IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 September 2024, accepted 9 October 2024, date of publication 14 October 2024, date of current version 22 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3479409

RESEARCH ARTICLE

Performance and Architectural Tradeoffs in Scalable Cell-Free Massive MIMO

MUTEEN MUNAWAR, (Student Member, IEEE), MAMOUN GUENACH, (Senior Member, IEEE), AND INGRID MOERMAN, (Senior Member, IEEE)
Interuniversity Microelectronics Centre (IMEC), 3001 Leuven, Belgium
IDLab, Ghent University, 9000 Ghent, Belgium

Corresponding author: Muteen Munawar (muteen.munawar@imec.be)

This work was supported by the European Community's Research Foundation Flanders (FWO) under Grant A2582000101.

ABSTRACT The massive number of APs is often perceived as a complexity bottleneck for the scalable deployment of Cell-free (CF) systems. In this context, we propose various system-level results and valuable insights on some of the multidimensional design challenges: the energy efficiency (EE) gap between cellular and the worst-case scenario of scalable CF systems, the interplay between different split processing options, the fronthauling bandwidth, and the offered low-resolution hardware implementations. We discuss the need for physical layer optimizations to trade off performance versus complexity. As a first result, we reveal significant fronthaul bandwidth savings through joint power control and access point scheduling, and proper dimensioning of resolutions for the converters and fronthaul data. In the context of the EE gap analysis, we first provide a novel generalized system model framework that depicts the possibility of all levels of processing with a single system model and allows multiple transmit antennas at each AP and user in the pool of M APs and K users and multistream transmission per user. We formulate a novel optimization framework to maximize the EE where the non-convex fractions corresponding to user performance are coupled with the log-sum function due to the necessity of selecting the optimal number of data streams for each user. To solve this problem, we propose a solution to determine the optimal number of data streams, power allocations, and transmit/receive digital filters. Based on these solutions, we introduce a novel four-step alternating optimization algorithm. Regarding the EE gap analysis, in the worst-case scenario of scalable CF networks, which is of practical interest, CF remains roughly twice as energy-efficient as cellular networks. To facilitate future comparative studies, we also provide a detailed complexity analysis.

INDEX TERMS Scalable cell-free systems, multistream transmission, split processing, fronthaul bandwidth, alternating optimization, energy efficiency, complexity analysis.

I. INTRODUCTION

The ever-increasing demand for high data rates and reliable wireless communication networks is shaping radio access architecture. Future wireless communication networks must support a wide range of heterogeneous applications, hence increasing the pressure on radio access requirements. 6G requirements will be even more challenging and diverse than those in 5G, pushing the extremes in high throughput, ultra-low latency and jitter, coverage, and massive machine-type communications. 6G will also target a wider landscape of challenging applications and verticals [1], requiring careful design choices for network building blocks. To meet these stringent requirements of future-generation communication systems, the concept of radio access has evolved from the conventional cellular network with a cell-centric design, where each user equipment (UE) is connected to one access point (AP), to a user-centric approach referred to as cell-free (CF) massive MIMO (mMIMO) (Fig. 1). In CF, each UE can, in principle, communicate with all APs, which are connected to one or more central processing units (CPUs) for further coordination, allowing

The associate editor coordinating the review of this manuscript and approving it for publication was Adamu Murtala Zungeru.

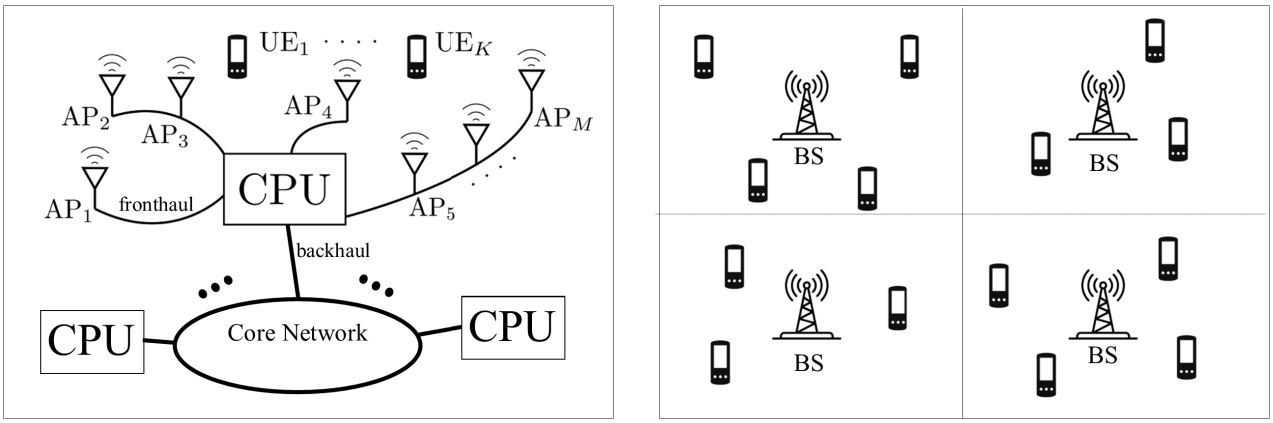

FIGURE 1. General CF massive MIMO architecture (left) and cellular massive MIMO architecture (right).

macro diversity gains and advanced interference management processing [2], [3].

However, several design bottlenecks hinder the deployment of a scalable and cost-effective CF mMIMO architecture, as depicted in Fig. 1. An underlying assumption of this architecture is that the transceiver processing chain will be shared between the CPU and the APs, commonly referred to as split processing [4], [5], [6]. A key design bottleneck is thus the interplay of the split processing between the CPU and the APs and the limited bandwidth of the fronthaul (FH) transporting the traffic between the CPU and its associated APs. Placing the split closer to the radio frequency side, i.e., a more centralized processing, will increase the bandwidth on the FH but will ease coordination between APs. If more of the processing steps are performed at the AP, coordination capabilities will be limited, hence the promised gain of the distributed radio access will be diminished. Several earlier contributions consider high functional splits, including distributed beamforming (e.g., [7], [8]), which, in collocated mMIMO, reduces the required FH bandwidth. However, this solution does not hold in the distributed scenario as will be shown in Section III. We also want the overall system design to minimize the total dissipated power through the use of simple signal processing and ensure higher energy efficiency (EE). Other nontrivial challenges include powering and tight synchronization between the large-scale distributed APs.

Improving the EE of CF mMIMO has spurred quite some research in the scientific community by modeling the problem as a resource allocation problem. For instance, [7] proposed joint AP scheduling to maximize the downlink (DL) EE wherein a simplified power model for some of the hardware (HW) components (namely the AP amplifier and the FH) is developed for this purpose. In [9] and [10], the authors proposed to jointly optimize the DL power control and the ON/OFF mode of the APs to minimize the total transceiver power consumption while achieving a target spectral efficiency. However, EE should not be restricted to some building blocks of the transceivers but should also cover the energy consumption of the whole end-to-end radio infrastructure, including supporting functions such as the cooling infrastructure, and any centralized computing resources in, e.g., cloud-edge data centers. Nevertheless, a theoretical analysis of some isolated system designs in the end-to-end link, such as the optimization of beamforming, power allocation, and EE to name a few, is a necessary first step. However, the end-to-end system co-optimization to efficiently design and interconnect the massive distributed APs is by nature a complex and heterogeneous multidimensional design space. It requires multidisciplinary research crossing different disciplines, beyond the current state of the art, in order to build up scalable CF mMIMO solutions.

Current literature on the study of EE maximization and its comparison with cellular systems is limited to single-stream transmission power per user [8], [11], [12], [13], [14], [15], [16]. However, it should be noted that modern communication devices are equipped with multiple antennas at both the AP and user sides. Hence, it is necessary to develop algorithms that can provide an optimal number of data streams for each user based on an objective function. Additionally, multiple receive antennas at each user device can help achieve better interference management, and thus improve EE compared to single receive antenna or single stream transmission per user [17]. Considering an EE gap analysis between such generalized CF and cellular systems can provide better insights into how much we can expect from CF over cellular in a worst-case scenario. Here, the worst case of CF means a scenario where scalability is implemented in the worst possible way, i.e., only one AP per user is selected using a method such as large-scale fading. Note that various scalability methods, combined with such a generalized CF setup, would result in several flavors of CF with complexity, performance, and cost tradeoffs. Hence, it is interesting to observe that given multiple receive antennas per user, each with multistream transmission and better interference cancellation capabilities in both cellular and CF, CF can still provide more EE than cellular systems. In this context, and in contrast to existing literature [11], [12], [13], [14], [15], [16], we first provide a generalized system model for CF systems

where each AP and user can have an arbitrary number of transmit and receive antennas in the pool of M APs and K users, with the possibility of L_k number of data streams for the k-th user. Moreover, the provided system model depicts all levels of signal processing possibilities in the same model, such as centralized and distributed processing.

Based on such a generalized system model, we formulate an EE maximization problem where the number of data streams per user is a novel optimization variable in addition to power allocations and transmit-receive digital filters. Based on the solutions provided, we devise a novel multi-step alternating optimization (AO) framework. To facilitate future comparative studies, we also provide a detailed breakdown of the complexity analysis associated with the proposed per-user multistream transmission algorithm.

In the context of these aforementioned directions, the contribution of this paper is multidimensional. Specifically, in this paper, we

- study a novel EE maximization problem in the presence of per-user multistream transmission and propose an algorithm with detailed complexity analysis,
- provide insights into the EE gap between CF and cellular systems (both ideal and practical extreme cases of scalable CF) when multiple data streams per users are involved,
- review some key design challenges that must be addressed to develop scalable and cost-effective CF mMIMO systems,
- highlight the design trade-offs and the interplay between split processing, hence the underlying signal coordination capabilities, and FH bandwidth requirements through simple representative system-level calculations,
- present first performance results showing some offered system design relaxations of the FH bandwidth requirements through (i) joint power control and AP scheduling, and (ii) low digital-to-analog converter (DAC) resolution, and
- provide an estimation of the energy savings that can be performed based on power estimations of this system.

In Section II, we provide an EE gap analysis between CF and cellular systems. Section III discusses the interplay between split processing and the FH bandwidth requirement. In Section IV, we elaborate on the need for resource allocation, taking some of the system limitations into account. In Section V, we discuss hardware relaxation options that benefit from the massive number of deployed APs and could reduce the FH bandwidth requirements. Finally, we conclude in Section VI.

