Received 7 February 2024; accepted 29 February 2024. Date of publication 5 March 2024; date of current version 22 March 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3373170

Cell-Free Massive MIMO With Multi-Antenna Users and Phase Misalignments: A Novel Partially Coherent Transmission Framework

UNNIKRISHNAN KUNNATH GANESAN [1] (Graduate Student Member, IEEE), TUNG THANH VU [2] (Member, IEEE), AND ERIK G. LARSSON [1] (Fellow, IEEE)

[1] Department of Electrical Engineering, Linköping University, 581 83 Linköping, Sweden

[2] School of Engineering, Macquarie University, Sydney, NSW 2113, Australia

CORRESPONDING AUTHOR: T. T. VU (e-mail: thanhtung.vu@mq.edu.au)

This work was supported in part by ELLIIT; in part by KAW Foundation; and in part by the REINDEER Project of the European Union's Horizon 2020 Research and Innovation Programme under Grant 101013425.

ABSTRACT CELL-FREE massive multiple-input multiple-output (MIMO) is a promising technology for next-generation communication systems. This work proposes a novel partially coherent (PC) transmission framework to cope with the challenge of phase misalignment among the access points (APs), which is important for unlocking the full potential of cell-free massive MIMO technology. With the PC operation, the APs are only required to be phase-aligned within clusters. Each cluster transmits the same data stream towards each user equipment (UE), while different clusters send different data streams. We first propose a novel algorithm to group APs into clusters such that the distance between two APs is always smaller than a reference distance ensuring the phase alignment of these APs. Then, we propose new algorithms that optimize the combining at UEs and precoding at APs to maximize the downlink sum data rates. We also propose a novel algorithm for data stream allocation to further improve the sum data rate of the PC operation. Numerical results show that the PC operation using the proposed framework with a sufficiently small reference distance can offer a sum rate close to the sum rate of the ideal fully coherent (FC) operation that requires network-wide phase alignment. This demonstrates the potential of PC operation in practical deployments of cell-free massive MIMO networks.

INDEX TERMS Cell-free massive MIMO, downlink, coherent transmission, non-coherent transmission, partially coherent transmission, precoding, combining, data stream allocation.

I. INTRODUCTION

CELL-FREE massive multiple-input multiple-output (MIMO) is an innovative technology for next-generation communication systems, where user equipments (UEs) are served by a large number of access points (APs) distributed over an extensive geographic area [1], [2], [3], [4], [5], [6], [7], [8]. It inherits the multi-antenna benefits of cellular massive MIMO [9], [10] and provides extraordinary macro diversity gains [3]. In cell-free massive MIMO networks, inter-cell interference is eliminated due to the absence of rigid cell boundaries, providing uniformly great service to all UEs.

Precise phase alignment of the APs is crucial to unlock the full potential of cell-free massive MIMO [11], [12], [13], [14]. With phase alignment, all the APs can coordinate together and transmit the same data stream toward each UE, leading to a high beamforming/array gain, and hence, a high data rate. In this case, the APs are working in a fully coherent (FC) operation of cell-free massive MIMO systems. Note that the FC operation represents an upper bound on the performance of cell-free networks.

However, ensuring a full phase synchronization of all the APs in the preferable FC operation of cell-free massive MIMO networks is practically very challenging [11], [15].

The APs are driven by independent local oscillators (LOs) that may drift independently with time, which results in phase mismatches between them. This requires frequent re-alignment of the phase between different APs [14], which is infeasible for large-scale networks. This raises the need for finding solutions to operate cell-free massive MIMO networks in the presence of phase alignment errors, to unlock all the benefits of these networks.

There are two operations to cope with the problem of phase misalignment in cell-free massive MIMO networks. The first is a fully non-coherent (FNC) operation, where the APs are not phase-aligned and send independent data streams toward UEs [16], [17], [18], [19], [20] (see Section I-A for more details). The second is a partially coherent (PC) operation, allowing both coherent and non-coherent transmissions of APs in the same cell-free massive MIMO network [21]. With PC operation, there are multiple clusters, in each of which the APs are aligned in phase. The APs in the same cluster send the same data stream toward each UE, while those in different clusters send different data streams. The PC operation is more practical compared to the FC operation as phase alignment is only performed among APs within a relatively small area. Such PC operation can offer higher data rates compared to the FNC operation. The PC operation does not require network-wide synchronization, which makes it practically feasible and attractive for realizing future cell-free massive MIMO communication systems.

In this paper, we propose an innovative framework for the PC operation in cell-free massive MIMO networks. First, an algorithm is proposed to group APs into multiple phase-aligned clusters, which is fundamentally different from the current AP clustering algorithms where the phase misalignment is not taken into account. Then, novel signal processing and data stream allocation algorithms are developed to significantly improve the performance, compared to the FNC operation. Our AP clustering method is based on the insights from recent advances in phase alignment techniques proposed in [11], [12], [13]. Even though network-wide phase alignment might be infeasible, these works show that it is possible to synchronize the phases of APs that are close to each other in the network. This leads to two research questions: (Q1) What is the maximum "reference" distance between the APs such that they are considered to be in a phase-aligned cluster? (Q2) Given this reference distance, how to cluster the APs, and how to improve the performance of the system? In this work, we focus on answering the question (Q2). Answering question (Q1) requires detailed studies from field measurements, which is out of the scope of this paper.

A. RELATED LITERATURE AND DISCUSSIONS

A large literature on the FC operation of cell-free massive MIMO systems is available, assuming all the APs are operating coherently without considering the problem of phase misalignment. Comprehensive surveys of the field are provided in [4], [15]. The advantages of cell-free massive MIMO in terms of energy and cost efficiency are considered in [8]. The survey paper [15] provides a comprehensive study on the centralized and distributed operations of cell-free massive MIMO. Uplink performance analysis is studied in [22] and under limited fronthaul and hardware impairments are studied in [23]. The paper [24] studies the user-centric operation of cell-free massive MIMO to reduce the fronthaul overhead. To maximize the uplink sum rate of UEs, a max-min approach under power constraints is studied in [25]. A centralized precoding and power control strategy is proposed in [26] when users are served by overlapping clusters of APs. The paper [27] studies a scalable implementation of distributed massive MIMO systems exploiting dynamic cooperation of clusters and considers initial access, pilot assignment, cooperation cluster formation, precoding, and combining strategies. These existing works have shown that cell-free massive MIMO with perfect phase alignment offers significant gains in terms of spectral and energy efficiencies, compared to cellular massive MIMO networks.

The FNC operation of cell-free networks, which do not fully exploit the multi-antenna benefits of cell-free massive MIMO, have been studied in [16], [17], [18], [19], [20]. The asynchronous arrival of signals at the UEs increases the interference and degrades the system performance. The work [16] proposes a rate-splitting strategy by splitting the messages into common and private parts to improve the data rate with non-coherent operation between the APs. The paper [17] investigates the non-coherent joint transmission and poses the beamforming vector design as an optimization problem, with the objective of maximizing the sum rate of the system. This optimization problem is non-convex and NP-hard and different schemes have been proposed in the literature to achieve near-optimal solutions [17], [18], [19]. The work [17] develops an algorithm based on the alternating direction method of multipliers with an inner approximation technique to achieve a near-optimal solution. The paper [18] proposes an algorithm based on multi-agent reinforcement learning to maximize the sum rate with low computational complexity. The paper [19] uses tools from fractional programming, block coordinate descent, and compressed sensing to construct a smooth non-decreasing algorithm to optimize the beamforming vectors. The spectral efficiency of cell-free massive MIMO systems under Rician fading channels and phase shifts of the line-of-sight (LoS) path is studied in [20]. These works have shown the significant importance of optimizing beamforming in improving the performance of cell-free massive MIMO networks with FNC operation.

The research on the PC operation of cell-free massive MIMO networks is still in its infancy, and we are only aware of one related paper [21]. The work [21] proposed a mixed coherent and non-coherent transmission scheme for cell-free systems with single-antenna UEs. Here, there are multiple central processing units (CPUs), and the APs connect to a CPU and operate coherently within a cluster. Different CPUs or different clusters work in a non-coherent fashion.

Reference [21] showed that a mixed approach performs in between the coherent and non-coherent approaches. However, [21] assumed that the AP clusters are already known, and did not take into account the practical problem of phase misalignment. Moreover, [21] considered fixed beamforming and did not consider the aspect of optimizing the beamforming to achieve maximum beamforming gains and data rates for the PC operation.

Clustering in cell-free massive MIMO has been studied in [4], [7], [19], [27], [28], [29], [30]. The purpose of the clustering methods in these papers is to maximize the spectral efficiency or energy efficiency with limited-capacity fronthaul. None of them takes into account the phase misalignment problem. Moreover, the works on user clustering and cooperative transmissions in cell-free networks assume network-wide synchronization of the APs. In practice, network-wide full synchronization is not feasible. Using state-of-the-art methods for synchronization described in [11], [12], [13], it is possible to synchronize APs within a small area. How such areas and clusters of APs can be determined has not been studied so far. In this work, we propose a PC framework to overcome the challenge of phase misalignment. It consists of an AP clustering algorithm to determine the APs within a reference distance that can be fully synchronized.

The studies on the systems with multi-antenna UEs have been studied primarily for cellular massive MIMO [31], [32], [33], and recently for cell-free massive MIMO in [34], [35], [36]. Optimizing combining, precoding, and data stream allocation is important for these systems. It is normally preferred to achieve higher multiplexing gains by using multiple data streams at the same time. However, multiplexing more data streams towards one UE can bring more inter-stream interference, which degrades the data rates and requires optimized combining/precoding techniques to mitigate. A proper method to allocate data streams can strike the best trade-off between high multiplexing gains and low inter-stream interference. The work [31] proposes a maximum-ratio (MR) precoding scheme to enhance the channel hardening effect in systems with limited hardening capabilities in cellular massive MIMO. A joint precoding and combiner design in the presence of reciprocity calibration errors is studied in [32]. Data stream allocation for multi-antenna UEs for a cellular massive MIMO system is studied in [33]. Downlink spectral efficiency (SE) of cell-free massive MIMO is derived in [34] considering imperfect channel state information (CSI), non-orthogonal pilots and power control. An iterative weighted minimum-mean-square-error (WMMSE) precoding scheme for uplink cell-free massive MIMO is proposed in [35] which shows a higher SE with a large number of UE antennas. An eigenbasis-based uplink precoding scheme is proposed in [36] to improve the SE. These existing works focus on the FC operation and do not consider the PC operation and data stream allocation for cell-free massive MIMO networks.

B. RESEARCH GAP AND MAIN CONTRIBUTIONS

PC is an innovative and practical way of unlocking the full potential of cell-free massive MIMO networks. To the best of our knowledge, there is no systematic study of this operation available in the literature. The designs of AP clustering based on the reference distance, precoding/combining optimization, and data stream allocation significantly impact the data rate performance of the PC operation in cell-free massive MIMO systems. By proposing an innovative framework that involves these aspects, the paper makes the following contributions:

- We propose an AP clustering algorithm for cell-free massive MIMO systems, providing a set of non-overlapping phase-aligned clusters for PC operation. The proposed algorithm groups the APs into phase-aligned clusters based on both the reference distance of phase alignment and the channel conditions, which is different from that in [21].
- We develop an algorithm for optimizing the combining and precoding matrices to maximize the downlink sum rate. The formulated optimization problem involves per-AP power constraints and varying-size variable matrices. It is also a non-convex and NP-hard problem, which is very challenging to solve for global optimality. We propose an algorithm that tailors the WMMSE framework to obtain a sub-optimal solution to the formulated optimization problem. Importantly, the developed algorithm only requires calculating closed-form expressions for the combining and precoding matrices in each iteration, which is computationally efficient for large-scale networks. Note that the WMMSE algorithms proposed in existing works [37], [38] cannot be directly applied to solve the formulated problem. We show that our proposed scheme performs significantly better than the fixed beamforming techniques in [21].
- We propose a greedy data stream allocation algorithm for multi-antenna UEs, which further maximizes the sum rate of the network. The data stream of each UE is iteratively allocated to have both strong channel gains and low interference strengths.
- We analyze numerically the performance of the PC system with the proposed algorithms. We also make detailed comparisons in terms of sum data rates between the PC and the traditional operations FC and FNC. Numerical results show that the PC operation with phase-aligned AP clusters clustered based on an appropriate reference distance offers similar performance as that of the FC operation. This means that it is possible to approach the ideal performance of the FC operation by using PC operation, without the strict requirement of network-wide phase alignment. This highlights the importance of the PC operation in the practical deployment of cell-free massive MIMO networks. The results also show that data stream allocation plays an important role in improving the performance of the PC.

The rest of the paper is organized as follows. Section II provides a motivational example to explain why studying the operation of non-phase-aligned APs is important. Section III introduces the system model and problem formulation. The proposed AP clustering algorithm is presented in Section IV. Section V provides algorithms to optimize precoding and combining matrices to maximize the sum rate of the network for a PC system, while Section VI provides the greedy data-stream allocation, and Section VII discusses the complexities of the the proposed algorithms. Numerical results are provided in Section VIII and concluding remarks are given in Section IX.

Notations: Bold lowercase letters are used to denote vectors and bold uppercase letters are used to denote matrices. $\mathbb{C}$ denotes the set of complex numbers. For a matrix $\mathbf{A}$, $\mathbf{A}^*$, $\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{H}}$ denote conjugate, transpose and conjugate transpose of the matrix $\mathbf{A}$ respectively. $\mathcal{CN}(0,\sigma^2)$ denotes a circularly symmetric complex Gaussian random variable with zero mean and variance equal to σ^2. $\mathrm{Tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$. The identity matrix of size $N \times N$ is denoted by $\mathbf{I}_N$.

II. MOTIVATING EXAMPLE

In this section, we consider a small example to understand the motivation of the PC operation and the importance of beamforming in operating APs with phase misalignment. Consider two APs, each with M antennas, and a single UE with N antennas. Let $\mathbf{G}_1$ and $\mathbf{G}_2$ be the $N \times M$ channel between the APs 1 and 2 to the UE, respectively. We assume that the channels are known to every entity. Let ρ be the operating signal-to-noise ratio (SNR) (normalized transmit power) at each AP.

