSE algorithm can still be high, especially for MIMO systems with massive transmit antennas. Moreover, the centralized WMMSE algorithm requires each AP i to communicate the channel matrix $\{\mathbf{H}_{i,k}, k \in \mathcal{U}_i\}$ to the CUs, and the CUs to communicate the beamforming matrix $\{\mathbf{P}_{i,k}^*, k \in \mathcal{U}_i\}$ back to each AP i, causing high interaction between the APs and the CUs. These motivate us to explore low-complexity and low-interaction beamforming algorithms.

In the following, we draw some theoretical insights on the structures of the BCD solutions to problem (5), which will be leveraged later to further reduce the algorithm complexity and interaction.

1) Properties of Beamforming for WSR Maximizatoin: We now look through the inherent properties of the WSR maximization problem (5) to study the structure of $\mathbf{P}_{i,k}^*$. For ease of expression, we concat the channel matrices between AP i and all the UEs as follows

$$\bar{\mathbf{H}}_i \triangleq \left[\mathbf{H}_{i,1}^H, \dots, \mathbf{H}_{i,K}^H\right]^H \in \mathbb{C}^{\sum_{k \in \mathcal{U}} N_k \times M_i}. \quad (18)$$

The following **Theorem 1** shows if $\bar{\mathbf{H}}_i$ is of full column rank, the beamformer $\mathbf{P}_{i,k}^*$ for WSR maximization consumes full transmit power.

Theorem 1. (Full Power Property): For any AP i, if $M_i \geq \sum_{k \in \mathcal{U}} N_k$ and $\bar{\mathbf{H}}_i^H$ consists of $\sum_{k \in \mathcal{U}} N_k$ linearly independent vectors of size $M_i \times 1$, the local optimal beamformer $\mathbf{P}_{i,k}^*$ always satisfies the full transmit power constraint, i.e. $\sum_{k \in \mathcal{U}_i} \left\| \mathbf{P}_{i,k}^* \right\|_{\mathrm{F}}^2 = P_{\max,i}$.

Proof: See Appendix A. ∎

The following proposition shows another property on the power consumption of the beamformer $\mathbf{P}_{i,k}^*$, i.e., the column vectors of $\mathbf{P}_{i,k}^*$ that are orthogonal to the channel matrix $\bar{\mathbf{H}}_i$ may consume power but do not contribute to the WSR.

Proposition 1. (Null Space Property): The part of the beamformer $\mathbf{P}_{i,k}^*$ in the null space of $\bar{\mathbf{H}}_i^H$ consumes power but does not contributes to the WSR.

Proof: See Appendix B. ∎

The following theorem shows that each beamformer $\mathbf{P}_{i,k}^*$ for WSR maximization has a low-dimensional substitution.

Theorem 2. (Low-Dimension Substitution): Any local optimal solution $\{\mathbf{P}_{i,k}^* \in \mathbb{C}^{M_i \times D_{i,k}}, k \in \mathcal{U}_i\}$ to problem (12) must lie in the column space of $\bar{\mathbf{H}}_i^H$, i.e.,

$$\mathbf{P}_{i,k}^* = \bar{\mathbf{H}}_i^H \mathbf{X}_{i,k}^*, \forall k \qquad (19)$$

where $\mathbf{X}_{i,k}^* \in \mathbb{C}^{\sum_{k \in \mathcal{U}} N_k \times D_{i,k}}$ is a low-dimension substitution of $\mathbf{P}_{i,k}^*$.

Proof: See Appendix C. ∎

Remark 3. (Low-Complexity and Low-Interaction Properties): By investigating the structure of the beamformer $\mathbf{P}_{i,k}^*$ in (19) of **Theorem 2**, solution to $\mathbf{P}_{i,k}^*$ in (17) with the size of $M_i \times D_{i,k}$ can be equivalently reduced to $\mathbf{X}_{i,k}^*$ with the size of $\sum_{k \in \mathcal{U}} N_k \times D_{i,k}$. In addition, the interaction between the APs and their connected CUs is changed from $\mathbf{P}_{i,k}^*$ to $\mathbf{X}_{i,k}^*$. By solving for $\mathbf{X}_{i,k}^*$ instead of $\mathbf{P}_{i,k}^*$, both the computation cost and the communication overhead can be substantially reduced, especially for the networks with a large number of transmit antennas $M_i \gg \sum_{k \in \mathcal{U}} N_k$.

Remark 4. The full power and the low-dimension subspace properties presented in **Theorem 1** and **Theorem 2** are substantially different from the results in [28] in the following aspects. First, the findings in [28] are limited to CJT, which cannot be directly applied to NCJT in this work. Second, the results in [28] are limited to the *single*-AP case, while we study the more general cell-free networks with *multiple* APs. Note that the proof of the full power property in [28] relies on the contradiction of one single Lagrangian multiplier in the Karush-Kuhn-Tucker (KKT) conditions, which cannot be directly applied to the case of multiple APs. Finally, our proofs based on the linear independence of columns in $\bar{\mathbf{H}}_i^H$ avoids the tedious discussion on the KKT conditions in [28].

2) Problem Reformulation: We now utilize the low-dimensional substitution property in **Theorem 2** to reformulate problem (12). From the low-dimension substitution (19) and the definition of $\mathbf{P}_k$ in (11), we have

$$\mathbf{P}_k = \tilde{\mathbf{H}}_{\mathcal{I}_k}^H \tilde{\mathbf{X}}_{\mathcal{I}_k} \qquad (20)$$

where $\tilde{\mathbf{H}}_{\mathcal{I}_k} \triangleq \mathrm{blkdiag}\left(\bar{\mathbf{H}}_{i_1}, \ldots, \bar{\mathbf{H}}_{i_{|\mathcal{I}_k|}}\right) \in \mathbb{C}^{|\mathcal{I}_k| \sum_{k \in \mathcal{U}} N_k \times M^k}$ and $\tilde{\mathbf{X}}_{\mathcal{I}_k} \triangleq \mathrm{blkdiag}\left(\mathbf{X}_{i_1,k}, \ldots, \mathbf{X}_{i_{|\mathcal{I}_k|},k}\right) \in \mathbb{C}^{|\mathcal{I}_k| \sum_{k \in \mathcal{U}} N_k \times D^k}$ are the concatenations of the channel matrices $\{\bar{\mathbf{H}}_i, i \in \mathcal{I}_k\}$ and the low-dimension substitutions $\{\mathbf{X}_{i,k}, i \in \mathcal{I}_k\}$.

Substituting (20) into the WSR maximization problem (12), we have a low-dimension substitution problem given by

$$\max_{\{\mathbf{X}_{i,k}\}} \sum_{k \in \mathcal{U}} \alpha_k \log \det\left(\mathbf{I} + \mathbf{H}_k \tilde{\mathbf{H}}_{\mathcal{I}_k}^H \tilde{\mathbf{X}}_{\mathcal{I}_k} \tilde{\mathbf{X}}_{\mathcal{I}_k}^H \tilde{\mathbf{H}}_{\mathcal{I}_k} \mathbf{H}_k^H \mathbf{N}_k^{-1}\right) \quad (21a)$$

$$\mathrm{s.\,t.} \quad \sum_{k \in \mathcal{U}_i} \mathrm{Tr}\left(\bar{\mathbf{H}}_i^H \mathbf{X}_{i,k} \mathbf{X}_{i,k}^H \bar{\mathbf{H}}_i\right) \leq P_{\max,i}, \ \forall i. \quad (21b)$$

Following the WMMSE transformation in **Lemma 1**, problem (21) can be equivalently transformed to

$$\max_{\substack{\{\mathbf{X}_{i,k}\}, \{\mathbf{U}_k\} \\ \{\mathbf{W}_k \succ \mathbf{0}\}}} \sum_{k \in \mathcal{U}} \alpha_k (\log \det(\mathbf{W}_k) - \mathrm{Tr}\left(\mathbf{W}_k \mathbf{E}_k(\mathbf{U}_k, \mathbf{X})\right)) \qquad (22)$$

$$\mathrm{s.\,t.} \quad (21b)$$

where $\mathbf{X} \triangleq \{\mathbf{X}_{i,k}\}$ is the set of low-dimension beamformer substitution, and the MSE matrix is defined as

$$\mathbf{E}_k(\mathbf{U}_k, \mathbf{X}) \triangleq \left(\mathbf{I} - \mathbf{U}_k^H \mathbf{H}_k \tilde{\mathbf{H}}_{\mathcal{I}_k}^H \tilde{\mathbf{X}}_{\mathcal{I}_k}\right)\left(\mathbf{I} - \mathbf{U}_k^H \mathbf{H}_k \tilde{\mathbf{H}}_{\mathcal{I}_k}^H \tilde{\mathbf{X}}_{\mathcal{I}_k}\right)^H + \mathbf{U}_k^H \mathbf{N}_k \mathbf{U}_k. \qquad (23)$$

