

Cluster-Wise Processing in Fronthaul-Aware Cell-Free Massive MIMO Systems

Zahra Mobini, *Senior Member, IEEE*, Ahmet Hasim Gokceoglu, Li Wang, *Senior Member, IEEE*, Gunnar Peters, Hyundong Shin, *Fellow, IEEE*, and Hien Quoc Ngo, *Fellow, IEEE*

Abstract—We exploit a general cluster-based network architecture for a fronthaul-limited user-centric cell-free massive multiple-input multiple-output (CF-mMIMO) system under different degrees of cooperation among the access points (APs) to achieve scalable implementation. In particular, we consider a CF-mMIMO system wherein the available APs are grouped into multiple processing clusters (PCs) to share channel state information (CSI), ensuring that they have knowledge of the CSI for all users assigned to the given cluster for the purposes of designing resource allocation and precoding. We utilize the sum pseudo-SE metric, which accounts for intra-cluster interference and inter-cluster-leakage, providing a close approximation to the true sum achievable SE. For a given PC, we formulate two optimization problems to maximize the cluster-wise weighted sum pseudo-SE under fronthaul constraints, relying solely on local CSI. These optimization problems are associated with different computational complexity requirements. The first optimization problem jointly designs precoding, user association, and power allocation, and is performed at the small-scale fading time scale. The second optimization problem optimizes user association and power allocation at the large-scale fading time scale. Accordingly, we develop a novel application of modified weighted minimum mean square error (WMMSE)-based approach to solve the challenging formulated non-convex mixed-integer problems. Numerical results show that (a) the proposed cluster-wise processing solutions significantly outperform the heuristic approaches under both the statistical and instantaneous CSI-based designs, while statistical CSI-based design is good enough in some network configurations since it provides a better performance/implementation complexity trade-off; (b) under limited fronthaul, our proposed cluster-wise processing frameworks achieve sum-spectral efficiency (SE) that is competitive with state-of-the-art network-wide processing solutions, while avoiding the high computational complexity of processing and the heavy overhead of CSI acquisition.

Index Terms—Cell-free massive multiple-input multiple-output

The work of H. Shin was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under RS-2025-00556064 and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2021-II212046) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). (*Corresponding authors: Hyundong Shin; Hien Quoc Ngo*).

Z. Mobini is with the Department of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K., and also with the Centre for Wireless Innovation (CWI), Queen's University Belfast, BT3 9DT Belfast, U.K. (email: zahra.mobini@manchester.ac.uk).

A. gokceoglu1, L. Wang, and G. Peters are with the Huawei's Sweden Research Center, Stockholm, Sweden (e-mail: {ahmet.hasim.gokceoglu1, leo.li.wang, gunnar.peters}@huawei.com).

H. Shin is with the Department of Electronics and Information Convergence Engineering, Kyung Hee University, 1732 Deogyong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Republic of Korea (e-mail: hshin@khu.ac.kr).

H. Q. Ngo is with the Centre for Wireless Innovation (CWI), Queen's University Belfast, BT3 9DT Belfast, U.K., and is also with the Department of Electronic Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Republic of Korea (email: hien.ngo@qub.ac.uk).

(CF-mMIMO), cluster-wise processing, fronthaul, resource allocation.

I. INTRODUCTION

Cell-free massive multiple-input multiple-output (CF-mMIMO) is a cutting-edge wireless technology developed to address the extensive connectivity requirements and escalating data traffic demands of next-generation wireless networks [1]. In CF-mMIMO, a large number of access points (APs) are distributed over a wide area to simultaneously and coherently serve multiple users. This technology integrates massive MIMO, network MIMO (also known as coordinated multipoint joint transmission, CoMP-JT), and cooperative networks [2], capitalizing on their strengths to manage interference, provide high macro/micro-diversity gains, and achieve high array gains, ensuring ubiquitous connectivity.

Since the seminal work on canonical CF-mMIMO [2], this network topology has become a focal point of extensive research [3]. Over the years, a substantial body of literature has emerged, focused on improving the practicality, spectral efficiency (SE), and energy efficiency of CF-mMIMO systems through advanced signal processing techniques and resource allocation strategies [4]. Early CF-mMIMO configurations assumed that i) all APs possessed network-wide channel state information (CSI) through a centralized processing unit (CPU) connected by fronthaul links, and ii) all APs transmitted/received information signals to/from all users during the downlink/uplink data transmission phases. While this configuration optimized system performance, it required a high degree of coordination among APs, introduced considerable computational complexity, and necessitated extensive fronthaul/backhaul signaling for CSI and data exchange. These factors posed significant challenges to scalability as the network size increased, whether in terms of the number of APs or users. To address these scalability challenges, a user-centric approach has been proposed in [5]–[7], wherein each user is served by a subset of APs rather than the entire network. This user-centric CF-mMIMO offers a more scalable solution for CF-mMIMO implementation as network sizes grow, while delivering performance that is close to the canonical CF-mMIMO in terms of achievable SE [4].

Although the theoretical frameworks for CF-mMIMO are well-established, the practical implementation of this technology still faces significant challenges. These challenges include practical and scalable signal processing, resource allocation (power control and user association), CSI knowledge, and fronthaul requirements, particularly as network size increases [4], [8]. Signal processing and resource allocation

in CF-mMIMO systems can generally be implemented using either a centralized approach at the CPU or a distributed approach at each AP. Centralized processing has been widely utilized in the CF-mMIMO literature for tasks such as precoding/combining design, including centralized minimum mean square error (MMSE) with varying levels of coordination in [9] and centralized zero-forcing (ZF) in [10], [11], as well as for resource allocation in [6], [12]–[16]. While these methods optimize system performance, they also require extensive CSI sharing via fronthaul/backhaul links, which can heavily burden network resources. Moreover, the computational load of performing optimization at the CPU becomes increasingly complex due to the high dimensionality of the aggregated channels. These combined factors make centralized processing both impractical and difficult to scale in real-world applications.

To address these limitations, distributed processing approaches have been proposed as a more scalable alternative. For example, the authors in [17] introduced a framework for cooperative precoding design in CF-mMIMO systems, which eliminates the need for backhaul signaling for CSI exchange by employing over-the-air (OTA) signaling mechanisms to acquire the necessary information at APs. Furthermore, a distributed precoding design known as Team MMSE was introduced in [18]. Moreover, distributed yet simple and sub-optimal power allocation algorithms have been proposed in [9], [10]. Scalable fractional power control strategies along with conjugate precoding for downlink CF-mMIMO was proposed in [19], while distributed max-min power control by training the neural network with only local CSI at each AP was proposed in [20]. While distributed processing offers better scalability and lower complexity by relying only on local channel estimates at each AP, it also has performance limitations compared to centralized methods. This is due to the fact that distributed approaches lack network-wide coordination between APs, which can lead to suboptimal resource allocation and low efficient interference management. The absence of network-wide control means that distributed approaches may not be able to fully utilize the available diversity gains or cancel interference as effectively as centralized approaches. Consequently, system performance may reduce in scenarios where coordination between APs is critical for obtaining optimal performance. Therefore, finding a balance between system performance and the degree of cooperation among APs becomes a key challenge for ensuring the scalability of CF-mMIMO systems.

To this end, partially centralized processing was considered as a promising solution that ensures scalability within the system architecture [4]. In particular, the authors in [21] employed a semi-distributed version of partial ZF, which partially mitigates inter-AP interference by sharing a limited amount of CSI over the fronthaul network. However, this comes at the cost of increased computational complexity and fronthaul overhead. Furthermore, the authors of [22] harnessed a centralized ZF-based precoding for a subset of APs and distributed maximum ratio transmission precoding for the remaining APs. However, the significant drawback of these studies is that they only focused on either precoding design

or power control. Therefore, the study of how to efficiently perform partially centralized processing for more complicated methods such as joint precoding, user association, and power control in a CF-mMIMO system under imperfect fronthaul network is extremely timely and important.

A. Key Contributions

To address the need for scalable CF-mMIMO in practical fronthaul-limited scenarios, we are inspired by the novel paradigm of cluster-wise processing [23]. Cluster-wise processing involves performing signal processing and resource allocation within a cluster (group) of APs, requiring only partial CSI sharing and potentially reducing the computational requirements due to the smaller dimensionality of the aggregated channels. This approach strikes a balance between having some coordination among APs and using a fully centralized system. Cluster-wise processing not only supports scalable implementation but also aids in managing inter-cluster interference by localizing coordination [24]. One fundamental challenge in designing cluster-wise schemes is the signal-to-interference-plus-noise ratio (SINR) metric, which is inter-dependent across all users and APs in different clusters. Therefore, any optimization framework based on the SINR maximization criterion may not be suitable for cluster-wise resource allocation. An alternative is to use leakage-based metrics [23], [25]. Considering the notion of hybrid signal-to-leakage-and-intra-cluster-interference-and-noise ratio (SLINR), which considers the desired signal, the intra-cluster interference, and the leakage interfering with other users in other clusters, the authors in [23], developed cluster-wise resource allocation approaches for user-centric cell-free systems for two scenarios: system optimization is carried out at the APs or at multiple central units (CUs) controlling a subset (a cluster) of APs. However, this work did not account for fronthaul constraints and relied solely on instantaneous CSI. While instantaneous CSI-based optimization enables dynamic user association, precoding, and power allocation, it comes with high computational and signaling costs. The need for frequent CSI acquisition increases with the number of antennas, sub-carriers, and users. A more scalable alternative is statistical CSI-based design, which takes advantage of channel hardening to allow efficient resource allocation based on large-scale fading, rather than the fast-varying small-scale fading. This approach significantly reduces computational complexity, as system parameters remain stable over longer time scales, and minimizes the need for real-time downlink CSI estimation, making it particularly suitable for large-scale CF-mMIMO deployments. However, in certain propagation environments or for some system setups channel hardening deteriorates [26], rendering statistical CSI-based power allocation suboptimal. As a result, instantaneous CSI-based power allocation remains necessary in some CF-mMIMO configurations. Therefore, a comparative analysis of statistical versus instantaneous CSI-based designs remains unexplored for cluster-wise fronthaul-limited CF-mMIMO.

In this context, we adopt the hybrid SLINR metric for a fronthaul-limited CF-mMIMO system, where APs are grouped

into multiple processing clusters (PCs). We formulate two distinct optimization problems, each operating at a different time scale: 1) Instantaneous CSI-based optimization: Jointly optimizes precoding, user association, and power allocation at the small-scale fading time scale, enabling adaptation to rapid channel variations. 2) Statistical CSI-based optimization: Jointly optimizes power control and user association based on large-scale fading statistics, leveraging the channel hardening property of CF-mMIMO systems. To this end, we introduce a hardening-based pseudo-SE, which simplifies resource allocation while maintaining performance in large-scale networks. For efficient cluster-wise processing, we develop two novel applications of a modified weighted minimum mean square error (WMMSE)-based approach. The key motivation behind using modified WMMSE algorithms is their ability to jointly optimize user association, power allocation, and/or precoding vectors within a unified framework. Furthermore, each iteration of the WMMSE algorithm is computationally efficient, making it feasible for large-scale CF-mMIMO systems without imposing excessive overhead.

The key contributions of this paper are summarized as follows:

- We provide a cluster-based network architecture for a fronthaul-limited user-centric CF-mMIMO system with multiple-antenna APs. The proposed framework is very general and can cover different CF-mMIMO implementations. In particular, it can be degenerated to different special cases such as CF-mMIMO with network-wide fully centralized operation or fully distributed operation.
- We formulate two optimization problems for maximizing the cluster-wise weighted sum pseudo-SE/hardening-based pseudo-SE under per-AP transmit power and fronthaul constraints which can be carried out at two different time scales. Two novel modified WMMSE-based algorithms are then proposed to solve the challenging formulated non-convex mixed-integer problems.
- Numerical results show that the proposed cluster-wise processing solutions significantly outperform the heuristic approaches for both the statistical and instantaneous CSI-based designs. They also confirm that CF-mMIMO can be efficiently deployed by utilising our proposed cluster-wise processing without significant loss in performance but with much lower fronthaul requirements compared to centralized system. In fact, they show that cluster-wise processing is highly preferable compared to network-wide alternatives, either in the regime of a high number of APs or high user loads. In addition, in various simulation setups, e.g., large values of transmit antennas at APs or stringent fronthaul constraints, the statistical CSI based design leads to a negligible performance loss, compared to instantaneous CSI-based design. Therefore, statistical CSI-based design is enough since it provides a better performance/implementation complexity trade-off.