Notations: Scalars are denoted by italic letters, whereas vectors and matrices are denoted by bold-face lower- and upper-case letters, respectively. The subscripts/superscripts v and h denote vertical and horizontal polarization, respectively. For a complex-valued vector $\mathbf{v}$ of length N, $\mathbf{v}^T$ denotes the transpose, $v_{(n)}/\mathbf{v}_{(n)}$ denotes the nth element of $\mathbf{v}$, $\mathbf{v}^*$ denotes the complex conjugate of each element, $\mathbf{1}_{\mathbf{v}}$ denotes a vector of size $\mathbf{v}$ with all entries 1, $\mathbf{v}^H$ denotes the conjugate transpose, $\|\mathbf{v}\|$ denotes the Euclidean norm, diag $(\mathbf{v})$ denotes a diagonal matrix with each diagonal element being the corresponding element in $\mathbf{v}$, arg$(\mathbf{v})$ denotes a vector with each element being the phase of the corresponding element in $\mathbf{v}$, $\mathbf{v} > 0$ denotes the number of non zero elements in $\mathbf{v}$, logsum $(\mathbf{v})$ means $\sum_{k=1}^{K} \log(\mathbf{v}_k)$, and sum $(\mathbf{v})$ means $\sum_{k=1}^{K} (\mathbf{v}_k)$.

II. EE GAP BETWEEN CELLULAR AND SCALABLE CF

The general CF mMIMO architecture, as sketched in Fig. 1, consists of one or more CPUs connected to the core network through a backhaul link and a massive number of APs connected to the CPUs through one or a few wired FH links. Given a total of M APs, each equipped with $N_{t,m}$ transmit antennas, where $m = 1, \cdots, M$, and a total of K users, each equipped with $N_{r,k}$ receive antennas and L_k data streams, where $k = 1, \cdots, K$, the DL channel between AP m and user k is denoted by $\mathbf{H}_{k,m}^H \in {}^{Nr,k \times Nt,m}$ (standard block fading model), resulting in the total channel from all APs to user k as $\mathbf{H}_k^H \in {}^{N_{r,k} \times N_t}$, i.e., $\mathbf{H}_k^H = [\mathbf{H}_{k,1}, \mathbf{H}_{k,2}, \cdots, \mathbf{H}_{k,M}]^H$, and $N_t = \sum_{m=1}^{M} N_{t,m}$.

Let $s_{k,l}$ and $p_{k,l}/q_{k,l}$ denote the data symbol and DL/uplink (UL) power allocation for the l-th data stream of user k. The DL/UL signal model for user k can be developed as (1)/(2), shown at the bottom of the next page, where $\bar{\mathbf{w}}_{k,m,l} \in {}^{N_{t,m} \times 1}$ denotes the transmit digital filter (beamforming vector) at AP m for the l-th data stream of user k, $\mathbf{f}_{k,l}^H \in {}^{1 \times N_{r,k}}$ denotes the receive combining vector for the l-th data stream of user k, and $\mathbf{z}$ denotes the noise vector.

To analyze the EE gap between cellular and CF, we follow the strategy outlined below: we specify a target SE performance for each user, consistent in both cellular and CF; We derive the corresponding linear optimal transmit/receive digital filters, i.e., $\bar{\mathbf{w}}_{k,m,l}$, $\mathbf{f}_{k,l}$, optimize the number of data streams, L_k, for each user, and calculate the minimum transmit power needed to achieve those targets; finally, we calculate the EE as the ratio of achieved SEs and consumed total power.

A. PROBLEM FORMULATION

To maximize the EE, we need to achieve the pre-specified target spectral efficiency performance with the minimum possible transmit power; hence, the problem can be formulated as shown in (P1), shown at the bottom of the next page. The meaning of each expression in (P1) is as follows:

$$\left| \mathbf{f}_{k,l} \sum_{m=1}^{M} \mathbf{H}_{k,m}^H \mathbf{w}_{k,m,l} \right|^2$$

depicts the intended signal strength for the l-th data stream of user k,

$$\sum_{j \neq l}^{L_k} \left| \mathbf{f}_{k,l} \sum_{m=1}^{M} \mathbf{H}_{k,m}^H \mathbf{w}_{k,m,j} \right|^2$$

depicts the interference for the l-th data stream of the k-th user from other data streams of the same user, whereas

$$\sum_{i\neq k}^{K}\sum_{j=1}^{L_i}\left|\mathbf{f}_{k,l}\sum_{m=1}^{M}\mathbf{H}_{i,m}^{H}\mathbf{w}_{i,m,j}\right|^2$$

shows the interference for the l-th data stream of user k by the data streams of other users. The log-sum expression in (P1) shows the coupling of the SE expression for a user k with its number of data streams.

$$\hat{\mathbf{s}}_{k,\mathrm{DL}} = \left[\mathbf{f}_{k,1},\mathbf{f}_{k,2},\cdots,\mathbf{f}_{k,L_k}\right]^H \begin{bmatrix}\mathbf{H}_{k,1}\\ \mathbf{H}_{k,2}\\ \vdots\\ \mathbf{H}_{k,m}\\ \vdots\\ \mathbf{H}_{k,M}\end{bmatrix}^H \begin{bmatrix}\bar{\mathbf{w}}_{1,1,1} & \bar{\mathbf{w}}_{1,1,2} & \cdots & \bar{\mathbf{w}}_{k,1,l} & \cdots & \bar{\mathbf{w}}_{K,1,L_K}\\ \bar{\mathbf{w}}_{1,2,1} & \bar{\mathbf{w}}_{1,2,2} & \cdots & \bar{\mathbf{w}}_{k,2,l} & \cdots & \bar{\mathbf{w}}_{K,2,L_K}\\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots\\ \bar{\mathbf{w}}_{1,m,1} & \bar{\mathbf{w}}_{1,m,2} & \cdots & \bar{\mathbf{w}}_{k,m,l} & \cdots & \bar{\mathbf{w}}_{K,m,L_K}\\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots\\ \bar{\mathbf{w}}_{1,M,1} & \bar{\mathbf{w}}_{1,M,2} & \cdots & \bar{\mathbf{w}}_{k,M,l} & \cdots & \bar{\mathbf{w}}_{K,M,L_K}\end{bmatrix}$$
$$\begin{bmatrix}\sqrt{p_{1,1}} & & & & & \\ & \sqrt{p_{1,2}} & & & & \\ & & \ddots & & & \\ & & & \sqrt{p_{k,l}} & & \\ & & & & \ddots & \\ & & & & & \sqrt{p_{K,L_K}}\end{bmatrix}\begin{bmatrix}s_{1,1}\\ s_{1,2}\\ \vdots\\ s_{k,l}\\ \vdots\\ s_{K,L_K}\end{bmatrix} + \left[\mathbf{f}_{k,1},\mathbf{f}_{k,2},\cdots,\mathbf{f}_{k,L_k}\right]^H\mathbf{z}. \tag{1}$$

$$\hat{\mathbf{s}}_{k,\mathrm{UL}} = \begin{bmatrix}\bar{\mathbf{w}}_{k,1,1} & \bar{\mathbf{w}}_{k,1,2} & \cdots & \bar{\mathbf{w}}_{k,1,l} & \cdots & \bar{\mathbf{w}}_{k,1,L_k}\\ \bar{\mathbf{w}}_{k,2,1} & \bar{\mathbf{w}}_{k,2,2} & \cdots & \bar{\mathbf{w}}_{k,2,l} & \cdots & \bar{\mathbf{w}}_{k,2,L_k}\\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots\\ \bar{\mathbf{w}}_{k,m,1} & \bar{\mathbf{w}}_{k,m,2} & \cdots & \bar{\mathbf{w}}_{k,m,l} & \cdots & \bar{\mathbf{w}}_{k,m,L_k}\\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots\\ \bar{\mathbf{w}}_{k,M,1} & \bar{\mathbf{w}}_{k,M,2} & \cdots & \bar{\mathbf{w}}_{k,M,l} & \cdots & \bar{\mathbf{w}}_{k,M,L_k}\end{bmatrix}^H \sum_{i=1}^{K}\begin{bmatrix}\mathbf{H}_{i,1}\\ \mathbf{H}_{i,2}\\ \vdots\\ \mathbf{H}_{i,m}\\ \vdots\\ \mathbf{H}_{i,M}\end{bmatrix}\left[\mathbf{f}_{i,1},\mathbf{f}_{i,2},\cdots,\mathbf{f}_{i,L_i}\right]\begin{bmatrix}\sqrt{q_{i,1}} & & & & & \\ & \sqrt{q_{i,2}} & & & & \\ & & \ddots & & & \\ & & & \sqrt{q_{i,l}} & & \\ & & & & \ddots & \\ & & & & & \sqrt{q_{i,L_i}}\end{bmatrix}$$
$$\begin{bmatrix}s_{i,1}\\ s_{i,2}\\ \vdots\\ s_{i,l}\\ \vdots\\ s_{i,L_i}\end{bmatrix} + \begin{bmatrix}\bar{\mathbf{w}}_{k,1,1} & \bar{\mathbf{w}}_{k,1,2} & \cdots & \bar{\mathbf{w}}_{k,1,l} & \cdots & \bar{\mathbf{w}}_{k,1,L_k}\\ \bar{\mathbf{w}}_{k,2,1} & \bar{\mathbf{w}}_{k,2,2} & \cdots & \bar{\mathbf{w}}_{k,2,l} & \cdots & \bar{\mathbf{w}}_{k,2,L_k}\\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots\\ \bar{\mathbf{w}}_{k,m,1} & \bar{\mathbf{w}}_{k,m,2} & \cdots & \bar{\mathbf{w}}_{k,m,l} & \cdots & \bar{\mathbf{w}}_{k,m,L_k}\\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots\\ \bar{\mathbf{w}}_{k,M,1} & \bar{\mathbf{w}}_{k,M,2} & \cdots & \bar{\mathbf{w}}_{k,M,l} & \cdots & \bar{\mathbf{w}}_{k,M,L_k}\end{bmatrix}^H \mathbf{z}. \tag{2}$$

$$\min_{\mathbf{w}_{k,m,l},\mathbf{f}_{k,l},L_k,p_{k,m,l}} \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L_k} p_{k,m,l}$$

$$\text{s.t.} \sum_{l=1}^{L_k}\log_2\left(1+\frac{\left|\mathbf{f}_{k,l}\sum_{m=1}^{M}\mathbf{H}_{k,m}^{H}\mathbf{w}_{k,m,l}\right|^2}{\sum_{j\neq l}^{L_k}\left|\mathbf{f}_{k,l}\sum_{m=1}^{M}\mathbf{H}_{k,m}^{H}\mathbf{w}_{k,m,j}\right|^2+\sum_{i\neq k}^{K}\sum_{j=1}^{L_i}\left|\mathbf{f}_{k,l}\sum_{m=1}^{M}\mathbf{H}_{i,m}^{H}\mathbf{w}_{i,m,j}\right|^2+\sigma^2}\right)\geq\lambda_k,\forall k,$$

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L_k} p_{k,m,l}\leq P_{\max},$$

$$\left\|\mathbf{w}_{k,m,l}\right\|^2\leq p_{k,m,l},\forall k,m,l, \tag{P1}$$

More specifically, note that in (P1), the SE of user k and the objective function, i.e.,

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L_k} p_{k,m,l}$$

are coupled with three variable dimensions, i.e., AP index m, data stream index l, and user index k. Hence, to minimize the total transmit power, we need to ensure that the power allocated to L_k data streams of user k at all APs is minimal so that EE can be maximized. In other words, we also need to optimize the number of data streams L_k for all users.