Similar to the BCD updates in (15) and (16), with fixed $\mathbf{X}$ we update $\{\mathbf{U}_k^*\}$ and $\{\mathbf{W}_k^*\}$ by replacing $\mathbf{P}_k$ with its low-dimension substitution in (20) as

$$\mathbf{U}_k^* = \left(\mathbf{N}_k + \mathbf{H}_k \tilde{\mathbf{H}}_{\mathcal{I}_k}^H \tilde{\mathbf{X}}_{\mathcal{I}_k} \tilde{\mathbf{X}}_{\mathcal{I}_k}^H \tilde{\mathbf{H}}_{\mathcal{I}_k} \mathbf{H}_k^H\right)^{-1} \mathbf{H}_k \tilde{\mathbf{H}}_{\mathcal{I}_k}^H \tilde{\mathbf{X}}_{\mathcal{I}_k}, \forall k, \quad (24)$$

and

$$\mathbf{W}_k^* = \left(\mathbf{I} - (\mathbf{U}_k^*)^H \mathbf{H}_k \tilde{\mathbf{H}}_{\mathcal{I}_k}^H \tilde{\mathbf{X}}_{\mathcal{I}_k}\right)^{-1}, \ \forall k. \quad (25)$$

Fixing $\{\mathbf{U}_k^*\}$ and $\{\mathbf{W}_k^*\}$, optimization problem (22) can be reduced to

$$\min_{\{\mathbf{X}_{i,k}\}} \sum_{k \in \mathcal{U}} \alpha_k \mathrm{Tr}\left(\mathbf{W}_k \mathbf{E}_k\left(\mathbf{U}_k^*, \mathbf{X}\right)\right) \qquad (26)$$

$$\mathrm{s.\,t.} \quad (21b).$$

Dropping the content terms in the objective of problem (26), we have

$$\min_{\{\mathbf{X}_{i,k}\}} \sum_{k \in \mathcal{U}} \alpha_k \sum_{i \in \mathcal{I}_k} \mathrm{Tr}\left(\mathbf{X}_{i,k}^H \bar{\mathbf{H}}_i \mathbf{H}_{i,k}^H \mathbf{A}_k \mathbf{H}_{i,k} \bar{\mathbf{H}}_i^H \mathbf{X}_{i,k}\right)$$
$$- 2 \sum_{k \in \mathcal{U}} \alpha_k \sum_{i \in \mathcal{I}_k} \mathrm{Re}\left\{\mathrm{Tr}\left(\mathbf{\Xi}_{i,k} \mathbf{W}_k^* (\mathbf{U}_k^*)^H \mathbf{H}_{i,k} \bar{\mathbf{H}}_i^H \mathbf{X}_{i,k}\right)\right\}$$
$$+ \sum_{k \in \mathcal{U}} \alpha_k \mathrm{Tr}\left(\mathbf{A}_k \sum_{j \in \mathcal{I}_l} \sum_{l \in \mathcal{U}_{-k}} \mathbf{H}_{j,k} \bar{\mathbf{H}}_j^H \mathbf{X}_{j,l} \bar{\mathbf{H}}_j^H \mathbf{X}_{j,l}^H \mathbf{H}_{j,k}^H\right) \qquad (27)$$
$$\mathrm{s.\,t.} \quad (21b)$$

where $\mathbf{A}_k \triangleq \mathbf{U}_k \mathbf{W}_k \mathbf{U}_k^H \in \mathbb{C}^{N_k \times N_k}$ and $\mathbf{\Xi}_{i,k} \in \mathbb{B}^{D_{i,k} \times D^k}$ is a binary matrix with the $(i-1) \times D_{i,k} + 1$ to $i \times D_{i,k}$ columns of $\mathbf{\Xi}_{i,k}$ being $\mathbf{I}_{D_{i,k}}$ and the rest of elements being zero when $i \in \mathcal{I}_k$, otherwise $\mathbf{\Xi}_{i,k} = \mathbf{0}$ for $i \notin \mathcal{I}_k$.

We note here that the updates of $\{\mathbf{X}_{i,k}\}$ in (27) can be decoupled across APs, each AP i solves its own optimization problem

$$
\begin{aligned}
\min_{\{\mathbf{X}_{i,k}, k \in \mathcal{U}_i\}} & \sum_{l \in \mathcal{U}} \sum_{k \in \mathcal{U}_i} \alpha_l \mathrm{Tr}\left(\mathbf{A}_l \mathbf{H}_{i,l} \bar{\mathbf{H}}_i^H \mathbf{X}_{i,k} \mathbf{X}_{i,k}^H \bar{\mathbf{H}}_i \mathbf{H}_{i,l}^H\right) \\
& -2 \sum_{k \in \mathcal{U}_i} \alpha_k \mathrm{Re}\left\{\mathrm{Tr}\left(\boldsymbol{\Xi}_i \mathbf{W}_k^* (\mathbf{U}_k^*)^H \mathbf{H}_{i,k} \bar{\mathbf{H}}_i^H \mathbf{X}_{i,k}\right)\right\} \quad (28)
\end{aligned}
$$

s. t.   (21b).

Different from the sequential beamforming updates in [8] and [14], the APs can update their own beamformer $\mathbf{P}_{i,k}$ (WMMSE) or $\mathbf{X}_{i,k}$ (RWMMSE) by solving problem (28) in parallel.

Problem (28) is convex quadratic w.r.t. $\mathbf{X}_{i,k}$. Using the Lagrangian multiplier method, the optimal solution $\{\mathbf{X}_{i,k}^*\}$ to problem (28) is given by

$$
\begin{aligned}
\mathbf{X}_{i,k}^* = & \left(\sum_{l \in \mathcal{U}} \alpha_l \bar{\mathbf{H}}_i \mathbf{H}_{i,l}^H \mathbf{A}_l \mathbf{H}_{i,l} \bar{\mathbf{H}}_i^H + \lambda_i \bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H\right)^{-1} \\
& \times \alpha_k \bar{\mathbf{H}}_i \mathbf{H}_{i,k}^H \mathbf{U}_k^* \mathbf{W}_k^* \boldsymbol{\Xi}_{i,k}^H
\end{aligned} \quad (29)
$$

where $\lambda_i \geq 0$ is the Lagrangian multiplier that can also be obtained by the bisection method. The detailed RWMMSE process is described in **Algorithm** 2.

Remark 5. (Low-Interaction Execution of the RWMMSE Algorithm): When executing the RWMMSE algorithm, each AP i first transmits $\{\bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H \in \mathbb{C}^{\sum_{k \in \mathcal{U}} N_k \times \sum_{k \in \mathcal{U}} N_k}\}$ (with lower dimension than the interaction $\{\bar{\mathbf{H}}_i\}$ in the WMMSE algorithm) to the connected CUs. The CUs then transmit $\{\mathbf{X}_{i,k}^* \in \mathbb{C}^{M_i \times D_{i,k}}, k \in \mathcal{U}_i\}$ back to their connected APs, while the WMMSE approach need to communicate complete beamformer $\{\mathbf{P}_{i,k}^*, k \in \mathcal{U}_i\}$ with higher dimension. Finally, each AP i calculates their beamforming matrices via (19) locally.

Remark 6. The RWMMSE algorithm is equivalent to the centralized WMMSE algorithm in Section IV-A. Due to the fact that $(\mathbf{I} + \mathbf{AB})^{-1} \mathbf{A} = \mathbf{A}(\mathbf{I} + \mathbf{BA})^{-1}$, the WMMSE beamformer update $\mathbf{P}_{i,k}^*$ in (17) can be rewritten to follow the low-dimension substitute property in **Theorem** 2. Specifically, the concatenation of $\{\mathbf{P}_{i,k}^*, k \in \mathcal{U}\}$ can be expressed as

$$
\begin{aligned}
\mathbf{P}_i^* = & \bar{\mathbf{H}}_i^H \underbrace{\mathbf{U}\left(\boldsymbol{\Omega}\mathbf{U}^H \bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H \mathbf{U} + \mu_i \mathbf{W}^{-1}\right)^{-1} \boldsymbol{\Omega} \bar{\boldsymbol{\Xi}}_i^H}_{\mathbf{X}_i} \\
= & \left[\mathbf{P}_{i,1}^*, \mathbf{P}_{i,2}^*, \ldots, \mathbf{P}_{i,K}^*\right] \in \mathbb{C}^{M_i \times \sum_{k \in \mathcal{U}} D_{i,k}}
\end{aligned} \quad (30)
$$

where we define $\mathbf{U} \triangleq \mathrm{blkdiag}\left(\mathbf{U}_1^*, \ldots, \mathbf{U}_K^*\right) \in \mathbb{C}^{\sum_{k \in \mathcal{U}} N_k \times \sum_{k \in \mathcal{U}} D^k}$, $\mathbf{W} \triangleq \mathrm{blkdiag}\left(\mathbf{W}_1^*, \ldots, \mathbf{W}_K^*\right) \in \mathbb{C}^{\sum_{k \in \mathcal{U}} D^k \times \sum_{k \in \mathcal{U}} D^k}$, $\boldsymbol{\Omega} \triangleq \mathrm{blkdiag}\left(\alpha_1 \mathbf{I}_{D^1}, \ldots, \alpha_K \mathbf{I}_{D^K}\right) \in \mathbb{R}^{\sum_{k \in \mathcal{U}} D^k \times \sum_{k \in \mathcal{U}} D^k}$, and $\bar{\boldsymbol{\Xi}}_i \triangleq \mathrm{blkdiag}\left(\boldsymbol{\Xi}_{i,1}, \ldots, \boldsymbol{\Xi}_{i,K}\right) \in \mathbb{R}^{\sum_{k \in \mathcal{U}} D_{i,k} \times \sum_{k \in \mathcal{U}} D^k}$.

