Scalable User-Centric Distributed Massive MIMO Systems with Limited Processing Capacity

Marx M. M. Freitas*, Daynara D. Souza*, A. L. P. Fernandes*, Daniel B. da Costa†
André Mendes Cavalcante‡, Luca Valcarenghi⋆, and João C. Weyl Albuquerque Costa*
*Federal University of Pará, Belém, Brazil
†Technology Innovation Institute, Abu Dhabi, UAE
‡Ericsson Telecommunications S.A., Indaiatuba, Brazil
⋆Scuola Superiore Sant'Anna, Pisa, Italy
marx@ufpa.br, daynara@ufpa.br, andrelpf@ufpa.br daniel.costa@tii.ae,
andre.mendes.cavalcante@ericsson.com, luca.valcarenghi@santannapisa.it, jweyl@ufpa.br

***Abstract*—This paper investigates the performance of scalable user-centric (UC) distributed massive multiple-input multiple-output (D-mMIMO) systems, widely known in the literature as cell-free mMIMO, with limited processing capacity. Specifically, it is assumed that the computational complexity (CC) of performing channel estimation and precoding signals does not increase with the number of access points (APs). In this regard, it is considered that each user equipment (UE) can only be associated with a finite number of APs. Moreover, a method is proposed for adjusting the AP clusters according to the network implementation, i.e., centralized or distributed. We compare the proposed approaches with a scalable UC system that does not perform AP cluster adjustment and does not prevent the processing demands from growing with the number of APs. Simulation results reveal that UC systems can keep the spectral efficiency (SE) under minor degradation even if the processing capacity is limited, reducing the CC by up to 96%. Besides, the proposed method for adjusting the AP cluster leads to further reductions in CC.**

***Index Terms*—AP selection, cell-free networks, computational complexity, distributed massive MIMO, user-centric approach.**

I. Introduction

User-centric (UC) distributed massive multiple-input multiple-output (D-mMIMO) systems, also referred to as cell-free (CF) mMIMO, have been envisaged as one of the most promising technologies for future mobile communication networks (6G and beyond) [1], [2]. In these systems, several access points (APs) are spread out in the coverage area, and the user equipment (UE) is served by a subset of APs, called AP cluster, providing a more uniform service and a better coverage probability than cell-based systems due to the enhanced macro-diversity and reduction of AP-UE distances [3], [4]. Despite the benefits, computational complexity (CC) can still be a drawback in these systems [5], [6].

Several baseline solutions consider that the complexity of UC systems grows with the number of UEs and APs, which is not practical [3], [4], [7]. In this regard, [8]–[10] proposed a framework to provide scalability to UC systems. Essentially, it limits the number of UEs each AP can serve. Consequently, the network resources (i.e., processing requirement, signaling on fronthaul/backhaul, and total power) remain finite even if the number of UEs goes to infinity, which is scalable. The authors showed that scalable UC systems can still provide uniform coverage with negligible spectral efficiency (SE) losses compared to the case when the UEs are served by all APs. The conclusions are valid for both centralized and distributed network implementations, in which the processing tasks are executed in central processing units (CPUs) or distributively in APs [10]–[12].

However, the current definition of scalability has addressed only one of the issues of UC systems. More specifically, although the network resources become independent of the number of UEs, the signal processing complexity can still grow with the number of APs [10]. This happens because the total number of complex multiplications required to perform channel estimation and precoding is proportional to the number of APs serving the UE [8]. Thus, a deeper investigation into this topic is necessary, as the literature regularly assumes that there are more APs than UEs in the network. Another inconsistency of UC systems is that the AP selection processes are not adapted to the network implementations. They generally only intend to improve some performance metrics, such as effective channel gain [13], reduce pilot contamination [10], and others [14], [15]. Consequently, AP clusters may benefit one implementation over another. For instance, AP clusters with a large number of APs can degrade the energy efficiency (EE) and CC of UC systems operating in distributed implementation while they can improve the SE for the centralized ones.

This paper investigates the performance of scalable UC D-mMIMO systems whose CC to perform channel estimation and precoding signals does not grow with the number of APs. In particular, it is considered a UC system where the UE is associated only with a finite number of APs, i.e., the CPU keeps the UEs connected only with the APs having the strongest channel gains. To the best of the authors' knowledge, this is the first paper to propose an approach that limits the CC of UC systems from growing with the number of APs. Moreover, it is proposed a method to adjust the AP clusters according to the network implementation. The proposed method works in UC systems with and without processing capacity limitations,

This work was supported by the Innovation Center, Ericsson Telecomunicações S.A, Brazil, the National Council for Scientific and Technological Development (CNPq), and the Coordination for the Improvement of Higher Education Personnel (CAPES). The research has also been partly supported by the project CLEVER (project number 101097560). The project is supported by the Key Digital Technologies Joint Undertaking and its members (including top-up funding by the Italian Ministry of Research and University (MUR)).

and it can be used as an alternative solution for reducing CC in UC systems without processing capacity limitation. As far as the authors are aware, this is also the first work that proposes a method for adjusting the AP clusters according to the network implementation in UC systems. Simulation results demonstrate that it is possible to keep the SE under minor degradation even if the CC is reduced by up to 96%. Nonetheless, the centralized implementation may require more processing capacity than the distributed one to avoid significant losses in the SE. It is also shown that the proposed strategy to adjust the AP clusters can reduce CC and potentially increase EE.