Notation: We use bold upper (lower) case letters to denote matrices (vectors). The superscripts $(\cdot)^T$ and $(\cdot)^\dagger$ stand for the transpose and conjugate-transpose, respectively. A zero-mean circular symmetric complex Gaussian distribution having a

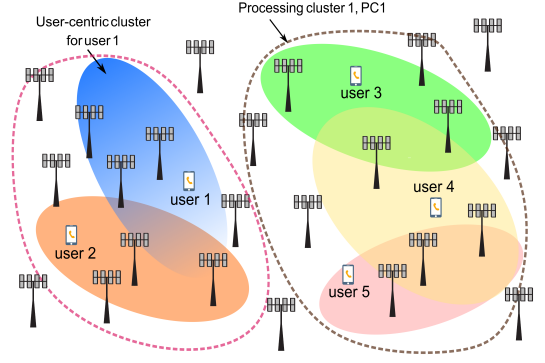


Fig. 1: User-centric CF-mMIMO with cluster-wise processing.

variance of σ^2 is denoted by $\mathcal{CN}(0, \sigma^2)$, while $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. $\mathbb{E}\{\cdot\}$ denotes the statistical expectation. Finally, operations $\text{Re}(\cdot)$ and $|\cdot|$ represent the real part of a complex number and the size (or cardinality) of a set.

II. GENERIC CLUSTER-WISE PROCESSING-BASED NETWORK ARCHITECTURE

We consider a CF-mMIMO system comprising M L -antenna APs and K single-antenna users. The sets of APs and users are denoted by $\mathcal{M} = \{1, \dots, M\}$ and $\mathcal{K} = \{1, \dots, K\}$, respectively. The baseband unit (BBU) functionality is split into two entities: baseband low (BBL) and baseband high (BBH). The BBH handles processing tasks such as precoding, encoding, and radio resource management, while the BBL is responsible for tasks like weight applications, error correction, and modulation. Each BBH is connected to its associated BBL via limited-capacity fronthaul links to transmit information such as precoding vectors, information signals, and power allocation coefficients. In addition, BBHs are interconnected through backhaul links to facilitate information exchange between APs. It is notable that in this work, we do not impose explicit backhaul capacity constraints to focus on fronthaul limitations, which are usually the main bottleneck in CF-mMIMO systems, especially when fronthaul links are wireless or limited. Backhaul links—typically high-capacity fiber or Ethernet—are assumed sufficient to support intra-cluster signalling [27]. Nevertheless, cluster-wise processing for CF-mMIMO systems under joint fronthaul and backhaul constraints is an interesting direction worthy of future research.

We assume a frequency-flat slow fading channel model for each orthogonal frequency-division multiplexing (OFDM) subcarrier. Here, we assume that the cyclic prefix length is greater than the maximum delay spread, ensuring the orthogonality between OFDM subcarriers. For notational simplicity, the subcarrier index will be omitted. Let $\mathbf{g}_{mk} \in \mathbb{C}^{L \times 1}$ represent the complex channel vector between the m -th AP and the k -th user. This channel vector can be modeled as

$$\mathbf{g}_{mk} = \beta_{mk}^{1/2} \mathbf{h}_{mk}, \quad (1)$$

where β_{mk} denotes the large-scale fading coefficient that includes path-loss and shadowing effects, while $\mathbf{h}_{mk} \in \mathbb{C}^{L \times 1}$ is the small-scale fading vector whose entries are independent

and identically distributed (i.i.d.) $\mathcal{CN}(0, 1)$ RVs. Large-scale fading coefficients change slowly and may be constant in range of many small-scale fading coherence intervals (over time and frequency bands). Hence, it is assumed that these coefficients are priori known at each BBL/BBH.

We consider the concept of cluster-wise processing in conjunction with user-centric association. Specifically, in cluster-wise processing, the available APs are divided into S disjoint PCs represented by the set of sets $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_S\}$. APs within each PC $\mathcal{C}_s$ share CSI via intra-cluster backhaul links, ensuring knowledge of the CSI for all users served by APs in that cluster to facilitate resource allocation and precoding¹.

Each user k is associated with only one PC and can be coherently served by the APs, or a subset of APs, within that cluster. Fig. 1 shows a typical cluster-wise processing-based user-centric CF-mMIMO architecture including five users served by their associated APs and two PCs. The colored regions indicate which set of APs serve which users, while the purple and brown regions show the PCs.

Remark 1. When $S = 1$, there is a single PC consisting of all the APs in the network. This corresponds to the CF-mMIMO system with fully centralized (network-wide) processing. In contrast, when $S = M$, the system operates with fully distributed processing.

Remark 2. Unlike traditional cellular networks, where users are typically associated with a single AP, our model allows each user to be coherently served by multiple APs within its associated PC. This preserves the joint transmission and reception characteristics of CF-mMIMO, enabling macro-diversity and effective interference suppression. While we introduce cluster-wise processing to manage fronthaul limitations and computational complexity, these clusters do not enforce fixed cell boundaries. Instead, they support flexible user-AP associations and maintain the core CF-mMIMO principle of user-centric service. In contrast to clustered CoMP, where coordination is often limited, massive MIMO properties are not applied, and user association is predefined by fixed cell layout, our approach allows dynamic and optimized cooperation among distributed APs within each cluster.

A. Uplink Training for Channel Estimation

In the uplink training phase, all users send pilot signals to the APs. This full-pilot model is adopted to enable user-centric coordination and coherent joint transmission, in line with foundational CF-mMIMO studies [4]. Alternative strategies, such as efficient pilot reuse or partial CSI acquisition, may be employed to reduce overhead while preserving acceptable system performance.

Accordingly, each AP can estimate the corresponding channels to all users using the obtained pilot signal. Note that at each AP, the channel estimation is performed at its BBH. We

consider orthogonal pilot assignment². This requires $\tau_u \geq K$, where τ_u is the uplink training duration. AP m uses the MMSE estimation technique to estimate the channels to all users. The MMSE channel estimate of $\mathbf{g}_{mk}$, $\hat{\mathbf{g}}_{mk}$, includes L i.i.d. $\mathcal{CN}(0, \gamma_{mk})$ elements, where $\gamma_{mk} = \frac{\tau_u \varrho_u \beta_{mk}^2}{\tau_u \varrho_u \beta_{mk}^2 + 1}$, and ϱ_u is the normalized transmit power of each pilot symbol.

B. Downlink Payload Data Transmission

For each user k , we choose a PC $\mathcal{C}_s$, where PC selection can be based on diverse criteria, including the value of large-scale fading coefficient β_{mk} . Our proposed PC selection strategy will be discussed in section V. Also, we define the set $\mathcal{U}_s \subset \mathcal{K}$ representing the users that need to be served by PC $\mathcal{C}_s$, while $\mathcal{U}_{-s}$ represents the set of users that have no association with APs in set $\mathcal{C}_s$, i.e., $\mathcal{U}_{-s} = \mathcal{K} \setminus \mathcal{U}_s$. It is notable that each AP m in PC $\mathcal{C}_s$ communicates only with a subset of users in $\mathcal{U}_s$. In other words, each user k will be served by a subset of APs (not all APs) within PC $\mathcal{C}_s$, which is referred to as the user-centric cluster for user k . We use the binary variable a_{mk} to show the user assignment for each AP m , so that

$$a_{mk} \triangleq \begin{cases} 1, & \text{if AP } m \text{ associates with user } k, \\ 0, & \text{Otherwise,} \end{cases} \quad \forall m, k. \quad (2)$$

Therefore, the $L \times 1$ signal transmitted by the m -th AP in $\mathcal{C}_s$ ³ can be expressed as

$$\mathbf{s}_m = \sum_{k \in \mathcal{U}_s} a_{mk} \sqrt{\eta_{mk}} \mathbf{q}_{mk} x_k, \quad \forall m \in \mathcal{C}_s, \quad (3)$$

where $\mathbf{q}_{mk}$ denotes the downlink precoding vector constructed for the m -th AP to the k -th user. Moreover, x_k is the information symbol intended for the k -th user, $\mathbb{E}\{x_k x_k^H\} = 1$, and η_{mk} represents the k -th user power control coefficient. Each AP m is required to meet the power constraint

$$\sum_{k \in \mathcal{U}_s} \|a_{mk} \sqrt{\eta_{mk}} \mathbf{q}_{mk}\|_2^2 \leq P, \quad \forall m \in \mathcal{C}_s, \quad (4)$$

where P is the maximum transmit power of each AP. The k -th user receives signal contributions from the APs in $\mathcal{C}_s$; the observable signal is given by

$$\begin{aligned} r_k = & \underbrace{\sum_{m \in \mathcal{C}_s} a_{mk} \sqrt{\eta_{mk}} \mathbf{g}_{mk}^H \mathbf{q}_{mk} x_k}_{\text{Desired signal}} + \\ & \underbrace{\sum_{m \in \mathcal{C}_s} \sum_{\substack{k' \in \mathcal{U}_s \\ k' \neq k}} a_{mk'} \sqrt{\eta_{mk'}} \mathbf{g}_{mk'}^H \mathbf{q}_{mk'} x_{k'}}_{\text{Intra-cluster interference}} + \\ & \underbrace{\sum_{m \in \mathcal{C}_{-s}} \sum_{k' \in \mathcal{U}_{-s}} a_{mk'} \sqrt{\eta_{mk'}} \mathbf{g}_{mk'}^H \mathbf{q}_{mk'} x_{k'}}_{\text{Inter-cluster interference}} + n_k, \end{aligned} \quad (5)$$

²Orthogonal pilot assignment assumption is applicable in many scenarios, particularly when the coherence interval is sufficiently long and/or the number of users is not excessively large. In practice, environments with medium or low mobility often experience a relatively long coherence interval [12], [28], [29]. However, in high-mobility or dense network settings, non-orthogonal pilots are necessary, which require more advanced techniques, such as pilot contamination mitigation.

³In what follows, for ease of notation, we focus on PC $\mathcal{C}_s$ and do not include the PC index in some system parameters.

¹While this coordination introduces additional overhead, detailed modeling of this overhead is beyond the scope of this paper and is left for protocol-level and cross-layer investigations.

where $\mathcal{C}_{-s} = \mathcal{C} \setminus \mathcal{C}_s$ and $n_k \sim (0, \sigma^2)$ is the additive white Gaussian noise. The first term in (5) represents the desired signal part, while the second term represents the intra-cluster interference (all the signal components intended for user $k' \in \mathcal{U}_s, k' \neq k$ from the APs in cluster $\mathcal{C}_s$), and the third term is inter-cluster interference (all the signal components intended for user $k' \in \mathcal{U}_{-s}$ from the APs in other clusters than $\mathcal{C}_s$).

C. Fronthaul Requirements

We formulate the total fronthaul requirement at each AP m for transmission from BBH m to BBL m which consists of two parts: 1) fronthaul requirement for downlink data transmission, $\text{FH}_{m,\text{da}}$, and 2) fronthaul requirement for sending the precoding vectors, $\text{FH}_{m,\text{pr}}$. To this end, the per-AP fronthaul capacity constraint can be written as

$$\text{FH}_{m,\text{da}} + \text{FH}_{m,\text{pr}} \leq \text{FH}_{\max}, \quad \forall m \in \mathcal{C}_s, \quad (6)$$

where $\text{FH}_{\max}$ is the per-AP maximum fronthaul capacity. More specifically, for each AP m , the fronthaul consumption for transmitting information symbols to its associated users with $a_{mk} = 1$, utilizing packed-based evolved common public radio interface (eCPRI) for the fronthaul transmission, is given by

$$\text{FH}_{m,\text{da}} = \frac{\log_2(M_{\text{mo}}) N_{\text{sub}} N_o \sum_{k \in \mathcal{U}_s} a_{mk}}{\varepsilon_{\text{cp}} \delta_{\text{da}}}, \quad (7)$$

where M_{mo} is the modulation cardinality, N_{sub} is the number of OFDM subcarriers, while N_o is the number of OFDM symbols. Moreover, δ_{da} shows the transmit delay for the data, and ε_{cp} is the efficiency of the CPRI. In addition, the required fronthaul requirement for sending the precoding vectors can be written as

$$\text{FH}_{m,\text{pr}} = \frac{2L \sum_{k \in \mathcal{U}_s} a_{mk} N_{\text{bits}} N_{\text{Gran}}}{\varepsilon_{\text{cp}} \delta_{\text{pr}}}, \quad (8)$$

where N_{bits} denotes the number of quantization bits, while N_{Gran} shows the precoding granularity, δ_{pr} shows the transmit delay of the precoding weights. One intuitive observation from (7) and (8) is that the fronthaul consumption scales with the total number of users that each AP serves. Accordingly, in a fronthaul limited CF-mMIMO system with the fronthaul constraint (6), we must constrain the downlink traffic bandwidth to be fronthauled by restricting the number of users each AP $m \in \mathcal{C}_s$ serves as

$$\sum_{k \in \mathcal{U}_s} a_{mk} \leq \left\lfloor \frac{\text{FH}_{\max}}{(\alpha_1 \log_2(M_{\text{mo}}) + \alpha_2)} \right\rfloor \triangleq K_{\max}, \quad (9)$$

where $\lfloor \cdot \rfloor$ is the floor function, $\alpha_1 \triangleq \frac{N_{\text{sub}} N_o}{\varepsilon_{\text{cp}} \delta_{\text{da}}}$, and $\alpha_2 \triangleq \frac{2L N_{\text{bits}} N_{\text{Gran}}}{\varepsilon_{\text{cp}} \delta_{\text{pr}}}$.