C. Fully Distributed Local EZF Algorithm

Motivated by the unique property of NCJT that signals transmitted from the serving APs of a UE are independent, we extend the EZF method for CJT that requires CSI interaction across the network [28], to a fully distributed EZF method

---

Algorithm 2 Distributed RWMMSE Algorithm

1: Each AP i calculates $\{\bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H\}$;
Interact: Each AP i sends $\{\bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H\}$ to the connected CUs;
2: Initialize $\{\mathbf{X}_{i,k}\}$ in CUs;
3: **Repeat** ←[CUs]
4:     Update $\{\mathbf{U}_k^*\}$ via (15);
5:     Update $\{\mathbf{W}_k^*\}$ via (16);
6:     Update $\{\mathbf{X}_{i,k}^*\}$ via (29);
7: **Until** Converge
Interact: CUs transmit $\{\mathbf{X}_{i,k}^*, k \in \mathcal{U}_i\}$ to the connected AP i;
Output: Each AP i calculates $\{\mathbf{P}_{i,k}^* = \bar{\mathbf{H}}_i^H \mathbf{X}_{i,k}^*, k \in \mathcal{U}_i\}$.

---

for NCJT without any CSI interaction. Our proposed fully distributed EZF algorithm can be used to initialize the iteration of the RWMMSE algorithm for better system performance.

1) Local EZF Method: The idea of the EZF method is to perform singular value decomposition (SVD) on the channel matrix first, and then does zero-forcing (ZF) on the effective channel matrix composed of only the right singular vectors.

Specifically, each AP i performs SVD on $\{\mathbf{H}_{i,k}, k \in \mathcal{U}_i\}$ to get its right singular vectors

$$
\mathbf{H}_{i,k} = \bar{\mathbf{U}}_{i,k} \bar{\boldsymbol{\Sigma}}_{i,k} \bar{\mathbf{V}}_{i,k}^H \quad (31)
$$

where $\bar{\mathbf{U}}_{i,k} \in \mathbb{C}^{N_k \times N_k}$ and $\bar{\mathbf{V}}_{i,k} \in \mathbb{C}^{M_i \times N_k}$ are unitary matrices of the left and right singular vectors, and $\bar{\boldsymbol{\Sigma}}_{i,k} \in \mathbb{C}^{N_k \times N_k}$ is a diagonal matrix with descending singular values. The effective channel of AP i $\bar{\mathbf{V}}_i \in \mathbb{C}^{M_i \times \sum_{k \in \mathcal{U}_i} D_{i,k}}$ can be formed by concating the first $D_{i,k}$ columns of $\{\bar{\mathbf{V}}_{i,k}, k \in \mathcal{U}_i\}$ as

$$
\bar{\mathbf{V}}_i = \left[\bar{\mathbf{V}}_{i,k_1}(:, 1:D_{i,k_1}), \ldots, \bar{\mathbf{V}}_{i,k_{|\mathcal{U}_i|}}\left(:, 1:D_{i,k_{|\mathcal{U}_i|}}\right)\right]. \quad (32)
$$

Finally, each AP i performs ZF and power scaling on $\bar{\mathbf{V}}_i$ in parallel. The Local EZF beamformer of AP i is given by

$$
\begin{aligned}
\mathbf{P}_i^{\text{Local EZF}} &= \frac{\left(\bar{\mathbf{V}}_i^H\right)^\dagger}{\left\|\left(\bar{\mathbf{V}}_i^H\right)^\dagger\right\|_{\mathrm{F}}} \sqrt{P_{\max,i}} \\
&= \left[\mathbf{P}_{i,k_1}, \mathbf{P}_{i,k_2}, \ldots, \mathbf{P}_{i,k_{|\mathcal{U}_i|}}\right], \forall i.
\end{aligned} \quad (33)
$$

We present our Local EZF method for NCJT in **Algorithm** 3.

2) Initialization of the RWMMSE Algorithm: We can show that the Local EZF beamformer also conforms **Theorem** 2, i.e. $\{\mathbf{P}_{i,k}^{\text{Local EZF}}, k \in \mathcal{U}_i\}$ are in the column space of $\bar{\mathbf{H}}_i$. Therefore, the local EZF beamformer also has its low-dimension substitution, which can be used to initialize the RWMMSE algorithm. Specifically, the Local EZF beamformer $\mathbf{P}_i^{\text{Local EZF}}$ in (33) can be rewritten as

$$
\mathbf{P}_i^{\text{Local EZF}} = \tilde{\mathbf{H}}_i^H \bar{\mathbf{X}}_i^{\text{Local EZF}} \quad (34)
$$

where $\tilde{\mathbf{H}}_i = \left[\mathbf{H}_{i,k_1}^H, \mathbf{H}_{i,k_2}^H, \ldots, \mathbf{H}_{i,k_{|\mathcal{U}_i|}}^H\right]^H \in \mathbb{C}^{\sum_{k \in \mathcal{U}_i} N_k \times M_i}$ denotes the channel matrix between AP i and its serving UEs, and $\tilde{\mathbf{H}}_i^H$ can be expressed as

$$
\tilde{\mathbf{H}}_i^H = \left[\bar{\mathbf{V}}_{i,k_1} \bar{\boldsymbol{\Sigma}}_{i,k_1}^H \bar{\mathbf{U}}_{i,k_1}^H, \ldots, \bar{\mathbf{V}}_{i,k_{|\mathcal{U}_i|}} \bar{\boldsymbol{\Sigma}}_{i,k_{|\mathcal{U}_i|}}^H \bar{\mathbf{U}}_{i,k_{|\mathcal{U}_i|}}^H\right], \quad (35)
$$

Algorithm 3 Fully Distributed Local EZF Algorithm

1: At each AP i, **do** the following:
2: Perform SVD on $\{\mathbf{H}_{i,k}, k \in \mathcal{U}_i\}$;
3: Concat $\bar{\mathbf{V}}_i$ via (32) as its effective channel;
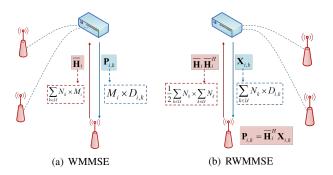4: Calculate the Local EZF beamformer $\mathbf{P}_i^{\text{Local EZF}}$ via (33).

(a) WMMSE (b) RWMMSE

Fig. 2. An exemplary illustration of the interaction between the APs and the CUs.

and the block diagonal matrix $\bar{\mathbf{X}}_i^{\text{Local EZF}}$ is a low-dimension substitution of $\mathbf{P}_i^{\text{Local EZF}}$, given by

$$
\begin{aligned}
&\bar{\mathbf{X}}_i^{\text{Local EZF}} \\
&= \text{blkdiag}\left(\bar{\mathbf{U}}_{i,k_1}\left(\bar{\mathbf{\Sigma}}_{i,k_1}^H\right)^{-1}, \dots, \bar{\mathbf{U}}_{i,k_{|\mathcal{U}_i|}}\left(\bar{\mathbf{\Sigma}}_{i,k_{|\mathcal{U}_i|}}^H\right)^{-1}\right)\mathbf{\Gamma}_i
\end{aligned} \tag{36}
$$

with $\mathbf{\Gamma}_i = \dfrac{\left(\bar{\mathbf{V}}_i^H \bar{\mathbf{v}}_i\right)^{-1}}{\left\|\left(\bar{\mathbf{V}}_i^H\right)^\dagger\right\|_{\text{F}}} \sqrt{P_{\max,i}}.$

D. Interaction and Computational Complexity Analysis

In this subsection, we analyze the interaction and computational complexity of the WMMSE algorithm, the low-interaction RWMMSE algorithm, and the Local EZF algorithm. As mentioned in **Remark** 2, the current lowest computational complexity for WSR maximization in NCJT is $\mathcal{O}(M_i^3)$. Considering the general status of modern communication networks, we assume that the number of transmit antennas on AP i is greater than the total number of receive antennas, i.e., $M_i \geq \sum_{k \in \mathcal{U}} N_k$.

1) Interaction: In the WMMSE algorithm, each AP i needs to transmit their channel matrices $\{\bar{\mathbf{H}}_i \in \mathbb{C}^{\sum_{k \in \mathcal{U}} N_k \times M_i}\}$ to their connected CUs, and CUs need to transmit the beamformers $\{\mathbf{P}_{i,k}^* \in \mathbb{C}^{M_i \times D_{i,k}}\}$ back to their controlling APs. In the RWMMSE algorithm, the APs transmit $\{\bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H \in \mathbb{C}^{\sum_{k \in \mathcal{U}} N_k \times \sum_{k \in \mathcal{U}} N_k}\}$ to their connected CUs. Since $\bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H$ is a symmetric matrix, only its upper triangular part needs to be transmitted. The CUs then transmit the low-dimension substitution $\{\mathbf{X}_{i,k}^* \in \mathbb{C}^{\sum_{k \in \mathcal{U}} N_k \times D_{i,k}}\}$ back to their controlling APs. The Local EZF algorithm requires only local CSI, thereby eliminating the need for any interaction across networks. In Fig. 2, we illustrate the interaction of the proposed WMMSE and RWMMSE algorithms.

