IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 February 2023, accepted 7 March 2023, date of publication 10 March 2023, date of current version 21 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3255828

RESEARCH ARTICLE

Study of Clustering Solutions for Scalable Cell-Free Massive MIMO

DANAISY PRADO-ALVAREZ[1], DANIEL CALABUIG[1], (Member, IEEE), JOSE F. MONSERRAT[1], (Senior Member, IEEE), SAMER BAZZI[2], AND WEN XU[2], (Senior Member, IEEE)

[1]Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), Universitat Politècnica de València, 46022 Valencia, Spain
[2]Huawei Technologies Duesseldorf GmbH, 80992 Munich, Germany

Corresponding author: Danaisy Prado-Alvarez (dapraal@upv.edu.es)

ABSTRACT In response to the requirements of 5G and beyond, cell-free systems have emerged to ensure uniform service throughout the area. However, the idea of having no cells leads to a considerably large monolithic system, which is not scalable. In this context, common sense suggests the creation of clusters of Access Points (APs) and User Equipments (UEs) that allow the system to be managed locally to some extent. The clustering procedure has to ensure certain performance for the global system and, at the same time, ensure its scalability. In this paper, first, scalability is analyzed. Subsequently, we study and compare two clustering approaches for uplink and downlink. Finally, we propose a combination of clustering and resource allocation techniques that outperforms the rest of the analyzed state-of-the-art solutions.

INDEX TERMS Cell-free massive multiple-input multiple-output (MIMO), scalability, clustering.

I. INTRODUCTION

The performance of systems based on a cell-centric design has always been limited by the interference experienced by UEs at cell borders. New applications for Fifth Generation (5G) and beyond require seamless deployments to ensure uniform service over the entire area. A natural solution is the use of user-centric cell-free systems. In this context and considering massive Multiple-Input Multiple-Output (MIMO) as one of the enabling technologies of next-generation systems, cell-free massive MIMO [1] was proposed to overcome the high interference levels at the cell borders.

Normally, when talking about cell-free massive MIMO, and due to the large number of antennas at the network side, the APs acquire the Channel State Information (CSI) through the use of uplink pilot signals sent by the UEs, and then this CSI is used in the design of precoders. To be able to use this CSI in the downlink, Time Division Duplexing (TDD) mode is considered. In general, at the APs, the signal is multiplexed/de-multiplexed, converted by the analog/digital and digital/analog converters, and processed by the base-band (BB) unit. On the other hand, a Central Processing Unit (CPU) is in charge of the coordination among all APs.

Some challenges arise when implementing the canonical form of cell-free massive MIMO described before. First, since all APs must be connected to the same CPU via front-haul, the system has to withstand a considerable amount of traffic and signaling, which increases the computational requirements of the CPU and the front-haul capacity demands. This problem can be tackled via a user-centric approach, so that a given UE is served by a subset of APs close to it that share a CPU, helping to relax the front-haul performance requirements significantly [2]. In [3], the authors described four network deployments going from the conventional cellular network to a cell-free massive MIMO network considering a user-centric approach. In this work, several practical aspects are analyzed such as the cost and complexity of the deployment, limited capacity of back/front-haul connections, and network synchronization. It is important to remember that the objective of cell-free massive MIMO is to reduce the interference at the cell borders of traditional cellular systems. Therefore, the subset of APs for each UE has to be selected taking this objective into account. In this line, it is possible to identify two different clustering

The associate editor coordinating the review of this manuscript and approving it for publication was Pietro Savazzi.

approaches depending on how interference is reduced. The first approach creates a different cluster for each UE, locating the UEs at the center of their clusters. In this case, the cluster border is distanced as much as possible. The second approach includes many UEs in the same cluster and uses interference cancellation techniques to reduce the interference.

In this work, we will analyze these two approaches for uplink and downlink under different scenarios in order to figure out which would be the most suitable in terms of achievable data rate under certain circumstances. Based on the obtained results, we will propose a combination of clustering and resource allocation techniques that outperforms the current solutions in the literature.

The reminder of this paper is organized as follows. In Section II, the system model is presented. Afterward, the scalability concept is analyzed in Section III. Section IV describes the two previously mentioned clustering approaches. The performance of particular clustering solutions from the two approaches with the corresponding resource allocation techniques is evaluated in Section V for the uplink, and in Section VI for the downlink. Finally, Section VII draws the conclusion.

II. SYSTEM MODEL

In this work, a cell-free massive MIMO system with clusters is considered. In particular, we define the system $\mathcal{S} = \{\{\mathcal{L}_m\}_{m=1}^{M}, \{\mathcal{K}_m\}_{m=1}^{M}\}$ as a system with M clusters of single-antenna APs and UEs. The APs and UEs in the m-th cluster are those in $\mathcal{L}_m$ and $\mathcal{K}_m$, respectively. We assume that the UEs are exactly in one of those clusters, although APs can be in several clusters, i.e., $\mathcal{K}_m \cap \mathcal{K}_n = \emptyset$ for all $m \neq n$, but this is not necessarily true for any pair of sets in $\{\mathcal{L}_m\}_{m=1}^{M}$. The amount of APs and UEs in the system is $L = |\cup_{m=1}^{M} \mathcal{L}_m|$ and $K = |\cup_{m=1}^{M} \mathcal{K}_m|$, respectively.