As a first step, owing to the centralized processing, we can define

$$\mathbf{w}_{k,l} = \left[\, \mathbf{w}_{k,1,l}^T \; \mathbf{w}_{k,2,l}^T \cdots \mathbf{w}_{k,M,l}^T \,\right]^T$$

and

$$\mathbf{H}_k^H = \left[\, \mathbf{H}_{k,1}^T \; \mathbf{H}_{k,2}^T \cdots \mathbf{H}_{k,M}^T \,\right]^H$$

which depict the total digital filter from all APs to the l-th data stream of user k and the total channel from all APs to user k, respectively. We can reshape our problem (P1) as follows:

$$\begin{aligned}
&\min_{\bar{\mathbf{w}}_{k,l},\mathbf{f}_{k,l},L_k,p_{k,l}} \sum_{k=1}^{K}\sum_{l=1}^{L_k} p_{k,l} \\
&\text{s.t.} \sum_{l=1}^{L_k} \log_2\left(1+\frac{p_{k,l}e_{k,l}}{r_{k,l}+q_{k,l}+\sigma^2}\right) \geq \lambda_k, \quad \forall k \\
&\quad e_{k,l} = \left|\mathbf{f}_{k,l}\mathbf{H}_k^H\bar{\mathbf{w}}_{k,l}\right|^2, \\
&\quad r_{k,l} = \sum_{j\neq l}^{L_k}\left|\mathbf{f}_{k,l}\mathbf{H}_k^H\sqrt{p_{k,j}}\bar{\mathbf{w}}_{k,j}\right|^2, \\
&\quad q_{k,l} = \sum_{i\neq k}^{K}\sum_{j=1}^{L_i}\left|\mathbf{f}_{k,l}\mathbf{H}_i^H\sqrt{p_{i,j}}\bar{\mathbf{w}}_{i,j}\right|^2, \\
&\quad \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L_k} p_{k,m,l} \leq P_{\max}.
\end{aligned} \tag{P1.1}$$

Note that in (P1.1), $p_{k,m,l}$ can be easily extracted from $p_{k,l}$ by selecting the proper indexing corresponding to APs in the expression of $\sqrt{p_{k,l}}\bar{\mathbf{w}}_{k,l}$. For instance, the norm of the first $N_{t,m}$ elements of $\sqrt{p_{k,l}}\bar{\mathbf{w}}_{k,l}$ determines the power allocated to the l-th data stream of user k at AP m, i.e., $p_{k,m,l}$. Similarly, the power allocated to the l-th data stream of user k at other APs can also be retrieved by proper indexing. To solve (P1.1), in the following subsections, we propose solutions and introduce a four-step AO algorithm.

B. PROPOSED SOLUTION

Assuming an initial setup where the number of data streams for user k, denoted by L_k, and the corresponding power allocations, p_k (where $p_k = \sum_{l=1}^{L_k} p_{k,l}$), for all users are configured to $\min\left(N_t, N_{r,k}\right)$ and $\frac{P_{\max}}{L_k}$, respectively, the optimal linear MMSE filters for the transmit and receive digital filters, $\bar{\mathbf{w}}_{k,l}$ and $f_{k,l}$, can be derived through interdependent optimization:

$$\bar{\mathbf{w}}_{k,l} = \frac{\hat{e}_{\max}\left(\mathbf{S}_{k,l}^{\mathbf{f}_{k,l}}, \mathbf{T}_{k,l}^{\mathbf{f}_{k,l}}\right)}{\left\|\hat{e}_{\max}\left(\mathbf{S}_{k,l}^{\mathbf{f}_{k,l}}, \mathbf{T}_{k,l}^{\mathbf{f}_{k,l}}\right)\right\|} \tag{3}$$

$$\mathbf{f}_{k,l} = \frac{\hat{e}_{\max}\left(\mathbf{S}_{k,l}^{\bar{\mathbf{w}}_{k,l}}, \mathbf{T}_{k,l}^{\bar{\mathbf{w}}_{k,l}}\right)}{\left\|\hat{e}_{\max}\left(\mathbf{S}_{k,l}^{\bar{\mathbf{w}}_{k,l}}, \mathbf{T}_{k,l}^{\bar{\mathbf{w}}_{k,l}}\right)\right\|}, \tag{4}$$

where $\mathbf{S}_{k,l}^{\mathbf{f}_{k,l}} = \mathbf{H}_k\mathbf{f}_{k,l}\mathbf{f}_{k,l}^H\mathbf{H}_k^H$ represents the correlation matrix for the received signal strength, while $\mathbf{T}_{k,l}^{\mathbf{f}_{k,l}}$ is given by $\mathbf{T}_{k,l}^{\mathbf{f}_{k,l}} = \sum_{i\neq k}^{K}\sum_{j=1}^{L_i}\rho_{i,j}\mathbf{H}_i\mathbf{f}_{i,j}\mathbf{f}_{i,j}^H\mathbf{H}_i^H + \sum_{j\neq l}^{L_k}\rho_{k,j}\mathbf{H}_k\mathbf{f}_{k,j}\mathbf{f}_{k,j}^H\mathbf{H}_k^H + \sigma^2\mathbf{I}$ denotes the matrix of interference plus noise, assuming $\mathbf{f}_{k,l}$ is fixed. Similarly, $\mathbf{S}_{k,l}^{\mathbf{w}_{k,l}} = \mathbf{H}_k^H\mathbf{w}_{k,l}\mathbf{w}_{k,l}^H\mathbf{H}_k$ is the correlation matrix for the received signal strength when $\mathbf{w}_{k,l}$ is fixed, and $\mathbf{T}_{k,l}^{\bar{\mathbf{w}}_{k,l}} = \sum_{i\neq k}^{K}\sum_{j=1}^{L_i}\rho_{i,j}\mathbf{H}_k^H\bar{\mathbf{w}}_{i,j}\bar{\mathbf{w}}_{i,j}^H\mathbf{H}_k + \sum_{j\neq l}^{L_k}\rho_{k,j}\mathbf{H}_k^H\bar{\mathbf{w}}_{k,j}\bar{\mathbf{w}}_{k,j}^H\mathbf{H}_k + \sigma^2\mathbf{I}$ represents the interference plus noise matrix under the same condition. Further details on optimizing $\overline{w}_{k,l}$ and $\mathbf{f}_{k,l}$ are provided in Algorithm 1.

Given $\overline{w}_{k,l}$ and $\mathbf{f}_{k,l}$ for all k and l from (3) and (4), respectively, the problem (P1.1) involving L_k for all k and the corresponding p_k for all k can be formulated as

$$\begin{aligned}
&\min_{L_k,p_{k,l}} \sum_{k=1}^{K}\sum_{l=1}^{\min(N_t,N_{r,k})} p_{k,l} \\
&\text{s.t.} \sum_{l=1}^{\min(N_t,N_{r,k})} \log_2\left(1+\frac{p_{k,l}e_{k,l}}{r_{k,l}+q_{k,l}+\sigma^2}\right) \geq \lambda_k, \quad \forall k \\
&\quad e_{k,l} = \left|\mathbf{f}_{k,l}^H\mathbf{H}_k^H\bar{\mathbf{w}}_{k,l}\right|^2, \\
&\quad r_{k,l} = \sum_{j\neq l}^{\min(N_t,N_{r,k})}\left|\mathbf{f}_{k,l}^H\mathbf{H}_k^H\sqrt{p_{k,j}}\bar{\mathbf{w}}_{k,j}\right|^2, \\
&\quad q_{k,l} = \sum_{i\neq k}^{K}\sum_{j=1}^{\min(N_t,N_{r,j})}\left|\mathbf{f}_{k,l}^H\mathbf{H}_i^H\sqrt{p_{i,j}}\bar{\mathbf{w}}_{i,j}\right|^2, \\
&\quad \sum_{k=1}^{K}\sum_{l=1}^{\min(N_t,N_{r,k})} p_{k,l} \leq P_{\max}
\end{aligned} \tag{P1.2}$$

To solve (P1.2), we need to handle the fractions that are coupled with log-sum expressions. To address the log-sum expressions, we define $\min\left(N_t, N_{r,k}\right)$ optimization variables, denoted as $\lambda_{k,l}$ for all k and l, and replace λ_k with $\sum_{l=1}^{L_k}\lambda_{k,l}$ for

all k. Thus, we have

$$\begin{aligned}
&\min_{L_k, p_{k,l}, \lambda_{k,l}} \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \\
&\text{s.t.} \quad \sum_{l=1}^{\min(N_t, N_{r,k})} \log_2\left(1 + \frac{p_{k,l} e_{k,l}}{r_{k,l} + q_{k,l} + \sigma^2}\right) \\
&\qquad \geq \sum_{l=1}^{\min(N_t, N_{r,k})} \lambda_{k,l} = \lambda_k, \quad \forall k, \\
&\qquad \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \leq P_{\max}. \qquad \text{(P1.3)}
\end{aligned}$$

Introducing the constraint $\sum_{l=1}^{L_k} \lambda_{k,l} = \lambda_k$ for all k, we can modify (P1.3) as

$$\begin{aligned}
&\min_{L_k, p_{k,l}, \lambda_{k,l}} \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \\
&\text{s.t.} \quad \log_2\left(1 + \frac{p_{k,l} e_{k,l}}{r_{k,l} + q_{k,l} + \sigma^2}\right) \geq \lambda_{k,l}, \quad \forall k, l, \\
&\qquad \sum_{l=1}^{\min(N_t, N_{r,k})} \lambda_{k,l} = \lambda_k, \quad \forall k, \\
&\qquad \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \leq P_{\max}. \qquad \text{(P1.4)}
\end{aligned}$$

Note that there is now a one-to-one correspondence between the left and right expressions in the first constraint of (P1.4). Hence, it is straightforward to show that the convex formulation of (P1.4) can be derived as

$$\begin{aligned}
&\min_{L_k, p_{k,l}, \lambda_{k,l}} \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \\
&\text{s.t.} \quad p_{k,l} e_{k,l} - \left(2^{\lambda_{k,l}} - 1\right)\left(r_{k,l} + q_{k,l} + \sigma^2\right) \geq 0, \\
&\qquad \forall k, l, \\
&\qquad \sum_{l=1}^{\min(N_t, N_{r,k})} \lambda_{k,l} = \lambda_k, \quad \forall k, \\
&\qquad \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \leq P_{\max}. \qquad \text{(P1.5)}
\end{aligned}$$

Note that (P1.5) is a convex formulation and hence can be solved using existing general-purpose convex solvers, such as CVX [18]. More specifically, following the guidelines provided in [18], (P1.5) can be written in CVX, which return optimized values for $\lambda_{k,l}$ and $p_{k,l}$ for all k and l. Given these optimized $\lambda_{k,l}$ and $p_{k,l}$ from CVX, the optimal values for L_k and p_k for all k can be computed as $\left[p_{k,1}, p_{k,2}, \ldots, p_{k,\min(N_t, N_{r,k})}\right] > 0$ and $\sum_{l=1}^{\min(N_t, N_{r,k})} \rho_{k,l}$, respectively, for all k.