A. PHASE ALIGNMENT SCENARIO

When both APs are phase-aligned, they can form a virtual large massive MIMO array with $2M$ elements and operate coherently together. This is analogous to a point-to-point MIMO system and the achievable rate is $\log_2|\mathbf{I}+\rho\,\mathbf{H}\mathbf{K}\mathbf{H}^{\mathrm{H}}|$ [10, eq. (C.28)], where $\mathbf{H} = [\,\mathbf{G}_1\,\mathbf{G}_2\,]$ is the combined MIMO channel and

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \tag{1}$$

is the covariance matrix of the signal transmitted from the virtual large massive MIMO array. Each AP has a maximum transmit power constraint given by

$$\mathrm{Tr}(\mathbf{K}_{11}) \le 1, \quad \mathrm{Tr}(\mathbf{K}_{22}) \le 1. \tag{2}$$

The achievable rate under per-AP power constraints is determined by the following maximization problem:

$$R_{\text{aligned}} = \max_{\mathbf{K}} \ \log_2\left|\mathbf{I}+\rho\,\mathbf{H}\mathbf{K}\mathbf{H}^{\mathrm{H}}\right| \tag{3a}$$

$$\text{subject to } \mathrm{Tr}(\mathbf{K}_{11}) \le 1, \quad \mathrm{Tr}(\mathbf{K}_{22}) \le 1. \tag{3b}$$

This problem can be solved to global optimality using standard convex programming software packages. R_{aligned} will be used as the ideal data rate for comparisons with the data rates in practical scenarios of phase misalignment.

B. PHASE MISALIGNMENT SCENARIO

When both APs are not phase-aligned, the relative phase between the two APs is unknown. This scenario is likely to happen in practice due to the drift of the clocks between the APs unless a strict phase-synchronization protocol is implemented. To cope with this phase misalignment problem, several transmission strategies can be used as follows.

1) TRANSMIT ONLY FROM THE BEST AP

The first strategy is to simply let UE pick the AP that gives the maximum rate and receives its data stream from that AP only. The UE achievable rate in this case is

$$R_{\text{best AP}} = \max\left\{R_{\text{AP 1 only}}, R_{\text{AP 2 only}}\right\}, \tag{4}$$

where $R_{\text{AP i only}}$ is the data rate when only AP $i, \forall i \in \{1,2\}$, is in operation is

$$R_{\text{AP i only}} = \max_{\mathbf{K}_{ii}} \ \log_2\left|\mathbf{I}+\rho\,\mathbf{G}_i\,\mathbf{K}_{ii}\,\mathbf{G}_i^{\mathrm{H}}\right| \tag{5a}$$

$$\text{subject to } \mathrm{Tr}(\mathbf{K}_{ii}) \le 1. \tag{5b}$$

2) TRANSMIT INDEPENDENTLY CODED DATA STREAMS WITH OPTIMAL BEAMFORMING AND SUCCESSIVE INTERFERENCE CANCELLATION

Another strategy is to transmit independently coded data from the two APs. The transmitted signals from APs 1 and 2 have the covariance matrices $\mathbf{K}_{11}$ and $\mathbf{K}_{22}$, respectively. The UE applies successive interference cancellation (SIC) and the achievable rate is given by [39, Sec. 8.3.3]

$$R_{\text{SIC}} = \max_{\mathbf{K}_{11},\mathbf{K}_{22}} \ \log_2\left|\mathbf{I}+\rho\,\mathbf{G}_1\,\mathbf{K}_{11}\,\mathbf{G}_1^{\mathrm{H}}+\rho\,\mathbf{G}_2\,\mathbf{K}_{22}\,\mathbf{G}_2^{\mathrm{H}}\right| \tag{6a}$$

$$\text{subject to } \mathrm{Tr}(\mathbf{K}_{11}) \le 1, \ \ \mathrm{Tr}(\mathbf{K}_{22}) \le 1. \tag{6b}$$

This approach can be used when we have multi-point to single-point communication. However, in scenarios with multiple users, this approach requires all users to know the channels of all other users, and hence, becomes infeasible in practice. Thus, it is necessary to look for sub-optimal strategies.

3) TRANSMIT INDEPENDENTLY CODED DATA STREAMS WITH SUB-OPTIMAL BEAMFORMING

Another strategy is to find sub-optimal solutions for beamforming from each AP and combining at the UE. For example, consider a simple case where APs beamform a rank-one signal each in the direction $\mathbf{w}_1$ and $\mathbf{w}_2$, respectively, satisfying the power constraints at each AP. If a zero-forcing (ZF) combiner is used at the UE, the rate obtained is given by

$$R_{\text{non-aligned}} = \log_2\left(1+\frac{\rho}{\sigma_1^2}\right)+\log_2\left(1+\frac{\rho}{\sigma_2^2}\right), \tag{7}$$

where $\sigma_i^2 = [(\mathbf{H}^{\mathrm{H}}\mathbf{H})^{-1}]_{ii}, \forall i \in \{1,2\}$, and $\mathbf{H} = [\,\mathbf{G}_1\,\mathbf{w}_1 \ \ \mathbf{G}_2\,\mathbf{w}_2\,] \in \mathbb{C}^{N\times 2}$ is the overall effective channel. Then, the remaining problem is to select the beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ to maximize the rate under the power constraints. A natural approach is to select the beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$

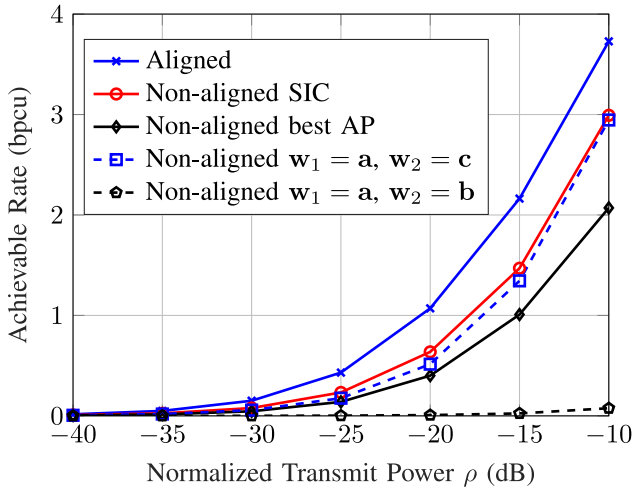


FIGURE 1. Achievable rates with different beamforming designs and transmission schemes for the example considered in (8) and (9).

to be the dominant right singular vectors of $\mathbf{G}_1$ and $\mathbf{G}_2$, respectively, to maximize the per-stream SNR from each AP. The selection of beamformers must also consider the angle between $\mathbf{G}_1\mathbf{w}_1$ and $\mathbf{G}_2\mathbf{w}_2$ as well.

To gain more insights on the above concept, we consider an example with $N = 2$ such that the UE receives signals from two directions and

$$\mathbf{G}_1 = \mathbf{g}\mathbf{a}^{\mathrm{H}} \tag{8}$$

$$\mathbf{G}_2 = \mathbf{g}\mathbf{b}^{\mathrm{H}} + \alpha\mathbf{f}\mathbf{c}^{\mathrm{H}}. \tag{9}$$

Here, $\|\mathbf{g}\| = 1$, $\|\mathbf{f}\| = 1$, $\|\mathbf{a}\| = 1$, $\|\mathbf{b}\| = 1$, $\|\mathbf{c}\| = 1$ and $|\alpha| < 1$ is some constant, and such that $\mathbf{g}^{\mathrm{H}}\mathbf{f} = 0$ and $\mathbf{c}^{\mathrm{H}}\mathbf{b} = 0$. The above channels correspond to a situation where the signal to the UE arrives from AP 1 from a single direction (direction $\mathbf{g}$) and signal from AP 2 arrives from two directions: (i) from direction $\mathbf{g}$ (same direction as AP-1); (ii) a weaker signal from direction $\mathbf{f}$. Selecting dominant right singular vectors as the beamformers, i.e., $\mathbf{w}_1 = \mathbf{a}$ and $\mathbf{w}_2 = \mathbf{b}$, gives $\mathbf{G}_1\mathbf{w}_1 = \mathbf{g}$ and $\mathbf{G}_2\mathbf{w}_2 = \mathbf{g}$. This makes the effective channel $\mathbf{H}$ non-invertible, creating huge interference during decoding. Instead, choosing $\mathbf{w}_2 = \mathbf{c}$, the second-strongest singular vector of $\mathbf{G}_2$, we have $\mathbf{G}_1\mathbf{w}_1 = \mathbf{g}$ and $\mathbf{G}_2\mathbf{w}_2 = \alpha\mathbf{f}$. Thus the second option is better, even though the signal from AP 2 in the direction of $\mathbf{c}$ is weaker than that in the direction of $\mathbf{b}$.

Fig. 1 shows the achievable rates in all transmission strategies for the above example. For this plot, we consider $M = 16$, $N = 2$, and $\alpha = 0.7$. The key observations are:

1) APs with phase-alignment give the maximum rate. Hence, the phase of APs should be aligned whenever possible to improve data rates.
2) The gap between optimal and sub-optimal beamforming significantly varies based on the designs of beamforming and transmission schemes.

The above observations give motivations for proposing and optimizing the PC operation in a general network that contains many APs and UEs, which will be considered in the rest of the paper. Although not all APs can be phase-aligned, it is possible to synchronize the APs in clusters, and hence, there are phase-aligned clusters in the network. A cluster can operate coherently, while different clusters operate non-coherently, thus yielding a PC scenario. Maximizing the sum rate of the network with the PC operation will strongly require optimizing the beamforming. Also, multi-antenna UEs require optimizing the combining and data stream allocation to cope with inter-stream interference as discussed in Section I-A.

III. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a cell-free massive MIMO system with L APs and K UEs. We assume that each AP is equipped with M antennas and each UE is equipped with N antennas. All APs are connected to a CPU through high-capacity fronthaul links. Let $\mathbf{G}_{kl} \in \mathbb{C}^{N\times M}$ be the channel matrix between the UE $k \in \mathcal{K} \triangleq \{1, \ldots, K\}$ and the AP $l \in \mathcal{L} \triangleq \{1, \ldots, L\}$.

We consider a network with the PC operation, where the APs are grouped into phase-aligned clusters. Each cluster forms a virtual large antenna array to transmit data symbols coherently. Different clusters are not phased-aligned, and hence, their transmissions are performed in a non-coherent way. Let there be $L_c \leq L$ clusters. There are no common AP among any pair of clusters. We will propose an AP clustering algorithm to cluster APs in Section IV.

The considered PC system model is a general system model for the networks with the traditional operations FC and FNC. In particular, $L_c = L$ means that there are no clusters and all APs operate independently, which is the FNC operation. Similarly, $L_c = 1$ means that all the APs are in a single phase-aligned cluster, which gives the FC operation.

Let $\mathcal{C}_c$ be the set of APs in cluster $c \in \{1, 2, \ldots, L_c\}$. Let $\mathbf{q}_{kc} \in \mathbb{C}^{d_{kc}\times 1}$ be the data symbol vector transmitted by cluster c to UE k, where $d_{kc} \leq \min(M|\mathcal{C}_c|, N)$ is the number of data streams, $\mathbb{E}\{\mathbf{q}_{kc}\} = \mathbf{0}$, $\mathbb{E}\{\mathbf{q}_{kc}\mathbf{q}_{kc}^{\mathrm{H}}\} = \mathbf{I}_{d_{kc}}$, and $\mathbb{E}\{\mathbf{q}_{kc}\mathbf{q}_{k'c'}^{\mathrm{H}}\} = \mathbf{0}, \forall k' \neq k, c' \neq c$. Note that APs within each cluster send the same data stream to achieve high beamforming gain. The APs from different clusters transmit independent and different data streams. Cluster c transmits a data stream toward UE k with the collective precoding matrix

$$\overline{\mathbf{W}}_{kc} = \begin{bmatrix} \mathbf{W}_{kl_{c,1}} \\ \mathbf{W}_{kl_{c,2}} \\ \vdots \\ \mathbf{W}_{kl_{c,|\mathcal{C}_c|}} \end{bmatrix} \in \mathbb{C}^{M|\mathcal{C}_c|\times d_{kc}}, \tag{10}$$

where $\mathbf{W}_{kl_{c,j}} \in \mathbb{C}^{M\times d_{kc}}$ is the precoding matrix at AP $l_{c,j}$ in cluster $\mathcal{C}_c$, and $j \in \{1, \ldots, |\mathcal{C}_c|\}$. The transmitted signal from cluster c is

$$\mathbf{x}_c = \sqrt{\rho}\sum_{k=1}^{K}\overline{\mathbf{W}}_{kc}\mathbf{q}_{kc}, \tag{11}$$

where ρ is the normalized maximum transmit power at each AP. The per-AP maximum power constraint is expressed as

$$\sum_{k=1}^{K} \mathrm{Tr}\left(\mathbf{W}_{kl}\,\mathbf{W}_{kl}^{\mathrm{H}}\right) \leq 1, \quad \forall l. \tag{12}$$

The received signal $\mathbf{y}_k \in \mathbb{C}^{N\times 1}$ at UE k is given by

$$\begin{aligned}\mathbf{y}_k &= \sum_{c=1}^{L_c} \overline{\mathbf{G}}_{kc}\mathbf{x}_c + \mathbf{n}_k \\ &= \sqrt{\rho}\sum_{c=1}^{L_c} \overline{\mathbf{G}}_{kc} \sum_{k'=1}^{K} \overline{\mathbf{W}}_{k'c}\mathbf{q}_{k'c} + \mathbf{n}_k,\end{aligned} \tag{13}$$

where $\overline{\mathbf{G}}_{kc} = [\mathbf{G}_{kl_{c,1}}, \ldots, \mathbf{G}_{kl_{c,|\mathcal{C}_c|}}] \in \mathbb{C}^{N\times M|\mathcal{C}_c|}$ is the collective channel matrix of the APs in cluster $\mathcal{C}_c$ to UE k and $\mathbf{n}_k$ is an additive white Gaussian noise (AWGN) vector with independent and identically distributed (i.i.d.) $\mathcal{CN}(0, 1)$ entries. To estimate data symbol $\mathbf{q}_{kc}$, UE k applies the combining matrix $\overline{\mathbf{V}}_{kc} \in \mathbb{C}^{N\times d_{kc}}$ as

$$\begin{aligned}\hat{\mathbf{q}}_{kc} &= \overline{\mathbf{V}}_{kc}^{\mathrm{H}}\mathbf{y}_k \\ &= \sqrt{\rho}\overline{\mathbf{V}}_{kc}^{\mathrm{H}}\overline{\mathbf{G}}_{kc}\overline{\mathbf{W}}_{kc}\mathbf{q}_{kc} + \sqrt{\rho}\overline{\mathbf{V}}_{kc}^{\mathrm{H}} \sum_{c'=1,c'\neq c}^{L_c} \overline{\mathbf{G}}_{kc'}\overline{\mathbf{W}}_{kc'}\mathbf{q}_{kc'} \\ &\quad + \sqrt{\rho}\overline{\mathbf{V}}_{kc}^{\mathrm{H}} \sum_{c'=1}^{L_c} \overline{\mathbf{G}}_{kc'} \sum_{k'=1,k'\neq k}^{K} \overline{\mathbf{W}}_{k'c'}\mathbf{q}_{k'c'} + \overline{\mathbf{V}}_{kc}^{\mathrm{H}}\mathbf{n}_k.\end{aligned} \tag{14}$$

The desired signal for the data stream from cluster c to UE k is the first term of (14), while the inter-cluster and inter-UE interference are the second and the third terms of (14), respectively.