Notation: Boldface lowercase and uppercase letters denote column vectors and matrices, respectively, the superscript $(\cdot)^{\mathrm{H}}$ denotes the conjugate-transpose operation, the $N \times N$ identity matrix is $\mathbf{I}_N$, and the cardinality of the set $\mathcal{A}$ is represented by $|\mathcal{A}|$. The trace, euclidean norm and expectation operator are denoted as $\mathrm{tr}(\,.\,)$, $\|\,.\,\|$ and $\mathbb{E}\{\,.\,\}$, respectively, and the notation $\mathcal{N}_{\mathbb{C}}\left(\mu, \sigma^2\right)$ stands for a complex Gaussian random variable with mean μ and variance σ^2.

II. System Model

We consider a D-mMIMO network composed of L APs and K single-antenna UEs, where $L > K$. Each AP is equipped with N antennas, and the total number of antennas considering all APs is $M = NL$. The APs connect to the CPUs through fronthaul links, while the CPUs are linked to each other through backhaul ones. The system operates on time-division duplex (TDD) mode and it is assumed reciprocity for the uplink (UL) and downlink (DL) channels. The channel vector $\mathbf{h}_{kl} \in \mathbb{C}^{N\times 1}$ between the AP l and UE k undergoes an independent correlated Rician fading, being defined as

$$\mathbf{h}_{kl} = \underbrace{\sqrt{\frac{\kappa_{kl}}{1+\kappa_{kl}}}\mathbf{h}_{kl}^{\mathrm{LOS}}\, e^{j\theta_{kl}}}_{\overline{\mathbf{h}}_{kl}\, e^{j\theta_{kl}}} + \underbrace{\sqrt{\frac{1}{1+\kappa_{kl}}}\mathbf{h}_{kl}^{\mathrm{NLOS}}}_{\widetilde{\mathbf{h}}_{kl}}, \quad (1)$$

where $\theta_{kl} \sim \mathcal{U}[0, 2\pi)$ denotes random phase shifts, $\overline{\mathbf{h}}_{kl}\, e^{j\theta_{kl}} \in \mathbb{C}^{N\times 1}$ means the deterministic line-of-sight (LOS) component, and κ_{kl} stands for the Rician factor. The term $\widetilde{\mathbf{h}}_{kl} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}_N, \widetilde{\mathbf{R}}_{kl}) \in \mathbb{C}^{N\times 1}$ is the small-scale fading component with covariance matrix $\widetilde{\mathbf{R}}_{kl} = \mathbb{E}\{\widetilde{\mathbf{h}}_{kl}\widetilde{\mathbf{h}}_{kl}^{\mathrm{H}}\} \in \mathbb{C}^{N\times N}$.

A. Uplink Training and Channel Estimation

Each coherence block comprises τ_c samples, where τ_p samples are dedicated for UL pilot signals and τ_d for DL data. In the UL training phase, the UEs send pilot sequences of τ_p-length to the APs. Then, the UL channels are estimated using phase-unaware linear minimum mean square error (LMMSE) estimation. The pilot signals are orthogonal to each other, and a pilot t_k can be reused by some UEs if $K > \tau_p$. Let $\mathcal{P}_k \subset \{1, \ldots, K\}$ denote the subset of the UEs assigned to the pilot t_k, including the UE k. The received pilot signal at AP l can be expressed as [8]

$$\mathbf{y}_{t_k l}^{\mathrm{pilot}} = \sum_{i\in\mathcal{P}_k} \sqrt{\tau_p \eta_i}\, \mathbf{h}_{il} + \mathbf{n}_{t_k l}, \quad (2)$$

where $\mathbf{n}_{t_k l} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}_N, \sigma_{ul}^2 \mathbf{I}_N\right)$ denotes the noise and η_i is the power that the UE i transmits in the UL direction. The LMMSE channel estimate is given by

$$\widehat{\mathbf{h}}_{kl} = \sqrt{\tau_p \eta_k}\mathbf{R}_{kl}\boldsymbol{\Psi}_{t_k l}^{-1}\mathbf{y}_{t_k l}^{\mathrm{pilot}}, \quad (3)$$

where $\mathbf{R}_{kl} = \mathbb{E}\{\mathbf{h}_{kl}\mathbf{h}_{kl}^{\mathrm{H}}\} = (\overline{\mathbf{h}}_{kl}\overline{\mathbf{h}}_{kl}^{\mathrm{H}} + \widetilde{\mathbf{R}}_{kl})$ and $\boldsymbol{\Psi}_{t_k l} = \mathbb{E}\{(\mathbf{y}_{t_k l}^{\mathrm{pilot}})(\mathbf{y}_{t_k l}^{\mathrm{pilot}})^{\mathrm{H}}\} = \sum_{i\in\mathcal{P}_k} \eta_i \tau_p (\overline{\mathbf{h}}_{il}\overline{\mathbf{h}}_{il}^{\mathrm{H}} + \widetilde{\mathbf{R}}_{il}) + \sigma_{\mathrm{ul}}^2 \mathbf{I}_N$.

B. Downlink Data Transmission

In UC systems, each UE is associated with a subset of APs called AP cluster, represented by $\mathcal{M}_k \subset \{1, \ldots, L\}$. The connections between the UE k and APs are denoted by a diagonal matrix $\mathbf{D}_{kl} \in \mathbb{N}^{N\times N}$, being defined as

$$\mathbf{D}_{kl} = \begin{cases} \mathbf{I}_N & \text{if } l \in \mathcal{M}_k \\ \mathbf{0}_N & \text{if } l \notin \mathcal{M}_k. \end{cases} \quad (4)$$