Although (P1.5) can be solved using CVX, we note that general-purpose convex solvers are usually computationally expensive. Hence, we discuss a relatively low-complexity approach to solve (P1.5) below.

To achieve this, we decompose (P1.5) further into two sub-problems, i.e., a divide-and-conquer strategy, corresponding to $\lambda_{k,l}$ and $p_{k,l}$. First, we write (P1.5) for $\lambda_{k,l}$ by assuming that $p_{k,l}$ for all k and l are initialized by equal power allocations, i.e., $\frac{P_{\max}}{L_k}$, similar to what was done in the context of $\overline{w}_{k,l}$ and $\mathbf{f}_{k,l}$ for all k and l in (3) and (4), respectively. Given $p_{k,l}$ for all k and l, (P1.5) for $\lambda_{k,l}$ is reduced to a feasibility check problem:

$$\begin{aligned}
&\text{Find } \lambda_{k,l} \\
&\text{s.t.} \quad p_{k,l} e_{k,l} - \left(2^{\lambda_{k,l}} - 1\right)\left(r_{k,l} + q_{k,l} + \sigma^2\right) \geq 0, \quad \forall k, l, \\
&\qquad \sum_{l=1}^{\min(N_t, N_{r,k})} \lambda_{k,l} = \lambda_k, \quad \forall k \qquad \text{(P1.6)}
\end{aligned}$$

where $e'_{k,l} = p_{k,l} e_{k,l}$. We note that (P1.6) is convex and less complex than (P1.5) for the following reasons: 1) Fewer constraints, and 2) Fewer optimization variables, i.e., only $\lambda_{k,l}$ for all k and l. Now, before we decompose (P1.5) for $p_{k,l}$ for all k and l, we note that the solution of (P1.6) can be further accelerated by introducing an auxiliary objective function as follows:

$$\begin{aligned}
&\max_{\lambda_{k,l}} \sum_{k=1}^{K} \sum_{l=1}^{L_k} \alpha_{k,l} \\
&\text{s.t.} \quad p_{k,l} e_{k,l} - \left(2^{\lambda_{k,l}} - 1\right)\left(r_{k,l} + q_{k,l} + \sigma^2\right) \geq \alpha_{k,l}, \quad \forall k, l, \\
&\qquad \sum_{l=1}^{\min(N_t, N_{r,k})} \lambda_{k,l} = \lambda_k, \quad \forall k \qquad \text{(P1.7)}
\end{aligned}$$

where $\sum_{l=1}^{L_k} \alpha_{k,l}$ for all k and l is an introduced auxiliary variable that helps (P1.7) narrow down its feasibility region and thus accelerates the solution process. More details on such methodologies can be found in [18]. Nevertheless, given $\lambda_{k,l}$ for all k and l, we can write (P1.5) for $p_{k,l}$ as follows:

$$\begin{aligned}
&\min_{p_{k,l}} \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \\
&\text{s.t.} \quad p_{k,l} e_{k,l} - \lambda'_{k,l}\left(r_{k,l} + q_{k,l} + \sigma^2\right) \geq 0, \quad \forall k, l, \\
&\qquad \sum_{k=1}^{K} \sum_{l=1}^{\min(N_t, N_{r,k})} p_{k,l} \leq P_{\max} \qquad \text{(P1.8)}
\end{aligned}$$

Again, note that (P1.8) is convex. More specifically, the first constraint in (P1.8) sets $p_{k,l}$ for all k and l with equality

constraints, i.e.,

$$p_{k,l}e_{k,l} - \lambda'_{k,l}\left(r_{k,l} + q_{k,l} + \sigma^2\right) = 0, \quad \forall k, l,$$

resulting in the minimum power, i.e., the objective function. Additionally, constraint 1 of (P1.8) consists of $\sum_{k=1}^{K} L_k$ linear equations with $\sum_{k=1}^{K} L_k$ unknowns. Hence, it can be easily handled as a linear system of equations. To convert it into a matrix-vector formulation, i.e., a linear system of equations, we define:

1) The vector of unknowns

$$\mathbf{p} = [\rho_{1,1}, \rho_{1,2}, \cdots, \rho_{1,L_1}, \cdots, \rho_{K,\min(N_t,N_{r,k})}]^T$$

2) A diagonal matrix that represents the target weights provided by (P1.7) divided by the effective channel gains per data stream

$$\mathbf{J} = \text{diag}\left(\frac{2^{\lambda_{1,1}}-1}{\left|\mathbf{f}_{1,1}^H \mathbf{H}_1^H \overline{\mathbf{w}}_{1,1}\right|^2}, \frac{2^{\lambda_{1,2}}-1}{\left|\mathbf{f}_{1,2}^H \mathbf{H}_2^H \overline{\mathbf{w}}_{1,2}\right|^2}, \cdots, \frac{2^{\lambda_{K,L_k}}-1}{\left|\mathbf{f}_{K,L_K}^H \mathbf{H}_K^H \overline{\mathbf{w}}_{K,L_K}\right|^2}\right),$$

3) A non-diagonal matrix containing the total interference terms in rows for each data stream

$$\boldsymbol{\omega}_{(r,c)} = \begin{cases} \left|\mathbf{f}_{k,r}^H \mathbf{H}_k^H \overline{\mathbf{w}}_{k,c}\right|^2, & \text{if } r \neq c, \\ 0, & \text{if } r = c, \end{cases} \quad \forall k.$$

Based on these definitions, constraint 1 of (P1.8) can be written in matrix-vector form as follows:

$$\frac{\mathbf{p}}{\sigma^2} + \omega \mathbf{1}_{K'} = \frac{\mathbf{1}_{K'}}{\mathbf{J}}, \tag{5}$$

where $K' = \sum_{l=1}^{K} L_k$. From (5), it is not difficult to show that the optimal power allocations can be computed as follows:

$$\mathbf{p} = \sigma^2 \left(\mathbf{J}^{-1} - \boldsymbol{\omega}\right) \mathbf{1}_{K'}. \tag{6}$$

Note that (6) provides the optimal power allocation, i.e., because of the equality constraints, in a closed-form solution, and is therefore much less complex than solving (P1.5) using a general-purpose solver like CVX [18]. Given $\lambda_{k,l}$ and $p_{k,l}$ for all k and l from (P1.7) and (6), respectively, as discussed earlier, the optimal values for L_k and p_k for all k can be computed as $\left[p_{k,1}, p_{k,2}, \ldots, p_{k,\min(N_t,N_{r,k})}\right] > 0$ and $\sum_{l=1}^{\min(N_t,N_{r,k})} \rho_{k,l}$, respectively, for all k.

While solving (6) is computationally efficient, it does not take into account the second constraint of (P1.8). The second constraint of (P1.8) requires that the sum of all power allocations remains less than or equal to the total available power budget, i.e.,

$$\sum_{k=1}^{K} \sum_{l=1}^{\min(N_t,N_{r,k})} p_{k,l} \leq P_{\max}.$$

This constraint restricts us to a feasibility check problem. For instance, given a communication setup and simulation parameters, i.e., the number of transmit antennas, receive antennas, users, channel models, target SE, and maximum power budget, this constraint determines if the problem is feasible or not. To account for this constraint, we must perform a feasibility check before starting to solve (P1.8) using (6). More specifically, in this context, if $C < 1$, where

$$C = \max \min_{1\leq k\leq K,\, 1<l<L_k} \frac{\left(2^{\lambda_{k,l}}-1\right)\cdot\left|\mathbf{f}_{k,l}^H \mathbf{H}_k^H \mathbf{w}_{k,l}\right|^2}{\sum_{j\neq l}^{L_k}\left|\mathbf{f}_{k,l}^H \mathbf{H}_k^H \mathbf{w}_{k,j}\right|^2 + \sum_{i\neq k}^{K}\sum_{m=1}^{L_i}\left|\mathbf{f}_{k,l}^H \mathbf{H}_k^H \mathbf{w}_{i,m}\right|^2} \text{ s.t.}$$

$\sum_{k=1}^{K}\sum_{l=1}^{L_k} \rho_{kl} \leq P_{\max}$, the feasibility of problem (P1) may be compromised due to insufficient available power to meet the desired performance target. In such cases, additional measures may be required to address the issue. These measures could include user exclusion, adjustment of target SEs, or increasing the power limit $P_{\max}$.

Nevertheless, with a reasonable number of transmit antennas, users in the system, and sufficient power resources, such circumstances can generally be avoided [19].

A more detailed explanation of the optimization process for $\mathbf{f}_{k,l}$, $\overline{\mathbf{w}}_{k,l}$, L_k, and p_k is provided through a systematic four-step AO algorithm, as described in Algorithm 1.

C. OVERALL ALGORITHM

The proposed algorithm sequentially tackles (3), (4), (P1.5), and (6) until the convergence of the objective function defined in (P1.5) is achieved. Algorithm 1 provides a detailed description of the complete algorithm.

We note that the joint optimization in Algorithm 1 is complex, as detailed in Appendix. However, to achieve our goal, i.e., to observe the performance between extreme scenarios of scalability, we push the design from both performance and complexity points of view. Nevertheless, the complexity analysis framework provided in Appendix can be exploited by future comparative studies to find the performance and complexity trade-offs.

D. EE GAP ANALYSIS

In Fig. 2, we analyze the EE gap between cellular and CF systems for both UL and DL when multistream transmission per user is involved, in both the ideal and worst cases of scalable CF systems[2] The simulation parameters are as follows: a 500 × 500m communication area, $M = 50$, $N_{t,m} = 4$ for all m, $K = 40$, $N_{r,k} = 4$ for all k. In the case of cellular, four BSs each have 50 collocated transmit antennas 1. Unless otherwise specified, all other simulation parameters, setups, and channel models are similar to those in [11], and details are omitted here for brevity.

Specifically, in Fig. 2, we present the cumulative distribution function of the EE of randomly located users in the system. Solid curves represent DL and dotted curves are for

[2]A similar approach to that used for downlink can be applied to devise an uplink algorithm; however, details are omitted for brevity.