Treating the interference terms as Gaussian noise, the achievable rate R_{kc} for the data stream from cluster c to UE k as [40], [41]

$$R_{kc} = \log_2\left|\mathbf{I}_{d_{kc}} + \rho\overline{\mathbf{H}}_{kc}^{\mathrm{H}}\overline{\mathbf{Q}}_{kc}^{-1}\overline{\mathbf{H}}_{kc}\right|, \tag{15}$$

where

$$\overline{\mathbf{H}}_{kc} = \overline{\mathbf{V}}_{kc}^{\mathrm{H}}\overline{\mathbf{G}}_{kc}\overline{\mathbf{W}}_{kc}, \tag{16}$$

$$\begin{aligned}\overline{\mathbf{Q}}_{kc} = \overline{\mathbf{V}}_{kc}^{\mathrm{H}}\Bigg(&\rho \sum_{c'=1,c'\neq c}^{L_c} \overline{\mathbf{G}}_{kc'}\overline{\mathbf{W}}_{kc'}\overline{\mathbf{W}}_{kc'}^{\mathrm{H}}\overline{\mathbf{G}}_{kc'}^{\mathrm{H}} \\ &+ \rho\sum_{c'=1}^{L_c} \overline{\mathbf{G}}_{kc'}\left(\sum_{k'=1,k'\neq k}^{K} \overline{\mathbf{W}}_{k'c'}\overline{\mathbf{W}}_{k'c'}^{\mathrm{H}}\right)\overline{\mathbf{G}}_{kc'}^{\mathrm{H}} + \mathbf{I}_N\Bigg)\overline{\mathbf{V}}_{kc}.\end{aligned} \tag{17}$$

Since the data streams for each UE are independently coded among the clusters, the rate at UE k is the sum of the achievable rates of the data streams from all clusters to this UE, i.e.,

$$R_k = \sum_{c=1}^{L_c} R_{kc}. \tag{18}$$

We aim to optimize precoding and combining matrices as well as the number of data streams per UE to maximize the total sum rate under the per-AP maximum transmit power constraints. This problem can be formulated as

$$\begin{aligned}&\underset{\left\{\overline{\mathbf{W}}_{kc}\right\},\left\{\overline{\mathbf{V}}_{kc}\right\},\{d_{kc}\}}{\text{maximize}} \sum_{k=1}^{K}\sum_{c=1}^{L_c} R_{kc}\left(\left\{\overline{\mathbf{W}}_{kc}\right\}, \overline{\mathbf{V}}_{kc}, \{d_{kc}\}\right) \\ &\text{subject to (12).}\end{aligned} \tag{19}$$

Problem (19), even for the fixed values of $\{d_{kc}\}$, has a similar mathematical structure as that of the problem [42], which was shown therein to be NP-hard. Therefore, it is challenging to obtain a globally optimal solution to Problem (19). In this paper, we propose sub-optimal solutions to find beamforming vectors and data stream allocation. These solutions are discussed in Sections V and VI, respectively.

In this work, the challenge of phase misalignment is overcome by using the proposed PC framework. The PC framework involves three parts: (i) AP clustering to group APs into phase-aligned clusters based on reference distance; (ii) optimizing the precoding and combining to maximize the sum rate for given AP clusters; (iii) allocating data streams to further improve the sum rate performance. Specifically, the AP clustering in part (i) makes sure the APs are phase-aligned in their clusters to achieve higher beamforming gains. Compared to the ideal FC operation (where all the APs are phase-aligned), the PC framework only requires APs to be phase-aligned in clusters, which is more feasible in practical deployments. In return, the PC operation brings inter-cluster interference to the system because the APs clusters are not mutually phase-aligned. This is the price to pay when overcoming the challenge of phase misalignment using the PC framework. The optimization of precoding/combining and the data stream allocation in parts (ii) and (iii) efficiently manage inter-cluster interference to improve the system performance. Problem (19) belongs to part (ii), which aims to manage the inter-cluster interference for maximizing the sum rates for given AP clusters obtained by part (i). The sum rates are functions of inter-cluster interference as shown in (15). Thus, the problem of phase misalignment is taken into account by managing the inter-cluster interference.

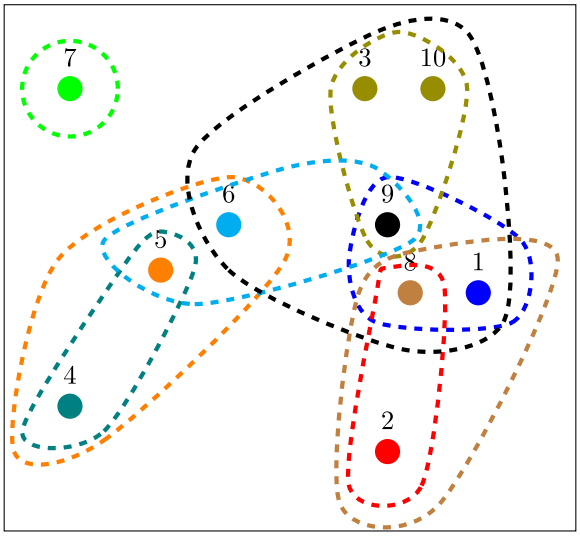
IV. AP CLUSTERING ALGORITHM

This section discusses how phase-aligned AP clusters can be formed for coherent transmission and high data rate. Synchronizing the phases of all the APs in the network is practically very challenging as discussed in Section I. Despite that, a certain set of APs that are near to one another can be phase-aligned as shown in [11], [12]. Two APs are guaranteed with an acceptable phase alignment if their distance is within a reference distance D, which is assumed to be known.

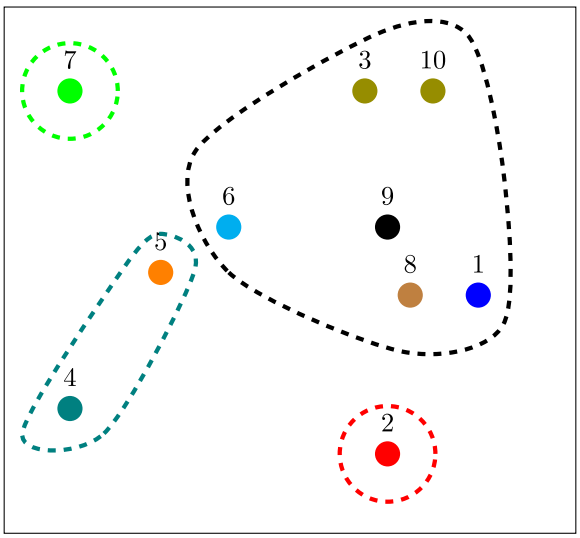
Let $D_{ll'}$ be the distance between APs l and l'. We say that AP l' is a neighbor of AP l if $D_{ll'} \leq D$. Let

$$\mathcal{Z}_l = \left\{l' | D_{ll'} \leq D, l' \in \{1, 2, \ldots, L\}\right\} \tag{20}$$

be the phase-aligned *zone* of AP l, which includes AP l itself and its neighboring APs. An example of how zones



(a) Phase-aligned *zones*. The zone of an AP is marked according to the color of that AP.



(b) Non-overlapping phase-aligned *clusters*.

FIGURE 2. **AP Clustering.**

are formed for each AP is provided in Fig. 2(a). For each AP zones are marked according to the color of the AP in the figure. Note that, all the APs included in the zone of an AP are within the reference distance D and hence, can be phase synchronized. With zones, we are considering only neighboring APs of an AP and hence, the zones can overlap with each other.

The APs in a phase-synchronized zone can perform coherent transmission to a UEs. Here, the terminology "*zone*" instead of "*cluster*" is used. This is because some APs might share the same phase-synchronized zones and some phase-synchronized zones can overlap with each other, while the clusters are non-overlapping. An example of phase-aligned zones and clusters is given in Fig. 2.

Algorithm 1 AP Clustering in PC Transmission in Cell-Free Massive MIMO Systems

Input: Channel matrices $\mathbf{G}_{kl}$, $\forall k, l$.
Initialize: $n = 1$, $\mathcal{Z}_l^{(1)} = \mathcal{Z}_l$, $\forall l$, $\mathcal{C} = \emptyset$

Update $L_{\max}^{(1)}$, $\mathcal{S}_{\max}^{(1)}$ using (21), (22)
while $L_{\max}^{(n)} > 1$ **do**
  if $|\mathcal{S}_{\max}^{(n)}| = 1$ **then**
    Update $\mathcal{C} = \mathcal{C} \cup \mathcal{S}_{\max}^{(n)}$.
  else
    for all $\mathcal{Z}_{l_i}^{(n)} \in \mathcal{S}_{\max}^{(n)}$ **do**
      Update $\overline{\mathbf{G}}_{kl_i} = \left[\mathbf{G}_{kl_{i,1}}, \mathbf{G}_{kl_{i,2}}, \ldots, \mathbf{G}_{kl_{i,L_{\max}^{(n)}}}\right]$
    end for
    Update $\mathcal{C} = \mathcal{C} \cup \mathcal{Z}_{l_i^\star}^{(n)}$,
    where $l_i^\star = \underset{l_i}{\operatorname{argmax}} \sum_{k \in \mathcal{K}} \|\overline{\mathbf{G}}_{kl_i}\|_F^2$
  end if
  Update $n := n + 1$
  Update $L_{\max}^{(n)}$, $\mathcal{S}_{\max}^{(n)}$, $\mathcal{Z}_l^{(n)}$ using (21)–(23)
end while
for all l such that $\left|\mathcal{Z}_l^{(n)}\right| \neq 0$ **do**
  Update $\mathcal{C} = \mathcal{C} \cup \mathcal{Z}_l^{(n)}$
end for
Output: Set $\mathcal{C}$ of non-overlapping AP clusters

We propose an AP clustering approach to find the non-overlapping phase-aligned AP clusters among the phase-aligned zones, which is provided in Algorithm 1. Since the phase of APs should be aligned whenever possible for high data rates as discussed in Section II, the number of APs in a phase-aligned cluster should be as large as possible. Therefore, the key idea of Algorithm 1 is to select a cluster as the phase-aligned zone that has the largest number of APs in each iteration.

Let $\mathcal{Z}_l^{(n)}$ be the zone of AP-l at iteration-n. For the first iteration, we initialize $\mathcal{Z}_l^{(1)} = \mathcal{Z}_l$. Denote by

$$L_{\max}^{(n)} = \max_l \left|\mathcal{Z}_l^{(n)}\right|, \tag{21}$$

the largest number of APs in all the zones in iteration n. Let

$$\mathcal{S}_{\max}^{(n)} = \left\{\mathcal{Z}_l^{(n)} \;\middle|\; \left|\mathcal{Z}_l^{(n)}\right| = L_{\max}^{(n)}\right\}, \tag{22}$$

be the set of zones having the same size of $L_{\max}^{(n)}$. For iteration 1 for Fig. 2(a), $L_{\max}^{(1)} = 6$ and $\mathcal{S}_{\max}^{(1)} = \mathcal{Z}_9^{(1)} = \{9, 3, 10, 6, 8, 1\}$, the zone of AP-9. If $|\mathcal{S}_{\max}^{(n)}| > 1$, a cluster is selected as the zone in $\mathcal{S}_{\max}^{(n)}$ that has the largest desired signal strength to the UEs. Hence, the proposed AP clustering algorithm focuses on increasing APs in a cluster considering both the reference distance (largest zone) and the UE positions (largest signal strength).

Let $\mathcal{Z}_{l^\star}^{(n)}$ be the chosen phase-aligned zone in iteration n, where $l^\star$ is the index of the corresponding AP. For the example in Fig. 2(a), there are no other zones with size 6, and hence, $\mathcal{C}_1 = \mathcal{Z}_9^{(1)}$ is chosen as cluster in the first

iteration. Then, the set of phase-aligned zones is updated by removing the APs of the selected zones from the zones that are not selected, i.e.,

$$\mathcal{Z}_l^{(n)} = \mathcal{Z}_l^{(n-1)} \setminus \left\{ l' \mid l' \in \mathcal{Z}_l^{(n-1)} \cap \mathcal{Z}_{l^\star}^{(n-1)} \right\}. \tag{23}$$

Thus, for the example, we have

$$\mathcal{Z}_4^{(2)} = \{4, 5\}, \quad \mathcal{Z}_7^{(2)} = \{7\}, \quad \mathcal{Z}_2^{(2)} = \{2\}.$$

The AP clustering algorithm continues with the next iteration with the updated zones. With the clustering algorithm, we have

$$\mathcal{C}_2 = \{4, 5\}, \quad \mathcal{C}_3 = \{2\}, \quad \mathcal{C}_4 = \{7\}.$$

Algorithm 1 terminates when there are no zones in which the numbers of APs are larger than 1. The output of Algorithm 1 is a set $\mathcal{C}$ of non-overlapping phase-aligned AP clusters with $L_c = |\mathcal{C}|$.