On the other hand, a higher M_{mo} is essential to achieve the high SE in the system. However, from (7), fronthaul consumption for sending information symbols increases with M_{mo} , thus, there is a trade-off between the SE and fronthaul requirement. From the information-theoretic perspective, modulation-constrained achievable SE, $\mathcal{R}_{\text{mo}}$, is limited by the AWGN channel capacity, $\bar{\mathcal{C}}$, such that $\mathcal{R}_{\text{mo}} < \bar{\mathcal{C}}$. Moreover, $\mathcal{R}_{\text{mo}}$ cannot exceed the entropy of the modulation constellation, i.e.,

$\mathcal{R}_{\text{mo}} < \log_2(M_{\text{mo}})$. Accordingly, $\mathcal{R}_{\text{mo}}$ can be upper-bounded as $\mathcal{R}_{\text{mo}} \leq \min(\bar{\mathcal{C}}, \log_2(M_{\text{mo}}))$ [30]. Therefore, we consider the following achievable SE constraint

$$R_k \leq \log_2(M_{\text{mo}}), \quad (10)$$

where R_k denotes the achievable SE for user k , in addition to the fronthaul constraint given in (9).

III. INSTANTANEOUS CSI-BASED CLUSTER-WISE SUM PSEUDO-SE MAXIMIZATION

Here, we assume that the users perfectly know the channels. These channels can be estimated by the users using downlink pilots. We would like to highlight that we also assume the channel estimation error at both users and APs is very small, i.e., the estimated channel $\hat{\mathbf{g}}_{mk}$ is almost identical to the true channel $\mathbf{g}_{mk}$, $\forall m, k$. This implies that the current results for the instantaneous CSI-based design serve as an upper bound. Given the instantaneous CSI, the achievable SE for user k can be written as:

$$R_k = \log_2(1 + \text{SINR}_k), \quad (11)$$

where

$$\text{SINR}_k = \frac{|\sum_{m \in \mathcal{M}} a_{mk} \sqrt{\eta_{mk}} \mathbf{g}_{mk}^H \mathbf{q}_{mk}|^2}{\sum_{k' \in \mathcal{K}} |\sum_{m \in \mathcal{M}} a_{mk'} \sqrt{\eta_{mk'}} \mathbf{g}_{mk'}^H \mathbf{q}_{mk'}|^2 + \sigma^2}. \quad (12)$$

A. Hybrid Leakage-Intra-Cluster-Interference

Based on (5) the strength of the desired signal component for user $k \in \mathcal{U}_s$ is given by

$$\text{DS}_k = \left| \sum_{m \in \mathcal{C}_s} a_{mk} \sqrt{\eta_{mk}} \mathbf{g}_{mk}^H \mathbf{q}_{mk} \right|^2. \quad (13)$$

while the strength of the intra-cluster interference at user k can be written as

$$\text{ICI}_k = \sum_{k' \in \mathcal{U}_s, k' \neq k} \left| \sum_{m \in \mathcal{C}_s} a_{mk'} \sqrt{\eta_{mk'}} \mathbf{g}_{mk'}^H \mathbf{q}_{mk'} \right|^2. \quad (14)$$

Now, we define the quantity, called *leakage* interference experienced by the users in $\mathcal{U}_{-s}$ from APs in $\mathcal{C}_s$ by serving user k as

$$\text{L}_k = \sum_{k' \in \mathcal{U}_{-s}} \left| \sum_{m \in \mathcal{C}_s} \sqrt{t_{k'}} a_{mk} \sqrt{\eta_{mk}} \mathbf{g}_{mk}^H \mathbf{q}_{mk} \right|^2, \quad (15)$$

where the binary parameter $t_{k'}$ represents the association assumption about user $k' \in \mathcal{U}_{-s}$. In particular, $t_{k'} = 1$ if user k' is scheduled by at least one AP in $\mathcal{C}_{-s}$; $t_{k'} = 0$ otherwise. We thus define hybrid expressions in terms of so-called *SLINR* that account for both the leakage interference and intra-cluster interference as

$$\text{SLINR}_k = \frac{\text{DS}_k}{\text{ICI}_k + \text{L}_k + \sigma^2}. \quad (16)$$

The SLINR expression in (16) depends only on locally constructed precoding vector and local CSI in each PC. Accordingly, a pseudo-SE between each PC $\mathcal{C}_s$ and its user $k \in \mathcal{U}_s$ is defined as

$$\zeta_k = \log_2(1 + \text{SLINR}_k). \quad (17)$$

B. Problem Formulation

In this section, we aim to jointly optimize precoding, down-link transmit powers, and user association for the maximization of the system sum pseudo-SE, subject to per-AP fronthaul capacity and maximum transmit power constraints for the given PC $\mathcal{C}_s$. Accordingly, in what follows we formulate joint optimization problem for a given small-scale fading coherence time. We would like to highlight that the sum pseudo-SE optimization is motivated by two main factors. First, it enables the development of a cluster-wise processing by relying solely on local variables (i.e., precoding vectors, power allocation coefficients, and CSI) in the given PC. In fact, in actual SE-based designs, the APs should have global CSI and precoding knowledge. In practice, it is not easy/scalable to measure/obtain the CSI between different APs and users at each AP, especially when there are high number of users or APs in the network. Second, pseudo-SE criterion emphasizes the importance of maximizing the useful signal while minimizing leakage and intra-cluster interference. By minimizing leaked interference to other users, users in other PCs improve their SINR by reducing the interference they experience. Additionally, addressing intra-cluster interference benefits users within the same PC. Balancing leakage and intra-cluster interference prevents uniform scaling of the AP's beam power in both signal and leakage terms during optimization [23].

We note that, for calculating the pseudo-SE of user $k \in \mathcal{C}_s$ as given in (17), PC $\mathcal{C}_s$ requires the parameter $t_{k'}$, which accounts for the association decisions of users in other PCs, i.e., $\mathcal{C}_{-s}$. In centralized resource allocation, each AP can have knowledge about the association of all the users. However, in cluster-wise processing, a PC might not know the user associations in other PCs. To this end, for the cluster-wise resource allocation, it is reasonable to assume that $t_{k'} = 1$, i.e., user k' is scheduled by at least one of the APs in its serving cluster.

For convenience, let $\mathbf{a}_s = \{a_{mk} : m \in \mathcal{C}_s, k \in \mathcal{U}_s\}$ denote the user-association control variable for PC $\mathcal{C}_s$, $\boldsymbol{\eta}_s$ denote the set of power control coefficients, $\boldsymbol{\eta}_s = \{\eta_{mk} : m \in \mathcal{C}_s, k \in \mathcal{U}_s\}$. Also, let $\mathbf{q}_s = \{\mathbf{q}_{mk} : m \in \mathcal{C}_s, k \in \mathcal{U}_s\}$ denote the collective precoding vector from APs in PC $\mathcal{C}_s$ to user k . Accordingly, for the PC $\mathcal{C}_s$, the joint optimization problem can be formulated as

$$\max_{\boldsymbol{\eta}_s, \mathbf{a}_s, \mathbf{q}_s} \sum_{k \in \mathcal{U}_s} w_k \zeta_k(\boldsymbol{\eta}_s, \mathbf{a}_s, \mathbf{q}_s) \quad (18a)$$

$$\text{st.} \quad \eta_{mk} \geq 0, \quad m \in \mathcal{C}_s, k \in \mathcal{U}_s, \quad (18b)$$

$$R_k \leq \log_2(M_{\text{mo}}), \quad k \in \mathcal{U}_s, \quad (18c)$$

$$\sum_{k \in \mathcal{U}_s} a_{mk} \leq K_{\text{max}}, \quad \forall m \in \mathcal{C}_s, \quad (18d)$$

$$\sum_{k \in \mathcal{U}_s} \|a_{mk} \sqrt{\eta_{mk}} \mathbf{q}_{mk}\|_2^2 \leq P, \quad \forall m \in \mathcal{C}_s. \quad (18e)$$

where w_k presents the priority weight associated with user k .

Fronthaul constraint (18c) includes the actual achievable SE R_k , $\forall k \in \mathcal{U}_s$, which inherently couples all users and APs across different PCs. This strong interdependence makes it challenging to incorporate the constraint directly within

a decentralized or cluster-wise optimization framework. As such, solving the optimization with this constraint in place would render the proposed scalable, distributed algorithm design infeasible. To address this issue, we investigated three strategies: 1) Relaxation approach: we solve the cluster-wise optimization without enforcing constraint (18c) during the iterative algorithm, and enforce it in a post-processing step on the resulting power allocation; 2) Approximation approach: we approximate R_k by an upper bound R_k^{up} , which considers only intra-cluster interference and ignores inter-cluster effects, thus making it a local function. This allows replacing the actual constraint with $R_k^{\text{up}} \leq K_{\text{max}}$; and 3) Replacement approach: we replace R_k with the pseudo-SE ζ_k , which captures both intra-cluster interference and inter-cluster leakage. Our simulation results show that the approximation and replacement approaches lead to performance degradation—approximately 23% and 10% reductions, respectively, in the system's sum SE (evaluated for $L = 20$, $K = 8$, $M = 10$, $\text{FH}_{\text{max}} = 10$ Gbps, and $M_{\text{mo}} = 32$), compared to the relaxation approach. Importantly, the R_k obtained from the optimization still represents the *achievable SE*. When the post-processing step caps R_k to $\log_2(M_{\text{mo}})$, the system transmits at a SE lower than the achievable SE, which provides an additional safety margin against channel impairments and estimation errors, thereby increasing the transmission reliability. Therefore, although it simplifies the methodology, we adopt the relaxation approach due to its better performance and ease of deployment. Notably, our post-processing step ensures that the final solution satisfies the modulation order and fronthaul constraints, ensuring feasibility in practical implementations. In particular, we consider the optimization problem

$$\max_{\boldsymbol{\eta}_s, \mathbf{a}_s, \mathbf{q}_s} \sum_{k \in \mathcal{U}_s} \omega_k \zeta_k(\boldsymbol{\eta}_s, \mathbf{a}_s, \mathbf{q}_s) \quad (19a)$$

$$\text{st.} \quad \eta_{mk} \geq 0, \quad \forall m \in \mathcal{C}_s, k \in \mathcal{U}_s, \quad (19b)$$

$$\sum_{k \in \mathcal{U}_s} a_{mk} \leq K_{\text{max}}, \quad \forall m \in \mathcal{C}_s, \quad (19c)$$

$$\sum_{k \in \mathcal{U}_s} \|a_{mk} \sqrt{\eta_{mk}} \mathbf{q}_{mk}\|_2^2 \leq P, \quad \forall m \in \mathcal{C}_s, \quad (19d)$$

and then impose the fronthaul constraint as a post-processing step, given by

$$R_{k,\text{post}} = \min(R_k, \log_2(M_{\text{mo}})). \quad (20)$$

Optimization problem (19) is obviously mixed-integer and non-convex due to the user association control variables a_{mk} , $m \in \mathcal{C}_s, k \in \mathcal{U}_s$, the presence of unknown optimization variables $\mathbf{q}_{mk}$ and η_{mk} , $m \in \mathcal{C}_s, k \in \mathcal{U}_s$, appearing as products in both nominator and denominator of ζ_k , and the non-convex constraints (19c) and (19d). To this end, we first define $\bar{\mathbf{q}}_{mk} \triangleq a_{mk} \sqrt{\eta_{mk}} \mathbf{q}_{mk}$. Hence, constraint (19d) is transformed to a convex one as

$$\sum_{k \in \mathcal{U}_s} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq P. \quad (21)$$