2) Computational Complexity: Each iteration in the WMMSE algorithm and the RWMMSE algorithm calculates matrix inverse of $\left(\mathbf{N}_k + \mathbf{H}_k \mathbf{P}_k \mathbf{P}_k^H \mathbf{H}_k^H\right) \in \mathbb{C}^{N_k \times N_k}$ and $\mathbf{E}_k^* \in \mathbb{C}^{D^k \times D^k}$ for $\mathbf{U}_k$ and $\mathbf{W}_k$ updates, respectively. Besides, in order to solve $\{\mathbf{P}_{i,k}^*\}$ (in the WMMSE algorithm) and $\{\mathbf{X}_{i,k}^*\}$ (in the RWMMSE algorithm), matrices inverse of $\left(\sum_{l \in \mathcal{U}} \alpha_l \mathbf{H}_{i,l}^H \mathbf{A}_l \mathbf{H}_{i,l} + \mu_i \mathbf{I}\right)$ and

	Interaction	Complexity
Local EZF	0	$\mathcal{O}\left(\left(\sum_{k \in \mathcal{U}_i} N_k\right)^2 M_i\right)$
WMMSE	$\left(\sum_{k \in \mathcal{U}} N_k + \sum_{k \in \mathcal{U}_i} D_{i,k}\right) M_i$	$\mathcal{O}\left(M_i^3\right)$
RWMMSE	$\left(\frac{1}{2}\sum_{k \in \mathcal{U}} N_k + \sum_{k \in \mathcal{U}_i} D_{i,k}\right)\sum_{k \in \mathcal{U}} N_k$	$\mathcal{O}\left(\left(\sum_{k \in \mathcal{U}} N_k\right)^2 M_i\right)$

$\left(\sum_{l \in \mathcal{U}} \alpha_l \bar{\mathbf{H}}_i \mathbf{H}_{i,l}^H \mathbf{A}_l \mathbf{H}_{i,l} \bar{\mathbf{H}}_i^H + \lambda_i \bar{\mathbf{H}}_i \bar{\mathbf{H}}_i^H\right)$ with dimensions $M_i \times M_i$ and $\sum_{k \in \mathcal{U}} N_k \times \sum_{k \in \mathcal{U}} N_k$ are required. The Local EZF method needs to perform SVD on the channel matrix $\mathbf{H}_{i,k} \in \mathbb{C}^{N_k \times M_i}$, and calculate pseudo-inverse of $\bar{\mathbf{V}}_i^H \in \mathbb{C}^{\sum_{k \in \mathcal{U}_i} D_{i,k} \times M_i}$.

See Table I for a summary of both the interaction and the computational complexity of the three proposed algorithms. We can see that the Local EZF method achieves the lowest interaction and computational complexity. Both the interaction and the computational complexity of the RWMMSE algorithm is lower than its equivalent WMMSE algorithm.

V. JOINT BEAMFORMING AND STREAM ALLOCATION

In this section, we investigate the joint beamforming and stream allocation problem (4) for WSR maximization in user-centric cell-free MIMO networks with NCJT. Due to the mix-integer and non-convexity nature, problem (4) is NP-hard and thus is challenging to solve in general. By studying the beamforming structure with varying data streams, we propose a low-complexity joint beamforming and linear stream allocation algorithm, termed as RWMMSE-LSA.

A. Decoupling Beamforming Design and Stream Allocation

For ease of illustration, we rewrite problem (4) in a concise form using the definitions of $\mathbf{H}_k$ and $\mathbf{P}_k$ as follows

$$
\max_{\{\mathbf{P}_{i,k}\},\{D_{i,k}\}} \sum_{k \in \mathcal{U}} \log\det\left(\mathbf{I} + \mathbf{H}_k \mathbf{P}_k \mathbf{P}_k^H \mathbf{H}_k^H \mathbf{N}_k^{-1}\right) \tag{37a}
$$

$$
\text{s. t.} \quad \text{(4b), (4c).} \tag{37b}
$$

The main difficulty of solving problem (37) roots in the coupled decision variables, i.e., the number of streams $D_{i,k}$ and the beamformer $\mathbf{P}_{i,k} \in \mathbb{C}^{M_i \times D_{i,k}}$. In particular, $D_{i,k}$ determines the size of the beamformer $\{\mathbf{P}_{i,k}\}$.

In order to decouple the decision variables $\{\mathbf{P}_{i,k}\}$ and $\{D_{i,k}\}$, we introduce a binary diagonal stream indicator matrix $\mathbf{L}_{i,k} \in \mathbb{B}^{N_k \times N_k}$ for transmit stream allocation and a virtual beamformer $\bar{\mathbf{P}}_{i,k} \in \mathbb{C}^{M_i \times N_k}$ with fixed dimension assuming UE k receives the maximum number of streams N_k from AP i. The actual total number of streams that AP i transmits to UE k is $\text{Tr}(\mathbf{L}_{i,k}) = D_{i,k}$. The columns in $\bar{\mathbf{P}}_{i,k}$ that are selected as the actual beamformer $\mathbf{P}_{i,k}$ (where the corresponding diagonal elements of $\mathbf{L}_{i,k}$ are one) can be represented by $\bar{\mathbf{P}}_{i,k}\mathbf{L}_{i,k}$.

Replacing $\mathbf{P}_{i,k}$ with $\bar{\mathbf{P}}_{i,k}\mathbf{L}_{i,k}$, the achievable data rate of UE k in (2) can be rewritten as

$$
\log\det\left(\mathbf{I} + \left(\sum_{i \in \mathcal{I}_k} \mathbf{H}_{i,k}\bar{\mathbf{P}}_{i,k}\mathbf{L}_{i,k}\mathbf{L}_{i,k}^H\bar{\mathbf{P}}_{i,k}^H\mathbf{H}_{i,k}^H\right)\bar{\mathbf{N}}_k^{-1}\right) \tag{38}
$$

where $\bar{\mathbf{N}}_k = \sum_{l \in \mathcal{U}_{-k}} \sum_{j \in \mathcal{I}_l} \mathbf{H}_{j,k}\bar{\mathbf{P}}_{j,l}\mathbf{L}_{j,l}\mathbf{L}_{j,l}^H\bar{\mathbf{P}}_{j,l}^H\mathbf{H}_{j,k}^H + \sigma_k^2\mathbf{I}.$

Problem (37) can then be equivalently transformed to

$$\max_{\substack{\{\bar{\mathbf{P}}_{i,k}\},\\ \{\mathbf{L}_{i,k}\}}} \quad \sum_{k\in\mathcal{U}} \alpha_k \log\det\left(\mathbf{I} + \mathbf{H}_k\bar{\mathbf{P}}_k\mathbf{L}_k\mathbf{L}_k\bar{\mathbf{P}}_k^H\mathbf{H}_k^H\bar{\mathbf{N}}_k^{-1}\right) \quad (39a)$$

$$\text{s.t.} \quad (4b), \quad (39b)$$

$$\sum_{i\in\mathcal{I}_k}\text{Tr}(\mathbf{L}_{i,k}) \le N_k, \ \forall k, \quad (39c)$$

$$\mathbf{L}_{i,k} = \text{diag}\left(l_{i,k}^{(1)},\ldots,l_{i,k}^{(N_k)}\right), \ \forall k, \ \forall i, \quad (39d)$$

$$l_{i,k}^{(m)} \in \{0,1\}, \ \forall k, \ \forall i, \ m = 1,\ldots,N_k \quad (39e)$$

where $\bar{\mathbf{P}}_k \triangleq \text{blkdiag}\left(\bar{\mathbf{P}}_{i_1,k},\ldots,\bar{\mathbf{P}}_{i_{|\mathcal{I}_k|},k}\right) \in \mathbb{C}^{M^k\times|\mathcal{I}_k|N_k}$, and $\mathbf{L}_k \triangleq \text{blkdiag}\left(\mathbf{L}_{i_1,k},\ldots,\mathbf{L}_{i_{|\mathcal{I}_k|},k}\right) \in \mathbb{B}^{|\mathcal{I}_k|N_k\times|\mathcal{I}_k|N_k}$. We see that the original optimization problem (37) with coupled variables $\{\mathbf{P}_{i,k}\}$ and $\{D_{i,k}\}$ is now equivalently transformed to problem (39) with independent variables $\{\bar{\mathbf{P}}_{i,k}\}$ and $\{\mathbf{L}_{i,k}\}$.

B. WMMSE Transformation and Problem Reformulation

It is well known that mixed-integer and nonconvex problems such as (39) are NP-hard and finding their global optimum is generally difficult. The classic branch-and-bound method has exponential complexity [29]. In [8], [30], each integer variable is rewritten as an ℓ_0-norm and then approximated by a weighted ℓ_1-norm to relax the original integer variables to continuous variables. However, the initialization of the above algorithms in each iteration can be difficult due to the coupled variables, while we have decoupled them in problem (39).