V. PRECODING AND COMBINING OPTIMIZATION ALGORITHM FOR SUM RATE MAXIMIZATION

For given values of $\{d_{kc}\}$, the optimization problem of maximizing the sum rate is given by

$$\begin{aligned} &\underset{\{\overline{\mathbf{W}}_{kc}\},\{\overline{\mathbf{V}}_{kc}\}}{\text{maximize}} && \sum_{k=1}^{K}\sum_{c=1}^{L_c} R_{kc}\left(\left\{\overline{\mathbf{W}}_{kc}\right\}, \overline{\mathbf{V}}_{kc}\right) \\ &\text{subject to} && (12). \end{aligned} \tag{24}$$

In this section, we develop a sub-optimal solution to problem (24) that is workable but still provides good performance. Specifically, we propose a novel algorithm to optimize the design of precoding and combining matrices for a given number of data streams, by tailoring the WMMSE framework [37], [38] to solve (24). The algorithm exploits the relationship between the mean square error (MSE) and data rate to transform the problem into a more tractable form, which can be solved by a block coordinate descent (BCD) method. Note that the power constraints (12) are per-AP, rather than per-cluster constraints. Moreover, the cluster sizes differ from cluster to cluster, causing the size of the optimization variable matrices in the objective function to have varying dimensions. Therefore, the WMMSE framework in [37], [38] cannot be directly applied to solve (24).

There are two key properties of the objective function of problem (24) that motivate the use of the WMMSE framework [37], [38]. These properties are stated in the following propositions.

Proposition 1: For given $\{\overline{\mathbf{W}}_{kc}\}$, the instantaneous rate $R_{kc}(\overline{\mathbf{V}}_{kc}) = \log_2 |\mathbf{I}_{d_{kc}} + \rho \overline{\mathbf{H}}_{kc}^{\mathrm{H}} \overline{\mathbf{Q}}_{kc}^{-1} \overline{\mathbf{H}}_{kc}|$ is maximized by the minimum mean square error (MMSE) combining matrix

$$\overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}} = \sqrt{\rho}\left(\rho \sum_{c'=1,c'\neq c}^{L_c} \overline{\mathbf{G}}_{kc'} \overline{\mathbf{W}}_{kc'} \overline{\mathbf{W}}_{kc'}^{\mathrm{H}} \overline{\mathbf{G}}_{kc'}^{\mathrm{H}} + \rho \sum_{c'=1}^{L_c}\sum_{k'=1}^{K} \overline{\mathbf{G}}_{kc'} \overline{\mathbf{W}}_{k'c'} \overline{\mathbf{W}}_{k'c'}^{\mathrm{H}} \overline{\mathbf{G}}_{kc'}^{\mathrm{H}} + \mathbf{I}_N\right)^{-1} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc}, \tag{25}$$

which results in the achievable rate for UE k in a MIMO broadcast channel given by (26) on the bottom of the page.

Proof: See the Appendix. ■

Proposition 2: For given $\overline{\mathbf{V}}_{kc} = \overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}}$, the instantaneous rate $R_{kc}(\{\overline{\mathbf{W}}_{kc}\})$ in (26) can be written as

$$\widetilde{R}_{kc}\left(\left\{\overline{\mathbf{W}}_{kc}\right\}\right) = d_{kc} + \max_{\overline{\mathbf{C}}_{kc}\succ\mathbf{0}, \overline{\mathbf{U}}_{kc}} \log_2\left|\overline{\mathbf{C}}_{kc}\right| - \mathrm{Tr}\left[\overline{\mathbf{C}}_{kc}\overline{\mathbf{E}}_{kc}\left(\overline{\mathbf{U}}_{kc}, \left\{\overline{\mathbf{W}}_{kc}\right\}\right)\right], \tag{27}$$

where $\overline{\mathbf{C}}_{kc} \in \mathbb{C}^{d_{kc}\times d_{kc}}$, $\overline{\mathbf{U}}_{kc} \in \mathbb{C}^{d_{kc}\times d_{kc}}$ are additional variables, and

$$\begin{aligned} &\overline{\mathbf{E}}_{kc}\left(\overline{\mathbf{U}}_{kc}, \left\{\overline{\mathbf{W}}_{kc}\right\}\right) \\ &= \left(\mathbf{I}_{d_{kc}} - \sqrt{\rho}\overline{\mathbf{U}}_{kc}^{\mathrm{H}}\overline{\mathbf{G}}_{kc}\overline{\mathbf{W}}_{kc}\right)\left(\mathbf{I}_{d_{kc}} - \sqrt{\rho}\overline{\mathbf{U}}_{kc}^{\mathrm{H}}\overline{\mathbf{G}}_{kc}\overline{\mathbf{W}}_{kc}\right)^{\mathrm{H}} \\ &\quad + \overline{\mathbf{U}}_{kc}^{\mathrm{H}}\left(\rho \sum_{c'=1,c'\neq c}^{L_c} \overline{\mathbf{G}}_{kc'} \overline{\mathbf{W}}_{kc'} \overline{\mathbf{W}}_{kc'}^{\mathrm{H}} \overline{\mathbf{G}}_{kc'}^{\mathrm{H}} \right. \\ &\quad \left. + \rho \sum_{c'=1}^{L_c}\sum_{k'=1,k'\neq k}^{K} \overline{\mathbf{G}}_{kc'} \overline{\mathbf{W}}_{k'c'} \overline{\mathbf{W}}_{k'c'}^{\mathrm{H}} \overline{\mathbf{G}}_{kc'}^{\mathrm{H}} + \mathbf{I}_N\right)\overline{\mathbf{U}}_{kc} \end{aligned} \tag{28}$$

has the same form of the MSE with combining matrix $\overline{\mathbf{U}}_{kc}$. The optimal values of $\overline{\mathbf{U}}_{kc}$ and $\overline{\mathbf{C}}_{kc}$ are

$$\overline{\mathbf{U}}_{kc}^{\mathrm{opt}} = \overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}}, \tag{29}$$

$$\overline{\mathbf{C}}_{kc}^{\mathrm{opt}} = \left(\mathbf{I}_{d_{kc}} - \sqrt{\rho}\left(\overline{\mathbf{U}}_{kc}^{\mathrm{opt}}\right)^{\mathrm{H}}\overline{\mathbf{G}}_{kc}\overline{\mathbf{W}}_{kc}\right)^{-1}. \tag{30}$$

Proof: The proof follows [38, Lemma 4.1], and hence, is omitted. ■

The function $\log_2|\overline{\mathbf{C}}_{kc}| - \mathrm{Tr}[\overline{\mathbf{C}}_{kc}\overline{\mathbf{E}}_{kc}(\overline{\mathbf{U}}_{kc}, \{\overline{\mathbf{W}}_{kc}\})]$ is concave with respect to each variable $\overline{\mathbf{C}}_{kc}$, $\overline{\mathbf{U}}_{kc}$, and $\{\overline{\mathbf{W}}_{kc}\}$, and hence, is a tractable function. Thus, from Propositions 1 and 2, the MMSE combining matrix $\overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}}$ is shown not only to maximize the rate of UE k from cluster c, but also to transform the rate function into a more tractable form (27) with respect to $\{\overline{\mathbf{W}}_{kc}\}$. Using the properties from

$$R_{kc} = \log_2\left|\mathbf{I}_{d_k} + \rho\,\mathbf{W}_k^{\mathrm{H}}\mathbf{G}_k^{\mathrm{H}}\left(\rho \sum_{c'=1,c'\neq c}^{L_c} \overline{\mathbf{G}}_{kc'} \overline{\mathbf{W}}_{kc'} \overline{\mathbf{W}}_{kc'}^{\mathrm{H}} \overline{\mathbf{G}}_{kc'}^{\mathrm{H}} + \rho \sum_{c'=1}^{L_c}\sum_{k'=1,k'\neq k}^{K} \overline{\mathbf{G}}_{kc'} \overline{\mathbf{W}}_{k'c'} \overline{\mathbf{W}}_{k'c'}^{\mathrm{H}} \overline{\mathbf{G}}_{kc'}^{\mathrm{H}} + \mathbf{I}_N\right)^{-1}\mathbf{G}_k\,\mathbf{W}_k\right|. \tag{26}$$

Propositions 1 and 2, we transform problem (24) into an equivalent problem as follows:

$$\underset{\left\{\overline{\mathbf{W}}_{kc}\right\},\left\{\overline{\mathbf{V}}_{kc}\right\},\left\{\overline{\mathbf{C}}_{kc}\right\}}{\text{minimize}} \sum_{k=1}^{K}\sum_{c=1}^{L_c}\left(\mathrm{Tr}\left(\overline{\mathbf{C}}_{kc}\,\mathbf{E}_{kc}\left(\overline{\mathbf{V}}_{kc},\left\{\overline{\mathbf{W}}_{kc}\right\}\right)\right) - \log_2\left|\overline{\mathbf{C}}_{kc}\right|\right) \quad (31)$$

subject to (12).

The objective function of problem (31) is the weighted sum-MSE with the weight matrices $\{\overline{\mathbf{C}}_{kc}\}$. It is jointly non-convex over all the block variables $(\{\overline{\mathbf{W}}_{kc}\},\{\overline{\mathbf{V}}_{kc}\},\{\overline{\mathbf{C}}_{kc}\})$, but convex over each block variable $\{\overline{\mathbf{W}}_{kc}\}$, $\{\overline{\mathbf{V}}_{kc}\}$ or $\{\overline{\mathbf{C}}_{kc}\}$. This motivates the use of the BCD method to iteratively minimize the weighted sum-MSE in the objective function of problem (31). In particular for fixed $(\{\overline{\mathbf{W}}_{kc}\},\{\overline{\mathbf{C}}_{kc}\})$, we update $\{\overline{\mathbf{V}}_{kc}\}$ based on the insights of Proposition 1 as follows:

$$\overline{\mathbf{V}}_{kc} = \overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}}, \quad \forall k, c. \quad (32)$$

For fixed $(\{\overline{\mathbf{V}}_{kc}\},\{\overline{\mathbf{W}}_{kc}\})$, we update $\{\overline{\mathbf{C}}_{kc}\}$ using the insights of Proposition 2 as

$$\overline{\mathbf{C}}_{kc} = \left(\mathbf{I}_{d_{kc}} - \sqrt{\rho}\overline{\mathbf{V}}_{kc}^{\mathrm{H}}\overline{\mathbf{G}}_{kc}\overline{\mathbf{W}}_{kc}\right)^{-1} \quad \forall k, c. \quad (33)$$

For fixed $(\{\overline{\mathbf{V}}_{kc}\},\{\overline{\mathbf{C}}_{kc}\})$, we update $\{\overline{\mathbf{W}}_{kc}\}$ by solving the problem (34) (see the bottom of the page). We summarize the steps to solve problem (31) (or problem (24)) in Algorithm 2.

Algorithm 2 Solving Problem (31)

Initialize: $\{\overline{\mathbf{W}}_{kc}\}$ that satisfies (12)

1: **repeat**
2: Update $\{\overline{\mathbf{V}}_{kc}\}$ as (32)
3: Update $\{\overline{\mathbf{C}}_{kc}\}$ as (33)
4: Update $\{\overline{\mathbf{W}}_{kc}\}$ by solving (34)
5: **until** convergence

Since $\overline{\mathbf{G}}_{kc}\overline{\mathbf{W}}_{kc} = \sum_{l\in\mathcal{C}_c}\mathbf{G}_{kl}\mathbf{W}_{kl}$, the problem (34) can be equivalently written as problem (35) (see the bottom of the page). In light of the per-AP power constraints in (12), $\{\mathbf{W}_{kl}\}$ can be considered as a block variable. Therefore, we apply the BCD approach with block variables $\{\mathbf{W}_{kl}\}$, for all $k \in \mathcal{K}, l \in \mathcal{L}$, to solve problem (35). Since problem (35) is convex with a single quadratic constraint, we can obtain a closed-form solution by dealing with its dual problem as follows.

The Lagrangian function associated with problem (35) with a Lagrangian multipliers $\{\lambda_l\}$ for the power constraints (12) is defined as

$$\mathcal{L}(\{\mathbf{W}_{kl}\},\{\lambda_l\}) = \tilde{f}(\{\mathbf{W}_{kl}\}) + \sum_{l_1=1}^{L}\lambda_{l_1}\left(\sum_{k_1=1}^{K}\mathrm{Tr}\left(\mathbf{W}_{k_1l_1}\mathbf{W}_{k_1l_1}^{\mathrm{H}}\right) - 1\right). \quad (36)$$

$$\begin{aligned}\underset{\{\mathbf{W}_{kc}\}}{\text{minimize}} \; & \sum_{k_1=1}^{K}\sum_{c_1=1}^{L_c}\mathrm{Tr}\left(\overline{\mathbf{C}}_{k_1c_1}\left(\mathbf{I}_{d_{k_1c_1}} - \sqrt{\rho}\overline{\mathbf{V}}_{k_1c_1}^{\mathrm{H}}\overline{\mathbf{G}}_{k_1c_1}\overline{\mathbf{W}}_{k_1c_1}\right)\left(\mathbf{I}_{d_{k_1c_1}} - \sqrt{\rho}\overline{\mathbf{V}}_{k_1c_1}^{\mathrm{H}}\overline{\mathbf{G}}_{k_1c_1}\overline{\mathbf{W}}_{k_1c_1}\right)^{\mathrm{H}}\right) \\ & + \sum_{k_1=1}^{K}\sum_{c_1=1}^{L_c}\mathrm{Tr}\left(\rho\overline{\mathbf{C}}_{k_1c_1}\sum_{c_2=1,c_2\neq c_1}^{L_c}\overline{\mathbf{V}}_{k_1c_1}^{\mathrm{H}}\overline{\mathbf{G}}_{k_1c_2}\overline{\mathbf{W}}_{k_1c_2}\overline{\mathbf{W}}_{k_1c_2}^{\mathrm{H}}\overline{\mathbf{G}}_{k_1c_2}^{\mathrm{H}}\overline{\mathbf{V}}_{k_1c_1}\right. \\ & \left. + \rho\overline{\mathbf{C}}_{k_1c_1}\sum_{c_2=1}^{L_c}\sum_{k_2=1,k_2\neq k_1}^{K}\overline{\mathbf{V}}_{k_1c_1}^{\mathrm{H}}\overline{\mathbf{G}}_{k_1c_2}\overline{\mathbf{W}}_{k_2c_2}\overline{\mathbf{W}}_{k_2c_2}^{\mathrm{H}}\overline{\mathbf{G}}_{k_1c_2}^{\mathrm{H}}\overline{\mathbf{V}}_{k_1c_1}\right) \\ & \text{subject to (12).}\end{aligned} \quad (34)$$