The subset of UEs served by an AP is denoted by $\mathcal{D}_l$, and it is restricted to $|\mathcal{D}_l| \leq \tau_p$ to ensure system scalability [10]. Let $s_k \in \mathbb{C}$ denote the symbol intended for the UE k. The DL received signal at the UE k can be expressed as

$$y_k^{dl} = \sum_{l=1}^{L} \underbrace{\mathbf{h}_{kl}^{\mathrm{H}}\mathbf{D}_{kl}\mathbf{w}_{kl}s_k}_{\text{Desired signal}} + \sum_{i=1, i\neq k}^{K}\sum_{l=1}^{L} \underbrace{\mathbf{h}_{kl}^{\mathrm{H}}\mathbf{D}_{il}\mathbf{w}_{il}s_i}_{\text{Interfering signals}} + \underbrace{n_k}_{\text{Noise}}, \quad (5)$$

where $\mathbf{x}_l = \sum_{k=1}^{K} \mathbf{D}_{kl}\mathbf{w}_{kl}s_k$ represents the data signal sent by the AP l, $\mathbf{w}_{kl}$ denotes the precoding vector, and $n_k \sim \mathcal{N}_{\mathbb{C}}\left(0, \sigma_{\mathrm{dl}}^2\right)$ is the receiver noise. The terms s_k and $\mathbf{w}_{kl}$ satisfy $\mathbb{E}\left\{\|s_k\|^2\right\} = 1$ and $\mathbb{E}\left\{\|\mathbf{w}_{kl}\|^2\right\} = \rho_{kl}$, with ρ_{kl} being the power allocated to the UE k regarding the AP l.

From (5), the achievable DL SE can be computed as [10]

$$\mathrm{SE}_k^{(\mathrm{dl})} = \frac{\tau_d}{\tau_c}\log_2\left(1 + \mathrm{SINR}_k^{(\mathrm{dl})}\right), \quad (6)$$

where $\mathrm{SINR}_k^{(\mathrm{dl})}$ denotes the DL signal-to-interference-plus-noise ratio (SINR), which is given by

$$\mathrm{SINR}_k^{(\mathrm{dl})} = \frac{\left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}\right|^2}{\sum_{i=1}^{K}\mathbb{E}\left\{\left|\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_i\mathbf{w}_i\right|^2\right\} - \left|\mathbb{E}\left\{\mathbf{h}_k^{\mathrm{H}}\mathbf{D}_k\mathbf{w}_k\right\}\right|^2 + \sigma_{\mathrm{dl}}^2}, \quad (7)$$

where $\mathbf{w}_k \in \mathbb{C}^{M\times 1}$ and $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ are, respectively, the collective vectors of $\mathbf{w}_{kl}$ and $\mathbf{h}_{kl}$. For instance, $\mathbf{w}_k = \left[\mathbf{w}_{k1}^{\mathrm{T}}, ..., \mathbf{w}_{kL}^{\mathrm{T}}\right]^{\mathrm{T}}$ for $l \in \{1, \cdots, L\}$. Besides, $\mathbf{D}_k = \mathrm{diag}\left(\mathbf{D}_{k1}, ..., \mathbf{D}_{kL}\right) \in \mathbb{N}^{M\times M}$ stands for the diagonal block matrix. Note that (6) represents the widely known hardening bound, which is a capacity lower bound valid for any choice of precoding vectors [10].

C. Network Implementations and Computational Complexity

In the centralized implementation, the CPUs perform channel estimation, precoding, and process the DL signals [10]. This implementation provides better interference cancellation, and the CPUs have access to channel statistics of the UEs. In the distributed one, the APs perform these tasks locally using local channel state information (CSI). The CPUs are responsible for encoding the DL data signals. This implementation

may require less signaling on the fronthaul/backhaul links. However, if $\tau_c/(\tau_c - \tau_p) \approx 1$ and $K \gg N$, the distributed implementation may require much more signaling [11].

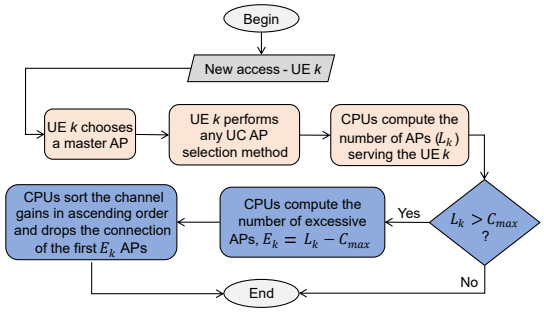


Fig. 1: Flowchart of the AP cluster size control.

The CC is calculated as in [8], accounting for the sum of the number of complex multiplications required from the network for a generic UE to perform channel estimation and generate the combining vectors in each coherence block. In UC D-mMIMO systems, the CC is a function of many parameters, such as the number of antennas N in each AP and the number of UL pilots τ_p, but it differs according to the network implementation and combining scheme. For the centralized implementation, the CC depends on the number of APs serving each UE ($|\mathcal{M}_k|$) and the number of UEs that are partially served by the same APs, being computed from [8, Table 5.1]. For the distributed implementation, the CC is a function of the number of UEs served by each AP ($|\mathcal{D}_l|$) and the number of APs serving each UE ($|\mathcal{M}_k|$), and is computed from [8, Table 5.3]. The CC considers the combining vectors since it is assumed that the precoding vectors are scaled versions of the combining ones, such as in [10].