Moreover, $\zeta_k(\boldsymbol{\eta}_s, \mathbf{a}_s, \mathbf{q}_s)$ in the objective function of (19) can be re-expressed in terms of $\bar{\mathbf{q}}_s \triangleq \{\bar{\mathbf{q}}_{mk} : m \in \mathcal{C}_s, k \in \mathcal{U}_s\}$ as (22) on top of the next page.

$$\zeta_k(\bar{\mathbf{q}}_s) = \log_2 \left(1 + \frac{\left| \sum_{m \in \mathcal{C}_s} \mathbf{g}_{mk}^H \bar{\mathbf{q}}_{mk} \right|^2}{\sum_{\substack{k' \in \mathcal{U}_s \\ k' \neq k}} \left| \sum_{m \in \mathcal{C}_s} \mathbf{g}_{mk'}^H \bar{\mathbf{q}}_{mk'} \right|^2 + \sum_{k' \in \mathcal{U}_{-s}} \left| \sum_{m \in \mathcal{C}_s} \sqrt{t_{k'}} \mathbf{g}_{mk'}^H \bar{\mathbf{q}}_{mk'} \right|^2 + \sigma^2} \right). \quad (22)$$

To deal with constraint (19c), we notice that each user k is served by AP m , $m \in \mathcal{C}_s$, if and only if its precoding vector $\mathbf{q}_{mk}$ (or equivalently its associated $\bar{\mathbf{q}}_{mk}$) is nonzero. In other words, we can characterize the user association by the indicator function

$$\mathbb{1} \left\{ \|\bar{\mathbf{q}}_{mk}\|_2^2 \right\} = \begin{cases} 0, & \text{if } \|\bar{\mathbf{q}}_{mk}\|_2^2 = 0 \\ 1, & \text{otherwise.} \end{cases} \quad (23)$$

Accordingly, the per-AP fronthaul constraint in (19c) can be casted as

$$\sum_{k \in \mathcal{U}_s} \mathbb{1} \left\{ \|\bar{\mathbf{q}}_{mk}\|_2^2 \right\} \leq K_{\max}, \quad \forall m \in \mathcal{C}_s. \quad (24)$$

In this way, the problem of determining user-AP association $\mathbf{a}_s$, precoding vectors $\mathbf{q}_s$, and power control coefficients η_s is integrated into a single task⁴ determining the precoding vector $\bar{\mathbf{q}}_{mk}$, $m \in \mathcal{C}_s, k \in \mathcal{U}_s$, for each user k . Accordingly, Problem (19) can be equivalently reformulated as

$$\max_{\{\bar{\mathbf{q}}_{mk} | m \in \mathcal{C}_s, k \in \mathcal{U}_s\}} \sum_{k \in \mathcal{U}_s} w_k \zeta_k(\bar{\mathbf{q}}_s) \quad (25a)$$

$$\text{st.} \quad \sum_{k \in \mathcal{U}_s} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq P, \quad \forall m \in \mathcal{C}_s, \quad (25b)$$

$$\sum_{k \in \mathcal{U}_s} \mathbb{1} \left\{ \|\bar{\mathbf{q}}_{mk}\|_2^2 \right\} \leq K_{\max}, \quad \forall m \in \mathcal{C}_s. \quad (25c)$$

C. Modified WMMSE-Based Approach

Problem (25) is still difficult to solve due to the non-convex objective function and fronthaul constraint (25c). To deal with this issue, we first approximate (25c) and then reformulate the objective function into an equivalent form by exploiting the WMMSE criterion. In particular, we equivalently expressed the indicator function in discrete constraint $\sum_{k \in \mathcal{U}_s} \mathbb{1} \left\{ \|\bar{\mathbf{q}}_{mk}\|_2^2 \right\} \leq K_{\max}$ as an ℓ_0 -norm of a scalar as

$$\mathbb{1} \left\{ \|\bar{\mathbf{q}}_{mk}\|_2^2 \right\} = \left\| \|\bar{\mathbf{q}}_{mk}\|_2^2 \right\|_0, \quad (26)$$

where ℓ_0 -norm is the number of nonzero elements in a vector. Then, we use the re-weighted ℓ_1 -norm approximation technique as $\|\mathbf{x}\|_0 \approx \sum_i \vartheta_i |x_i|$, where x_i denotes the i -th element of vector $\mathbf{x}$ and ϑ_i is the weight associated with x_i , to approximate a nonconvex ℓ_0 -norm by a convex ℓ_1 -norm. Therefore, the fronthaul constraint (25c) is reformulated as

$$\sum_{k \in \mathcal{U}_s} \vartheta_{mk} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq K_{\max}, \quad (27)$$

where ϑ_{mk} is a constant weight associated with the m -th AP and the k -th user and is updated iteratively based on

$$\vartheta_{mk} = \frac{1}{\|\bar{\mathbf{q}}_{mk}\|_2^2 + \epsilon}, \quad (28)$$

with ϵ is a small constant regularization factor. It prevents a zero-valued component in $\|\bar{\mathbf{q}}_{mk}\|_2^2$ from strictly blocking the

nonzero estimate in subsequent iterations. It is notable that the performance of ℓ_1 -norm approximation in (27) is not highly sensitive to the value of ϵ [31]. In addition, the weight update rule (28) is based on the fact that setting ϑ_{mk} to be inversely related to the transmit power level $\|\bar{\mathbf{q}}_{mk}\|_2^2$ ensures that APs with lower power transmission to user k are assigned higher weights. As a result, these APs are pushed to further decrease their transmit power to user k over consecutive iterations.

1) *Problem Transformation*: Now, we can reformulate the optimization problem

$$\max_{\{\bar{\mathbf{q}}_{mk} | m \in \mathcal{C}_s, k \in \mathcal{U}_s\}} \sum_{k \in \mathcal{U}_s} w_k \zeta_k(\bar{\mathbf{q}}) \quad (29a)$$

$$\text{st.} \quad \sum_{k \in \mathcal{U}_s} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq P, \quad \forall m \in \mathcal{C}_s, \quad (29b)$$

$$\sum_{k \in \mathcal{U}_s} \vartheta_{mk} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq K_{\max}, \quad \forall m \in \mathcal{C}_s, \quad (29c)$$

as an equivalent WMMSE problem and use the block coordinate descent method to reach a stationary point of (29). The equivalence between weighted sum SE maximization and WMMSE for MIMO interfering channels is established in [32]. By adopting a similar methodology to that in [32], it can be readily observed that the generalized WMMSE equivalence presented in [32] also applies to the problem defined in (29), which incorporates the introduced sum-pseudo-SE maximization objective and the weighted per-AP power constraint (29c). More precisely, the traditional WMMSE approach considers the mean square error (MSE) at each user k as

$$\hat{e}_k = \mathbb{E} \{ |\hat{u}_k r_k - x_k|^2 \}, \quad (30)$$

where $\hat{u}_k$ is the receiver weight, and takes the sum of the errors over all the users to get the final cost function. Now, we take another view point and consider the modified mean MSE at each user k as $e_k = \mathbb{E} \{ |u_k \bar{r}_k - x_k|^2 \}$ where $\bar{r}_k$ is the pseudo-received signal at user k , which includes the desired signal, intra-cluster interference, and the leakage interference experienced by the users in $\mathcal{U}_{-s}$ from APs in $\mathcal{C}_s$ by serving user k as

$$\begin{aligned} \bar{r}_k = & \underbrace{\sum_{m \in \mathcal{C}_s} a_{mk} \sqrt{\eta_{mk}} \mathbf{g}_{mk}^H \mathbf{q}_{mk} x_k}_{\text{Desired signal}} + \\ & \underbrace{\sum_{k' \in \mathcal{U}_s, k' \neq k} \sum_{m \in \mathcal{C}_s} a_{mk'} \sqrt{\eta_{mk'}} \mathbf{g}_{mk'}^H \mathbf{q}_{mk'} x_{k'}}_{\text{Intra-cluster interference}} + \\ & \underbrace{\sum_{k' \in \mathcal{U}_{-s}} \sum_{m \in \mathcal{C}_s} t_{k'} a_{mk} \sqrt{\eta_{mk}} \mathbf{g}_{mk}^H \mathbf{q}_{mk} \bar{x}_{k'}}_{\text{Leakage interference}}, \end{aligned} \quad (31)$$

where $\bar{x}_{k'}$ is the pseudo-information symbol for user k' . The motivation to look for an alternative MSE in (30) is to obtain an equivalent WMMSE-based approach for the maximization of sum-pseudo SE in (29). The equivalence is explicitly stated

⁴Incorporating user association within the precoding optimization framework does not increase the overall complexity of the algorithm, as it will be discussed in Subsection V-C.

in Proposition 1. Before proceeding, let us introduce the notations $\bar{\mathbf{q}}_{s,k}$ as a collective cluster-wide $\bar{\mathbf{q}}_{mk}$ vector from APs in $\mathcal{C}_s$ to user $k \in \mathcal{U}_s$, $\bar{\mathbf{q}}_{s,k} \triangleq [\bar{\mathbf{q}}_{mk} : m \in \mathcal{C}_s]$, and $\mathbf{g}_{s,k}$ as a collective channel vector from APs in $\mathcal{C}_s$ to user $k \in \mathcal{K}$.

Proposition 1. *The weighted sum-pseudo-SE maximization problem (29) has the same solution as the following WMMSE problem:*

$$\min_{\{\rho_k, u_k, \bar{\mathbf{q}}_{s,k} | k \in \mathcal{U}_s\}} \sum_{k \in \mathcal{U}_s} w_k (\rho_k e_k - \log \rho_k) \quad (32a)$$

$$\text{st.} \quad \sum_{k \in \mathcal{U}_s} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq P, \quad \forall m \in \mathcal{C}_s, \quad (32b)$$

$$\sum_{k \in \mathcal{U}_s} \vartheta_{mk} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq K_{\max}, \quad \forall m \in \mathcal{C}_s, \quad (32c)$$

where ρ_k represents the MSE weight for user k and e_k shows the corresponding MSE, which is given by

$$\begin{aligned} e_k &= \mathbb{E}\{|u_k \bar{r}_k - x_k|^2\} \\ &= u_k^2 \left(\sum_{j \in \mathcal{U}_s} \mathbf{g}_{s,k} \bar{\mathbf{q}}_{s,j} \bar{\mathbf{q}}_{s,j}^H \mathbf{g}_{s,k}^H + \sum_{j \in \mathcal{U}_{-s}} \mathbf{g}_{s,j} \bar{\mathbf{q}}_{s,k} \bar{\mathbf{q}}_{s,k}^H \mathbf{g}_{s,j}^H \right. \\ &\quad \left. + \sigma^2 \right) - 2 \text{Re} \{u_k \mathbf{g}_{s,k} \bar{\mathbf{q}}_{s,k}\} + 1. \end{aligned} \quad (33)$$

Proof. The proofs can be obtained by following similar steps as in [32] for the maximization of weighted sum SE and thus is omitted. $\square$

We highlight that a key advantage of reformulating the sum-pseudo-SE maximization problem (29) as the equivalent WMMSE problem (32a) is that (32a) exhibits convexity with respect to each individual optimization variable ρ_k , u_k , and $\bar{\mathbf{q}}_{mk}$ while holding others fixed. This convexity facilitates the efficient solution of (32a) through the block coordinate descent method, whereby the optimization over ρ_k , u_k , and $\bar{\mathbf{q}}_{mk}$ is conducted iteratively, as it be outlined as follows:

- For fixed values of $\bar{\mathbf{q}}_{s,k}$ and ρ_k , the objective function in (32a) can be minimized with respect to u_k by setting its first-order derivative to zero, resulting in the MMSE receiver

$$\begin{aligned} u_k^{\text{opt}} &= \left(\sum_{j \in \mathcal{U}_s} \mathbf{g}_{s,k} \bar{\mathbf{q}}_{s,j} \bar{\mathbf{q}}_{s,j}^H \mathbf{g}_{s,k}^H + \sum_{j \in \mathcal{U}_{-s}} \mathbf{g}_{s,j} \bar{\mathbf{q}}_{s,k} \bar{\mathbf{q}}_{s,k}^H \mathbf{g}_{s,j}^H \right. \\ &\quad \left. + \sigma^2 \right)^{-1} \mathbf{g}_{s,k} \bar{\mathbf{q}}_{s,k}, \quad \forall k \in \mathcal{U}_s. \end{aligned} \quad (34)$$