Due to the above discrepancies, we first utilize the WMMSE transformation to reformulate problem (39), and consider to leverage the low-dimension substitution property together with the BCD approach, to provide low-complexity beamforming and stream allocation solutions.

$$\max_{\substack{\{\mathbf{U}_k\},\{\mathbf{W}_k\succ\mathbf{0}\},\\ \{\bar{\mathbf{P}}_{i,k}\},\{\mathbf{L}_{i,k}\}}} \sum_{k\in\mathcal{U}} \alpha_k\big(\log\det\left(\mathbf{W}_k\right) - \text{Tr}\big(\mathbf{W}_k\mathbf{E}_k\big(\mathbf{U}_k,\bar{\mathbf{P}}\big)\big)\big) \quad (40a)$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{U}_i}\text{Tr}\left(\bar{\mathbf{P}}_{i,k}\mathbf{L}_{i,k}\mathbf{L}_{i,k}^H\bar{\mathbf{P}}_{i,k}^H\right) \le P_{\max,i}, \ \forall i, \quad (40b)$$

$$(39c), \ (39d), \ (39e). \quad (40c)$$

Considering the presence of the integer variables $\{\mathbf{L}_{i,k}\}$, directly using the classic BCD approach to problem (40) is basically challenging. An intuitive approach is to relax the integer constraints (39e) into continuous constraints, i.e., $l_{i,k}^{(m)} \in [0,\ 1]$. However, this approach cannot guarantee that the relaxed optimization problem converges to a solution that yields the original integer constraints (39e). In contrast, we propose a linear stream allocation approach to problem (40), based on the following unique observation on the identity of the stream indicator matrix $\mathbf{L}_{i,k}$.

Remark 7. (Quadratic-Linear Property): Since $\{\mathbf{L}_{i,k}\}$ are diagonal matrices with values of zero or one on the diagonal, we have:

$$\mathbf{L}_{i,k}\mathbf{L}_{i,k}^H = \mathbf{L}_{i,k}, \ \forall k, \ \forall i. \quad (41)$$

Combining **Remark** 7 and relaxing the 0-1 integer constraints (39e) to the continuous constraints (42d), problem (40) can be relaxed to

$$\max_{\substack{\{\mathbf{U}_k\},\{\mathbf{W}_k\succ\mathbf{0}\},\\ \{\bar{\mathbf{P}}_{i,k}\},\{\mathbf{L}_{i,k}\}}} \sum_{k\in\mathcal{U}} \alpha_k\left(\log\det\left(\mathbf{W}_k\right) + 2\text{Tr}\left(\mathbf{W}_k\mathbf{U}_k^H\mathbf{H}_k\bar{\mathbf{P}}_k\right)\right)$$

$$- \sum_{k\in\mathcal{U}} \alpha_k\text{Tr}\left(\mathbf{W}_k\mathbf{U}_k^H\mathbf{H}_k\bar{\mathbf{P}}_k\mathbf{L}_k\bar{\mathbf{P}}_k^H\mathbf{H}_k^H\mathbf{U}_k\right) \quad (42a)$$

$$- \sum_{k\in\mathcal{U}} \alpha_k\text{Tr}\left(\mathbf{W}_k\mathbf{U}_k^H\bar{\mathbf{N}}_k\mathbf{U}_k\right)$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{U}_i}\text{Tr}\left(\bar{\mathbf{P}}_{i,k}\mathbf{L}_{i,k}\bar{\mathbf{P}}_{i,k}^H\right) \le P_{\max,i}, \ \forall i, \quad (42b)$$

$$(39c), \ (39d), \ (39e). \quad (42c)$$

$$l_{i,k}^{(m)} \in [0,1], \ \forall k, \ \forall i, \ m = 1,\ldots,N_k. \quad (42d)$$

Note that problem (42) is linear w.r.t the decision variables $\{\mathbf{L}_{i,k}\}$, and is convex w.r.t the decision variables $\{\mathbf{W}_k\}$, $\{\mathbf{U}_k\}$ and $\{\bar{\mathbf{P}}_{i,k}\}$.

C. RWMMSE-LSA Algorithm

In the following, we use the low-dimension substitution property in **Theorem** 2 together with the BCD method to provide efficient solutions to problem (42). Using results in **Theorem** 2, we can rewrite $\bar{\mathbf{P}}_{i,k}$ as

$$\bar{\mathbf{P}}_{i,k} = \bar{\mathbf{H}}_i^H\bar{\mathbf{X}}_{i,k} \quad (43)$$

where $\bar{\mathbf{X}}_{i,k} \in \mathbb{C}^{\sum_{k\in\mathcal{U}} N_k\times N_k}$ is a low-dimension substitution of $\bar{\mathbf{P}}_{i,k}$. We define

$$\tilde{\mathbf{X}}_{i,k} \triangleq \bar{\mathbf{X}}_{i,k}\mathbf{L}_{i,k} \in \mathbb{C}^{\sum_{k\in\mathcal{U}} N_k\times N_k} \quad (44)$$

as the actual low-dimension beamformer substitution (non-zero columns) that carries the allocated data streams from AP i to UE k.

Following the BCD approach, we first fix $\{\bar{\mathbf{X}}_{i,k}\}$, $\{\mathbf{L}_{i,k}\}$ and update $\{\mathbf{W}_k^*\}$, $\{\mathbf{U}_k^*\}$ by substituting (44) into (24) and (25). With fixed $\{\mathbf{W}_k^*\}$ and $\{\mathbf{U}_k^*\}$, we update $\bar{\mathbf{X}}_{i,k}^*$ by

$$\tilde{\mathbf{X}}_{i,k}^* = \left(\sum_{l\in\mathcal{U}} \alpha_l\bar{\mathbf{H}}_i\mathbf{H}_{i,l}^H\mathbf{A}_l\mathbf{H}_{i,l}\bar{\mathbf{H}}_i^H + \lambda_i\bar{\mathbf{H}}_i\bar{\mathbf{H}}_i^H\right)^{-1}$$
$$\times \alpha_k\bar{\mathbf{H}}_i\mathbf{H}_{i,k}^H\mathbf{U}_k^*\mathbf{W}_k^*\mathbf{\Xi}_{i,k}^H\mathbf{L}_{i,k}. \quad (45)$$

Then fixing $\{\mathbf{W}_k^*\}$, $\{\mathbf{U}_k^*\}$ and $\{\bar{\mathbf{X}}_{i,k}^*\}$ (resp. $\{\bar{\mathbf{P}}_{i,k}^*\}$ in (43)), the linear stream allocation problem can be expressed as

$$\min_{\{\mathbf{L}_{i,k}\}} \sum_{k\in\mathcal{U}}\sum_{i\in\mathcal{I}_k} \alpha_k\text{Tr}\left((\bar{\mathbf{P}}_{i,k}^*)^H\mathbf{H}_{i,k}^H\mathbf{A}_k\mathbf{H}_{i,k}\bar{\mathbf{P}}_{i,k}^*\mathbf{L}_{i,k}\right)$$

$$-2\sum_{k\in\mathcal{U}}\sum_{i\in\mathcal{I}_k}\alpha_k\text{Re}\big\{\text{Tr}\big(\mathbf{\Xi}_{i,k}\mathbf{W}_k^*(\mathbf{U}_k^*)^H\mathbf{H}_{i,k}\bar{\mathbf{P}}_{i,k}^*\mathbf{L}_{i,k}\big)\big\}$$

$$+\sum_{k\in\mathcal{U}}\alpha_k\text{Tr}\left(\sum_{j\in\mathcal{I}_l}\sum_{l\in\mathcal{U}_{-k}}(\bar{\mathbf{P}}_{j,l}^*)^H\mathbf{H}_{j,k}^H\mathbf{A}_k\mathbf{H}_{j,k}\bar{\mathbf{P}}_{j,l}^*\mathbf{L}_{j,l}\right) \quad (46)$$

$$\text{s.t. } (39c), \ (39d), \ (42b), \ (42d).$$

By extracting $\mathbf{L}_k$ related items from (46), the stream allocation problem for each UE k can be written as

$$\min_{\mathbf{L}_k} \sum_{i\in\mathcal{I}_k}\sum_{l\in\mathcal{U}}\alpha_l\mathrm{Tr}\left((\bar{\mathbf{P}}_{i,k}^*)^H\mathbf{H}_{i,l}^H\mathbf{A}_l\mathbf{H}_{i,l}\bar{\mathbf{P}}_{i,k}^*\mathbf{L}_{i,k}\right)$$
$$-2\sum_{i\in\mathcal{I}_k}\alpha_k\mathrm{Re}\left\{\mathrm{Tr}\left(\boldsymbol{\Xi}_{i,k}\mathbf{W}_k^*(\mathbf{U}_k^*)^H\mathbf{H}_{i,k}\bar{\mathbf{P}}_{i,k}^*\mathbf{L}_{i,k}\right)\right\} \quad (47)$$

s. t. (39c), (39d), (42b), (42d).