Algorithm 1 Algorithm for Solving (P1)

1: Set iteration $i = 0$ and initialize $L_{i,k} = \min\left(N_t, N_{r,k}\right)$, $\bar{\mathbf{w}}_i = \frac{1}{\sqrt{N_t}}\left[\mathbf{1}, \mathbf{1}, \cdots, \mathbf{1}_{1,L_1}, \cdots, \mathbf{1}_{K,L_K}\right]^T$,[1] and $\mathbf{p}_i = \frac{P_{\max}}{\sum_{k=1}^{K} \min(N_t, N_{r,k})}\left[1, 1, \cdots, 1_{1,L_1}, \cdots, 1_{K,L_K}\right]$.
2: **repeat**
3: Given $\mathbf{p}_i$, $L_{k,i}$, and $\bar{\mathbf{w}}_{k,l,i}$, calculate $\mathbf{f}_{k,l,i}, \forall k, l$ using (4).
4: Given $\mathbf{f}_{k,l,i+1}, \forall k$, $L_{k,i}, \forall k$, and $\mathbf{p}_i$, update $\bar{\mathbf{w}}_{k,l,i}$ using (3).
5: Given $\mathbf{f}_{k,l,i+1}, \bar{\mathbf{w}}_{k,l,i+1}, \forall k, l$, and $\mathbf{p}_i$, solve (P1.7) and update $L_{k,i}$ as $\left[\lambda_{k,1,\cdots,}\lambda_{k,\min(N_t,N_{r,k})}\right] > 0, \forall k$.
6: Given $L_{k,i+1}, \forall k$, $\bar{\mathbf{f}}_{k,l,i+1}$, and $\bar{\mathbf{w}}_{k,l,i+1}, \forall k, l$, update $\mathbf{p}_i$ using (6).
7: Update $i = i + 1$.
8: **until** the objective value of (P1.5) converges or the maximum number of iterations is completed.
9: Retrieve $p_{k,m,l}$, $\overline{\mathbf{w}}_{k,m,l}$, and $\mathbf{f}_{k,m,l}$ from optimized $p_{k,l}$, $\overline{\mathbf{w}}_{k,l}$, and $\mathbf{f}_{k,l}, \forall m$, with indexing corresponding to the number of antennas of each AP, i.e., $N_{t,m}$.
10: Calculate the maximized EE per user k by dividing the achieved targeted SEs of k by the minimized corresponding power allocation, i.e., $\frac{\lambda_k}{p_k} = \frac{\lambda_k}{\sum_{l=1}^{L_k} p_{k,l}} = \frac{\lambda_k}{\sum_{m=1}^{M}\sum_{l=1}^{L_k} p_{k,m,l}}, \forall k$.

UL cases. Green curves represent ideal CF, i.e., when all APs communicate with all users. Red curves denote cellular systems. Finally, the black curves represent the worst scenario of scalable CF. Here, the worst scenario of scalable CF means that each user gets to communicate with only one AP, selected as a function of large-scale fading (a worst approach to scalability). Given these curves, Fig. 2 provides interesting insights. It shows, at 90 likelihood points, ideal CF is 20 times more energy-efficient than cellular. And in the worst scenario of scalable CF (which is of practical interest), CF is still roughly 2 times more energy-efficient than cellular. The performance of other flavors of CF e.g., with better ways of scalability, more APs per user, etc., will lie between these two extreme cases of CF, providing the corresponding tradeoffs of performance and FH bandwidths.[3]

E. COMPUTATIONAL COMPLEXITY ANALYSIS

To support future comparative studies, we provide a comprehensive breakdown of the complexity analysis related to solutions and algorithms in terms of floating-point operations (FLOPs) in Appendix.

[3]Remember that with distributed processing in CF, the EE gap will be even lower than shown in Fig. 2 (centralized processing), for both ideal and scalable CF scenarios.

III. SPLIT PROCESSING AND FH ARCHITECTURE

The general CF massive MIMO architecture, as illustrated in Fig. 1, consists of one or more CPUs connected to the core network via a backhaul link and a large number of APs connected to the CPUs through one or a few wired FH links. The type of processing at the CPUs and APs depends on the desired processing split, which is further determined by factors such as the target FH latency and throughput capabilities, the desired AP hardware complexity, and system dynamics, such as UE mobility.

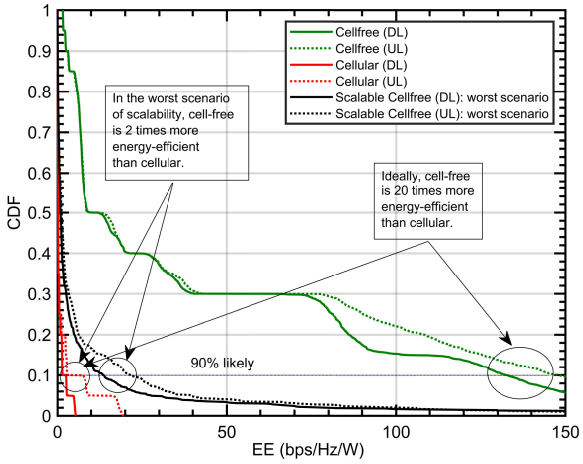

FIGURE 2. EE gap with extreme scenarios of scalability.

A. FH ARCHITECTURE

In general, every FH technology should (i) provide sufficient bandwidth, (ii) be reliable, (iii) introduce minimal additional latency crucial in time-critical indoor applications expected in e.g., industry 4.0, and (iv) allow synchronization between distributed APs. Simple calculations of the required FH bandwidth (see discussions regarding Figs. 3 and 5 later) show that the FH architecture needs to accommodate significantly more traffic than the net user data, significantly increasing the cost as this scales with the number of interconnected APs.

There are essentially two classes of FH architectures: the star-topology with point-to-point (P2P) FH links and the more general point-to-multipoint (P2MP), where each subset of APs is connected to a serial FH wire, as depicted in Fig. 1. A ring topology can be perceived as P2MP architecture where the serial FH loop is closed. Although the star topology offers the highest bandwidth, reliability, and lowest latency, it suffers from higher cost due to the number of cables that need to be installed and managed. The P2MP, with few serial FH links, offers a cost-effective solution that reduces the number of FH wires. From the deployment point of view and to reduce the chance that a single cable failure brings down the entire system, a number of serial FH wires, each accommodating a limited number of APs, should be foreseen. Furthermore, some redundancy can also be added if the likelihood of a wire breaking is high in some environments, for instance, by connecting the APs to more than one cable or

duplicating the serial FH wires. The radio-stripe architecture recently proposed in [20] is an example of such a promising architecture with a serial shared FH that offers cheap CF mMIMO deployments. In the proposed architecture, the APs are confined in cables (stripes), enabling, for instance, invisible installation in existing construction elements.

However, such a serial FH architecture is severely limited by its bandwidth constraints and different performance-complexity trade-offs still need to be investigated through, e.g., HW relaxations and the joint power control and AP scheduling.

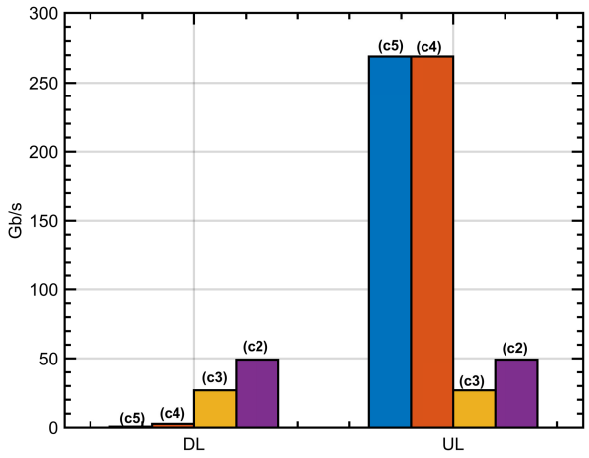

FIGURE 3. FH bandwidth: LTE use case.

TABLE 1. System parameters for the reference scenarios.

Parameters	Notation	5G	LTE	Units
Carrier bandwidth		200	20	MHz
Subcarrier spacing		60	15	kHz
FFT size	N_{FFT}	4096	2048	-
No. subcarriers	N_d	3300	1200	-
Cyclic prefix	N_{CP}	1 (245)	4.7 (144)	ms (samples)
Coding rate	R_c	3/4	3/4	-
Spectral efficiency	b	4	4	bits/s/Hz
I/Q resolution	b_X	8	8	bits
Number of APs	M	100	100	-
Number of UEs	K	10	10	-

B. SPLIT PROCESSING

In Fig. 4, we show some of the basic split processing options between the CPU(s) and the different distributed APs. As depicted in the figure, there are essentially two main classes of split processing architectures [21]:

- Analog-connected architecture, denoted as (c1), where all the digital baseband processing up to the DAC and the Analog-to-Digital Converter (ADC) included lies in the CPU.
- Digitally connected architecture with options (c2), (c3), (c4), and (c5), where the Analog Front-End (AFE) and a fraction of the digital baseband processing are distributed across the APs. The options (c2), (c3), (c4), and (c5) can be linked with the options (8), (7)-(1), (7)-2a), and (7.3) from the main functional split options of 5G [22].

To show the interplay for the different split options with the FH bandwidth requirements, we consider two reference scenarios with relatively low and high data rate requirements. The different system settings are summarized in Table 1. No system optimization is assumed at this stage, and the different (worst case) FH bandwidths of the data path for the different split processing options are summarized in Figs. 3 and 5 for LTE and 5G use cases, respectively. These do not include the additional (potentially limited) overhead of channel state information (CSI) and the digital beamformer coefficient exchange. Analytical expressions of the different required FH bandwidths are given in Fig. 4, where the bit resolution b_X denotes the resolution per in-phase/quadrature dimension of the different processing blocks in the digital baseband processing, namely X = encoder, decoder, mapper, demapper, digital beamformer, (inverse) Fourier transform, DAC, ADC. These bandwidths are proportional to the OFDM symbol rate (not shown in the figure), which is the inverse of the subcarrier spacing listed in Tab. 1.

Split option (c1), which is an analog interconnection between CPU and APs, is not considered. It requires a dedicated analog channel towards each AP. Although the bandwidth requirement per AP is restricted to the actual wireless bandwidth, which aggregates 2 GHz and 20 GHz for the LTE and 5G case respectively in both UL and DL, the required analog transceivers offer little flexibility, introduce noise and distortion, making the signal transmission sensitive to FH impairments and the cost prohibitive for large scale deployment [23].

Split option (c2) is similar to (c1), except ADC/DAC conversion is performed in the AP. As a result, a much more resilient digital FH can be employed. The aggregated bandwidth as depicted in Fig. 4 is equal in UL and DL and proportional to the OFDM symbol length ($N_{FFT} + N_{CP}$), the resolutions (b_{ADC}, b_{FFT}), and the number of APs (M).

Moving from (c2) to (c3), FH bandwidth can be easily reduced by performing up/downsampling, adding/removing CP and (I)FFT in the APs, as no additional information needs to be conveyed. This results, per transmit direction, in about 45% and 25% saving for respectively LTE and 5G, and the gain will proportionally increase with the up/down-sampling factor.

It is worth noting that the previously discussed functional splits and the underlying FH requirements are equivalent for distributed and collocated mMIMO scenarios. However, moving to (c4) and (c5), this equivalence does not hold anymore.