- The optimum value for ρ_k that minimizes the objective function of (32a) for the fixed u_k and $\bar{\mathbf{q}}_{s,k}$ is obtained by taking the first-order derivative and equating it to zero as

$$\rho_k^{\text{opt}} = e_k^{-1}, \quad \forall k \in \mathcal{U}_s. \quad (35)$$

- As we discuss earlier, the WMMSE method for maximizing the sum-pseudo SE involves iteratively updating one of the three sets of variables ρ_k , u_k , and $\bar{\mathbf{q}}_{mk}$ —while holding the others constant, to approach a local optimum. The update for u_k has been detailed in (34), and the update for ρ_k in (35). The next step is to determine the optimal values for $\bar{\mathbf{q}}_{mk}$, given the current values of u_k and ρ_k . To do this, we substitute the expression for e_k from (33) into the objective function in (32a). The resulting optimization problem for finding the optimal

Algorithm 1 Instantaneous CSI-Based Sum-Pseudo-SE Maximization With Modified WMMSE Approach at PC $\mathcal{C}_s$

- 1: **Initialize:** $\vartheta_{mk}^{(0)}$, $\bar{\mathbf{q}}_{mk}^{(0)}$, $\forall k \in \mathcal{U}_s, \forall m \in \mathcal{C}_s$, iteration index $i = 0$, convergency accuracy ξ .
- 2: **while** $\frac{|\sum_{k \in \mathcal{U}_s} w_k \zeta_k^{(i)} - \sum_{k \in \mathcal{U}_s} w_k \zeta_k^{(i-1)}|}{\sum_{k \in \mathcal{U}_s} w_k \zeta_k^{(i-1)}} < \xi$ **do**
- 3: $i = i + 1$;
- 4: Calculate $u_k^{(i)}$ according to (34) with $\bar{\mathbf{q}}_{mk}$ fixed, $k \in \mathcal{U}_s$
- 5: Calculate $e_k^{(i)}$ according to (33) with $\bar{\mathbf{q}}_{mk}$ and u_k fixed, $k \in \mathcal{U}_s$.
- 6: Update ρ_k according to (35).
- 7: Calculate the optimal transmit beamformer $\bar{\mathbf{q}}_{mk}$ with $\mathbf{u}_k$ and ρ_k fixed $\forall k \in \mathcal{U}_s, m \in \mathcal{C}_s$, by solving the problem (36a).
- 8: Calculate $\vartheta_{mk}^{(i)}$ according to (28).
- 9: **end while**

transmit beamformer $\bar{\mathbf{q}}_{mk}$ is a quadratically constrained quadratic programming (QCQP) problem as

$$\begin{aligned} \min_{\{\bar{\mathbf{q}}_{mk} | m \in \mathcal{C}_s, k \in \mathcal{U}_s\}} & \sum_{k \in \mathcal{U}_s} \bar{\mathbf{q}}_{s,k}^H \left(\sum_{j \in \mathcal{U}_s} w_j \rho_j u_j^2 \mathbf{g}_{s,j} \mathbf{g}_{s,j}^H + w_k \rho_k \right. \\ & \times \left. \sum_{j \in \mathcal{U}_{-s}} u_k^2 \mathbf{g}_{s,j} \mathbf{g}_{s,j}^H \right) \bar{\mathbf{q}}_{s,k} - 2 \sum_{k \in \mathcal{U}_s} w_k \rho_k \text{Re} \{u_k \mathbf{g}_{s,k}^H \bar{\mathbf{q}}_{s,k}\} \end{aligned} \quad (36a)$$

$$\begin{aligned} \text{st.} \quad & \sum_{k \in \mathcal{U}_s} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq P, \quad \forall m \in \mathcal{C}_s \\ & \sum_{k \in \mathcal{U}_s} \vartheta_{mk} \|\bar{\mathbf{q}}_{mk}\|_2^2 \leq K_{\max}, \quad \forall m \in \mathcal{C}_s, \end{aligned} \quad (36b)$$

which can be addressed using a standard convex optimization tool like CVX.

The solution of (32a) is summarized in **Algorithm 1**.⁵

IV. STATISTICAL CSI-BASED CLUSTER-WISE SUM HARDENING-BASED PSEUDO-SE MAXIMIZATION

In this section, each AP m uses the channel estimates $\hat{\mathbf{g}}_{mk}$ in Section II-A to precode the information signals before transmitting them to its assigned users. Moreover, each user k relies only on statistical CSI to detect x_k from the received signal in (5). This eliminates the need for users to know the instantaneous channel estimates, which in turn reduces the amount of information that must be exchanged. Accordingly, by utilizing the widely adopted hardening bounding technique [2], the achievable SE for user k can be expressed in closed form as [33]:

$$\tilde{R}_k = \log_2(1 + \widetilde{\text{SINR}}_k), \quad (37)$$

where

$$\widetilde{\text{SINR}}_k = \frac{(\mathbf{d}_k^T \mathbf{p}_k)^2}{\sum_{j \in \mathcal{K}} \mathbf{p}_j^T \mathbf{B}_{kj} \mathbf{p}_j - (\mathbf{d}_k^T \mathbf{p}_k)^2 + \sigma^2}, \quad (38)$$

⁵While this work employs a WMMSE-based optimization approach, using alternative optimization methods such as fractional programming or alternating direction method of multipliers (ADMM) approach for solving the QCQP subproblem could be explored in future research.

with

- $\mathbf{d}_k = [d_{1k}, \dots, d_{Mk}]^T \in \mathbb{R}^{M \times 1}$, with $d_{mk} = |\mathbb{E}\{\mathbf{g}_{mk}^H \mathbf{q}_{mk}\}|$, $m \in \mathcal{M} \in \mathbb{R}^{M \times 1}$.
- $\mathbf{p}_k = [p_{1k}, \dots, p_{Mk}]^T$, with $p_{mk} = a_{mk} \sqrt{\eta_{mk}}$, $m \in \mathcal{M}$.
- $\mathbf{B}_{kj} \in \mathbb{R}^{M \times M}$, with $[\mathbf{B}_{kj}]_{lm} = \text{Re}(\mathbb{E}\{\mathbf{g}_{lk}^H \mathbf{q}_{lj} \mathbf{q}_{mj}^H \mathbf{g}_{mk}\})$, $l, m \in \mathcal{M}$.

In addition, a closed-form expression for the hardening-based pseudo-SE at user k can be written as

$$\tilde{\zeta}_k = \log_2(1 + \widetilde{\text{SLINR}}_k), \quad (39)$$

where

$$\widetilde{\text{SLINR}}_k = \frac{(\tilde{\mathbf{d}}_k^T \tilde{\mathbf{p}}_k)^2}{\sum_{j \in \mathcal{U}_s} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{B}}_{kj} \tilde{\mathbf{p}}_j + \sum_{j \in \mathcal{U}_{-s}} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{F}}_{kj} \tilde{\mathbf{p}}_k - (\tilde{\mathbf{d}}_k^T \tilde{\mathbf{p}}_k)^2 + \sigma^2}, \quad (40)$$

- $\tilde{\mathbf{d}}_k = [\tilde{d}_{1k}, \dots, \tilde{d}_{|C_s|k}]^T \in \mathbb{R}^{|C_s| \times 1}$, with $\tilde{d}_{ik} = |\mathbb{E}\{\mathbf{g}_{lk}^H \mathbf{q}_{\ell k}\}|$, $\ell = C_s\{i\}$, $\forall i \in \{1, \dots, |C_s|\}$.
- $\tilde{\mathbf{p}}_k = [\tilde{p}_{1k}, \dots, \tilde{p}_{|C_s|k}]^T \in \mathbb{R}^{|C_s| \times 1}$, with $\tilde{p}_{ik} = a_{\ell k} \sqrt{\eta_{\ell k}}$, $\ell = C_s\{i\}$, $\forall i \in \{1, \dots, |C_s|\}$.
- $\tilde{\mathbf{B}}_{kj} \in \mathbb{R}^{|C_s| \times |C_s|}$, with $[\tilde{\mathbf{B}}_{kj}]_{lm} = \text{Re}(\mathbb{E}\{\mathbf{g}_{lk}^H \mathbf{q}_{lj} \mathbf{q}_{oj}^H \mathbf{g}_{ok}\})$, $\ell = C_s\{l\}$, $o = C_s\{m\}$, $\forall l, m \in \{1, \dots, |C_s|\}$.
- $\tilde{\mathbf{F}}_{kj} \in \mathbb{R}^{|C_s| \times |C_s|}$, with $[\tilde{\mathbf{F}}_{kj}]_{lm} = \text{Re}(\mathbb{E}\{\mathbf{g}_{lk}^H \mathbf{q}_{\ell k} \mathbf{q}_{ok}^H \mathbf{g}_{oj}\})$, $\ell = C_s\{l\}$, $o = C_s\{m\}$, $\forall l, m \in \{1, \dots, |C_s|\}$.

A. Problem Formulation and Modified WMMSE-Based Solution

We now propose a statistical CSI-based resource allocation design for the given precoding. In this approach, resource allocation is updated according to the large-scale fading time scale (statistical channel properties). For the optimization, any precoding design can be utilized, but in the simulation results we will consider cluster-wise MMSE precoding scheme. Let $\hat{\mathbf{G}}_s$ be an $L|C_s| \times |\mathcal{U}_s|$ collective channel estimation matrix for corresponding APs in set C_s . More specifically, $\hat{\mathbf{G}}_s$ consists of $|C_s| \times |\mathcal{U}_s|$ vectors of dimension $L \times 1$, $\hat{\mathbf{g}}_{ij}$, each corresponding to a particular AP i in set C_s and user j in set $\mathcal{U}_s$ as

$$\hat{\mathbf{G}}_s = [\hat{\mathbf{g}}_{ij} : i \in C_s, j \in \mathcal{U}_s]. \quad (41)$$

By using MMSE precoding scheme the whole precoding vector constructed for the APs in C_s can be expressed as

$$\tilde{\mathbf{Q}}_s = \hat{\mathbf{G}}_s \left((\tilde{\mathbf{P}}_s \circ \hat{\mathbf{G}}_s)^H \hat{\mathbf{G}}_s + \sigma^2 \mathbf{I}_{|\mathcal{U}_s|} \right)^{-1}, \quad (42)$$

where, $\tilde{\mathbf{P}}_s = [\tilde{p}_{ij} \mathbf{1}_{L \times 1} : i \in C_s, j \in \mathcal{U}_s]$, $\circ$ is the Hadamard or entry-wise product. Now, each BBH construct its precoding vector for transmission to user k by choosing the corresponding column vector. In particular, let AP m correspond to the i -th element of set C_s , $i \in \{1, \dots, |C_s|\}$, and user k correspond to the j -th element of set $\mathcal{U}_s$, $j \in \{1, \dots, |\mathcal{U}_s|\}$. Accordingly, the downlink precoding vector constructed for the m -th AP to the k -th user can be calculated as the vector of $\tilde{\mathbf{Q}}_s$ obtained by selecting the L rows a_{mk} to $\tilde{a}_{mk}$ from j -th column as $\mathbf{q}_{mk} = \frac{\tilde{\mathbf{q}}_{mk}}{\|\tilde{\mathbf{q}}_{mk}\|_2}$ with

$$\tilde{\mathbf{q}}_{mk} = [\tilde{\mathbf{Q}}_s]_{(a_{mk} : \tilde{a}_{mk}, j)}, \quad (43)$$

where $a_{mk} = (i-1) \times L + 1$ and $\tilde{a}_{mk} = a_{mk} + L - 1$.