III. Scalable UC D-mMIMO Systems with Limited Processing Capacity

In scalable D-mMIMO systems, the network complexity does not grow with the number of UEs since the number of UEs that each AP serve is limited, i.e., $K_l \leq \tau_p$, where $K_l = |\mathcal{D}_l|$. Therefore, the maximum number of UEs served by each AP remains finite even if the number of UEs K goes to infinity. However, the complexity of performing channel estimation and computing the precoding vectors can still grow with the number of APs [10]. That is, as L increases, the number of APs connected to the UE k (L_k) can also increase resulting in more processing complexity from the network, where $L_k = |\mathcal{M}_k|$. To circumvent this issue, we rely on a strategy where each UE can be associated only with a finite number of APs, denoted as C_{max}, with $L_k \leq \mathrm{C}_{max}$ [16]. We call this strategy maximum AP cluster size control. It is noteworthy that despite having a similar function, the C_{max} on this work is fundamentally different from the one presented in [16]. In this paper, C_{max} is a parameter that refers to the system processing capacity limitation that proportionates a new type of analysis for UC D-mMIMO systems.

The maximum AP cluster size control procedure is presented in Fig. 1 and it can be described as follows: when a new UE k enters the network, it measures the large-scale fading coefficients of the APs in its vicinity, which is calculated according to $\beta_{kl} = \mathrm{tr}\left(\mathbf{R}_{kl}\right)/N$ [10]. Then, it claims a master AP to ensure its connection with at least one AP. The master AP serves the UE k even if it has a poor channel condition [8]. To select a master AP, the UE k requests a connection to the available APs in its surroundings. Then, the available APs respond, and the UE k chooses the one with the strongest channel gain β_{kl} to be its master AP. Let $\mathcal{A}_l \subset \mathcal{D}_l$ denote the subset of UEs that the AP l is master. The available APs are the ones presenting $|\mathcal{A}_l| < \tau_p$, $\forall l \in \{1, \ldots, L\}$.

After selecting the master AP, the UE k performs any UC AP selection scheme. Then, the CPUs associated with the AP cluster of the UE k compute the number of APs (L_k) serving the UE k. If $L_k < \mathrm{C}_{max}$, no action is required. Otherwise, the CPUs will drop the connection of the UE k with the E_k APs presenting the weakest channel gains, where E_k denotes the number of APs that exceed C_{max}, which is calculated as $E_k = L_k - \mathrm{C}_{max}$. To drop the APs in excess, the CPUs associated with the UE k sorts the channel gains in ascending order, such that $\tilde{\beta}_{kl} \leq \tilde{\beta}_{kl} \leq \cdots \leq \tilde{\beta}_{kl}$, where $\tilde{\beta}_{kl}$ denotes the sorted version of β_{kl}, $\forall l \in \mathcal{M}_k$. Thus, the CPUs impose that $\mathbf{D}_{kl} = \mathbf{0}_N$ for the first E_k APs presenting the smallest channel gains after the sort operation. One can note that the APs in excess can belong to different CPUs. Therefore, each CPU drops the connection only of the APs linked to it by fronthaul. When all the CPUs associated with the E_k APs in excess perform these tasks, the E_k APs are finally dropped.

IV. AP Cluster Adjustment

In this section, it is proposed a heuristic method that adjusts the AP clusters according to the network implementation. Such method holds for any UC AP selection scheme, i.e., with and without processing capacity limitation. Besides, it is a heuristic strategy because only heuristic solutions are scalable [10]. In a nutshell, the UEs are associated with a subset of APs following any AP selection process. Then, the proposed method aims to simultaneously reduce the number of UEs served by each AP l (K_l) and the number of APs connected to each UE k (L_k) while keeping the SE under minor degradation. In this context, it is a novel way to reduce the CC and increase EE in scalable UC distributed D-mMIMO systems. Throughout the analysis, it is also assumed that each UE connects to a master AP.

A. AP Cluster Adjustment in the Distributed Implementation

In the distributed implementation, the proposed method exploits the local long-term CSI at each AP and intends to reduce K_l without causing significant SE degradation. When all APs are involved, the average value of L_k is also reduced. It is noteworthy that L_k is not directly reduced in distributed implementation, and neither could it be since it would require global long-term CSI at each AP.

The adjustment of the AP cluster relies on two metrics that we have proposed: (i) the partial channel strength indicator ($\bar{\beta}_{kl}$) and (ii) the total channel strength indicator ($\bar{\beta}_l$). We use these metrics to prevent the less fortunate UEs from being

easily dropped by the AP. Therefore, they do not directly represent the long-term CSI of the UEs that the AP serves. Instead, they are the long-term CSI raised to a normalization exponent, defined as λ_l, which provides a better balance between the channel gains of the most and less fortunate UEs served by the AP, such that $0 < \lambda_l < 1$. Without this normalization, the AP could easily drop a UE presenting a weaker channel gain if the AP was also serving UEs with stronger channel gains. However, these differences can be reduced when the channel gains are raised to a power lower than one and greater than zero, such as λ_l.

The partial channel strength indicator is given by $\bar{\beta}_{kl} = (\beta_{kl})^{\lambda_l}$, where $\lambda_l = \min_{k\in\mathcal{D}_l}(\beta_{kl})/\max_{k\in\mathcal{D}_l}(\beta_{kl})$. The second metric, called total channel strength indicator, is calculated as $\bar{\beta}_l = \sum_{k\in\mathcal{D}_l}\bar{\beta}_{kl}$. In the proposed method, the two metrics are used by each AP l to calculate $\bar{\beta}_{l,-k} = \bar{\beta}_l - \bar{\beta}_{kl}$, $\forall k \in \mathcal{D}_l$. The purpose of calculating $\bar{\beta}_{l,-k}$ is to evaluate how much $\bar{\beta}_l$ is affected by dropping the UE k from the AP l. The AP l keeps the connection of UE k only if

$$\mathbf{D}_{kl} = \begin{cases} \mathbf{I}_N & \text{if } \bar{\beta}_{l,-k} \leq \bar{\beta}_l^{mean} \\ \mathbf{I}_N & \text{if } k \in \mathcal{A}_l \\ \mathbf{0}_N & \text{otherwise,} \end{cases} \tag{8}$$