$$\begin{aligned}\underset{\{\mathbf{W}_{kl}\}}{\text{minimize}}\; \tilde{f}(\{\mathbf{W}_{kl}\}) = & -\sqrt{\rho}\sum_{k_1=1}^{K}\sum_{c_1=1}^{L_c}\sum_{l_1\in\mathcal{C}_{c_1}}\mathrm{Tr}\left(\overline{\mathbf{C}}_{k_1c_1}\overline{\mathbf{V}}_{k_1c_1}^{\mathrm{H}}\mathbf{G}_{k_1l_1}\mathbf{W}_{k_1l_1}\right) \\ & -\sqrt{\rho}\sum_{k_1=1}^{K}\sum_{c_1=1}^{L_c}\sum_{l_1\in\mathcal{C}_{c_1}}\mathrm{Tr}\left(\mathbf{W}_{k_1l_1}^{\mathrm{H}}\mathbf{G}_{k_1l_1}^{\mathrm{H}}\overline{\mathbf{V}}_{k_1c_1}\overline{\mathbf{C}}_{k_1c_1}\right) \\ & +\rho\sum_{k_1=1}^{K}\sum_{c_1=1}^{L_c}\sum_{c_2=1}^{L_c}\sum_{l_1\in\mathcal{C}_{c_2}}\sum_{l_2\in\mathcal{C}_{c_2}}\mathrm{Tr}\left(\mathbf{W}_{k_1l_2}^{\mathrm{H}}\mathbf{G}_{k_1l_2}^{\mathrm{H}}\overline{\mathbf{V}}_{k_1c_1}\overline{\mathbf{C}}_{k_1c_1}\overline{\mathbf{V}}_{k_1c_1}^{\mathrm{H}}\mathbf{G}_{k_1l_1}\mathbf{W}_{k_1l_1}\right) \\ & +\rho\sum_{k_1=1}^{K}\sum_{k_2=1,k_2\neq k_1}^{K}\sum_{c_1=1}^{L_c}\sum_{c_2=1}^{L_c}\sum_{l_1\in\mathcal{C}_{c_2}}\sum_{l_2\in\mathcal{C}_{c_2}}\mathrm{Tr}\left(\mathbf{W}_{k_2l_2}^{\mathrm{H}}\mathbf{G}_{k_1l_2}^{\mathrm{H}}\overline{\mathbf{V}}_{k_1c_1}\overline{\mathbf{C}}_{k_1c_1}\overline{\mathbf{V}}_{k_1c_1}^{\mathrm{H}}\mathbf{G}_{k_1l_1}\mathbf{W}_{k_2l_1}\right) \\ & \text{subject to (12).}\end{aligned} \quad (35)$$

Then, the dual problem of problem (35) is

$$\underset{\lambda_l}{\text{maximize}}\ \widetilde{\mathcal{L}}(\lambda_l) \tag{37a}$$

$$\text{subject to}\ \lambda_l \geq 0, \tag{37b}$$

where $\widetilde{\mathcal{L}}(\lambda_l)$ is the dual function, which is given by

$$\widetilde{\mathcal{L}}(\lambda_l) = \underset{\{\mathbf{W}_{kl}\}}{\text{minimize}}\ \mathcal{L}(\{\mathbf{W}_{kl}\}, \lambda_l). \tag{38}$$

Problem (38) is convex and its optimal $\mathbf{W}_{kl}$ can be obtained by taking the partial derivative of $\mathcal{L}(\{\mathbf{W}_{kl}\}, \{\lambda_l\})$ with respect to $\mathbf{W}_{kl}$ and setting it to zero. The optimal $\mathbf{W}_{kl}$ for a given λ_l is given by (39) (see the bottom of the page), where c^l is the index of the cluster where l-th AP belongs to.

The Lagrangian multiplier $\lambda_l \geq 0$ should be chosen to satisfy the complementary slackness conditions (of strong duality) corresponding to the power constraint (12), i.e.,

$$\lambda_l \left(\sum_{k=1}^{K} \text{Tr}\big(\mathbf{W}_{kl}^{\star}(\mathbf{W}_{kl}^{\star})^{\text{H}}\big) - 1 \right) = 0, \ \ \forall l. \tag{40}$$

Towards this, we analyze the transmission power constraint using the optimal value $\mathbf{W}_{kl}^{\star}(\lambda_l)$ for given λ_l as follows.

Problem (34) is convex quadratically constrained quadratic program (QCQP) and can be solved by a commercial interior-point optimization solver such as CVX [43]. However, using CVX to solve (34) is computationally demanding, especially when the numbers of APs and UEs are large. Therefore, in the following, we propose an algorithm to find the solution to (34) with closed-form expressions in each iteration. The proposed algorithm will require much lower computational resources in a large-scale system than using a commercial convex solver.

Using the singular value decomposition (SVD), (41) and (42) (see the bottom of the page) hold. From (39) and (41), the transmission power at AP l can be written in terms of λ_l as (43) (see the bottom of the page). Let

$$\mathbf{T}_{kl} = \mathbf{\Psi}_{kl}^{\text{H}} \, \mathbf{\Lambda}_{kl} \, \mathbf{\Lambda}_{kl}^{\text{H}} \mathbf{\Psi}_{kl} \, . \tag{44}$$

Then, from (43) and (44), it is true that

$$\begin{aligned} \sum_{k=1}^{K} \text{Tr}\big(\mathbf{W}_{kl}^{\star}(\mathbf{W}_{kl}^{\star})^{\text{H}}\big) &= \sum_{k=1}^{K} \text{Tr}\Big(\rho(\rho \, \mathbf{\Sigma}_{kl} + \lambda_l \, \mathbf{I}_M)^{-2} \, \mathbf{T}_{kl}\Big) \\ &= \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{\rho [\mathbf{T}_{kl}]_{mm}}{(\rho [\mathbf{\Sigma}_{kl}]_{mm} + \lambda_l)^2}. \end{aligned} \tag{45}$$

As seen from (45), the transmit power at AP l is a monotonically decreasing function in λ_l. Based on this property, we can choose λ_l that satisfies the slackness

$$\begin{aligned} \mathbf{W}_{kl}^{\star} = \Bigg(& \rho \sum_{k_1=1, k_1 \neq k}^{K} \sum_{c_1=1}^{L_c} \mathbf{G}_{k_1 l}^{\text{H}} \, \overline{\mathbf{V}}_{k_1 c_1} \overline{\mathbf{C}}_{k_1 c_1} \overline{\mathbf{V}}_{k_1 c_1}^{\text{H}} \, \mathbf{G}_{k_1 l} + \rho \sum_{c_1=1}^{L_c} \mathbf{G}_{kl}^{\text{H}} \, \overline{\mathbf{V}}_{k c_1} \overline{\mathbf{C}}_{k c_1} \overline{\mathbf{V}}_{k c_1}^{\text{H}} \, \mathbf{G}_{kl} + \lambda_l \, \mathbf{I}_M \Bigg)^{-1} \\ & \times \Bigg(\sqrt{\rho} \, \mathbf{G}_{kl}^{\text{H}} \, \overline{\mathbf{V}}_{k c^l} \overline{\mathbf{C}}_{k c^l} - \rho \sum_{c_1=1}^{L_c} \sum_{l_1 \in \mathcal{C}_{c^l} \setminus \{l\}} \mathbf{G}_{kl}^{\text{H}} \, \overline{\mathbf{V}}_{k c_1} \overline{\mathbf{C}}_{k c_1} \overline{\mathbf{V}}_{k c_1}^{\text{H}} \, \mathbf{G}_{k l_1} \, \mathbf{W}_{k l_1} \\ & \quad - \rho \sum_{k_1=1, k_1 \neq k}^{K} \sum_{c_1=1}^{L_c} \sum_{l_1 \in \mathcal{C}_{c^l} \setminus \{l\}} \mathbf{G}_{k_1 l}^{\text{H}} \, \overline{\mathbf{V}}_{k_1 c_1} \overline{\mathbf{C}}_{k_1 c_1} \overline{\mathbf{V}}_{k_1 c_1}^{\text{H}} \, \mathbf{G}_{k_1 l_1} \, \mathbf{W}_{k l_1} \Bigg) \end{aligned} \tag{39}$$

$$\mathbf{\Psi}_{kl} \, \mathbf{\Sigma}_{kl} \, \mathbf{\Psi}_{kl}^{\text{H}} = \rho \sum_{k_1=1, k_1 \neq k}^{K} \sum_{c_1=1}^{L_c} \mathbf{G}_{k_1 l}^{\text{H}} \, \overline{\mathbf{V}}_{k_1 c_1} \overline{\mathbf{C}}_{k_1 c_1} \overline{\mathbf{V}}_{k_1 c_1}^{\text{H}} \, \mathbf{G}_{k_1 l} + \rho \sum_{c_1=1}^{L_c} \mathbf{G}_{kl}^{\text{H}} \, \overline{\mathbf{V}}_{k c_1} \overline{\mathbf{C}}_{k c_1} \overline{\mathbf{V}}_{k c_1}^{\text{H}} \, \mathbf{G}_{kl} \tag{41}$$

$$\mathbf{\Lambda}_{kl} = \sqrt{\rho} \, \mathbf{G}_{kl}^{\text{H}} \, \overline{\mathbf{V}}_{k c^l} \overline{\mathbf{C}}_{k c^l} - \rho \sum_{c_1=1}^{L_c} \sum_{l_1 \in \mathcal{C}_{c^l} \setminus \{l\}} \mathbf{G}_{kl}^{\text{H}} \, \overline{\mathbf{V}}_{k c_1} \overline{\mathbf{C}}_{k c_1} \overline{\mathbf{V}}_{k c_1}^{\text{H}} \, \mathbf{G}_{k l_1} \, \mathbf{W}_{k l_1} \tag{42}$$

$$\begin{aligned} \sum_{k=1}^{K} \text{Tr}\big(\mathbf{W}_{kl}^{\star}(\mathbf{W}_{kl}^{\star})^{\text{H}}\big) &= \sum_{k=1}^{K} \text{Tr}\big((\mathbf{W}_{kl}^{\star})^{\text{H}} \mathbf{W}_{kl}^{\star}\big) \\ &= \sum_{k=1}^{K} \text{Tr}\Big[\rho \, \mathbf{\Lambda}_{kl}^{\text{H}} \big(\rho \, \mathbf{\Psi}_{kl} \, \mathbf{\Sigma}_{kl} \, \mathbf{\Psi}_{kl}^{\text{H}} + \lambda_l \, \mathbf{I}_M\big)^{-1} \big(\rho \, \mathbf{\Psi}_{kl} \, \mathbf{\Sigma}_{kl} \, \mathbf{\Psi}_{kl}^{\text{H}} + \lambda_l \, \mathbf{I}_M\big)^{-1} \, \mathbf{\Lambda}_{kl} \Big] \\ &= \sum_{k=1}^{K} \text{Tr}\Big[\rho (\rho \, \mathbf{\Sigma}_{kl} + \lambda_l \, \mathbf{I}_M)^{-2} \, \mathbf{\Psi}_{kl}^{\text{H}} \, \mathbf{\Lambda}_{kl} \, \mathbf{\Lambda}_{kl}^{\text{H}} \, \mathbf{\Psi}_{kl} \Big]. \end{aligned} \tag{43}$$

Algorithm 3 Bisection Search for Solving Problem (35)

Input: $\{\overline{\mathbf{V}}_{kc}\}$, $\{\overline{\mathbf{C}}_{kc}\}$, $\{\mathbf{W}_{kl}\}$, ϵ.
Initialize: $\{\mathbf{\Psi}_{kl}\}$ and $\{\mathbf{\Sigma}_{kl}\}$
for $l = 1 : L$ **do**
  Update $\{\mathbf{\Lambda}_{kl}\}$, $\forall\ k$ using (41)
  Update $\{\mathbf{T}_{kl}\}$, $\forall\ k$ using (44)
  Initialize $\lambda_{lb} = 0$ and $\lambda_{ub} = \sqrt{\sum_{k=1}^{K}\sum_{m=1}^{M}\rho[\,\mathbf{T}_{kl}\,]_{mm}}$
  while $|\lambda_{ub} - \lambda_{lb}| \leq \epsilon$ **do**
    Update $\lambda_l = \frac{\lambda_{lb}+\lambda_{ub}}{2}$
    if $\sum_{k'=1}^{K} \mathrm{Tr}(\mathbf{W}^{\star}_{k'l}(\lambda_l)(\mathbf{W}^{\star}_{k'l}(\lambda_l))^{\mathrm{H}}) \geq 1$ **then**
      set $\lambda_{lb} = \lambda_l$
    else
      set $\lambda_{ub} = \lambda_l$
    end if
  end while
  Set $\lambda^{\star}_l = \lambda_l$
  Update $\{\mathbf{W}^{\star}_{kl}\}$ for given $\lambda^{\star}_l$ as (39) $\forall k$.
end for

condition in (40) by a bisection search, i.e., choosing $\lambda_l \geq 0$ to ensure $\sum_{k=1}^{K} \mathrm{Tr}(\mathbf{W}^{\star}_{kl}(\mathbf{W}^{\star}_{kl})^{\mathrm{H}}) - 1 = 0$. In the bisection search method, λ_l is searched within the range $(\lambda_{lb}, \lambda_{ub})$. Here, we choose

$$\lambda_{lb} = 0 \tag{46}$$

$$\lambda_{ub} = \sqrt{\sum_{k=1}^{K}\sum_{m=1}^{M}\rho[\mathbf{T}_{kl}]_{mm}}, \tag{47}$$

so that

$$\sum_{k=1}^{K}\sum_{m=1}^{M}\frac{\rho[\mathbf{T}_{kl}]_{mm}}{(\rho[\mathbf{\Sigma}_{kl}]_{mm}+\lambda_l)^2} \leq \sum_{k=1}^{K}\sum_{m=1}^{M}\frac{\rho[\mathbf{T}_{kl}]_{mm}}{\lambda_l^2} \leq 1. \tag{48}$$

The bisection search algorithm to solve problem (35) is provided in Algorithm 3. Algorithm 3 converges to a stationary solution to problem (35) (hence, problem (34)). The proof of the convergence of Algorithm 3 follows a similar proof as in [37], and hence, is omitted.