Now, for PC C_s we formulate the following hardening-based sum pseudo-SE maximization problem

$$\max_{\boldsymbol{\eta}_s, \mathbf{a}_s} \sum_{k \in \mathcal{U}_s} \tilde{w}_k \tilde{\zeta}_k(\boldsymbol{\eta}_s, \mathbf{a}_s) \quad (44a)$$

$$\text{st. } \eta_{mk} \geq 0, \quad \forall m \in C_s, k \in \mathcal{U}_s, \quad (44b)$$

$$\sum_{k \in \mathcal{U}_s} a_{mk} \leq K_{\max}, \quad \forall m \in C_s, \quad (44c)$$

$$\sum_{k \in \mathcal{U}_s} \tilde{p}_{mk}^2 \leq P, \quad \forall m \in C_s. \quad (44d)$$

where $\tilde{w}_k$ presents the priority weight associated with user k . Following the same approach as in the previous section, we can formulate the equivalent WMMSE problem to optimize power control and user association as

$$\min_{\{\tilde{\rho}_k, u_k, \tilde{\mathbf{p}}_k | k \in \mathcal{U}_s\}} \sum_{k \in \mathcal{U}_s} \tilde{w}_k (\tilde{\rho}_k \tilde{e}_k - \log \tilde{\rho}_k) \quad (45a)$$

$$\text{st. } \sum_{k \in \mathcal{U}_s} \tilde{p}_{mk}^2 \leq P, \quad \forall m \in C_s, \quad (45b)$$

$$\sum_{k \in \mathcal{U}_s} \tilde{\vartheta}_{mk} \tilde{p}_{mk}^2 \leq K_{\max}, \quad \forall m \in C_s, \quad (45c)$$

where

$$\tilde{\vartheta}_{mk} = (\tilde{p}_{mk}^2 + \epsilon)^{-1}, \quad (46)$$

and

$$\tilde{e}_k = \tilde{u}_k^2 \left(\sum_{j \in \mathcal{U}_s} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{B}}_{kj} \tilde{\mathbf{p}}_j + \sum_{j \in \mathcal{U}_{-s}} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{F}}_{kj} \tilde{\mathbf{p}}_k + \sigma^2 \right) - 2\tilde{u}_k \tilde{\mathbf{d}}_k^T \tilde{\mathbf{p}}_k + 1. \quad (47)$$

For fixed $\tilde{\mathbf{p}}_k$ and $\tilde{\rho}_k$, the objective function in (45a) can be minimized with respect to $\tilde{u}_k$ by setting its first-order derivative to zero, resulting in

$$\tilde{u}_k^{\text{opt}} = \frac{\tilde{\mathbf{d}}_k^T \tilde{\mathbf{p}}_k}{\sum_{j \in \mathcal{U}_s} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{B}}_{kj} \tilde{\mathbf{p}}_j + \sum_{j \in \mathcal{U}_{-s}} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{F}}_{kj} \tilde{\mathbf{p}}_k + \sigma^2}. \quad (48)$$

Accordingly, for fixed $\tilde{u}_k, \tilde{\mathbf{p}}_k$, the optimal MSE weight is

$$\tilde{\rho}_k^{\text{opt}} = (\tilde{e}_k)^{-1}. \quad (49)$$

Finally, with fixed $\tilde{u}_k, \tilde{\rho}_k$, we obtain a QCQP problem with respect to $\tilde{\mathbf{p}}_k$ as

$$\min_{\tilde{\mathbf{p}}} \tilde{f}(\tilde{\mathbf{p}}) \quad (50a)$$

$$\text{st. } \sum_{k \in \mathcal{U}_s} \tilde{p}_{mk}^2 \leq P, \quad \forall m \in C_s, \quad (50b)$$

$$\sum_{k \in \mathcal{U}_s} \tilde{\vartheta}_{mk} \tilde{p}_{mk}^2 \leq K_{\max}, \quad \forall m \in C_s, \quad (50c)$$

where

$$\tilde{f}(\tilde{\mathbf{p}}) = \sum_{k \in \mathcal{U}_s} \tilde{w}_k \tilde{\rho}_k \tilde{u}_k^2 \left(\sum_{j \in \mathcal{U}_s} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{B}}_{kj} \tilde{\mathbf{p}}_j + \sum_{j \in \mathcal{U}_{-s}} \tilde{\mathbf{p}}_j^T \tilde{\mathbf{F}}_{kj} \tilde{\mathbf{p}}_k + \sigma^2 \right) - 2\tilde{w}_k \tilde{\rho}_k \tilde{u}_k \tilde{\mathbf{d}}_k^T \tilde{\mathbf{p}}_k \quad (51)$$

and $\tilde{\mathbf{p}} \triangleq \{\tilde{\mathbf{p}}_k | k \in \mathcal{U}_s\}$.

Both the objective function and the constraints in (50a) are convex and standard convex optimization tools can be used to solve it. The solution of (50a) is shown in **Algorithm 2**.

Algorithm 2 Statistical CSI-Based Sum-Pseudo-SE Maximization With Modified WMMSE Approach at PC $\mathcal{C}_s$

- 1: **Initialize:** $\tilde{v}_{mk}^{(0)}, \tilde{\mathbf{p}}_k^{(0)}, \forall k \in \mathcal{U}_s, \forall m \in \mathcal{C}_s$, iteration index $i = 0$, convergency accuracy ξ .
 - 2: **while** $\frac{|\sum_{k \in \mathcal{U}_s} \tilde{w}_k \tilde{\zeta}_k^{(i)} - \sum_{k \in \mathcal{U}_s} \tilde{w}_k \tilde{\zeta}_k^{(i-1)}|}{\sum_{k \in \mathcal{U}_s} \tilde{w}_k \tilde{\zeta}_k^{(i-1)}} < \xi$ **do**
 - 3: $i = i + 1$;
 - 4: Calculate $\tilde{u}_k^{(i)}$ according to (48) with $\tilde{\mathbf{p}}_k$ fixed, $k \in \mathcal{U}_s$
 - 5: Calculate $\tilde{e}_k^{(i)}$ according to (47) with $\tilde{\mathbf{p}}_k$ and $\tilde{u}_k$ fixed, $k \in \mathcal{U}_s$.
 - 6: Update $\tilde{\rho}_k$ according to (49).
 - 7: Calculate the optimal power control $\tilde{\mathbf{p}}_k$ with $\tilde{u}_k$ and $\tilde{\rho}_k$ fixed $\forall k \in \mathcal{U}_s, m \in \mathcal{C}_s$, by solving the problem (50a)
 - 8: Calculate $\tilde{v}_{mk}^{(i)}$ according to (46)
 - 9: **end while**
-

V. NUMERICAL RESULTS

We consider a CF-mMIMO system where the APs and users are randomly distributed within a 1×1 km², with wrapped-around edges to eliminate boundary effects. In cluster-wise processing, the available APs are divided into S disjoint PCs, which can be based on factors such as AP location and interference relationships. In our simulations, we group APs into PCs based on their geographical locations. Moreover, for each user k , we choose a cluster $\mathcal{C}_s$ comprising the APs providing the highest sum average received power at user k . In particular, we select cluster $\mathcal{C}_s$ for user k with

$$s = \arg \max_{s \in \{1, \dots, S\}} \sum_{m \in \mathcal{C}_s} \beta_{mk}. \quad (52)$$

Each AP can serve up to $K_{\max}$ users out of a set of K users. The value of parameter $K_{\max}$ depends on the system and fronthaul parameters and is determined based on (9). We evaluate the performance of CF-mMIMO system with cluster-wise processing under limited fronthaul capacity and transmit power constraint relying on our proposed Algorithm 1 and Algorithm 2. We compare the performance of the following cases:

- Network-wide (centralized) processing with $S = 1$: In this case, there is one PC consisting of all the APs in the network. The performance of this case can be considered the fronthaul-limited upper bound.
- Cluster-wise (decentralized) processing with $S > 1$ PCs: In this case, APs are divided into $S \in \{2, 4\}$ PCs. In each PC, we resort to the proposed instantaneous CSI-based Algorithm 1 (statistical CSI-based Algorithm 2) to design precoding, power allocation, and user association (power allocation and user association).

A. Parameters and Setup

The maximum transmit power for training pilot sequences and for the transmit power at each AP is set to 100 mW. The noise power is $\sigma_n^2 = -92$ dBm, while the fronthaul parameters are chosen based on Table I. In addition, we consider 100 MHz bandwidth with 30 kHz subcarrier spacing which corresponds to $N_{\text{sub}} = 3264$ as in Table I. In addition, we set $\tau_u = 2000$

TABLE I: Fronthaul Parameters

Parameter	Value	Parameter	Value
N_{sub}	3264	N_{Gran}	136
ε_{cp}	0.85	$\delta_{\text{pr}}(\delta_{\text{da}})$	$2 \times 10^{-4} (5 \times 10^{-4})$ s
N_o	14 sym	N_{bits}	16

samples, which corresponds to a coherence bandwidth of 200 KHz and a coherence time of 10 ms. The large-scale fading and the path-loss between AP m and user k is modeled as

$$\beta_{mk} = \text{PL}_{mk} 10^{\frac{\sigma_{sh} y_{mk}}{10}}, \quad (53)$$

where PL_{mk} denotes the path loss and $10^{\frac{\sigma_{sh} y_{mk}}{10}}$ represents the shadow fading with standard deviation $\sigma_{sh} = 4$ dB and $y_{mk} \sim \mathcal{CN}(0, 1)$. To model PL_{mk} , we consider the popular three-slope model as described in [2].

B. Results and Discussions

1) *Performance of the Proposed Cluster-wise Weighted Pseudo-SE-Maximization Approaches:* In Figs. 2a and 2b we evaluate the performance of the proposed cluster-wise pseudo-SE-maximization approaches in Algorithm 1 and Algorithm 2 in the fronthaul-aware CF-mMIMO system, respectively. We consider the following benchmark schemes for comparisons:

- Benchmark 1: In this scheme, we consider successive instantaneous CSI-based cluster-wise processing design, where user association, precoding, and power allocation are successively implemented over the small-scale fading time scale. In particular, for user association, each AP sorts the instantaneous channel gains in descending order and independently selects $K_{\max}$ users with the strongest channel gains. For the given user association, the power control coefficients are determined based on the low-complexity heuristic scheme proposed in [10], which has shown excellent performance and serves as a reliable baseline. Using this scheme, the power coefficient used by AP m for transmission to user k is calculated as

$$\eta_{mk} = \frac{1}{\max_m (\sum_{k \in \mathcal{K}} a_{mk})}, \quad m \in \mathcal{C}_s, k \in \mathcal{U}_s. \quad (54)$$

In addition, precoding vectors are designed based on cluster-wise MMSE scheme.

- Benchmark 2: In this scheme, we consider successive statistical CSI-based cluster-wise processing design, where user association and power allocation are successively implemented over the large-scale fading time scale. In particular, for user association, each AP m sorts the large-scale fading coefficients, β_{mk} , in descending order and independently selects $K_{\max}$ users with the strongest channel gains. For power control, given the user association, we use heuristic scheme [10], while precoding vectors are designed based on (43) for the given user association and power control.

The main observations that follow from these simulations are as follows:

- The proposed cluster-wise processing solutions enhance the system performance significantly for both the statistical and instantaneous CSI-based designs. In particular,

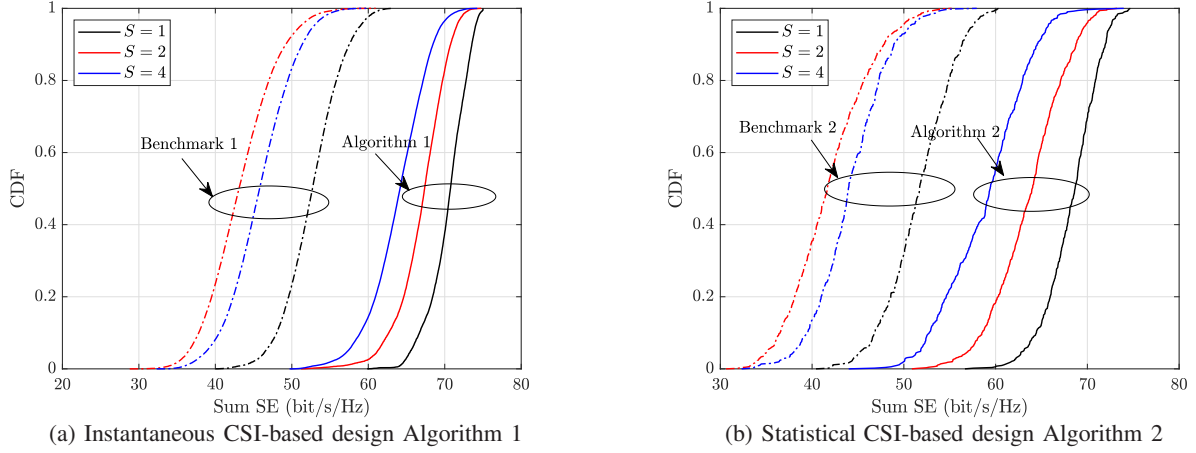


Fig. 2: Comparison among the sum-SE achieved by the proposed Algorithms and benchmark schemes, where $L = 24$, $K = 15$, $M = 10$, $FH_{\max} = 10$ Gbps, and $M_{\text{mo}} = 32$.

when $S = 2$, the joint precoding, power allocation, and user association Algorithm 1 provides a performance gain of up to 56% compared to *Benchmark 1*. Meanwhile, the statistical CSI-based design in Algorithm 2 yields a performance gain of up to 52% compared to *Benchmark 2*. This highlights the advantage of our proposed solutions over the heuristic benchmarks.