where $\bar{\beta}_l^{mean} = \sum_{k\in\mathcal{D}_l}\bar{\beta}_{l,-k}/K_l$ is a threshold value and $\mathcal{A}_l \subset \mathcal{D}_l$ denotes the subset of UEs that the AP l is master. One can note that the term $\bar{\beta}_{l,-k}$ has to be smaller than $\bar{\beta}_l^{mean}$, because $\bar{\beta}_{l,-k}$ will be small if the UE k has a large partial channel strength indicator $\bar{\beta}_{kl}$, since $\bar{\beta}_{l,-k} = \bar{\beta}_l - \bar{\beta}_{kl}$. Meanwhile, $\bar{\beta}_{l,-k}$ will be large if the UE k adds only a marginal gain to $\bar{\beta}_l$. That is, if $\bar{\beta}_{kl}$ represents a considerable percentage of $\bar{\beta}_l = \sum_{k\in\mathcal{D}_l}\bar{\beta}_{kl}$, the term $\bar{\beta}_l$ will be significantly reduced if the UE k is disconnected from the AP l.

B. AP Cluster Adjustment in the Centralized Implementation

In the centralized implementation, the long-term CSI of APs and UEs is available at the CPUs [10], [11]. Hence, the proposed method exploits the global long-term CSI to reduce L_k. At first, reducing L_k may appear counter-intuitive since the centralized implementation has a better interference suppression capability. However, since CC grows with the number of APs serving the UE (recall that $L_k = |\mathcal{M}_k|$), the AP cluster expansion will not always be beneficial, and reducing L_k may be necessary even in this implementation.

The partial channel strength indicator is now calculated as $\bar{\beta}_{kl} = (\beta_{kl})^{\lambda_k}$, where λ_k introduces a balance between the serving APs presenting the smallest and highest channel gain to the UE k. It is assumed that $\lambda_k = \min_{l\in\mathcal{M}_k}(\beta_{kl})/\max_{l\in\mathcal{M}_k}(\beta_{kl})$ and the total channel strength indicator is computed as $\bar{\beta}_k = \sum_{l\in\mathcal{M}_k}\bar{\beta}_{kl}$. Then, the CPUs calculates the contribution that each AP brings to $\bar{\beta}_k$ as $\bar{\beta}_{k,-l} = \bar{\beta}_k - \bar{\beta}_{kl}$, $\forall l \in \mathcal{M}_k$. Therefore, a CPU keeps the connection of AP l only if

$$\mathbf{D}_{kl} = \begin{cases} \mathbf{I}_N & \text{if } \bar{\beta}_{k,-l} \leq \bar{\beta}_k^{mean} \\ \mathbf{I}_N & \text{if } k \in \mathcal{A}_l \\ \mathbf{0}_N & \text{otherwise,} \end{cases} \tag{9}$$

where $\bar{\beta}_k^{mean} = \sigma_{si}/2 + \sum_{l\in\mathcal{M}_k}\bar{\beta}_{k,-l}/L_k$ and σ_{si} denotes the standard deviation of $\bar{\beta}_{k,-l}$, $\forall l \in \mathcal{M}_k$. The term σ_{si} is utilized to make the CPUs drop fewer APs from the AP cluster of UE k to exploit the centralized implementation's capacity in improving SE. It is worth noting that only the CPUs associated with the AP cluster of the UE run the proposed method. Besides, each CPU disconnects the UE from the subset of APs linked to it by fronthaul.

C. Pros and Cons of the two AP clusters Adjustments

The utilization of the proposed method on a distributed implementation proportionates a fronthaul signaling reduction since the number of data flows on the fronthaul is proportional to K_l [10], [11]. Besides, it allows the AP to carry out fewer operations while attaining the same SE performance, increasing the system's EE. Moreover, the reduction of L_k also reduces the signaling on fronthaul links as fewer APs forward the data signal of UE k. The utilization of the proposed method in a centralized implementation also allows significant savings in CC resources and fronthaul signaling. In a centralized implementation, the number of data flows on the fronthaul is not proportional to K_l as in the distributed implementation. However, the quantization level required to avoid loss of SE increases with K_l. It is worth noting that in this paper, we have considered that the AP cluster adjustment is only activated when λ_l and λ_k are lesser than a threshold Γ to avoid excessive adjustments, where Γ is a project parameter. We have set $\Gamma = 10^{-2}$ and $\Gamma = 10^{-3}$ for the distributed and centralized implementations, respectively.

V. Numerical Results

We consider a D-mMIMO network consisting of K single antenna UEs and L APs, each equipped with N antennas. The values of L, N, and K vary and are specified throughout the results. The K UEs are uniformly distributed over a square area of 1×1 km, and the distribution of the APs follows a hard core point process (HCPP)[1]. The simulations focus on DL channels and it is assumed that $\tau_c = 200$, $\tau_p = 10$, and $\tau_d = 190$. The total transmission powers of the UEs and APs are $100\,\mathrm{mW}$ and $200\,\mathrm{mW}$, respectively. We perform Monte-Carlo simulations to account for different locations and channel realizations and different AP/UE locations. The wrap-around technique is also utilized to provide a better balance regarding the amount of interference affecting each AP.