In collocated scenarios, the required FH bandwidth is further reduced in splits (c4) and (c5), where the beamforming is performed at the AP, and the proportionality of the number of antennas on the FH bandwidth requirement is interchanged to the number of UEs (or MIMO layers). In distributed

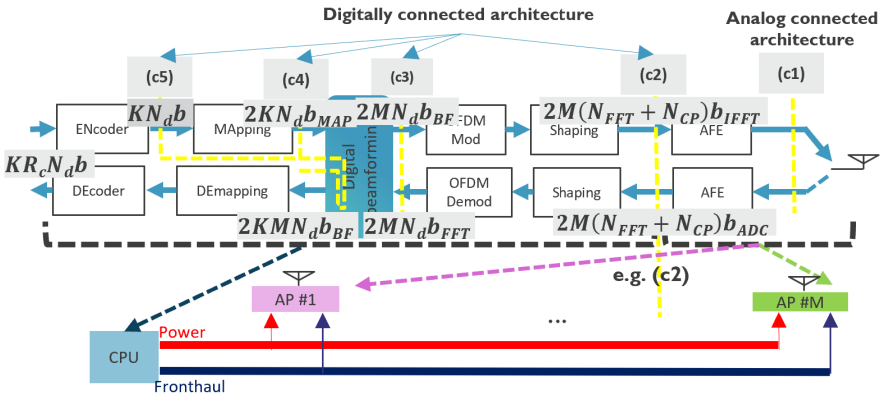

FIGURE 4. Typical functional splits in CF mMIMO.

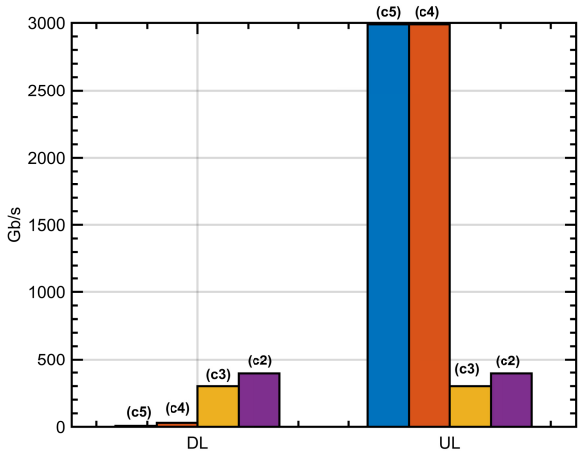

FIGURE 5. FH bandwidth: 5G use case.

scenarios, however, the (c4) and (c5) functional splits can be challenging depending on the type of beamforming and the availability of the CSI, as explained next.

The channel acquisition can, in principle, be distributed through the use of a time division duplexing (TDD) protocol that allows the CSI to be acquired at the APs during UL training segment using orthogonal pilot sequences. The DL channels can then be inferred for precoding by the virtue of the UL and DL channel reciprocity, assuming the AP front-end non-reciprocities are estimated and compensated for [24]. Different problems arise for UL versus DL, depending on the level of signal coordination.

In the DL, distributed signal coordination can be carried out at each AP by using only local CSI, without the need for any extra overhead with the CPU. Furthermore, with respect to the data transmission from the CPU to the APs, a split at (c4) or (c5) results, in the same way as in the collocated case, in a required FH bandwidth proportional to the number of UEs. As shown in Figs. 3 and 5, this provides the smallest possible load on the FH bandwidth. However, if the same split is also used in the UL (which corresponds to level 3 and level 2 combining in [11]), a severe bandwidth expansion is observed since (in the most simplistic implementation) each AP has to send the equalized data per UE to the CPU. Furthermore, if complex beamforming schemes such as zero forcing are used, then the corresponding CSI is also required at the CPU, which further increases the required FH bandwidth. To that end, it becomes more economical to forward the demodulated (i.e., unequalized) data at each AP directly to the CPU, where it is centrally processed. This corresponds to a split at (c3) or (c2) in the UL (or level 4 combining in [11]), and reduces the required FH bandwidth by an order of magnitude when compared to (c4) and (c5), since the amount of data becomes independent of the number of UEs (cf. Fig. 4).

It hence should not come as a surprise that the FH requirements are significantly increased by roughly a factor of 10 in the 5G use case compared to the LTE, corresponding to the increase in wireless bandwidth. Also, in DL split processing, (c5) and (c4) are attractive as the beamforming is distributed without any exchange of the CSI between the APs and the CPU when combined with the TDD protocol. In UL, however, the bandwidth becomes proportional to the number of scheduled APs and the number of users, hence the need for proper AP scheduling [25] (see discussion later in section IV, where it is shown that only a limited number of APs are required per UE). Split processing (c3) can reduce the FH bandwidth of the payload by a factor proportional to the number of active UEs which, when combined with the AP scheduling, can further relax the required FH bandwidth [26].

As noted earlier, the bandwidth requirements reported in Figs. 3 and 5 do not include the exchange of the CSI and/or the digital beamforming coefficients incurring additional significant FH bandwidth. For instance, if the channels are constant over a frame of 14 symbols, the exchange of the channel coefficients when the channel estimation is distributed exhibits roughly 20 Gb/s and 200 Gb/s for LTE and 5G use cases, respectively. To counteract this prohibitive additional overhead of exchanging the CSI, it is desirable to use asymmetrical functional split for UL and DL, at least

when using simple distributed signal coordination such as conjugate beamforming or local zero-forcing. In this case, the functional splits (c4) and (c5) can still be used in DL enabling distributed beamforming hence no additional CSI exchange as the channel estimation is distributed. In UL, the splits (c3) and (c4)/(c5) with centralized and distributed beamforming are enabled as the channel estimation can be also performed centrally with limited additional overhead due to the exchange of the UL training segment(s) of the TDD frame, which amounts roughly to 2 Gb/s and 22 Gb/s for the LTE and 5G use cases.

It is worth noting that process and forward techniques over a serial FH as proposed in [20] could reduce the FH bandwidth for distributed beamforming. However, latency becomes a limiting factor, especially for indoor time-sensitive applications in e.g. industry 4.0 [27].

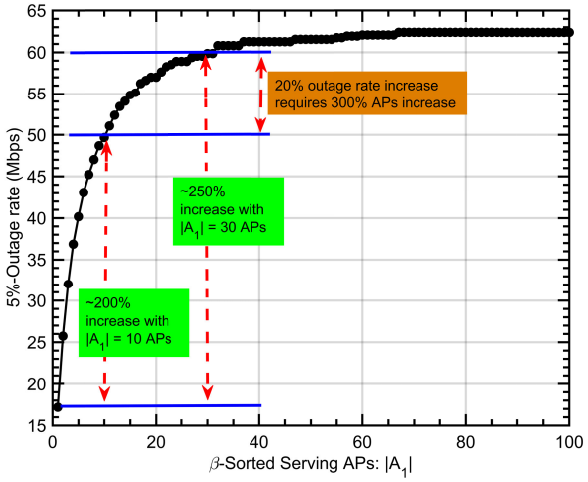

FIGURE 6. UL single user outage data rate.

IV. POWER CONTROL AND AP SCHEDULING

In serial FH architectures, AP scheduling is mandatory to take into account the FH-constrained bandwidth. In this context, in Fig. 6, we show that, in practice, a given reference UE will not need to use all APs to reach a certain performance target; hence, it may favor FH bandwidth. To illustrate this, we consider a piazza topology [11], wherein 100 APs are placed along the perimeter of a 100 m $\times$ 100 m square. The power control coefficients are optimized for the worst-case sum spectral efficiency, subject to a maximum transmit power constraint per AP. The single-user UL 5% outage data rate (this is a lower bound on the achieved data rate in 95% of the cases) in Fig. 6 shows that most of the data rate comes from 20-30% of the APs, provided that they are properly selected, for instance using large-scale fading. Hence, significant performance-complexity tradeoffs can be achieved by optimizing the UEs and AP associations in UL (and DL), allowing the system to be operated at a much lower FH bandwidth compared to the worst-case numbers reported in Figs. 3 and 5. Obviously, real-time scheduling aspects should also be considered using, for instance, simple heuristics [25] or deep neural network-based solutions [28].

V. ADDITIONAL TRADEOFFS THROUGH HW RELAXATION

The use of a large number of APs opens another avenue to relax the HW design. HW non-idealities of the APs translate into additive noise, which can be approximated as Gaussian, thanks to the law of large numbers. This fact has been exploited in, e.g., [29], to work with low (mixed) ADC resolutions. We show that, in our case, this can help relax the quantization used on the FH links, hence further decreasing the required FH bandwidth reported in Figs. 3 and 5.

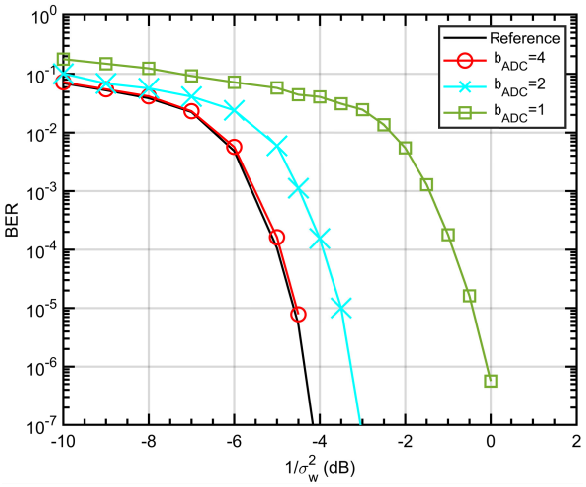

FIGURE 7. Bit error probability for different DAC resolutions.

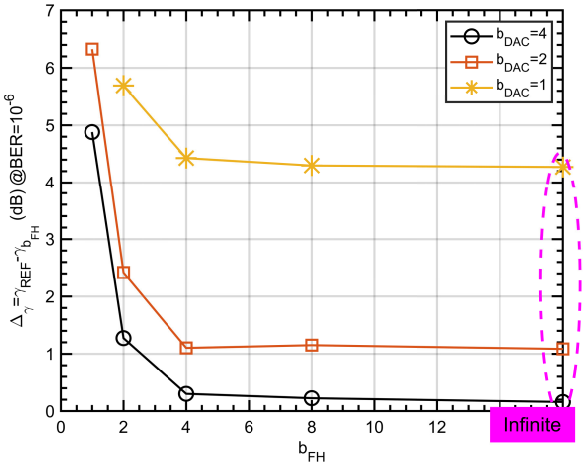

FIGURE 8. SNR loss w.r.t. b_{DAC} and b_{FH} (c3 split processing).

We consider next a DL CF mMIMO with 128 APs placed along the perimeter of a 100 m $\times$ 100 m square, and 10 UEs with OFDM parameters of the LTE use case listed in Table. 3, assuming QPSK and (c3) functional split. The power control coefficients are optimized for the worst-case sum spectral efficiency subject to a maximum transmit power constraint per AP.