Lemma 2. The individual power constraints (42b) in the stream allocation problem (47) are always fulfilled.

Proof: See Appendix D. ∎

According to **Lemma 2**, constraints (42b) can be excluded from (47). Thus, we can easily obtain a closed-form optimal solution to problem (47), which readily satisfies the 0-1 integer constraints (39e). The derivation is as follows. Problem (47) can be equivalently rewritten as

$$\min_{\mathbf{L}_k} \mathrm{Tr}(\boldsymbol{\Psi}_k\mathbf{L}_k)$$
$$\text{s. t. } \mathbf{L}_k(l)\in[0,1], \ \forall l=1,\ldots,|\mathcal{I}_k|N_k, \quad (48)$$
$$\mathrm{Tr}(\mathbf{L}_k)\leq N_k$$

where $\mathbf{L}_k(l)$ denotes the lth diagonal element of $\mathbf{L}_k$, $\boldsymbol{\Psi}_k = \mathrm{blkdiag}\left(\mathrm{diag}(\boldsymbol{\Psi}_{i_1,k}),\ldots,\mathrm{diag}(\boldsymbol{\Psi}_{i_{|\mathcal{I}_k|},k})\right)$ with $\boldsymbol{\Psi}_{i,k} \triangleq \sum_{l\in\mathcal{U}}\alpha_l(\bar{\mathbf{P}}_{i,k}^*)^H\mathbf{H}_{i,l}^H\mathbf{A}_l\mathbf{H}_{i,l}\bar{\mathbf{P}}_{i,k}^*-2\alpha_k\mathrm{Re}\left\{\boldsymbol{\Xi}_{i,k}\mathbf{W}_k^*(\mathbf{U}_k^*)^H\mathbf{H}_{i,k}\bar{\mathbf{P}}_{i,k}^*\right\}$. The optimal solution to problem (48) is selecting the smallest $\pi_k\leq N_k$ diagonal elements of $\boldsymbol{\Psi}_k$ that are less than zero, i.e., the optimal $\mathbf{L}_k^*$ to problem (47) is given by

$$\mathbf{L}_k^*(l)=\begin{cases}1,\text{for }l\in\boldsymbol{\Pi}_k,\\0,\text{o.w.}\end{cases} \quad (49)$$

where $\boldsymbol{\Pi}_k$ denotes the indexes of the π_k smallest diagonal elements of $\boldsymbol{\Psi}_k$ that $\boldsymbol{\Psi}_k(l)<0$.

In **Algorithm** 4, we summarize the proposed RWMMSE-LSA algorithm. Different from the complicated initialization process in [8], the stream indicator matrix $\mathbf{L}_{i,k}$ in our RWMMSE-LSA process can be randomly initialized to meet the maximum number of receive streams constraints (4c). After initializing $\mathbf{L}_{i,k}$, we can use our proposed Local EZF method in Section IV-D to initialize $\bar{\mathbf{X}}_{i,k}$.

Proposition 2. (Convergence): Any limit point of the iterative sequence generated by the RWMMSE-LSA algorithm is a stationary point of problem (4).

Proof: See Appendix E. ∎

Remark 8. (User Scheduling): The proposed RWMMSE-LSA algorithm for stream allocation can be further extended to enable linear user scheduling. Specifically, if we set $\mathcal{I}_k = \mathcal{I}$, $\forall k$, the RWMMSE-LSA algorithm selects the serving APs to each UE k among all APs through stream allocation to maximize the WSR. The APs corresponding to the selected streams are then chosen as the serving APs $\mathcal{I}_k$ for UE k, which we call the RWMMSE-LUS algorithm.

VI. SIMULATION RESULTS

In this section, we present simulation results to illustrate the efficiency of the proposed algorithms.

Algorithm 4 Joint Beamforming and Stream Allocation Algorithm (RWMMSE-LSA)

Interact: Each AP i sends $\{\bar{\mathbf{H}}_i\bar{\mathbf{H}}_i^H\}$ to the connected CUs;
1: Initialize $\{\mathbf{L}_{i,k}\}$ and $\{\bar{\mathbf{X}}_{i,k}\}$ in CUs;
2: **Repeat** ←[CUs]
3:    Update $\{\mathbf{U}_k^*\}$ by substituting (44) into (24);
4:    Update $\{\mathbf{W}_k^*\}$ by substituting (44) into (25);
5:    Update $\{\tilde{\mathbf{X}}_{i,k}^*\}$ via (45);
6:    Update $\{\mathbf{L}_{i,k}^*\}$ via (49);
7: **Until** Converge
Interact: CUs transmit $\{\mathbf{X}_{i,k}^*, k\in\mathcal{U}_i\}$ to the connected AP i;
Output: Each AP i calculates $\{\mathbf{P}_{i,k}^*=\bar{\mathbf{H}}_i^H\mathbf{X}_{i,k}^*, k\in\mathcal{U}_i\}$.

TABLE II
SIMULATION PARAMETERS

	Parameter	Value
Network config.	I, K, M, N	4, 8, 64, 4
Power budget	$P_{\max}$	1W
Fairness weight	α_k	1
Size of serving set	L	2

A. Simulation Setup

We model the channel as Rayleigh channel with circularly symmetric standard complex normal distribution. We model the pathloss as $128.1 + 37.6\log_{10}(d)$[dB] [31], where $d \in [0.1, 0.3]$ in kilometers denotes the distance between the AP and the UE. Without loss of generality, all APs (and UEs) are assumed to have the same number of antennas, i.e., $M_i = M$ and $N_k = N$. The transmit power budget is set to be $P_{\max}$ for all APs, the fairness weight α_k and noise power $\sigma_k^2 = 10^{\frac{1}{K}\sum_k\log_{10}\frac{1}{N_kM_i}\|\mathbf{H}_k\|_F^2}\times 10^{-\frac{\mathrm{SNR}}{10}}$ are set equally for all UEs, where SNR is the average received SNR for all UEs without beamforming. In addition, the size of the serving AP set for each UE k is assumed to be same, i.e., $|\mathcal{I}_k| = L$, and the L nearest APs are selected to serve UE k. Monte Carlo simulations are performed over 100 randomly generated channel realizations. The zero-interaction Local EZF beamformer serves as the initial points of the WMMSE and the RWMMSE algorithms. Unless otherwise stated, the system parameters are summarized in Table II.

B. Comparison with the SCA Based Approach

In Fig. 3 and Fig. 4, we compare the WSR performance and the average CPU time between the WMMSE algorithm and the SCA-CVX algorithm in [8] under different SNR values. We set the number of receive antennas as $N = 1$ and the number of data streams as $D = 1$, since the SCA-CVX method is proposed for the single receive antenna case. We can see that the SCA-CVX algorithm in [8] and the WMMSE algorithm achieve the same WSR, but the SCA-CVX algorithm consumes 100x+ time than the WMMSE algorithm to converge. This behavior conforms to the complexity analysis in **Remark** 2 that the WMMSE approach has substantially lower complexity than the SCA based algorithms in [8], [17]. In the following, we choose the WMMSE algorithm as the baseline.
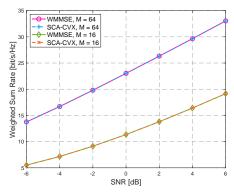
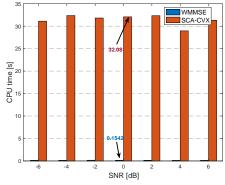Fig. 3. Comparison of WSR between the SCA-CVX algorithm in [8] and the WMMSE baseline.

Fig. 4. Comparison of average CPU time among FP based algorithm in [2], SCA-CVX algorithm in [8] and WMMSE baseline.
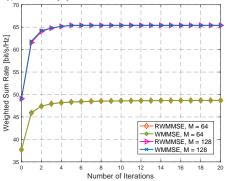

Fig. 5. Convergence performance of the WMMSE and the RWMMSE algorithms, where SNR = 0 [dB].

C. Comparison between WMMSE and RWMMSE

Fig. 5 depicts the convergence behavior of our proposed RWMMSE algorithm and the baseline WMMSE algorithm for the case of SNR = 0 [dB] with different numbers of transmit antennas. We observe that the low-interaction RWMMSE and the WMMSE algorithms converge smoothly to the same WSR, which is consistent with **Theorem** 2. Moreover, Fig. 6 shows that the baseline WMMSE algorithm generally requires more CPU time than the RWMMSE algorithm, particularly as the number of transmit antennas increases. The reason is that the WMMSE and the RWMMSE algorithms have linear and cubic complexity in M, respectively. These observed phenomena demonstrate that our proposed RWMMSE algorithm can achieve the same WSR performance as the WMMSE algorithm, but with significantly lower computational complexity.