VI. DATA STREAM ALLOCATION

This section will answer the next question of how to optimally allocate the data streams to the UEs. If $ML > KN$, the maximum number of orthogonal data streams that can be transmitted from the effective array constituted by the APs is KN, which is ideally achieved in a system with full-rank channels, appropriate combining and precoding designs. However, if $KN > ML$, then KN data streams cannot be made orthogonal over the channels, resulting in strong inter-stream interference. Therefore, it is important to efficiently allocate data streams to UEs, which helps to manage inter-stream interference and improve the sum rate.

In this section, we allocate data streams to the UEs in a greedy manner. This helps to avoid an exhaustive search over all possible UE data stream allocations, which is practically impossible when L, M, K, N are large. It has been shown that greedy UE scheduling algorithms can provide a performance that is close to the optimum in downlink MIMO systems with a block diagonalization precoding approach [44], [45]. Therefore, the greedy UE scheduling approach is expected to provide good performance.

Let $\mathbf{P}_{kc}$ be the orthonormal basis matrix of $\overline{\mathbf{G}}_{kc}$. Then, let

$$\tilde{\mathbf{G}}_{kc} = \sum_{k'=1,k'\neq k} \mathbf{P}_{kc}^{\mathrm{H}}\overline{\mathbf{G}}_{k'c} \tag{49}$$

$$\check{\mathbf{G}}_{kc} = \sum_{c'=1,c'\neq c}^{L_c} \mathbf{P}_{kc}^{\mathrm{H}}\overline{\mathbf{G}}_{kc'} \tag{50}$$

be the projections of the intra-cluster and inter-cluster interference channels onto the channel $\overline{\mathbf{G}}_{kc}$, respectively. Define a matrix $\mathbf{S} \in \mathbb{R}^{K\times L_c}$ whose elements, i.e.,

$$[\mathbf{S}]_{kc} = \frac{\|\overline{\mathbf{G}}_{kc}\|^2}{1 + \|\tilde{\mathbf{G}}_{kc}\|^2 + \|\check{\mathbf{G}}_{kc}\|^2}, \tag{51}$$

are channel-to-noise-plus-interference-channel ratio (CINR). Let $R(\{d_{kc}\})$ be the sum rate obtained from using Algorithm 2 to solve problem (24) for a fixed $\{d_{kc}\}$. A greedy data allocation algorithm for sum rate maximization in cell-free massive MIMO systems is provided in Algorithm 4.

The key idea behind the proposed greedy data stream allocation algorithm is to prioritize allocating data streams for the user-cluster pairs with the strongest CINRs. The data stream allocation Algorithm 4 takes the minimum required values of $\mathcal{D} = \{d_{kc}\}$ as input. To ensure seamless connectivity for all the UEs, it is normal to allocate at least one data stream to each UE. Then we initialize the CINR matrix according to (51). The algorithm runs for all cluster-user pairs, until data streams are allocated for all the pairs. In each iteration, we pick the cluster-user pair that has the largest CINR in $\mathbf{S}$. Denote this pair as $(\hat{k}, \hat{c})$. Then we compute the sum rate as per (24) for different values of data stream, i.e., $d \in \{1, \ldots, \min\{M, N\}\}$, for this cluster-user pair. Let

$$d^* = \underset{d\in\{1,\ldots,\min\{M,N\}\}}{\mathrm{argmax}} R\Big(\mathcal{D} \setminus \{d_{\hat{k}\hat{c}}\} \bigcup \{d_{\hat{k}\hat{c}} = d\}\Big), \tag{52}$$

be the allocation quantity that maximizes the sum rate. We update the data stream allocation quantity for this cluster-user pair to be d^*, if $d^* \geq d_{\hat{k},\hat{c}}$, otherwise we allocate the minimum required quantity. Then, we remove this cluster-user pair from the matrix $\mathbf{S}$, and the allocation proceeds with the next strongest pair in matrix $\mathbf{S}$.

VII. COMPLEXITY ANALYSIS

In this section, we provide a complexity analysis of the proposed algorithms.

A. AP CLUSTERING ALGORITHM

Algorithm 1 for AP clustering has low complexity because it involves only the computation of distances and Frobenius norms of channel matrices. The complexity for the computation of distances between APs has a complexity of $\mathcal{O}(L^2)$.

Algorithm 4 Greedy Data Stream Allocation for Sum Rate Maximization

Input: $\mathcal{D} = \{d_{kc}\}$, where $\{d_{kc}\}$ are minimal values required from the system.

Initialize: $\mathbf{S}$ according to (51)

for $t = 1 : KL_c$ **do**
  Update $\hat{k}, \hat{c} = \underset{k,c}{\text{argmax}}\ [\mathbf{S}]_{kc}$
  Compute d^* as per (52)
  if $d^* \geq d_{\hat{k}\hat{c}}$ **then**
    $d_{\hat{k}\hat{c}} = d^*$
  end if
  Update $[\mathbf{S}]_{\hat{k}\hat{c}} = 0$
end for

The system has L APs and K UEs and thus, the Frobenius norm involves $\mathcal{O}(KLMN)$ computations. Hence, the total complexity for Algorithm 1 is $\mathcal{O}(KLMN + L^2)$.

B. PRECODING AND COMBINING OPTIMIZATION ALGORITHM

Algorithm 2 for the precoding and combining optimization involves the computation of the matrices $\overline{\mathbf{V}}_{kc}$, $\overline{\mathbf{C}}_{kc}$, and $\overline{\mathbf{W}}_{kc}$ for each cluster-user pair in each iteration. The computation of the MMSE matrix $\overline{\mathbf{V}}_{kc}$ in (32) consists of an $N \times N$ matrix inversion and some matrix multiplications. The complexity of computing $\overline{\mathbf{V}}_{kc}$ involves computations of $\mathcal{O}(KL_c(N^3 + N^2ML + N^2d + NMLd))$, where d is the maximum number of data streams allocated. The complexity of computing $\overline{\mathbf{C}}_{kc}$ in (33) involves computations of $\mathcal{O}(KL_c(d^3 + N^2d + NMLd))$. To compute $\overline{\mathbf{W}}_{kc}$, we use Algorithm 3 with closed-form expressions. The computation of $\mathbf{\Lambda}_{kl}$ requires $\mathcal{O}(LMNd + LMd^2 + LM^2d)$ operations. Thus, for all UEs and APs, the total complexity is of$\mathcal{O}(KL^2MNd + KL^2Md^2 + KL^2M^2d))$. Similarly, the complexity of computing a$\{\mathbf{\Psi}_{kl}\}$ is $\mathcal{O}(KL(M^3 + KL_cMNd + KL_cMd^2 + KL_cM^2d))$. The bisection search completes in a few iterations and thus Algorithm 3 requires L times the computation of $\{\mathbf{\Lambda}_{kl}\}$ and $\{\mathbf{\Psi}_{kl}\}$. Thus, the total complexity of an iteration of Algorithm 2 is $\mathcal{O}(KL_cN^3 + KL_cN^2ML + KL_cN^2d + KL_cNMLd + KL_cd^3 + KL_cN^2d + KL_cNMLds + KL^3MNd + KL^3Md^2 + KL^3M^2d + KL^2M^3 + K^2L^2L_cMNd + K^2L^2L_cMd^2 + K^2L^2L_cM^2d)$.

C. DATA STREAM ALLOCATION ALGORITHM

Algorithm 4 for greedy data stream allocation involves the computation of the CINR matrix, which has the complexity of $\mathcal{O}(KLcMNL)$, and running Algorithm 2 whose complexity is discussed in the previous subsection.

VIII. NUMERICAL RESULTS

We consider a cell-free massive MIMO system, where the APs and UEs are randomly distributed in a 0.5 km $\times$ 0.5 km square area, whose edges are wrapped around to avoid boundary effects. The distances between adjacent APs are at least 50 m [46]. We set the bandwidth to $B = 50$ MHz and the noise figure to $F = 9$ dB. Thus, the noise power $\sigma_n^2 = k_BT_0BF$, where $k_B = 1.381 \times 10^{-23}$ Joules/^{o}K is Boltzmann's constant, while $T_0 = 290^o$K is the noise temperature. Let $P_d = 1$ W be the maximum transmit power of the APs. The normalized maximum transmit power ρ is calculated by dividing P_d by the noise power.

The channels are modeled as uncorrelated Rayleigh fading. More specifically, $\mathbf{G}_{kl} = \sqrt{\beta_{kl}}\widetilde{\mathbf{G}}_{kl}$, where β_{kl} is the large-scale fading coefficient, and $\widetilde{\mathbf{G}}_{kl}$ are i.i.d. $\mathcal{CN}(0, 1)$ variables representing the small-scale fading coefficients. We model the large-scale fading coefficients β_{kl} as [46]

$$\beta_{kl} = 10^{\frac{\text{PL}(d_{kl})}{10}} 10^{\frac{F_{kl}}{10}}, \tag{53}$$

where

$$\text{PL}(d_{kl})\ \text{(dB)} = -30.5 - 36.7 \log_{10}\left(\frac{d_{kl}}{1\,\text{m}}\right), \tag{54}$$

represents the path loss, and $F_{kl} \in \mathcal{N}(0, 4^2)$ (dB) represents the shadowing effect. The correlation among the shadowing terms from the AP $l, \forall l \in \mathcal{L}$ to different UEs $k \in \mathcal{K}$ is expressed as:

$$\mathbb{E}\{F_{kl}F_{k'l'}\} \triangleq \begin{cases} 4^2 2^{-\delta_{kk'}/9\,\text{m}}, & \text{if } l' = l \\ 0, & \text{otherwise}, \quad \forall l' \in \mathcal{L}, \end{cases} \tag{55}$$

where $\delta_{kk'}$ is the physical distance between UEs k and k'. In this work, we assume that the channel coefficients are perfectly known at the APs and CPU. In practice, the channels are estimated using pilots. For example, a system using time-division duplexing operation has two phases: (i) an uplink transmission phase for channel estimation and (ii) a downlink transmission phase. In the uplink transmission phase, all the UEs transmit their pilots to all the APs, and then the APs use the pilot signal to estimate the channels. When the coherence interval is sufficiently long, the channel estimation is nearly perfect, which is the case considered in this work. Here, the PC operation combining the coherent and non-coherent transmissions is performed in the downlink transmission phase, and hence, does not have any impact on the channel estimation.

For simulations, we consider $D = 200$ m such that there is approximately a 10% drop of the maximum sum rate compared with that of the ideal FC scheme. Note that this is an example value of D, while the exact value of D depends on used phase-synchronization technologies.

Fig. 3 shows a sum rate comparison between FC, PC, and FNC under different network settings. Our proposed PC scheme can achieve much better performance than the FNC scheme. Moreover, as the number of APs in the network increases, it can achieve performance close to the FC scheme. This is reasonable because the PC scheme has higher beamforming gains obtained from the phase-aligned clusters, compared to the FC operation without any phase alignment among the APs in the networks.

To show the superiority of our proposed AP clustering algorithm, we compare our algorithm with an even

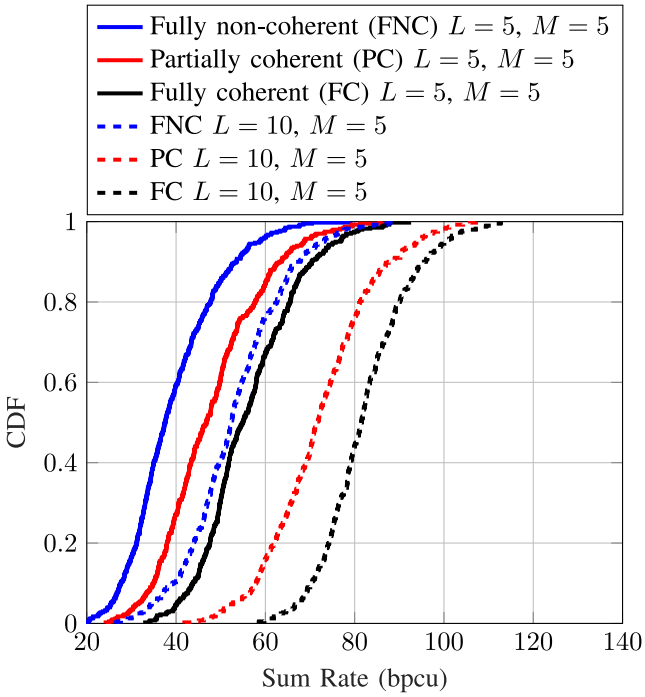


FIGURE 3. Comparison of FNC, PC, and FC under different network settings. Parameters for the plot: $K = 5$, $N = 2$, $D = 200$ m, and $d_{kc} = 2\ \forall k, c$.

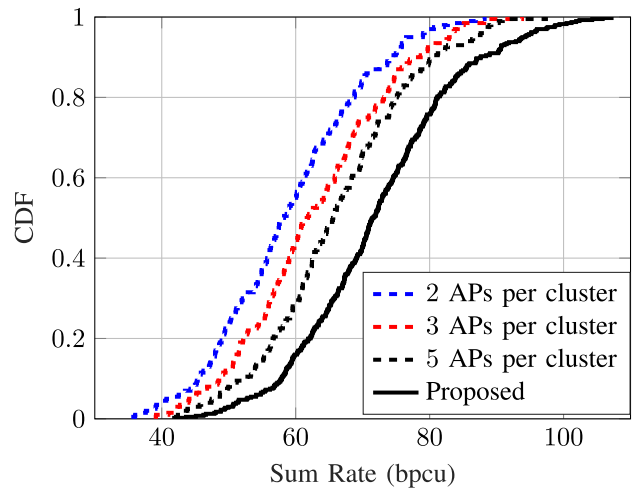


FIGURE 4. Performance of the proposed AP clustering algorithm. Parameters for the plot: $L = 10$, $K = 5$, $M = 5$, $N = 2$, $D = 200$ m, and $d_{kc} = 2\ \forall k, c$.

distribution of APs among the clusters; see Fig. 4. With even distribution, we cluster the APs based on the shortest distance between each other. Moreover, the APs are within the reference distance D of the phase-alignment. Those APs which are beyond the reference distance operate independently and are not assigned any cluster. From the figure, it can be seen that the proposed AP clustering algorithm outperforms the baseline schemes with an even clustering of APs. This is due to the fact that the proposed AP clustering algorithm maximizes the channel power in a cluster along with assigning a maximum number of APs per cluster.