- The performance gap between network-wide CF-mMIMO with $S = 1$ and CF-mMIMO with cluster-wise processing, decreases with our proposed solutions in Algorithm 1 and Algorithm 2. More precisely, the performance loss of CF-mMIMO with statistical CSI-based cluster-wise processing compared to the centralized case is around 7% and 16% when $S = 2$ and $S = 4$, respectively. These losses reduce to 4% and 9%, for instantaneous CSI-based design, respectively. This is an interesting result because it shows the importance of deploying multiple PCs in the CF-mMIMO system.

Figure 3 compares the performance of the joint optimization approach in Algorithm 1 for the CF-mMIMO system with $S = 2$ PCs against cases where only the power allocation coefficients or both power allocation and user association variables are optimized, denoted by OPA and OPA-OUA, respectively. It is observed that OPA yields a 28% performance gain over Benchmark 1, and OPA-OUA provides a 15% improvement over OPA by optimizing both power allocation and user association. Also, Algorithm 1 achieves an additional 8% gain over OPA-OUA by jointly optimizing user association, power allocation, and precoding. This demonstrates that the integrated optimization approach in Algorithm 1 significantly outperforms the individual optimizations.

2) *Impact of the Number of Antennas per AP*: Figure 4 presents the average sum-SE performance of the fronthaul-limited CF-mMIMO system with cluster-wise processing as a function of the number of transmit antennas at the AP. The main insights drawn from this figure are as follows.

- Increasing the number of transmit antennas at each AP impacts the sum-SE performance in two ways: (i) it boosts diversity and array gain, and (ii) reduces $K_{\max}$

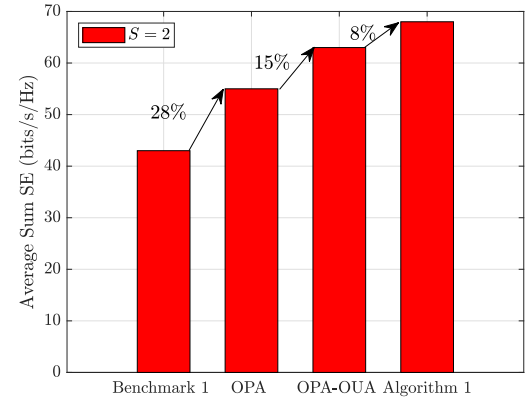


Fig. 3: Comparison among the sum-SE achieved by different optimization approaches, where $L = 24$, $K = 15$, $M = 10$, $FH_{\max} = 10$ Gbps, and $M_{\text{mo}} = 32$.

due to fronthaul limitations (9). However, the first effect dominates and results in a notable enhancement in sum-SE performance, especially for the statistical CSI-based design with higher number of PCs.

- It is evident that for a large number of transmit antennas, CF-mMIMO employing scalable cluster-wise processing along with the proposed WMMSE-based solutions incur only a minor performance loss for decentralization, e.g., around 10% and 12% for instantaneous and statistical CSI-based designs with $S = 4$ PCs, respectively. This is achieved while significantly reducing the computational complexity of cluster-wise processing and the overhead required for CSI acquisition across large processing sets. These results emphasize the benefit of our proposed scalable cluster-wise processing, namely minimizing the need for extensive network-wide processing and coordination among numerous APs, while still providing competitive performance.
- Herein, we also present the results for CF-mMIMO having infinite fronthaul capacity (represented as infinite FH) relying on the proposed WMMSE-based algorithms when

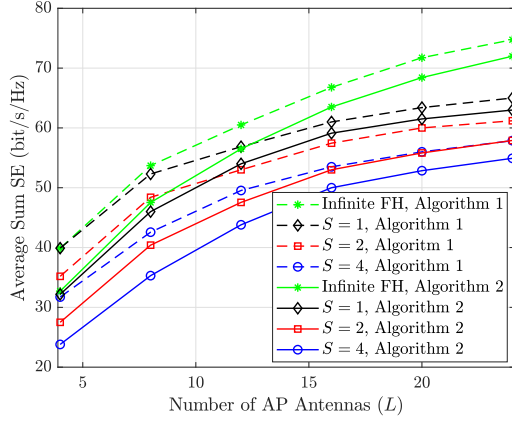


Fig. 4: Average sum SE versus number of AP antennas, L , where $K = 15$, $M = 8$, $\text{FH}_{\max} = 10$ Gbps, and $M_{\text{mo}} = 32$.

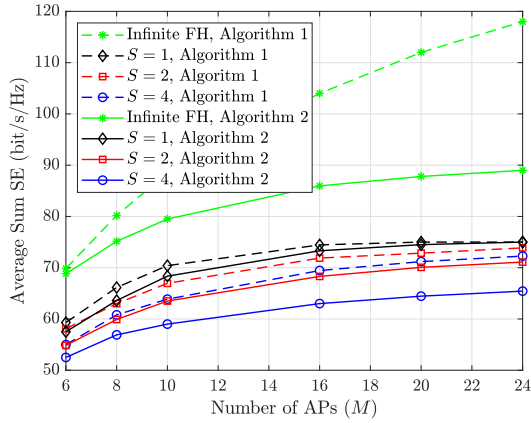


Fig. 5: Average sum SE versus number of APs, M , where $ML = 240$, $K = 15$, $\text{FH}_{\max} = 10$ Gbps, and $M_{\text{mo}} = 32$.

$S = 1$, $K_{\max} = K$, and there is no post-imposing fronthaul constraint (10), i.e., $R_{k,\text{post}} = R_k$. It is observed that the fronthaul constraints lead to the performance loss, which is more pronounced for the CF-mMIMO systems in the regime of large values of L . Nevertheless, our proposed optimization solutions in Algorithm 1 and 2 could potentially make CF-mMIMO with cluster-wise processing competitive compared to network-wide CF-mMIMO with infinite fronthaul capacity.

3) *Impact of the Number of APs*: Figure 5 presents the average sum SE achieved by CF-mMIMO system for different numbers of APs for systems having the same total numbers of service antennas, i.e., $LM = 240$, but different number of APs. The main observations that follow from these simulations are as follows.

- For all the schemes, distributing antennas results in better sum-SE performance due to the additional macro diversity gain.
- It can be observed that, as the number of APs increases, the sum-SE performance of the fronthaul-limited CF-mMIMO system remains relatively unchanged when $M \geq 20$, particularly for the $S = 1$ and $S = 2$ cases. This behaviour contrasts with the infinite fronthaul scenario,

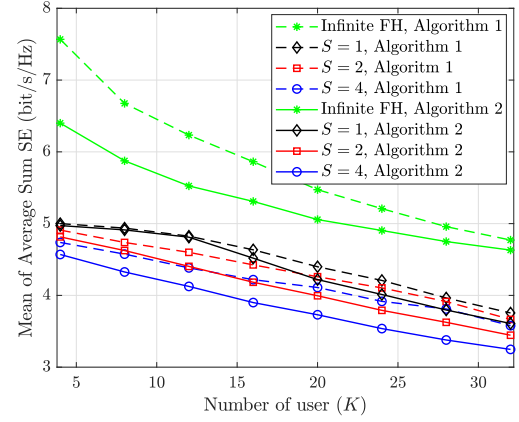


Fig. 6: Mean of average sum SE versus number of users, K , where $L = 24$, $M = 10$, $\text{FH}_{\max} = 10$ Gbps, and $M_{\text{mo}} = 32$.

where the SE increases notably with M . The limited improvement in the fronthaul-limited case is due to the fronthaul constraint in (10), which doesn't allow the SE to increase more than $\log_2(M_{\text{mo}})$. As a result, the fronthaul bottleneck prevents further SE gains as M increases. For this regime the performance gap between network-wise processing and cluster-wise processing under fronthaul constraint significantly reduces and cluster-wise processing is undoubtedly a better choice.

- The performance of CF-mMIMO system with cluster-wise processing relying on statistical CSI-based design Algorithm 2 is fairly close to CF-mMIMO system relying on instantaneous CSI-based design in Algorithm 1 for the large to medium range of number of antennas for different number of PCs. For example, when $L = 40$ or equivalently $M = 6$, the performance gap between statistical CSI-based and instantaneous CSI-based design is less than 5% and 7% when there is $S = 2$ PCs, and $S = 4$ PCs, respectively. This behaviour follows from the fact that the level of channel hardening remarkably increases for higher values of L and hence, using the mean of the effective gain instead of the true channel gains for cluster-wise processing in CF-mMIMO systems works very well. Even when it comes to the network-wide processing, the statistical CSI-based design in the fronthaul-limited CF-mMIMO system is capable to achieve 97% of the average sum-SE with instantaneous CSI-based design. It is an interesting observation, since statistical CSI-based designs provide better trade-offs between performance, complexity, and signaling overhead.

4) *Impact of the Number of Users*: Figure 6 shows the mean of average sum-SE, i.e., (average sum-SE)/ K , of a fronthaul-limited CF-mMIMO system, as a function of the number of users for different number of PCs. We observe that in the regime of small values of K , the performance gap between all the schemes are very small. On the other hand, by increasing K , the sum-SE performance of all cases deteriorates. Nevertheless, the CF-mMIMO system using the cluster-wise processing still yields excellent SE performance compared to network-wide processing. More specifically, when

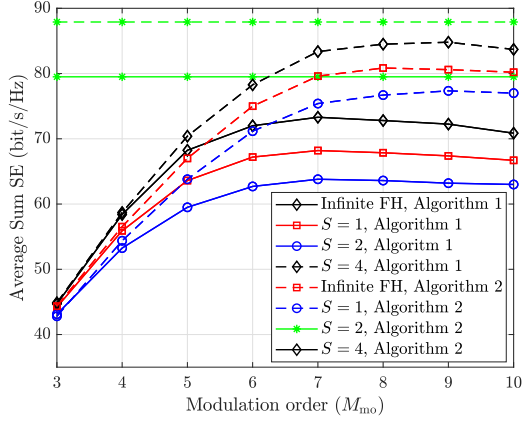


Fig. 7: Average sum SE versus modulation order, M_{mo} , where $L = 24$, $K = 15$, $M = 10$, $FH_{max} = 10$ Gbps.

$K = 28$, the performance loss of cluster-wise processing with $S = 2$ ($S = 4$) PCs compared to the network-wide processing is less than 1.2% (3%) for instantaneous CSI-based design. Importantly, this is the case with statistical CSI-based design; the performance loss of cluster-wise processing with $S = 2$ ($S = 4$) PCs is less than 3% (5%). These negligible performance losses verify the importance of an adequate cluster-wise processing along with statistical CSI-based design to provide a better performance/implementation complexity trade-off compared to its instantaneous CSI-based network-wide processing counterpart.

5) *Effect of the Modulation Order:* In Fig. 7 we investigate the average sum SE performance of the CF-mMIMO system with cluster-wise processing architecture for different number of PCs as a function of M_{mo} . It is observed that there exists an optimal value of M_{mo} for each scheme that maximizes the average sum SE performance. This is reasonable because, on one hand, the fronthaul consumption for transmitting information symbols increases with M_{mo} , which reduces K_{max} . On the other hand, a higher M_{mo} enables higher SE under the fronthaul constraint (10). Therefore, there is a trade-off between the SE and M_{mo} . In addition, the sum-SE performance gap for the systems relying on instantaneous CSI-based design and statistical CSI-based design increases with higher M_{mo} , while the performance gap between cluster-wise processing schemes and network-wide processing is relatively small when $M_{mo} \leq 4$. Therefore, for the application scenarios with lower M_{mo} , cluster-wise processing relying on statistical CSI-based design is undoubtedly a better choice.