D. Comparison of the Proposed Beamforming Algorithms

In Fig. 7, we compare the WSR of the RWMMSE algorithm and the fully distributed Local EZF algorithm with the baseline
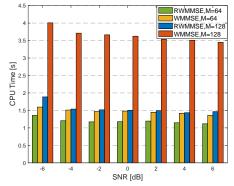

Fig. 6. Comparison of average CPU time of the WMMSE algorithm and the RWMMSE algorithm.
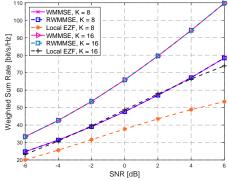

Fig. 7. Comparison of WSR between the fully distributed Local EZF method, the baseline WMMSE algorithm and the RWMMSE algorithm.
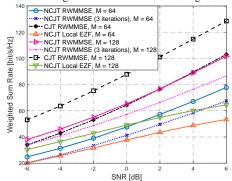

Fig. 8. Comparison of CJT and NCJT with different numbers of transmit antennas ($M = 64, 128$).

WMMSE algorithm, under different number of SNR values and UEs. We can see that the WMMSE algorithm and the RWMMSE algorithm yield almost the same WSR, both significantly outperforming the Local EZF method. In particular, when $K = 16$ and SNR = 6 [dB], the fully distributed Local EZF method only achieves 73.8 [bit/s/Hz], whereas the WMMSE/RWMMSE algorithm obtains a significantly higher WSR of 108.9 [bit/s/Hz], representing a 48% improvement over the Local EZF method. The reason for the lower WSR achieved by the Local EZF method is that it only suppresses inter-user interference but does not optimize transmit power for WSR maximization.

E. Comparison between NCJT and CJT

Fig. 8 exhibits the performance gaps between the CJT and the NCJT strategy under different numbers of transmit antennas. We can see that the CJT strategy provides approximately 30% more WSR than the NCJT strategy, since in CJT the APs cooperate as a virtue MIMO system and transmit the
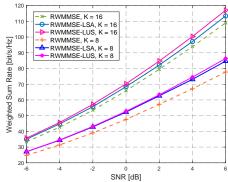
Fig. 9. WSR Performance of the proposed RWMMSE-LSA and RWMMSE-LUS algorithms.

same signals to their serving UEs. However, CJT requires strict synchronization among the APs, which can be hard to achieve in practical communication systems. In contrast, NCJT avoids synchronization overhead such as pilots at an acceptable cost of the WSR performance. In addition, our proposed RWMMSE algorithm achieves no less than 80% WSR performance in 3 iterations, showing great engineering prospect.

F. Stream Allocation Performance

We consider the RWMMSE algorithm with an average number of streams, i.e., $D = 2$, transmitted from the nearest APs $\mathcal{I}_k$ to UE k as our benchmark. In Fig. 9, we compare the WSR of the RMMMSE algorithm, the RWMMSE-LSA algorithm, and the RWMMSE-LUS algorithm, under different values of SNR and numbers of UEs. We observe that the RWMMSE-LUS algorithm achieves the highest WSR while the RWMMSE algorithm achieves the lowest WSR, showing the effectiveness of our proposed stream allocation algorithms. This is because that the RWMMSE algorithm allocates APs $\mathcal{I}_k$ to transmit an average number of streams to each UE k, the RWMMSE-LSA algorithm optimizes the number of data streams transmitted from APs $\mathcal{I}_k$ to maximize WSR, while the RWMMSE-LUS algorithm selects data streams among all the APs that contribute to the highest WSR.

VII. CONCLUSIONS

In this work, we have investigated joint beamforming and stream allocation algorithms for WSR maximization in user-centric cell-free MIMO networks with NCJT. We first propose a distributed low-interaction and low-complexity RWMMSE beamforming algorithm with closed-form updates in each iteration for the case of fixed streams. Our proposed RWMMSE algorithm exhibits the lowest known complexity and requires much lower interaction across the networks, with no compromise to WSR performance. We further propose a zero-interaction Local EZF method based only on the local CSI to initialize the RWMMSE iteration. Finally, by effectively decouple the beamforming and stream allocation variables, we develop a joint beamforming and linear stream allocation RWMMSE-LSA algorithm with linear stream allocation complexity for WSR maximization with varying streams. Simulation results demonstrate the significant advantages of our proposed algorithms over the current best alternatives, in terms of both convergence time and WSR performance.

APPENDIX A
PROOF OF THEOREM 1

We use contradiction to prove **Theorem** 1. Since for any AP i, the columns of $\bar{\mathbf{H}}_i^H$ are linearly independent, we define $\tilde{\mathbf{H}}_{i,k,n}$ consisting all columns of $\bar{\mathbf{H}}_i^H$ except the nth column $\mathbf{h}_{i,k,n}$, i.e.,

$$\tilde{\mathbf{H}}_{i,k,n} \triangleq \left[\mathbf{H}_{i,1}^H, \ldots, \mathbf{H}_{i,k-1}^H, \hat{\mathbf{H}}_{i,k}, \mathbf{H}_{i,k+1}^H, \ldots, \mathbf{H}_{i,K}^H \right] \quad (50)$$

where $\hat{\mathbf{H}}_{i,k} \triangleq [\mathbf{h}_{i,k,1}, \ldots, \mathbf{h}_{i,k,n-1}, \mathbf{h}_{i,k,n+1}, \ldots, \mathbf{h}_{i,k,N_k}]$, and the columns of $\hat{\mathbf{H}}_{i,k,n}$ are linearly independent.

Assume that for AP i, the local optimal beamformer is $\{\mathbf{P}_{i,k}^*, k \in \mathcal{U}_i\}$, and $\sum_{k \in \mathcal{U}_i} \left\| \mathbf{P}_{i,k}^* \right\|_F^2 < P_{\max,i}$. Then we introduce another set of beamformers $\{\hat{\mathbf{P}}_{i,k}, k \in \mathcal{U}_i\}$, and the dth column of $\hat{\mathbf{P}}_{i,k}$ equals to that of $\mathbf{P}_{i,k}^*$, i.e., $\hat{\mathbf{p}}_{i,k,d} = \mathbf{p}_{i,k,d}^*$ for all $d = 1, \ldots, D_{i,k}$ and $d \neq s$. The sth column of $\hat{\mathbf{P}}_{i,k}$ is defined as

$$\hat{\mathbf{p}}_{i,k,s} \triangleq \mathbf{p}_{i,k,s}^* + \beta e^{j\theta} \triangle \mathbf{p}_{i,k,s} \quad (51)$$

where $\beta > 0$ is a scaling factor such that $\sum_{k \in \mathcal{U}_i} \|\hat{\mathbf{P}}_{i,k}\|_F^2 = P_{\max,i}$, $\theta = \angle \left(\mathbf{h}_{i,k,n}^H, \mathbf{p}_{i,k,s}^* \right)$, and $\triangle \mathbf{p}_{i,k,s} = \prod_{\tilde{\mathbf{H}}_{i,k,n}}^{\perp} \mathbf{h}_{i,k,n}$. In addition, as $\prod_{\tilde{\mathbf{H}}_{i,k,n}}^{\perp} \succ \mathbf{0}$, then we have $\mathbf{h}_{i,k,n}^H \triangle \mathbf{p}_{i,k,s} > 0$, and $\mathbf{h}_{i,l,n}^H \triangle \mathbf{p}_{i,k,s} = 0$ for any $l \in \mathcal{U}_{-k}$.

Furthermore, the nth diagonal element of $\mathbf{H}_{i,l} \hat{\mathbf{P}}_{i,k} \hat{\mathbf{P}}_{i,k}^H \mathbf{H}_{i,l}^H$ is $|\mathbf{h}_{i,l,n}^H \sum_{d=1}^{D_{i,k}} \hat{\mathbf{p}}_{i,k,d}|^2$. And we have

$$\left| \mathbf{h}_{i,l,n}^H \hat{\mathbf{p}}_{i,k,d} \right| = \left| \mathbf{h}_{i,l,n}^H \mathbf{p}_{i,k,d}^* \right|, \text{if } d \neq s \text{ or } l \neq k, \quad (52)$$

and

$$\begin{aligned} \left| \mathbf{h}_{i,k,n}^H \hat{\mathbf{p}}_{i,k,s} \right| &= \left| \mathbf{h}_{i,l,n}^H \mathbf{p}_{i,k,s}^* + \beta e^{j\theta} \mathbf{h}_{i,l,n}^H \triangle \mathbf{p}_{i,k,s} \right| \\ &= \left| \mathbf{h}_{i,l,n}^H \mathbf{p}_{i,k,s}^* \right| + \beta \mathbf{h}_{i,l,n}^H \triangle \mathbf{p}_{i,k,s} \quad (53) \\ &> \left| \mathbf{h}_{i,l,n}^H \mathbf{p}_{i,k,s}^* \right|. \end{aligned}$$

Consequently, $R_k \left(\hat{\mathbf{P}}_{i,k} \right) > R_k \left(\mathbf{P}_{i,k}^* \right)$ and for $l \neq k$, $R_l \left(\hat{\mathbf{P}}_{i,k} \right) = R_l \left(\mathbf{P}_{i,k}^* \right)$. Plus, when $\beta = 0$, $\hat{\mathbf{P}}_{i,k}^* = \mathbf{P}_{i,k}^*$, as to $\beta \to 0$ and $\beta \neq 0$, $\hat{\mathbf{P}}_{i,k}^* \to \mathbf{P}_{i,k}^*$. That is to say, $\left\{ \hat{\mathbf{P}}_{i,k} \right\}$ contributes more rate then $\left\{ \mathbf{P}_{i,k}^* \right\}$, which contradicts to the assumption that $\{\mathbf{P}_{i,k}^*\}$ is the local optimal beamformer. Hence, to get the full WSR, $\sum_{k \in \mathcal{U}_i} \left\| \mathbf{P}_{i,k}^* \right\|_F^2 = P_{\max,i}$ always holds for any $i \in \mathcal{I}$. Then the proof is completed. ∎