It is utilized an AP clustering scheme that jointly performs the pilot assignment and AP selection [8]. In this one, the UEs can connect to master and non-masters APs. The non-masters serve only the UEs with the greatest channel gain in each pilot. The first τ_p UEs are assigned to mutually orthogonal pilots, and the remaining ones to the pilot causing the lowest pilot contamination. Hereafter, we name it as scalable cell-free (SCF) scheme. The 3GPP Urban Micro (UMi) path

[1]This method states that the distance between any two APs cannot be smaller than $d_{\min} = \sqrt{A/L}$, where A is the coverage area in square meters. The first step is to randomly drop the APs based on a homogeneous Poisson point process with mean rate $1/d_{\min}$, then randomly update the location of APs that do not meet the spacing requirement until it is fulfilled.

loss model is adopted for modeling the propagation channel, with LOS/non-line-of-sight (NLOS) conditions defined in the Technical Report (TR) 38.901 [17]. It is considered that the shadowing terms of an AP to different UEs are correlated, and the computation of correlation matrices $\mathbf{R}_{kl}$ follows the local scattering spatial correlation model [8]. Table I exhibits the parameters used in the UMi and $\mathbf{R}_{kl}$ models [8], [18].

TABLE I: Parameters assumed for the UMi path loss and local scattering spatial correlation model.

Parameter	Value
Shadow fading standard deviation, σ_{SF}	4 dB
AP/UE antenna height, $h_{\text{AP}}, h_{\text{UE}}$	11.65 m, 1.65 m
RX noise figure (NF)	8 dB
Carrier frequency, bandwidth	3.5GHz, 100MHz
Angular standard deviations (ASDs)	$\sigma_\varphi = \sigma_\theta = 15°$
Antenna spacing	1/2 wavelength distance

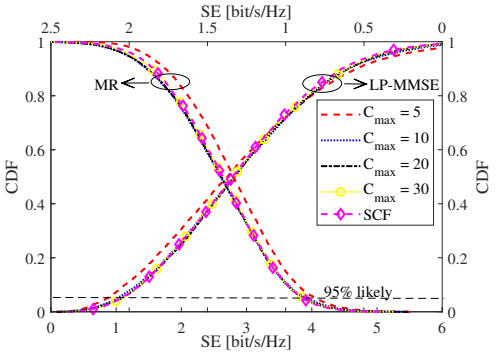
The power coefficients at AP l in the distributed implementation are set as $\rho_{kl} = \rho_d\sqrt{\beta_{kl}}/\sum_{k'\in\mathcal{D}_l}\sqrt{\beta_{k'l}}$, where ρ_d is the maximum DL transmit power per AP. For the centralized one, it is used the scalable fractional power control [8]. In order to compute the precoding vectors, it is employed the partial MMSE (P-MMSE) and partial regularized zero-forcing (P-RZF) for the centralized implementation. For the distributed, it is utilized the local partial MMSE (LP-MMSE) and maximum ratio (MR). These techniques were chosen due to their scalability features [10].

A. Impacts of Limiting the Processing Capacity

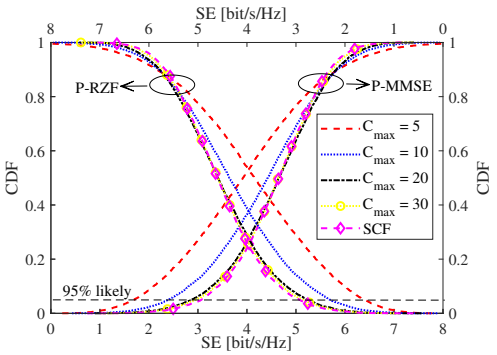
We start by evaluating a network composed of $K = 25$ UEs and $L = 100$ APs equipped with $N = 1$ antenna. Fig. 2 presents the cumulative distribution functions (CDFs) of the SE of UC systems with and without processing capacity limitation. It considers different processing capacity limitations, i.e., several values of C_{max}, and the system is compared with a traditional UC scheme (i.e., L_k is not restricted), which we have denoted as SCF.

In Fig. 2a, the SE is not as reduced by the variations of C_{max}. The SE even increases slightly for $10 \leq \mathrm{C}_{max} \leq 20$. This is because decreasing L_k also reduces K_l, helping precoding techniques such as LP-MMSE (of local processing) to mitigate interference. Still, this improvement has a limit since the SE decays about 9% when C_{max} goes from 40 to 5. In Fig. 2b, the SE can suffer significant losses when C_{max} is as small as 5. Hence, reducing the AP cluster sizes (L_k) may lead the centralized implementation to not exploit its full potential in mitigating interference and improving SE. Therefore, it is essential for this implementation to utilize more processing capacity, such as $\mathrm{C}_{max} \geq 20$.

Fig. 3 presents the SE and CC when the number of APs varies and by setting $K = 25$, and $\mathrm{C}_{max} = 20$. In Fig. 3a, the average SE grows with L for UC systems with and without processing capacity limitation. Despite this, limited systems have a significant advantage, as their CC does not always increase with L, starting to decay from $L = 60$. This behavior occurs because K_l reduces as L increases. Therefore, even

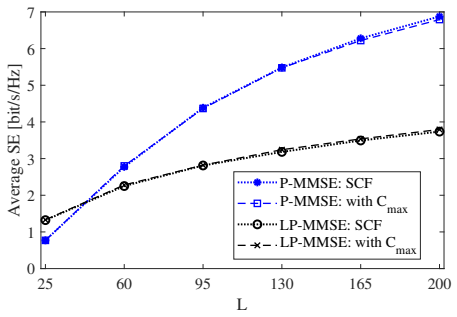


(a) Distributed implementation.

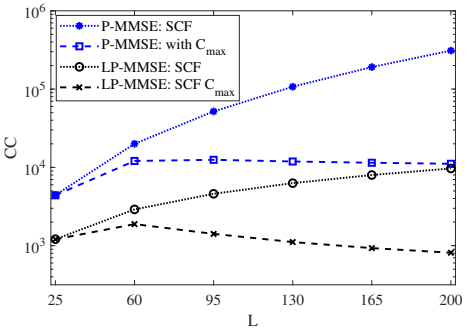


(b) Centralized implementation

Fig. 2: CDF of SE by varying C_{max} from 5 to 30. Parameters setting: $L = 100$, $K = 25$, and $N = 1$.