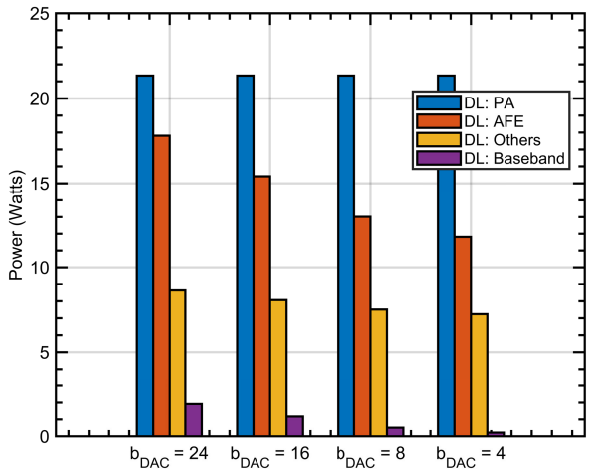

FIGURE 9. Power consumption w.r.t. DAC resolution.

To highlight the performance-complexity tradeoffs, we first show the average bit error probability (BER) in Fig. 7 for infinite resolution (see legend 'Reference') and different DAC resolutions of the APs denoted as b_{DAC}. The figure clearly shows that even with very low DAC resolution, the performance loss is limited. For instance, with 2-bits DAC resolution, the same BER diversity is obtained with roughly 1 dB signal-to-noise ratio (SNR) loss. This clearly demonstrates that the relaxation of the DAC specifications is promising to enable the deployment of cost-effective CF mMIMO.

The reduced DAC resolution is particularly interesting to further decrease the amount of the worst-case FH bandwidth depicted in Figs. 3 and 5 by working at a lower FH resolution b_{FH}. We plot in Fig. 8 the relative SNR loss Δ_γ at $BER = 10^{-6}$ with respect to the ideal SNR γ_{REF} when working with finite DAC and/or FH resolutions.

First of all, it can be noticed that the resolution of the DAC determines the best performance, and working at a relatively higher FH precision $b_{FH} > b_{DAC}$ does not necessarily improve the performance. For instance, the operating point given by $b_{DAC} = 2$ and $b_{FH} = 4$ yields $\Delta_\gamma \approx 1$ dB, which is the same as $b_{FH} = \infty$. However, operating at ($b_{DAC} = 2$, $b_{FH} = 2$), the SNR gap is about 2.5 dB, amounting to roughly 1.5 dB loss compared to ($b_{DAC} = 2, b_{FH} = 4$). This is mainly due to the fact that with the functional split (c3), the relatively low FH resolution ($b_{DAC} = 2$) is expected to have a higher impact on the signal integrity that will undergo the IFFT and pulse shaping processing before it can reach the DAC.

Furthermore, working with a smaller FH resolution than b_{DAC} incurs additional performance loss that amounts to e.g., $\approx$1.2 dB for ($b_{DAC} = 4$, $b_{FH} = 2$). Therefore, increasing the DAC resolution can help relax the FH BW requirements, which can be inferred by comparing e.g., the SNR losses obtained by the two designs, namely ($b_{DAC} = 4, b_{FH} = 2$) and ($b_{DAC} = 2$, $b_{FH} = 2$). For instance, by tolerating about 1 dB SNR loss when operating with $b_{DAC} = 4$ and $b_{FH} = 2$, the DL FH bandwidth of the (c3) functional split in Fig. 3 can be reduced by 75%. Similar conclusions can be drawn for other functional splits (figures not included). Therefore, significant complexity-performance tradeoffs can be explored in the design space of the DAC and FH resolutions.

While performance simulations can validate the operation at reduced quantization accuracy, as illustrated in Figs. 7 and 8, an estimation of the related energy savings can be performed based on power estimations of the system. Using the model of [30] enables the estimation of the system power consumption under different conditions. Fig. 9 illustrates the corresponding effect of reduced quantization resolution for a maximum transmit power of 20 dBm. While the power amplifier (PA) is not affected, reducing quantization accuracy strongly reduces the power consumption of digital computations (see legend *'BaseBand'*) as well as analog front-end (see legend *'AFE'*) power consumption via ADC and DAC components. Note that the legend *'others'* in the power model refers to power supply and scales proportionally to the sum of all the other components.

These examples have very low power consumption for the baseband part, which mainly comes from the limited bandwidth of 20 MHz of the LTE system. The picture will be different in future mm-wave systems with several GHz of bandwidth. Future studies can build on updated models such as [31] to assess the corresponding tradeoffs.

VI. CONCLUSION AND FUTURE DIRECTIONS

A. CONCLUSION

We first provide a generalized model for CF systems that depicts all levels of processing, allows for an arbitrary number of transmit and receive antennas at each AP and user, and accommodates the transmission of multiple data streams per user. Given this generalized per-user multistream transmission system model, we study a novel energy efficiency maximization problem where the number of data streams is an optimization variable, in addition to other variables such as power allocations, transmit digital filters at APs, and receive digital filters at users for all data streams. We propose a solution to optimize the number of data streams for each user. More specifically, to handle the log-sum expression coupled with the number of data streams, we introduce per-data-stream weights as optimization variables and create a one-to-one correspondence between performance and target SE. This approach allows us to apply fractional programming and convert our problem into a convex one. Based on these solutions, we provide a multistep AO algorithm framework and a detailed complexity analysis to facilitate future comparative studies. Our numerical analysis demonstrates that the worst-case scenario for CF systems is approximately two times more energy-efficient than cellular systems. This result provides an important driving factor when deciding on a theoretical architecture for CF implementation with

minimal hardware and computational costs while delivering more efficient performance than cellular systems.

Additionally, we provide system-level insights and reveal the main design and research challenges that the scientific community, practitioners, and industries should focus on to make CF mMIMO a reality. We have demonstrated that split processing options, FH topology, and hardware relaxation choices offer system designers promising optimization levers.

The massive number of APs can be seen as a complexity bottleneck. However, we have shown that this also presents significant complexity-performance trade-offs that can be explored during the design phase to optimize end-to-end radio access. Preliminary results for the considered setup reveal substantial FH bandwidth savings through joint power control and AP scheduling, as well as proper dimensioning of the distributed APs' DACs and FH resolution.

B. FUTURE DIRECTIONS

Considering power consumption at AP hardware, fronthaul, and CPU in the calculation of energy efficiency gap analysis compared to cellular systems can be easily incorporated into our proposed algorithm. However, such an analysis requires careful selection of power consumption values for both CF and cellular networks, as standards have not yet been defined. A possible approach is to consider a range of power consumption values for mobile-grade components in the CF architecture. Additionally, the formulated problem (P1) in the context of distributed processing remains an open and intriguing issue. Note that in distributed processing, each AP performs local processing for filters with minimal or no coordination with the CPU. Hence, it is interesting to analyze whether it is feasible to achieve multistream transmission per user in distributed processing or if a single stream per user is more practical, given that local processing by APs can cause interference with the same data stream for the same user.

Apart from these aspects, there are other system-level factors that require careful analysis, such as synchronization, financial feasibility of deploying massive distributed APs, total energy efficiency, and UEs compatibility. For instance, synchronization is crucial in cell-free networks, where all APs ideally need to be precisely synchronized. Practically, achieving this level of synchronization is challenging and involve sophisticated technologies and protocols, which result in performance and accuracy trade-offs.

Financially, deploying a cell-free network can be highly demanding due to the large number of APs required, the advanced equipment needed for synchronization, and the high-capacity backhaul and fronthaul links. This financial burden may limit deployment to areas with high economic viability or where there are incentives for advanced network infrastructure. Additionally, cell-free networks must be compatible with existing user equipment. This means the system needs to ensure that user devices can effectively interact with a large number of APs while supporting seamless handovers and network coordination.

These are open research directions that require careful consideration and further research to identify practical solutions.

APPENDIX
COMPUTATIONAL COMPLEXITY ANALYSIS

Here, we provide a computational complexity analysis of Algorithm 1. Specifically, we first determine the complexity associated with each proposed solution. By summing the complexities of all subproblems and multiplying by the number of iterations that Algorithm 1 performs, we can then determine the overall complexity of the algorithm.

- First, we address the computational complexity associated with the calculation of transmit and receive digital filters, i.e., $\mathbf{w}_{k,l}$ and $\mathbf{f}_{k,l}$. Specifically, in (3) and (4), we have provided the calculation of transmit and receive digital linear MMSE filters for each data stream using the strongest generalized eigenvector. Note that the matrices $\mathbf{S}$ and $\mathbf{T}$ for all k, l in (3) and (4) are rank-1 matrices. Therefore, the calculation of $\mathbf{w}_{k,l}$ and $\mathbf{f}_{k,l}$ for all k, l in (3) and (4) essentially solves the $\mathbf{Sv} = \lambda\mathbf{Tv}$ eigenvalue problem. Although we can simply exploit a MATLAB single-line command for this purpose, this problem is usually solved using specialized algorithms tailored for computing the dominant eigenvector of a matrix pair. One common approach to efficiently solving this problem is the Lanczos algorithm, an iterative method capable of computing a few of the largest or smallest eigenvalues and corresponding eigenvectors of a large sparse matrix. The computational complexity of the Lanczos algorithm is typically much lower than that of general-purpose algorithms for eigenvalue problems. For a rank-1 matrix pair $(\mathbf{S}, \mathbf{T})$, similar to our scenario, the Lanczos algorithm can be particularly efficient as it exploits the special structure of the problem. The computational complexity of the Lanczos algorithm for computing the dominant eigenvector of a matrix pair is generally on the order of $\mathcal{O}(mN)$, where m is the number of Lanczos iterations required for convergence and N is the size of the matrix. In our scenario, N is N_t and $N_{r,k}$ for $\mathbf{w}_{k,l}$ and $\mathbf{f}_{k,l}$, respectively, resulting in the computational complexity of $\mathcal{O}(mN_t)$ and $\mathcal{O}(mN_{r,k})$. In practice, the Lanczos algorithm often converges rapidly, especially for rank-1 matrix pairs, so the number of iterations required (and hence the computational complexity) can be relatively low compared to methods for more general eigenvalue problems. Other specialized algorithms, such as power iteration or inverse iteration, can also be used for computing the dominant eigenvector of a rank-1 matrix pair, and they may have similar or even lower computational complexity depending on the specific problem characteristics.
- Now, we discuss the computational complexity associated with solving (P1.5) for $\lambda_{k,l}$. To analyze the Big O

complexity of solving the (P1.5) using CVX, consider the following details:
The problem involves minimizing the sum of power

$$\sum_{k=1}^{K} \sum_{l=1}^{\min(N_t,N_{r,k})} p_{k,l}$$

subject to various constraints. Specifically, the constraints include:

- Inequality constraints of the form

$$p_{k,l}e_{k,l} - (2^{\lambda_{k,l}} - 1)(r_{k,l} + q_{k,l} + \sigma^2) \geq 0$$

  which apply to each k and l.
- Equality constraints requiring

$$\sum_{l=1}^{\min(N_t,N_{r,k})} \lambda_{k,l} = \lambda_k$$

  for each k.