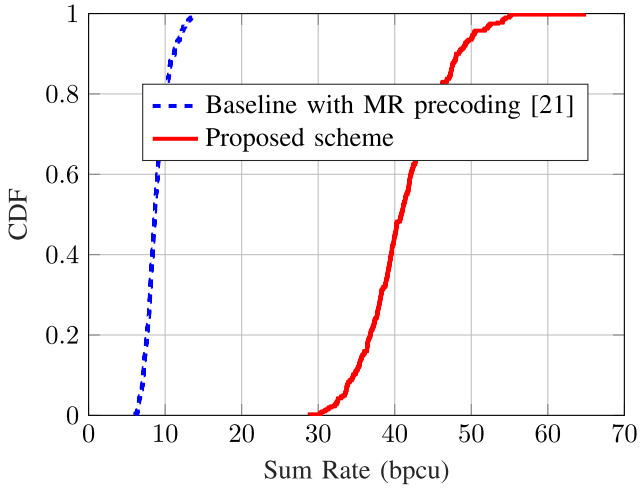


FIGURE 5. Performance of PC scheme compared to scheme in [21]. Parameters for the plot: $L = 10$, $M = 5$, $K = 5$, and $N = 1$.

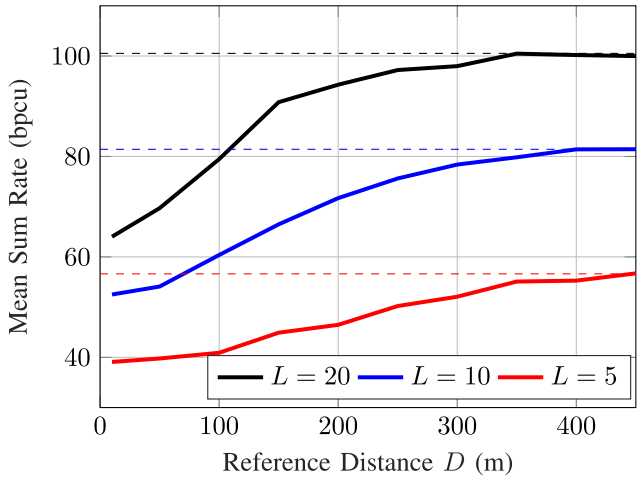


FIGURE 6. Sum rate of PC scheme with respect to the reference distance D. The dashed lines represent the sum rates of the FC operation. Parameters for the plot: $M = 5$, $K = 5$, $N = 2$, and $d_{kc} = 2\ \forall k, c$.

In Fig. 5, we compare the proposed scheme with a baseline using MR precoding scheme as in [21] for a PC system. For fair comparison with [21], in this figure, we consider single antenna UEs and allocate only a single data stream for every cluster-user pair. Here, the baseline uses the same AP clustering as that of the proposed PC scheme. Note that [21] assumes that the AP clusters are known, and does not take into account the practical problem of phase misalignment. Also, [21] considers MR precoding and does not consider the aspect of optimizing the beamforming to achieve maximum data rates for the PC operation. As discussed in Section II-B.3, a fixed beamforming technique without considering the channel of the whole network performs poorly as can be seen in the figure. Figure 5 shows that the PC scheme obtains a significantly higher sum rate than the baseline, which confirms the importance of optimizing precoding and combining in a PC system.

Fig. 6 shows the performance of the PC system with AP clustering under difference values of the reference

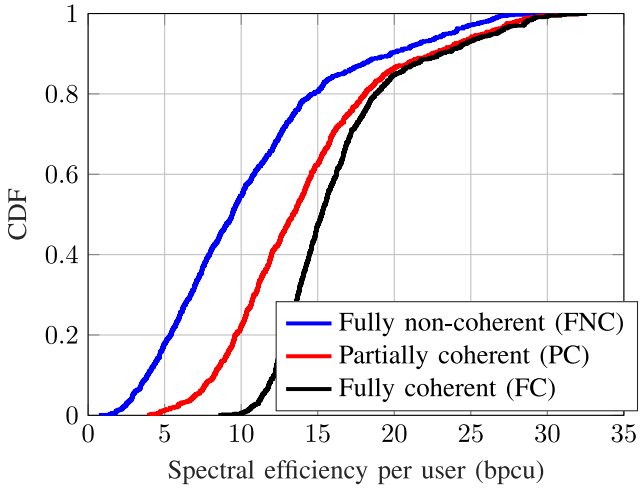


FIGURE 7. Spectral efficiency per user for different transmission schemes. Parameters for the plot: $L = 10$, $M = 5$, $K = 5$, $D = 200$ m, and $d_{kc} = 2\ \forall k, c$.

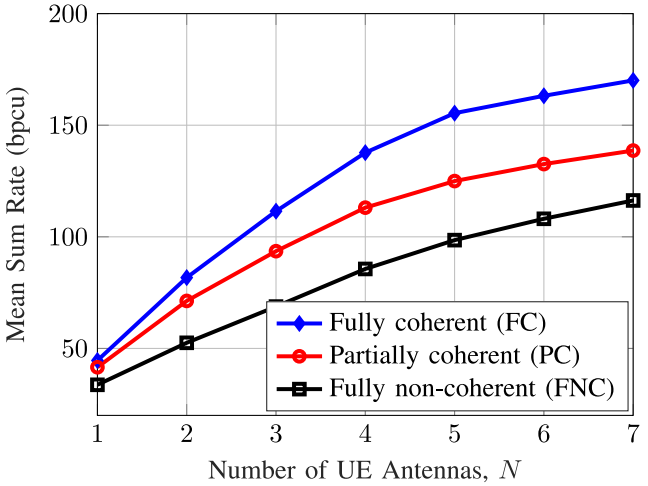


FIGURE 8. Comparison of FNC, PC and FC scheme with multiple antennas at the UEs. Parameters for the plot: $L = 10$, $M = 5$, $K = 5$, $D = 200$ m, and $d_{kc} = \min(M, N)\ \forall k, c$.

distance D. Note that the more D increases, the more PC system becomes a FC system. The larger the reference distance is, the less inter-cluster interference or the price of overcoming phase misalignment (by AP clustering) is. This is confirmed in Fig. 6, where the sum rate saturates after a certain value of the reference distance D. For a 500×500 m^2 system, it can be seen that forming phase-aligned clusters with approximately up to the reference distance of only 300 m is needed to achieve a sum rate that is significantly close to that of the system with the FC operation. Thus, from a system design perspective, it is possible to achieve a near-maximum sum rate of the network without the phase alignment of all APs as in the FC operation. A PC system with an appropriate reference distance as well as optimized precoding/combing and data stream allocation would suffice to achieve approximately the sum rate of the FC system.

The spectral efficiency per user for different schemes in the paper is shown in Fig. 7. The PC scheme performs significantly better than the FNC scheme and is close to the FC system. The PC system achieves better beamforming gain from the phase-aligned clusters and hence higher rates than the FNC system. Note that the optimization objective considered in this paper is the sum rate of the network. Therefore, the difference in the rates of UEs over the network can be large.

The sum rate performance of cell-free massive MIMO with multiple-antenna users is provided in Fig. 8. The sum rate increases significantly with the number of UE antennas as the number of data streams allocated to each user increases, until $d_{kc} = \min(M, N)\ \forall k, c$, beyond which the rate of increase is less. This is reasonable because higher beamforming gain is achieved with a larger number of data streams allocated to each user. However, a very large number of data streams allocated to each user leads to high inter-stream interference.

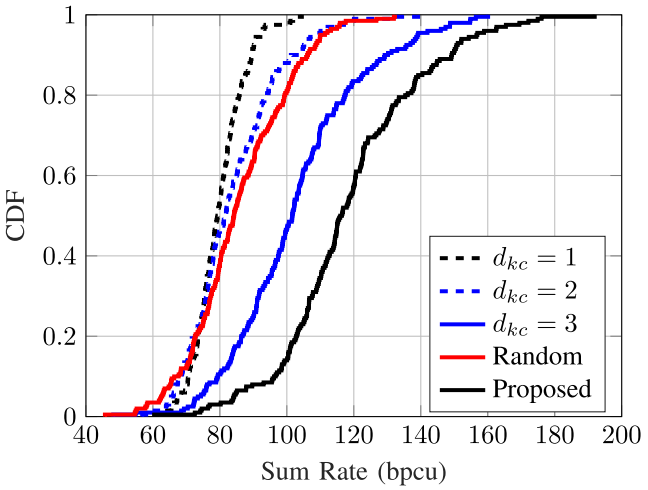


FIGURE 9. Performance of data stream allocation algorithm compared with other methods in a PC scenario. Parameters for the plot: $L = 10$, $M = 3$, $K = 10$, $N = 4$, and $D = 200$ m.

To show the effectiveness of our proposed data stream algorithm in PC scenario, we compare our algorithm with two additional heuristic baselines: (i) even data stream allocation where every user-cluster pair is allocated the same number of data streams, i.e., $d_{kc} = d,\ \forall k, c$; (ii) random data stream allocation where the numbers of data streams allocated for each user-cluster pair are selected randomly, i.e., $d_{kc} = \mathcal{U}(1, \min(M, N))$. The performance is plotted in Fig. 9. From the figure, it can be seen that our proposed algorithm performs significantly better than the other approaches.

The performance of greedy data stream allocation for all schemes is presented in Fig. 10. From the figure, it can be seen that the improvement in the performance of PC with data stream allocation as compared to fixed data allocation

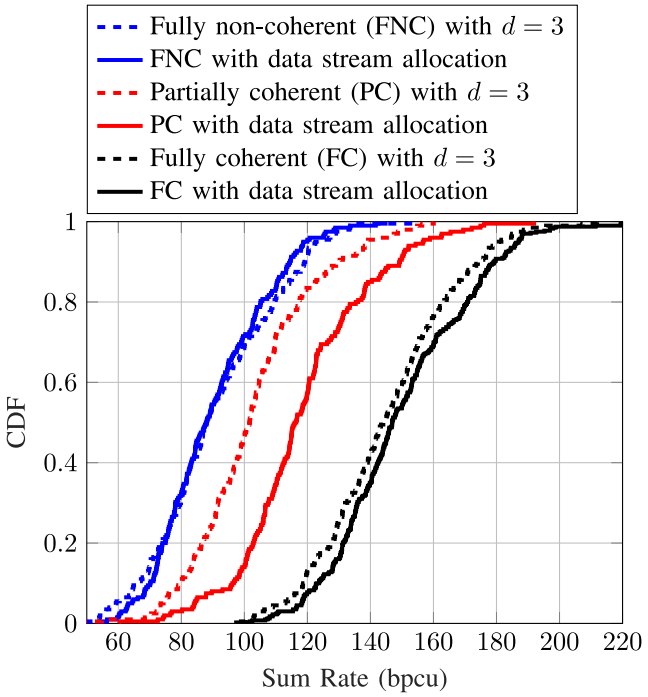


FIGURE 10. Performance comparison of FNC, PC, and FC schemes with data allocation. Parameters for the plot: $L = 10$, $M = 3$, $K = 10$, $N = 4$, and $D = 200$ m. The dashed curve represents the performance with data allocation and the solid curve represents the performance with fixed data allocation.

is much higher than for FNC and FC. This highlights the importance of data stream allocation in the PC operation.

IX. CONCLUSION

In this paper, we have proposed a novel framework for the PC operation of cell-free massive MIMO networks with multi-antenna UEs, when there is phase misalignment between the APs. In the proposed framework, the subsets of APs form phase-aligned clusters that work together in a non-coherent manner. We have proposed an algorithm for clustering APs based on the reference distance of phase alignment. We have also developed algorithms to optimize the precoding and combining matrices to maximize the network sum rate, for a given allocation of data streams. We also proposed a greedy data-stream allocation algorithm, which improves the sum rate of the PC operation. Numerical results showed that the PC scheme can obtain significantly higher rates than that of FNC operation. The proposed PC can offer a sum rate that is significantly close to that of the FC operation, without the need for network-wide phase alignment of the APs. This highlights that the PC operation is a promising solution to enable the practical deployment of cell-free massive MIMO technology in future communication systems.

APPENDIX

Let (17) be written as

$$\overline{\mathbf{Q}}_{kc} = \overline{\mathbf{V}}_{kc}^{\mathrm{H}} \mathbf{A}_{kc} \overline{\mathbf{V}}_{kc}, \tag{56}$$

where $\mathbf{A}_{kc}$ is the second-order interference term given by

$$\mathbf{A}_{kc} = \mathbf{I}_N + \rho \sum_{c'=1, c' \neq c}^{L_c} \overline{\mathbf{G}}_{kc'} \overline{\mathbf{W}}_{kc'} \overline{\mathbf{W}}_{kc'}^{\mathrm{H}} \overline{\mathbf{G}}_{kc'}^{\mathrm{H}} + \rho \sum_{c'=1}^{L_c} \overline{\mathbf{G}}_{kc'} \left(\sum_{k'=1, k' \neq k}^{K} \overline{\mathbf{W}}_{k'c'} \overline{\mathbf{W}}_{k'c'}^{\mathrm{H}} \right) \overline{\mathbf{G}}_{kc'}^{\mathrm{H}}. \tag{57}$$

It can be shown that the optimal combining matrix that maximizes the instantaneous rate R_{kc} is $\widetilde{\mathbf{V}}_{kc} = \sqrt{\rho}\, \mathbf{A}_{kc}^{-1} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc}$ [47, Appendix C.3.2]. Each column of $\widetilde{\mathbf{V}}_{kc}$ is the optimal combining vector for one data stream of UE k from cluster c. The structure of $\widetilde{\mathbf{V}}_{kc}$ can be intuitively explained as follows. To detect the data symbol $\mathbf{q}_{kc}$ from (13), we first whiten the inter-UE interference plus noise in the received signal $\mathbf{y}_k$ and obtain $\mathbf{A}_{kc}^{-1/2} \mathbf{y}_k$. After whitening, the desired signal is highest in the spatial direction $\mathbf{A}_{kc}^{-1/2} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc}$, while the interference plus noise is lowest in this direction. Following the maximum-ratio combining (MRC) approach, the optimal combining matrix for the whitened signal $\mathbf{A}_{kc}^{-1/2} \mathbf{y}_k$ is $\mathbf{A}_{kc}^{-1/2} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc}$. Therefore, the desired part of the signal at UE k from cluster c is

$$\left(\mathbf{A}_{kc}^{-\frac{1}{2}} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \right)^{\mathrm{H}} \mathbf{A}_{kc}^{-\frac{1}{2}} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc}\, \mathbf{q}_{kc} = \left(\mathbf{A}_{kc}^{-1} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \right)^{\mathrm{H}} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc}\, \mathbf{q}_{kc}, \tag{58}$$

which means that the optimal combining matrix is $\widetilde{\mathbf{V}}_{kc}$.