(a) SE



(b) CC

Fig. 3: Average DL SE (a) and CC (b) achieved by varying the number of APs L. Parameters setting: $K = 25$, $N = 1$, and $\mathrm{C}_{max} = 20$.

TABLE II: Average number of APs per UE (L_k) and UEs per AP (K_l) without and with AP cluster control. Parameters setting: $K = 25$, $N = 1$, and $C_{max} = 20$.

Method	$L = 95$		$L = 200$	
	K_l	L_k	K_l	L_k
SCF	10	38	10	80
With C_{max}	5.25	19.98	2.5	20

if L_k remains constant, there will be a reduction in K_l, as Table II demonstrates. Additionally, it is possible to observe that the CC decreases by about 96.4% when the processing capacity limitation is employed together with the P-MMSE for $L = 200$. However, a centralized implementation may require more processing capacity to be feasible compared to the distributed implementation. For instance, the P-MMSE scheme has a CC similar to LP-MMSE (without processing limitation) even limiting the processing capacity, when L is as large as 200.

Fig. 4 presents the EE achieved in the distributed implementation considering different values of C_{max} and a UC system without processing capacity limitation. Note that the processing capacity limitation can provide a considerable improvement in the EE, especially for small values of C_{max}. For instance, the processing capacity limitation guarantees an increase of about 10% in EE for $C_{max} < 30$. Besides, the EE grows by about 61% in the LP-MMSE and 36% in the MR, when C_{max} decreases from 40 to 5. This happens because reducing K_l also decreases the power consumption in each fronthaul link. Thus, even though the system presents SE losses when $C_{max} = 5$, the reduction of power consumption in each fronthaul link compensates them, increasing the EE.

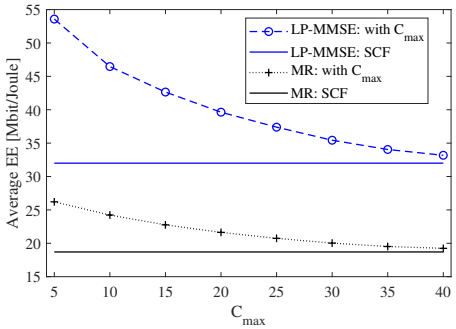
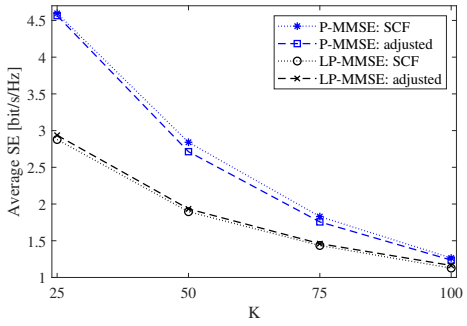


Fig. 4: Average EE achieved by varying C_{max}. Parameters setting: $L = 100$, $K = 25$, and $N = 1$.

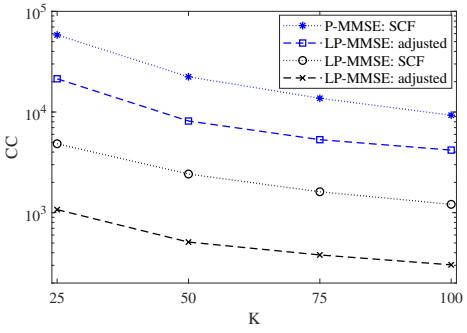
B. Impacts of AP Cluster Adjustment

From now on, we will investigate the impacts of adjusting the AP clusters in UC systems. We will focus on UC systems without processing capacity limitation to assess the full benefits of the AP cluster adjustment in reducing CC. Furthermore, we will consider only the P-MMSE and LP-MMSE schemes as they provide the best interference mitigation in centralized and distributed implementations.

Fig. 5 presents the average SE and CC versus the number of UEs K in a network composed of $L = 100$ APs equipped with $N = 1$ antenna. It can be noted that the proposed method causes a tiny reduction in the SE of P-MMSE. Despite this, the losses are not as expressive as in Fig. 2b. This is because the proposed method does not decrease L_k to a small value such as 5, as Table III indicates. One can also note that the proposed method causes a slight increase in the SE of LP-MMSE. Moreover, the AP cluster adjustment also reduces the CC of both network implementations, decreasing by up to 60% in the P-MMSE scheme for $K = 25$. Finally, the proposed method decreases K_l from 10 to 3.95 and L_k from 40 to 15.80, as illustrated in Table III, indicating that the proposed strategy can also increase the EE in distributed implementation.



(a) SE



(b) CC

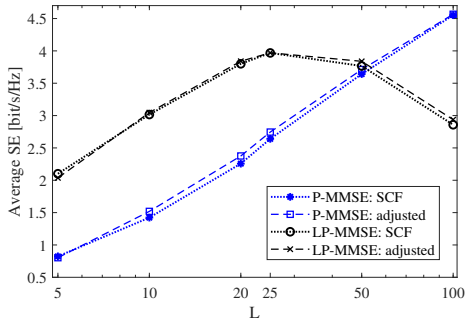
Fig. 5: Average DL SE (a) and CC (b) achieved by varying the number of UEs K, when the proposed AP cluster adjustment is employed. Parameters setting: $L = 100$ and $N = 1$.

TABLE III: Average number of APs per UE (L_k) and UEs per AP (K_l) without and with AP cluster adjustment. Parameters setting: $L = 100$ and $N = 1$.