6) *Impact of the Available Fronthaul Capacity:* Figure 8 illustrates the effect of the maximum available fronthaul capacity, FH_{max} , on the sum-SE performance of the CF-mMIMO system with the proposed cluster-wise processing. Different values of FH_{max} , correspond to different values of K_{max} , which are calculated based on (9). It is observed that the limited fronthaul capacity reduces the system performance. For example, under the instantaneous CSI-based design, when the fronthaul capacity $FH_{max} = 6$ Gbps (or equivalently $K_{max} = 5$), there is 22% performance loss due to fronthaul limitation for the centralized scheme with $S = 1$ compared to

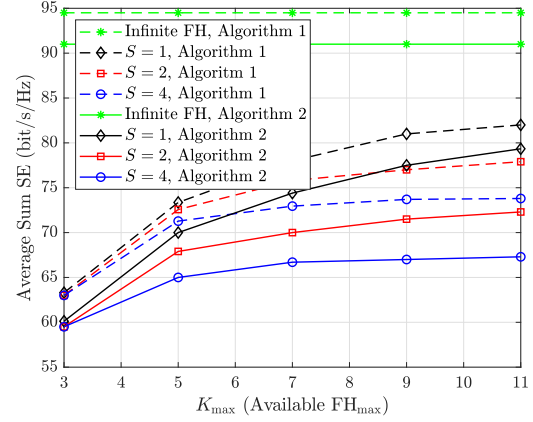


Fig. 8: Average sum SE versus K_{max} (FH_{max}), where $L = 24$, $K = 20$, $M = 8$, $M_{mo} = 32$.

the case of infinite fronthaul links. This loss slightly increase to 23% when $S = 2$. On the other hand, we observe that upon increasing K_{max} the sum-SE performance of all cases increases, specially for centralized schemes. Simulation results also confirm that multiple clusters is better suited for CF-mMIMO network architectures with low-capacity fronthaul links.

Finally, we would like to emphasize that fairness among users is inherently promoted through our optimization framework. Specifically, we adopt a proportional fairness strategy by weighting each user's pseudo-SE with a priority coefficient w_k , which is chosen as the inverse of the average achievable SE experienced by that user. For example, in the Monte Carlo evaluation of the instantaneous CSI-based **Algorithm 1**, w_k at small-scale fading realization n is computed as the inverse of the user's average achievable SE up to realization $n - 1$. Figure 9 presents the CDF of the per-user SE achieved by **Algorithm 1**. The 5th percentile SE (i.e., the SE value that 95% of users exceed) is reported as 1.7, 1.4, and 1 bit/s/Hz for $S = 1, 2, 4$, respectively. These results indicate that even the least-served users maintain a non-negligible SE, thereby confirming that our approach achieves a desirable balance between efficiency and fairness.

C. Computational Complexity

Here, we discuss the complexity of the proposed instantaneous CSI-based cluster-wise design in **Algorithm 1** and statistical CSI-based cluster-wise design in **Algorithm 2**. We note that in the instantaneous CSI-based design, both precoding and resource allocation for each PC are re-calculated based on the small-scale fading time scale (instantaneous channel conditions). It is important to note that small-scale fading coefficients fluctuate rapidly across time and frequency. Therefore, the required channel acquisition and computational complexity of **Algorithm 1** become prohibitive as the network size increases. However, in statistical CSI-based **Algorithm 2**, resource allocation is updated according to the large-scale fading time scale (statistical channel properties). Large-scale fading coefficients remain constant across frequencies and vary much more slowly over time compared to small-scale

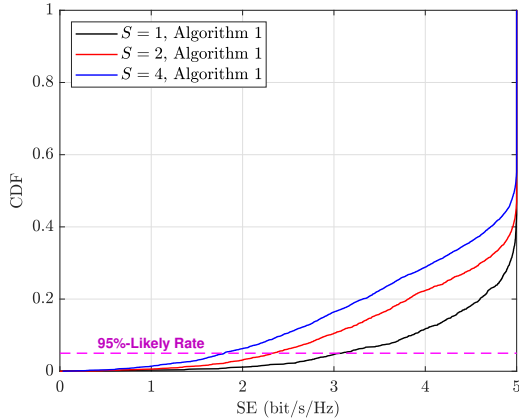


Fig. 9: CDF of the per-user SE achieved by Algorithm 1, where $L = 24$, $K = 15$, $M = 10$, $FH_{\max} = 10$ Gbps, and $M_{\text{mo}} = 32$.

fading. Consequently, with the statistical CSI-based designs, we need to perform **Algorithm 2** only 1 time, and use the results for all subcarriers and several frame duration [26]. To quantify these differences, the considered CF-mMIMO system with 100 MHz bandwidth and 136 kHz subcarrier spacing yields approximately 735 subcarriers. Assuming 10 transmit time intervals (TTIs) per 10 ms frame, **Algorithm 1** must be executed $735 \times 10 = 7350$ times per frame. In contrast, **Algorithm 2** is executed **once**, dramatically reducing overhead.

Moreover, both algorithms are implemented in a parallel cluster-wise manner, with each PC computing independently. Table II summarizes the observed average number of iterations and runtime per PC for Algorithm 1. The observed reduction in runtime as S increases confirms the scalability of the cluster-wise architecture. Finally, in what follows, we calculate the computational complexity per PC C_s and per iteration i . The computational complexity of Step 4 and Step 5 in **Algorithm 1** is $\mathcal{O}(|U_s|^2 |C_s| L)$, while the computational complexity for Step 6 is $\mathcal{O}(|U_s|)$. Step 7 of **Algorithm 1** involves solving a QCQP problem, which can be equivalently reformulated as a second-order cone programming (SOCP) problem. As discussed in [34], the complexity of solving an SOCP problem is $\mathcal{O}(N_{\text{so}} M_{\text{so}}^{3.5} + N_{\text{so}}^3 M_{\text{so}}^{2.5})$, where M_{so} is the number of second order cone constraints and N_{so} is the dimension of each. Problem (36a) contains $|C_s|$ transmit power constraints and $|C_s|$ fronthaul constraints with dimension $L|U_s|$. Therefore, the complexity of solving Problem (36a) is $\mathcal{O}(L|U_s| |C_s|^{3.5} + L^3 |U_s|^3 |C_s|^{2.5})$. Accordingly, the total computational complexity of **Algorithm 1** per iteration is $\mathcal{O}(L|U_s| |C_s|^{3.5} + L^3 |U_s|^3 |C_s|^{2.5})$. The computational complexity of **Algorithm 2** per iteration is the same as that of **Algorithm 1**.

VI. CONCLUSIONS

This paper has introduced a general cluster-wise processing network architecture for a fronthaul-limited CF-mMIMO system. We adopt the hybrid SLINR criterion and proposed two optimization approaches to maximize the cluster-

TABLE II: Average number of iterations and runtime per PC versus number of PCs, S .

S	Avg. Number of Iterations	Avg. Runtime (sec)
1	6.0	141
2	4.5	45
4	3.5	9.5

wise weighted sum pseudo-SE under per-AP transmit power and fronthaul constraints, namely 1) instantaneous CSI-based cluster-wise processing where precoding, user association, and power allocation are jointly optimized; 2) statistical CSI-based cluster-wise processing where user association and power allocation within a given cluster are jointly optimized. Two modified WMMSE-based algorithms were proposed to solve the challenging formulated non-convex mixed-integer problems. We investigated the trade-offs provided by the CF-mMIMO system with different number of PCs and highlighted the importance of the appropriate choice of cluster-wise processing relying on either instantaneous CSI-based or statistical CSI-based design for different system setups. Numerical results revealed that performance loss from increasing the number of processing clusters are primarily influenced by the number of APs, AP antennas, and fronthaul limitations. An interesting observation was that the proposed cluster-wise processing, relying only on local CSI, performs fairly close to network-wide alternative that relies on global CSI-knowledge for varying system setups. Investigating the scenarios involving pilot contamination and the loss of orthogonality between OFDM subcarriers in fronthaul-limited CF-mMIMO systems with cluster-wise processing is recommended for future studies.

REFERENCES

- [1] M. Mohammadi, Z. Mobini, H. Q. Ngo, and M. Matthaiou, "Next-generation multiple access with cell-free massive MIMO," *Proc. IEEE*, vol. 112, no. 9, pp. 1372–1420, Sept. 2024.
- [2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017.
- [3] M. Mohammadi, Z. Mobini, H. Q. Ngo, and M. Matthaiou, "Ten years of research advances in full-duplex massive MIMO," *IEEE Trans. Commun.*, Nov. 2024.
- [4] H. Q. Ngo, G. Interdonato, E. G. Larsson, G. Caire, and J. G. Andrews, "Ultradense cell-free massive MIMO for 6G: Technical overview and open questions," *Proc. IEEE*, vol. 112, no. 7, pp. 805–831, July 2024.
- [5] S. Buzzi and C. Dandrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.
- [6] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. and Networking*, vol. 2, no. 1, pp. 25–39, Mar. 2018.
- [7] 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 Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1482–1497, Mar. 2022.
- [8] Z. Mobini, A. H. Gokceoglu, L. Wang, G. Peters, and H. Q. Ngo, "Fronthaul-aware user-centric generalized cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2025.
- [9] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with WMMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.
- [10] E. Nayeibi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, July 2017.

- [11] P. Liu, K. Luo, D. Chen, and T. Jiang, "Spectral efficiency analysis of cell-free massive MIMO systems with zero-forcing detector," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 795–807, Feb. 2020.
- [12] 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. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.
- [13] M. Alonzo, S. Buzzi, A. Zappone, and C. D'Elia, "Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave," *IEEE Trans. Green Commun. and Networking*, vol. 3, no. 3, pp. 651–663, Sept. 2019.
- [14] M. Mohammadi, T. T. Vu, H. Q. Ngo, and M. Matthaiou, "Network-assisted full-duplex cell-free massive MIMO: Spectral and energy efficiencies," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 9, pp. 2833–2851, Sept. 2023.
- [15] C. Hao, T. T. Vu, H. Q. Ngo, M. N. Dao, X. Dang, C. Wang, and M. Matthaiou, "Joint user association and power control for cell-free massive MIMO," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 15 823–15 841, May 2024.
- [16] J. Fu, Z. Mobini, H. Q. Ngo, P. Zhu, and M. Matthaiou, "WMMSE-based processing in cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 14, no. 2, pp. 330–334, Feb. 2025.
- [17] I. Atzeni, B. Gouda, and A. Tölle, "Distributed precoding design via over-the-air signaling for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1201–1216, Feb. 2021.
- [18] L. Miretti, E. Björnson, and D. Gesbert, "Team MMSE precoding with applications to cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6242–6255, Aug. 2022.
- [19] G. Interdonato and S. Buzzi, "Conjugate beamforming with fractional-exponent normalization and scalable power control in cell-free massive MIMO," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2021, pp. 396–400.
- [20] S. Chakraborty, E. Björnson, and L. Sanguinetti, "Centralized and distributed power allocation for max-min fairness in cell-free massive MIMO," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 576–580.
- [21] G. Interdonato, M. Karlsson, E. Björnson, and E. G. Larsson, "Local partial zero-forcing precoding for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4758–4774, July 2020.
- [22] L. Du, L. Li, H. Q. Ngo, T. C. Mai, and M. Matthaiou, "Cell-free massive MIMO: Joint maximum-ratio and zero-forcing precoder with power control," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3741–3756, June 2021.
- [23] 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 Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3099–3115, May 2022.
- [24] L. Zhang, S. Yang, and Z. Han, "Pilot assignment for cell-free massive MIMO: A spectral clustering approach," *IEEE Commun. Lett.*, vol. 13, no. 1, pp. 243–247, Jan 2024.
- [25] M. Sadek and S. Aissa, "Leakage based precoding for multi-user MIMO-OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2428–2433, Aug. 2011.
- [26] Z. Mobini and H. Q. Ngo, "Massive multiple-input, multiple-output: Instantaneous versus statistical channel state information-based power allocation [lecture notes]," *IEEE Signal Process. Mag.*, vol. 42, no. 2, pp. 27–36, Mar. 2025.
- [27] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, 2014.
- [28] T. L. Marzetta and H. Yang, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
- [29] Q. Peng, H. Ren, C. Pan, N. Liu, and M. El Kashlan, "Resource allocation for uplink cell-free massive MIMO enabled URLLC in a smart factory," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 553–568, Jan. 2023.
- [30] M. Urlea and S. Loyka, "Simple closed-form approximations for achievable information rates of coded modulation systems," *J. Lightw. Technol.*, vol. 39, no. 5, pp. 1306–1311, Mar. 2021.
- [31] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Oct. 2014.
- [32] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.
- [33] S. Chakraborty, Ö. T. Demir, E. Björnson, and P. Giselsson, "Efficient downlink power allocation algorithms for cell-free massive MIMO systems," *IEEE Open J. Commun. Society*, vol. 2, pp. 168–186, Dec. 2021.
- [34] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Intelligent reflecting surface aided multigroup multicast MISO communication systems," *IEEE Trans. Signal Process.*, vol. 68, pp. 3236–3251, Apr. 2020.