Using the matrix inversion lemma and after some matrix manipulations [47], the relationship between the optimal combining $\widetilde{\mathbf{V}}_{kc}$ and the MMSE combining matrix $\overline{\mathbf{V}}_{kc} = \overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}}$ in (25) can be found. Consider the term

$$\mathbf{A}_{kc}^{-1} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} = \left(\mathbf{A}_{kc} + \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \overline{\mathbf{W}}_{kc}^{\mathrm{H}} \overline{\mathbf{G}}_{kc}^{\mathrm{H}} \right)^{-1} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \times \left(\mathbf{I}_{d_{kc}} + \overline{\mathbf{W}}_{kc}^{\mathrm{H}} \overline{\mathbf{G}}_{kc}^{\mathrm{H}} \mathbf{A}_{kc}^{-1} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \right)^{-1}. \tag{59}$$

Letting $\mathbf{B}_{kc} = (\mathbf{I}_{d_{kc}} + \overline{\mathbf{W}}_{kc}^{\mathrm{H}} \overline{\mathbf{G}}_{kc}^{\mathrm{H}} \mathbf{A}_{kc}^{-1} \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc})^{-1}$. It is true that $\widetilde{\mathbf{V}}_{kc} = \overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}} \mathbf{B}_{kc}$. Also, the maximum instantaneous rate R_{kc} obtained using $\widetilde{\mathbf{V}}_{kc}$ is

$$\begin{aligned} R_{kc}(\widetilde{\mathbf{V}}_{kc}) &= \log_2 \frac{\left| \widetilde{\mathbf{V}}_{kc}^{\mathrm{H}} \left(\mathbf{A}_{kc} + \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \overline{\mathbf{W}}_{kc}^{\mathrm{H}} \overline{\mathbf{G}}_{kc}^{\mathrm{H}} \right) \widetilde{\mathbf{V}}_{kc} \right|}{\left| \widetilde{\mathbf{V}}_{kc}^{\mathrm{H}} \mathbf{A}_{kc} \widetilde{\mathbf{V}}_{kc} \right|} \\ &= \log_2 \frac{\left| \mathbf{B}_{kc}^{\mathrm{H}} \overline{\mathbf{V}}_{kc}^{\mathrm{H}} \left(\mathbf{A}_{kc} + \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \overline{\mathbf{W}}_{kc}^{\mathrm{H}} \overline{\mathbf{G}}_{kc}^{\mathrm{H}} \right) \overline{\mathbf{V}}_{kc} \mathbf{B}_{kc} \right|}{\left| \mathbf{B}_{kc}^{\mathrm{H}} \overline{\mathbf{V}}_{kc}^{\mathrm{H}} \mathbf{A}_{kc} \overline{\mathbf{V}}_{kc} \mathbf{B}_{kc} \right|} \\ &= \log_2 \frac{\left| \overline{\mathbf{V}}_{kc}^{\mathrm{H}} \left(\mathbf{A}_{kc} + \overline{\mathbf{G}}_{kc} \overline{\mathbf{W}}_{kc} \overline{\mathbf{W}}_{kc}^{\mathrm{H}} \overline{\mathbf{G}}_{kc}^{\mathrm{H}} \right) \overline{\mathbf{V}}_{kc} \right|}{\left| \overline{\mathbf{V}}_{kc}^{\mathrm{H}} \mathbf{A}_{kc} \overline{\mathbf{V}}_{kc} \right|} \\ &= R_{kc}\left(\overline{\mathbf{V}}_{kc} \right) = R_{kc}\left(\overline{\mathbf{V}}_{kc}^{\mathrm{MMSE}} \right). \end{aligned} \tag{60}$$

The third equality in (60) holds due to the fact that $|\mathbf{X}\mathbf{Y}| = |\mathbf{X}||\mathbf{Y}|$ if $\mathbf{X}$ and $\mathbf{Y}$ are square matrices. Hence, from (60),

$\overline{\mathbf{V}}_{kc}^{\text{MMSE}}$ can achieve the maximum instantaneous rate R_{kc}. Substituting (25) in (15) gives (26), which completes the proof.

ACKNOWLEDGMENT

The computations/data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Linköpings Universitet partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

REFERENCES

[1] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and H. Yang, "Cell-free massive MIMO systems," in *Proc. 49th Asilomar Conf. Signals, Syst. Comput.*, 2015, pp. 695–699.

[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] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, p. 197, 2019.

[4] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Found. Trends® Signal Process.*, vol. 14, nos. 3–4, pp. 162–472, 2021.

[5] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. 16th IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2015, pp. 201–205.

[6] S. Elhoushy, M. Ibrahim, and W. Hamouda, "Cell-free massive MIMO: A survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 492–523, 1st Quart., 2022.

[7] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[8] 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. 99878–99888, 2019.

[9] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[10] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[11] J. Vieira and E. G. Larsson, "Reciprocity calibration of distributed massive MIMO access points for coherent operation," in *Proc. IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2021, pp. 783–787.

[12] U. K. Ganesan, R. Sarvendranath, and E. G. Larsson, "BeamSync: Over-the-air Synchronization for distributed massive MIMO systems," *IEEE Trans. Wireless Commun.*, early access, Nov. 30, 2023, doi: 10.1109/TWC.2023.3335089.

[13] E. G. Larsson, "Massive synchrony in distributed antenna systems," *IEEE Trans. Signal Process.*, vol. 72, pp. 855–866, 2024.

[14] R. Nissel, "Correctly modeling TX and RX chain in (distributed) massive MIMO—New fundamental insights on coherency," *IEEE Commun. Lett.*, vol. 26, no. 10, pp. 2465–2469, Oct. 2022.

[15] S. Chen, J. Zhang, J. Zhang, E. Björnson, and B. Ai, "A survey on user-centric cell-free massive MIMO systems," *Digit. Commun. Netw.*, vol. 8, no. 5, pp. 695–719, 2022.

[16] J. Zheng, J. Zhang, J. Cheng, V. C. M. Leung, D. W. K. Ng, and B. Ai, "Asynchronous cell-free massive MIMO with rate-splitting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1366–1382, May 2023.

[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 Trans. Wireless Commun.*, vol. 19, no. 9, pp. 5891–5907, Sep. 2020.

[18] S. Bai, Z. Gao, and X. Liao, "Distributed noncoherent joint transmission based on multi-agent reinforcement learning for dense small cell networks," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 851–863, Feb. 2023.

[19] 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. 2021.

[20] Ö. Özdogan, E. Björnson, and J. Zhang, "Performance of cell-free massive MIMO with Rician fading and phase shifts," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5299–5315, Nov. 2019.

[21] R. P. Antonioli, I. M. Braga, G. Fodor, Y. C. Silva, and W. C. Freitas, "Mixed coherent and non-coherent transmission for multi-CPU cell-free systems," in *Proc. IEEE Int. Conf. Commun.*, 2023, pp. 1068–1073.

[22] A. Papazafeiropoulos, P. Kourtessis, M. Di Renzo, S. Chatzinotas, and J. M. Senior, "Performance analysis of cell-free massive MIMO systems: A stochastic geometry approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3523–3537, Apr. 2020.

[23] H. Masoumi and M. J. Emadi, "Performance analysis of cell-free massive MIMO system with limited fronthaul capacity and hardware impairments," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1038–1053, Feb. 2020.

[24] S. Buzzi and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 706–709, Dec. 2017.

[25] M. Bashar, K. Cumanan, A. G. Burr, M. Debbah, and H. Q. Ngo, "On the uplink max–min SINR of cell-free massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2021–2036, Apr. 2019.

[26] A. Ö. Kaya and H. Viswanathan, "Dense distributed massive MIMO: Precoding and power control," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 756–763.

[27] E. Björnson and L. Sanguinetti, "A new look at cell-free massive MIMO: Making it practical with dynamic cooperation," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, 2019, pp. 1–6.

[28] Y. Al-Eryani, M. Akrout, and E. Hossain, "Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1028–1042, Apr. 2021.

[29] Q. N. Le, V.-D. Nguyen, O. A. Dobre, N.-P. Nguyen, R. Zhao, and S. Chatzinotas, "Learning-assisted user clustering in cell-free massive MIMO-NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12872–12887, Dec. 2021.

[30] B. Banerjee, R. C. Elliott, W. A. Krzymień, and M. Medra, "Access point clustering in cell-free massive MIMO using conventional and federated multi-agent reinforcement learning," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, pp. 107–123, 2023.

[31] J. A. Sutton, H. Q. Ngo, and M. Matthaiou, "Hardening the channels by precoder design in massive MIMO with multiple-antenna users," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4541–4556, May 2021.

[32] M. Kazemi, Ç. Göken, and T. M. Duman, "Robust joint precoding/combining design for multiuser MIMO systems with calibration errors," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5157–5169, Aug. 2023.

[33] E. Björnson, M. Kountouris, M. Bengtsson, and B. Ottersten, "Receive combining vs. multi-stream multiplexing in downlink systems with multi-antenna users," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3431–3446, Jul. 2013.

[34] T. C. Mai, H. Q. Ngo, and T. Q. Duong, "Downlink spectral efficiency of cell-free massive MIMO systems with multi-antenna users," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4803–4815, Aug. 2020.

[35] Z. Wang, J. Zhang, H. Q. Ngo, B. Ai, and M. Debbah, "Uplink precoding design for cell-free massive MIMO with iteratively weighted MMSE," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1646–1664, Mar. 2023.

[36] X. Li, J. Zhang, Z. Wang, B. Ai, and D. W. K. Ng, "Cell-free massive MIMO with multi-antenna users over Weichselberger Rician channels," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 12368–12373, Nov. 2022.

[37] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[38] Q. Shi, W. Xu, J. Wu, E. Song, and Y. Wang, "Secure beamforming for MIMO broadcasting with wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2841–2853, May 2015.

[39] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[40] C. Guthy, W. Utschick, R. Hunger, and M. Joham, "Efficient weighted sum rate maximization with linear precoding," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2284–2297, Apr. 2010.
[41] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
[42] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.
[43] M. Grant and S. Boyd. "CVX: MATLAB software for disciplined convex programming, version 2.2." Jan. 2020. [Online]. Available: http://cvxr.com/cvx
[44] A. Tolli and M. Juntti, "Scheduling for multiuser MIMO downlink with linear processing," in *Proc. 16th Int. Symp. Pers., Indoor Mobile Radio Commun.*, 2005, pp. 156–160.
[45] F. Boccardi and H. Huang, "A near-optimum technique using linear precoding for the MIMO broadcast channel," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 3, 2007, pp. 17–20.
[46] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.
[47] E. Björnson, J. Hoydis, and L. Sanguinetti, *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*. Hanover, MA, USA: Now Publ., Inc., 2017.

UNNIKRISHNAN KUNNATH GANESAN (Graduate Student Member, IEEE) received the Bachelor of Technology degree in electronics and communication engineering from the University of Calicut, India, in 2011, and the Masters in Engineering degree in telecommunication engineering from the Indian Institute of Science, Bengaluru, India, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Linköping University, Sweden. From 2014 to 2017, he worked as a Modem Systems Engineer with Qualcomm India Private Ltd., Bengaluru, and from 2017 to 2019, he worked as a Senior Firmware Engineer with Intel. His primary research interests includes MIMO wireless communications, signal processing, and information theory.

TUNG THANH VU (Member, IEEE) received the Ph.D. degree in wireless communications from The University of Newcastle, Australia, in 2021.

He is currently a Research Fellow with the School of Engineering, Macquarie University, Australia. His research interests include optimization, communication theories, and machine learning applications for 5G-and-beyond wireless networks, especially with massive MIMO, cell-free massive MIMO, federated learning, full-duplex communications, physical-layer security, and low-Earth orbit satellite communications.

Dr. Vu received the Best Poster Award at the AMSI Optimise Conference in 2018. He is serving as an Editor of *Physical Communication* (Elsevier). He has also served as a member of the technical program committee and the symposium/session chair at several IEEE international conferences, such as GLOBECOM, ICCE, and ATC. He was an IEEE WIRELESS COMMUNICATIONS LETTERS Exemplary Reviewer from 2020 to 2021 and an IEEE TRANSACTIONS ON COMMUNICATIONS Exemplary Reviewer in 2021.

ERIK G. LARSSON (Fellow, IEEE) received the Ph.D. degree from Uppsala University, Uppsala, Sweden, in 2002. received the Ph.D. degree from Uppsala University, Uppsala, Sweden, in 2002. He is currently Professor of Communication Systems at Linköping University (LiU) in Linköping, Sweden. He was with the KTH Royal Institute of Technology in Stockholm, Sweden, the George Washington University, USA, the University of Florida, USA, and Ericsson Research, Sweden. His main professional interests are within the areas of wireless communications and signal processing. He co-authored *Space-Time Block Coding for Wireless Communications* (Cambridge University Press, 2003) and *Fundamentals of Massive MIMO* (Cambridge University Press, 2016).

He served as chair of the IEEE Signal Processing Society SPCOM technical committee (2015–2016), chair of the *IEEE Wireless Communications Letters* steering committee (2014–2015), member of the *IEEE Transactions on Wireless Communications* steering committee (2019-2022), General and Technical Chair of the Asilomar SSC conference (2015, 2012), technical co-chair of the IEEE Communication Theory Workshop (2019), and member of the IEEE Signal Processing Society Awards Board (2017–2019). He was Associate Editor for, among others, the *IEEE Transactions on Communications* (2010-2014), the *IEEE Transactions on Signal Processing* (2006-2010), and the *IEEE Signal Processing Magazine* (2018-2022).

Prof. Larsson received the IEEE Signal Processing Magazine Best Column Award twice, in 2012 and 2014, the IEEE ComSoc Stephen O. Rice Prize in Communications Theory in 2015, the IEEE ComSoc Leonard G. Abraham Prize in 2017, the IEEE ComSoc Best Tutorial Paper Award in 2018, the IEEE ComSoc Fred W. Ellersick Prize in 2019, and the IEEE SPS Donald G. Fink Overview Paper Award in 2023. He is a member of the Swedish Royal Academy of Sciences (KVA), and Highly Cited according to ISI Web of Science.