In terms of the number of variables, there are $\sum_{k=1}^{K} \min(N_t, N_{r,k})$ variables for $\lambda_{k,l}$. The number of constraints includes $\sum_{k=1}^{K} \min(N_t, N_{r,k})$ inequality constraints and K equality constraints resulting in a total of approximately $2(\sum_{k=1}^{K} \min(N_t, N_{r,k})) + K$ constraints.
When solving such convex optimization problems using interior-point methods, which are commonly employed by CVX, the complexity is generally polynomial in the number of variables and constraints. The complexity of these methods is typically denoted as $O(n^{3.5})$, where n represents the number of variables and constraints [18]. Hence, given that the number of variables and constraints is $2\big(\sum_{k=1}^{K} \min(N_t, N_{r,k})\big) + K$ the complexity of solving the problem using CVX is $O\left(\left[2\big(\sum_{k=1}^{K} \min(N_t, N_{r,k})\big) + K\right]^{3.5}\right)$.

- Now, we discuss the computational complexity associated with the calculation of power allocations in (5). Note that we provide a closed-form solution to calculate the power allocations in (5). Given the closed-form solution, it is possible to calculate the exact FLOPs for this solution. First, let us examine the FLOPs of each expression $|\mathbf{f}_{k,l}^H \mathbf{H}_k^H \overline{\mathbf{w}}_{j,l}|^2$. Since the calculation of $\mathbf{J}^{-1}$ and Ω requires the calculation of these expressions, which are further used to calculate the power allocations, we can break down the computation step by step:
  1. Compute the matrix-vector multiplication $\mathbf{H}\overline{\mathbf{w}}$. This operation requires $N_{r,k} \times N_t$ complex multiplications and $N_{r,k} \times N_t - 1$ complex additions. 2. Compute the inner product $\mathbf{f}\,(\mathbf{H}\overline{\mathbf{w}})$. This operation requires $N_{r,k}$ complex multiplications and $N_{r,k} - 1$ complex additions. 3. Take the absolute value of the result, $|\mathbf{f}\mathbf{H}\overline{\mathbf{w}}|$. This requires one complex multiplication and one square root operation (or the equivalent of a complex absolute value operation). 4. Finally, squaring the absolute value gives $|\mathbf{f}\mathbf{H}\overline{\mathbf{w}}|^2$, which does not involve any additional FLOPs as it is just a multiplication operation.
  So, the total number of FLOPs is $2N_{r,k}N_t + 2N_{r,k} + 1$. This expression accounts for the FLOPs required for the matrix-vector multiplication, the inner product, and the absolute value operation.
  Now, there are $\sum_{k=1}^{K} L_k$ number of $|\mathbf{f}_{k,l}^H \mathbf{H}_k^H \overline{\mathbf{w}}_{j,l}|^2$ expressions in the calculation for J^{-1}, i.e., the diagonal matrix, and $\left(\sum_{k=1}^{K} L_k\right)^2 - \sum_{k=1}^{K} L_k$ expressions in Ω, i.e., the non-diagonal matrix. Hence, the total FLOPs for (5), i.e., power allocations, are $\left(2N_{r,k}N_t + 2N_{r,k} + 1\right) \times \left(\sum_{k=1}^{K} L_k\right)^2 + \left(\sum_{k=1}^{K} L_k\right)^2 + \sum_{k=1}^{K} L_k\,(N_t - 1) \sum_{k=1}^{K} L_k N_t$.
- Considering the complexities of each component and multiplying by the number of iterations, we obtain the total complexity of Algorithm 1:

$$i\Bigg[\mathcal{O}(mN_t) + \mathcal{O}(mN_r) + \left(2N_{r,k}N_t + 2N_{r,k} + 1\right) \times \left(\sum_{k=1}^{K} L_k\right)^2 + \left(\sum_{k=1}^{K} L_k\right)^2 + \sum_{k=1}^{K} L_k\,(N_t - 1) \sum_{k=1}^{K} L_k N_t + O\left(\left[2\left(\sum_{k=1}^{K} \min(N_t, N_{r,k})\right) + K\right]^{3.5}\right)\Bigg], \quad (7)$$

where i denotes the number of iterations in Algorithm 1.

REFERENCES

[1] N. Rajatheva et al., "White paper on broadband connectivity in 6G," 2020, *arXiv:2004.14247*.
[2] Y. He, G. Ren, and X. Dong, "Fixed cluster-based collision resolution random access for cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8405–8418, Aug. 2024.
[3] A. Lancho, G. Durisi, and L. Sanguinetti, "Cell-free massive MIMO for URLLC: A finite-blocklength analysis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 8723–8735, Dec. 2023.
[4] *Study of New Radio Access Technology: Radio Access Architecture and Interfaces*, document TR 38.801, Version 14.0.0, 3GPP, Aug. 2019.
[5] P. Marsch, Ö. Bulakci, O. Queseth, and M. Boldi, *5G System Design: Architectural and Functional Considerations and Long Term Research*. Hoboken, NJ, USA: Wiley, 2018.
[6] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description: Stage 2*, document TS 36.300, 3GPP, Jun. 2017.
[7] H. Q. Ngo, L.-N. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, "On the total energy efficiency of cell-free massive MIMO," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 25–39, Mar. 2018.
[8] 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.
[9] T. Van Chien, E. Björnson, and E. G. Larsson, "Joint power allocation and load balancing optimization for energy-efficient cell-free massive MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6798–6812, Oct. 2020.

[10] G. Femenias, N. Lassoued, and F. Riera-Palou, "Access point switch ON/OFF strategies for green cell-free massive MIMO networking," *IEEE Access*, vol. 8, pp. 21788–21803, 2020.
[11] 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.
[12] H. Yang and T. L. Marzetta, "Energy efficiency of massive MIMO: Cell-free vs. cellular," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.
[13] 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.
[14] 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.
[15] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "Distributed resource allocation optimization for user-centric cell-free MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3099–3115, May 2022.
[16] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1250–1264, Feb. 2020.
[17] M. Munawar and K. Lee, "Dual-polarized IRS-assisted MIMO network," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 2519–2532, Apr. 2024.
[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[19] P. Ma, W. Wang, X. Zhao, and K. Zheng, "Joint transmitter-receiver design for the downlink multiuser spatial multiplexing MIMO system," in *Proc. IEEE Int. Conf. Commun.*, Beijing, China, 2008, pp. 3526–3530, doi: 10.1109/ICC.2008.663.
[20] P. Frenger, J. Hederen, M. Hessler, and G. Interdonato, "Improved antenna arrangement for distributed massive MIMO," WO Patent 2018 103 897, Jan. 26, 2017.
[21] A. Puglielli, N. Narevsky, P. Lu, T. Courtade, G. Wright, B. Nikolic, and E. Alon, "A scalable massive MIMO array architecture based on common modules," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 1310–1315.
[22] C. Amatetti, A. Guidotti, and A. Vanelli-Coralli, "Non-terrestrial network support for massive machine-type communication: Architectural and radio channel model considerations," in *Integration of MTC and Satellites for IoT Toward 6G Era*. Hoboken, NJ, USA: Wiley, 2024, pp. 37–66.
[23] A. Delmade, C. Browning, A. Farhang, N. Marchetti, L. E. Doyle, R. D. Koilpillai, L. P. Barry, and D. Venkitesh, "Performance analysis of analog IF over fiber fronthaul link with 4G and 5G coexistence," *J. Opt. Commun. Netw.*, vol. 10, no. 3, pp. 174–182, Mar. 2018.
[24] A. Bourdoux, B. Come, and N. Khaled, "Non-reciprocal transceivers in OFDM/SDMA systems: Impact and mitigation," in *Proc. Radio Wireless Conf. (RAWCON)*, 2003, pp. 183–186.
[25] M. Guenach, A. A. Gorji, and A. Bourdoux, "Joint power control and access point scheduling in fronthaul-constrained uplink cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2709–2722, Apr. 2021.
[26] M. Guenach, A. Gorji, and A. Bourdoux, "Power control and node scheduling in uplink cell-free massive MIMO with centralized beamforming," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
[27] P. Popovski, C. Stefanović, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.
[28] M. Guenach, A. A. Gorji, and A. Bourdoux, "A deep neural architecture for real-time access point scheduling in uplink cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1529–1541, Mar. 2022.
[29] Y. Zhang, H. Cao, M. Zhou, S. Wu, and L. Yang, "Rate maximization for cell-free massive MIMO with low-resolution ADCs," in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput., Int. Conf Pervasive Intell. Comput., Int. Conf. Cloud Big Data Comput., Int. Conf Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Aug. 2019, pp. 897–900.
[30] C. Desset and B. Debaillie, "Massive MIMO for energy-efficient communications," in *Proc. 46th Eur. Microw. Conf. (EuMC)*, Oct. 2016, pp. 138–141.
[31] C. Desset, P. Wambacq, Y. Zhang, M. Ingels, and A. Bourdoux, "A flexible power model for mm-wave and THz high-throughput communication systems," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, Aug. 2020, pp. 1–6.

MUTEEN MUNAWAR (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering, in 2019 and 2021, respectively, and the second master's degree in electrical and information engineering from Seoul National University of Science and Technology, South Korea. He is currently pursuing the Ph.D. degree with UGent, Belgium. He was with the Communication and Signal Processing Laboratory, South Korea, from February 2023 to September 2023. He is also with IMEC, Leuven, as a Researcher. His research interests include wireless communications, applied machine learning, and digital multipliers. He received the Best Project Award at the IEEE Riphah Innovation Summit (IRIS), in 2019.

MAMOUN GUENACH (Senior Member, IEEE) received the degree from EMI, Morocco, in 1995, and the Ph.D. degree from UCL, Belgium, in 2002. He was a Post-Doctoral Researcher with Ghent University, from 2002 to 2006, where he has been a part-time Visiting Research Professor, since 2015. He was a member of Nokia Bell Labs Technical Staff, from 2006 to 2019. Since 2019, he has been with IMEC, Leuven, as a Research Scientist. His main research interests include coding, modulation, synchronization, and MIMO equalization for high-speed wired and wireless communication technologies.

INGRID MOERMAN (Senior Member, IEEE) received the degree in electrical engineering and the Ph.D. degree from Ghent University, in 1987 and 1992, respectively. She became a part-time Professor with Ghent University, in 2000. She is currently a Staff Member with IDLab, a core research group of IMEC, with research activities embedded with Ghent University and the University of Antwerp. She coordinates the research activities on intelligent wireless networking (iWiNe) with Ghent University, where she leads a team of more than 30 researchers. She is also the Program Manager of the deterministic networking track, part of the Connectivity Program at IMEC. In this role, she coordinates research on end-to-end wired/wireless networking solutions for business and mission-critical applications that must meet strict quality of service requirements in terms of throughput, bounded latency, and reliability in private wireless environments. She has extensive experience in running and coordinating national and EU research-funded projects.