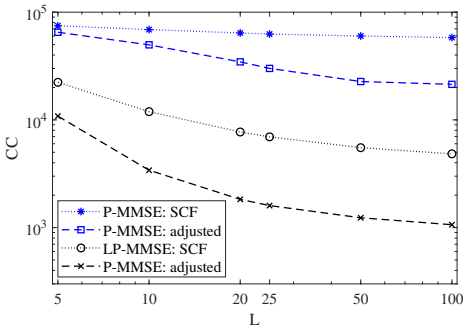
Method	$K = 25$		$K = 50$	
	K_l	L_k	K_l	L_k
SCF	10	40	10	20
Distributed adjustment	4.32	17.3	4.38	8.75
Centralized adjustment	6.23	24.92	6.17	12.35

Fig. 6 presents the average SE and CC versus the number of UEs L and N for a fixed total number of antennas $M = LN = 100$ and setting the number of UEs to be

$K = 25$. One can note that the same discussions about decreasing CC apply to this case. The difference is the SE behavior. When $L = 25$ and $N = 4$, the LP-MMSE scheme achieved the best balance regarding the amount of interference and desired signal, leading the average SE to its maximum value. Meanwhile, the P-MMSE presents better SE when the AP clusters are adjusted for $L < 100$. This is because the fewer APs in the coverage area, the further away the APs will be from the UE. Hence, the AP clusters can have many APs presenting poor channel gains. Therefore, disconnecting some of these APs will not impact the UE's performance. Additionally, it can be noticed that CC reduces as L increases and N decreases. The reduction is stronger in UC systems with AP cluster adjustment. At $L = 100$, the proposed method reduces the CC by about 63% and 78% for the P-MMSE and LP-MMSE schemes, respectively. Therefore the AP cluster adjustment can strongly reduce CC, especially for a large number of APs.



(a) SE



(b) CC

Fig. 6: Average DL SE (a) and CC (b) achieved by varying L and N, while keeping $M = 100$, when the proposed AP cluster adjustment is employed. Parameters setting: $K = 25$.

VI. Conclusions

This paper investigated the performance of scalable UC D-mMIMO systems whose processing capacity limitations do not increase with the number of APs. We analyzed UC systems whose AP clusters can have only a finite number of APs serving each UE. We also proposed a method that adjusts the AP clusters to the network implementation. The results demonstrated that restricting the network processing capacity can improve the EE by up to 61%. However, it can degrade the SE of centralized implementation when the maximum number of APs serving the UE is small. On the other hand, AP clusters comprising just a few APs almost do not harm the SE of the distributed implementation. Simulation results also reveal that the proposed AP cluster adjustment can slightly improve the SE of distributed implementation while reducing the CC in both network implementations. The CC can decrease by up to 96% in centralized implementation. These results open the way for future works to design practical UC systems with limited processing capacity and systems that intend to adjust the AP clusters to the network implementation. The cluster size control proposed in this work can inspire future publications related to UC systems with limited processing capacity as it works in any AP selection scheme.

References

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

[2] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133 995–134 030, Jul. 2020.

[3] 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.

[4] H. Q. Ngo, L. 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.

[5] R. Wang, M. Shen, Y. He, and X. Liu, "Performance of cell-free massive MIMO with joint user clustering and access point selection," *IEEE Access*, vol. 9, pp. 40 860–40 870, Feb. 2021.

[6] M. M. M. Freitas, D. D. Souza, D. B. da Costa, A. M. Cavalcante, L. Valcarenghi, G. S. Borges, R. Rodrigues, and J. C. W. A. Costa, "Reducing inter-CPU coordination in user-centric distributed massive MIMO networks," *IEEE Wireless Commun. Lett.*, pp. 1–5, 2023.

[7] 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.

[8] Ö. Demir, E. Björnson, and L. Sanguinetti, *Foundations of User-Centric Cell-Free Massive MIMO*. Foundations and Trends in Signal Processing Series, Now Publishers, 2021.

[9] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jul. 2019, pp. 1–6.

[10] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 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. 2019.

[12] F. Li, Q. Sun, X. Ji, and X. Chen, "Scalable cell-free massive MIMO with multiple CPUs," *Mathematics*, vol. 10, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2227-7390/10/11/1900

[13] H. T. Dao and S. Kim, "Effective channel gain-based access point selection in cell-free massive MIMO systems," *IEEE Access*, vol. 8, pp. 108 127–108 132, Jun. 2020.

[14] M. Freitas, D. Souza, D. B. d. Costa, G. Borges, A. M. Cavalcante, M. Marquezini, I. Almeida, R. Rodrigues, and J. C. W. A. Costa, "Matched-decision AP selection for user-centric cell-free massive MIMO networks," *IEEE Trans. Veh. Technol*, pp. 1–16, Jan. 2023.

[15] V. Ranasinghe, N. Rajatheva, and M. Latva-aho, "Graph neural network based access point selection for cell-free massive MIMO systems," in *Proc. IEEE Global Commun. Conf.*, Feb. 2021, pp. 01–06.

[16] D. D. Souza, M. M. M. Freitas, D. B. da Costa, G. S. Borges, A. M. Cavalcante, L. Valcarenghi, and J. C. Weyl Albuquerque Costa, "Effective channel DL pilot-based estimation in user-centric cell-free massive MIMO networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2022, pp. 705–710.

[17] 3GPP, *Study on channel model for frequencies from 0.5 to 100 GHz*, 2019, 3GPP TR 38.901 (Release 16).

[18] 3GPP, *NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone*, 2021, 3GPP TR 38.101-1 (Release 17).