APPENDIX B
PROOF OF PROPOSITION 1

By contradiction, suppose that $\mathbf{P}_{i,k}^*$ does not lie in the column space of $\bar{\mathbf{H}}_i^H$, then $\mathbf{P}_{i,k}^*$ can be expressed as

$$\mathbf{P}_{i,k}^* = \tilde{\mathbf{P}}_{i,k} + \hat{\mathbf{P}}_{i,k}, \forall k \in \mathcal{U}_i \quad (54)$$

where $\tilde{\mathbf{P}}_{i,k}$ and $\hat{\mathbf{P}}_{i,k}$ are located in the column space and null space of $\bar{\mathbf{H}}_i^H$, respectively. Note that $\tilde{\mathbf{P}}_{i,k}$ and $\mathbf{P}_{i,k}^*$ have the same WSR due to $\mathbf{H}_{i,k} \hat{\mathbf{P}}_{i,k} = \mathbf{0}$. In other words, removing the part of the beamforming matrix $\mathbf{P}_{i,k}^*$ in the null space of $\bar{\mathbf{H}}_i^H$ has no effect on the objective function value of (12). ∎

APPENDIX C
PROOF OF THEOREM 2

We prove by contradiction. According to **Proposition** 1, if the local optimum $\mathbf{P}_{i,k}^*$ does not lie in the column space of

$\bar{\mathbf{H}}_i^H$, then the new beamformer obtained by removing its part in the null space has the same WSR as $\mathbf{P}_{i,k}^*$. This contradicts the conclusion of full power property in Theorem 1. ∎

APPENDIX D
PROOF OF LEMMA 2

Since the beamformer $\bar{\mathbf{P}}_{i,k}^{*(r)}$ in the rth iteration is updated given the stream indicator matrix $\mathbf{L}_{i,k}^{*(r-1)}$ in the $(r-1)$th iteration, the $\mathbf{\Upsilon}_{i,k}$th columns of $\bar{\mathbf{P}}_{i,k}^{*(r)}$ are zero with $\mathbf{\Upsilon}_{i,k}$ being the indexes of the zero diagonal elements of $\mathbf{L}_{i,k}^{*(r-1)}$. Hence, we have $\sum_{k\in\mathcal{U}_i}\mathrm{Tr}\Big(\bar{\mathbf{P}}_{i,k}^{*(r)}\mathbf{L}_{i,k}^{*(r)}(\bar{\mathbf{P}}_{i,k}^{*(r)})^H\Big) \leq \sum_{k\in\mathcal{U}_i}\mathrm{Tr}\Big(\bar{\mathbf{P}}_{i,k}^{*(r)}\mathbf{L}_{i,k}^{*(r-1)}(\bar{\mathbf{P}}_{i,k}^{*(r)})^H\Big) \leq P_{\max,i}$.

APPENDIX E
PROOF OF PROPOSITION 2

The convergence of the RWMMSE-LSA to the stationary points of problem (42) is guaranteed by the classic convergence theory of the BCD method [29]. Moreover, by rewriting the 0-1 integer constraints (39e) as $l_{i,k}^{(m)}(l_{i,k}^{(m)}-1) = 0, \forall k, \forall i, m = 1,\ldots,N_k$, we derive the KKT condition of problem (40). Upon comparing the two KKT systems of problem (42) and (40), it can be observed that any limit point $\Big(\mathbf{U}_k^*, \mathbf{W}_k^*, \bar{\mathbf{P}}_{i,k}^*, \mathbf{L}_{i,k}^*\Big)$ of the iterative sequence generated by the RWMMSE-LSA adheres to the KKT condition of problem (40), by using the fact that the diagonal elements of $\mathbf{L}_{i,k}^*$ must be zero or one. Hence, $\Big(\mathbf{U}_k^*, \mathbf{W}_k^*, \bar{\mathbf{P}}_{i,k}^*, \mathbf{L}_{i,k}^*\Big)$ is both the stationary point of problem (40) and its equivalent original problem (4). ∎

REFERENCES

[1] H. Yang and T. L. Marzetta, "Capacity performance of multicell large-scale antenna systems," in *Proceedings of IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2013.

[2] 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*, vol. 21, pp. 1482–1497, 2022.

[3] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Vehicular Technology Magazine*, vol. 6, pp. 37–43, 2011.

[4] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99 878–99 888, 2019.

[5] A. Papazafeiropoulos, H. Q. Ngo, P. Kourtessis, S. Chatzinotas, and J. M. Senior, "Towards optimal energy efficiency in cell-free massive MIMO systems," *IEEE Transactions on Green Communications and Networking*, vol. 5, pp. 816–831, 2021.

[6] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Communications Magazine*, vol. 50, pp. 148–155, 2012.

[7] A. Davydov, G. Morozov, I. Bolotin, and A. Papathanassiou, "Evaluation of joint transmission comp in C-RAN based LTE-A HetNets with large coordination areas," in *Proceedings of IEEE Globecom Workshops (GC Wkshps)*, 2013.

[8] C. Pan, H. Ren, M. Elkashlan, A. Nallanathan, and L. Hanzo, "The non-coherent ultra-dense C-RAN is capable of outperforming its coherent counterpart at a limited fronthaul capacity," *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 2549–2560, 2018.

[9] J. Li, E. Björnson, T. Svensson, T. Eriksson, and M. Debbah, "Joint precoding and load balancing optimization for energy-efficient heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 5810–5822, 2015.

[10] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Transactions on Wireless Communications*, vol. 19, pp. 77–90, 2019.

[11] H. A. Ammar, Y. Nasser, and A. Kayssi, "Dynamic SDN controllers-switches mapping for load balancing and controller failure handling," in *Proceedings of International Symposium on Wireless Communication Systems (ISWCS)*, 2017.

[12] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 6384–6399, 2016.

[13] Ö. T. Demir, E. Björnson, L. Sanguinetti *et al.*, "Foundations of user-centric cell-free massive MIMO," *Foundations and Trends in Signal Processing*, vol. 14, pp. 162–472, 2021.

[14] H. Shao, H. Zhang, L. Sun, and Y. Qian, "Resource allocation and hybrid OMA/NOMA mode selection for non-coherent joint transmission," *IEEE Transactions on Wireless Communications*, vol. 21, pp. 2695–2709, 2022.

[15] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and user association optimization for massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 6384–6399, 2016.

[16] E. Björnson, M. Kountouris, and M. Debbah, "Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination," in *Proceeding of International Conference on Telecommunications (ICT)*, 2013.

[17] 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 Transactions on Wireless Communications*, vol. 19, pp. 5891–5907, 2020.

[18] J. Epstein, *Scalable VoIP mobility: Integration and deployment*. Newnes, 2009.

[19] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "User-centric cell-free massive MIMO networks: A survey of opportunities, challenges and solutions," *IEEE Communications Surveys & Tutorials*, vol. 24, pp. 611–652, 2021.

[20] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014.

[21] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Transactions on Signal Processing*, vol. 59, pp. 4331–4340, 2011.

[22] M. Alonzo, S. Buzzi, and A. Zappone, "Energy-efficient downlink power control in mmwave cell-free and user-centric massive MIMO," in *Proceedings of IEEE 5G World Forum (5GWF)*, 2018.

[23] S. Zhou, M. Zhao, X. Xu, J. Wang, and Y. Yao, "Distributed wireless communication system: A new architecture for future public wireless access," *IEEE Communications Magazine*, vol. 41, pp. 108–113, 2003.

[24] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Transactions on Communications*, vol. 68, pp. 4247–4261, 2020.

[25] H. A. Ammar and R. Adve, "Power delay profile in coordinated distributed networks: User-centric v/s disjoint clustering," in *Proceedings of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2019.

[26] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "Distributed resource allocation optimization for user-centric cell-free MIMO networks," *IEEE Transactions on Wireless Communications*, vol. 21, pp. 3099–3115, 2021.

[27] Q. Shi, W. Xu, J. Wu, E. Song, and Y. Wang, "Secure beamforming for MIMO broadcasting with wireless information and power transfer," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 2841–2853, 2015.

[28] X. Zhao, S. Lu, Q. Shi, and Z.-Q. Luo, "Rethinking WMMSE: Can its complexity scale linearly with the number of BS antennas?" *IEEE Transactions on Signal Processing*, vol. 71, pp. 433–446, 2023.

[29] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, pp. 334–334, 1997.

[30] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. Srinivas, "Resource allocation and scheduling in non-coherent user-centric cell-free MIMO," in *Proceeding of IEEE International Conference on Communications (ICC)*, 2021.

[31] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE transactions on wireless communications*, vol. 9, pp. 1748–1759, 2010.