Let $\tilde{m}(k)$ be the cluster the k-th UE belongs to. Without loss of generality, we index the UEs in the same cluster continuously, i.e., the UEs in the m-th cluster are those with indices $\tilde{k}(m-1)+1, \tilde{k}(m-1)+2, \ldots, \tilde{k}(m)$, where $\tilde{k}(m)$ is the largest index of the UEs in the m-th cluster, and $\tilde{k}(0) = 0$. The channel between the k-th UE and the APs of the m-th cluster is denoted by $\mathbf{H}_{km} \in \mathbb{C}^{|\mathcal{L}_m|}$, where $|\mathcal{L}_m|$ is the amount of APs in the m-th cluster.

Interference cancellation techniques are used for the set of UEs in each cluster. In particular, some interference cancellation is performed between the UEs $1, 2, \ldots, \tilde{k}(1)$, also between the UEs $\tilde{k}(1)+1, \ldots, \tilde{k}(2)$, etc.

As a performance measure of the system, the total sum data rate is calculated for both uplink and downlink. In the case of the uplink, Successive Interference Cancellation (SIC) [4] is considered in the decoding. This algorithm consists of (i) decoding the signal of the first UE considering the signal of the rest as interference, (ii) subtracting the signal decoded from the first UE from the received signal, (iii) decoding the signal of the second UE considering the signal from the third UE onwards as interference, and so on. This reception technique is known to be capacity-achieving if all the sources of interference have a known covariance. Specifically, since the decoding order does not affect the sum rate [5], we assume that the first signal to be decoded is the one corresponding to the UE with the highest index, i.e., the UE $\tilde{k}(m)$ of the m-th cluster, and then we follow the decreasing order of the indices. Therefore, the achievable data rate of the k-th UE in the uplink with an identity noise covariance is as in (1), shown at the bottom of the page, where P is the available power at the UEs. In (1), the interference has been divided into two terms according to its origin. In the numerator, the first term is the noise covariance; the second term includes the interference from UEs in clusters with indices $m < \tilde{m}(k)$, the interference from the UEs in the same cluster, and the signal of the k-th UE; and the third term includes the interference from the UEs in clusters with indices $m > \tilde{m}(k)$. In the denominator, the terms are similar except for the fact that the signal of the k-th UE is not included.

Analogously, in the case of the downlink, Dirty Paper Coding (DPC) [6] is considered. This coding technique consists of adapting the codebooks of the transmitter to a known interference in such a way that the interference does not affect the achievable rate. Therefore, the CPU of the AP cluster can (i) encode the signal of the first UE, (ii) treat this signal as a known interference for the second UE, and so on. In this case, the first UE sees the signal from the other UEs as interference, the second UE sees the signal from the third UE onwards as interference, etc. As in the case of SIC, this encoding technique is known to be capacity-achieving if all the sources of interference have a known covariance. Some works found an uplink-downlink duality when SIC and DPC are used [7]. In particular, the achievable rates are the same. In [7], it is shown that the DPC encoding order that achieves certain achievable rates is the reversed SIC decoding order of the dual uplink system. Following this principle, we assume that the encoding order is the reversed order considered for the uplink, i.e., the encoding follows the increasing order of the UEs' indices. Therefore, the achievable data rate of the k-th UE in the downlink with an identity noise covariance is as in (2), shown at the bottom of the next page, where $\mathbf{Q}_k$ is the transmit covariance matrix of the k-th UE. In (2), the interference has also been divided into two terms according to its origin, and in a similar manner as in (1). Since single-antenna UEs are considered, it is not needed to include determinants in the equation.

$$R_k^{\mathrm{UL}} = \log \frac{|I + P\sum_{i=1}^{k} \mathbf{H}_{i\tilde{m}(k)}\mathbf{H}_{i\tilde{m}(k)}^{*} + P\sum_{i=\tilde{k}(\tilde{m}(k))+1}^{K} \mathbf{H}_{i\tilde{m}(k)}\mathbf{H}_{i\tilde{m}(k)}^{*}|}{|I + P\sum_{i=1}^{k-1} \mathbf{H}_{i\tilde{m}(k)}\mathbf{H}_{i\tilde{m}(k)}^{*} + P\sum_{i=\tilde{k}(\tilde{m}(k))+1}^{K} \mathbf{H}_{i\tilde{m}(k)}\mathbf{H}_{i\tilde{m}(k)}^{*}|}. \quad (1)$$

III. ANALYSIS OF THE SCALABILITY

Cell-free massive MIMO systems in their canonical form rely on the cooperation of the L APs to serve the K UEs, i.e., $M = 1$. This implies that, as the network grows, so do the computational requirements of the network and the volume of data and signaling that has to be exchanged between each AP and each UE. Due to this, these systems are said to be non-scalable. It is therefore important to find solutions with superior performance and good scalability with the network size. To do this, it is important to agree on the main features that a scalable system should satisfy. Then, we could design systems with such features. In this section, we analyze a scalability definition from the state of the art, and subsequently, we propose a new definition.

A. STATE-OF-THE-ART DEFINITION

In [8], the authors formally defined that the network is scalable if the following tasks have finite complexity and resource requirements for each AP as the number of UEs, K, tends to ∞:

1) Signal processing for channel estimation;
2) Signal processing for data reception and transmission;
3) Front-haul signaling for data and CSI sharing;
4) Power control optimization.

In [3], a solution was proposed and considered to be fully scalable. The solution is based on the definition of fixed clusters of APs, each of them connected to a different CPU. Each UE selects the best-server AP and a set of serving APs, up to a maximum quantity of APs. The CPU of the best-server is in charge of the channel estimation and the reception/transmission processing. This processing and channel estimation is done independently for each UE and takes into account only the serving APs. Therefore, even in the worst case (all UEs require the maximum amount of serving APs), the total resource requirements of this solution increase linearly with the amount of UEs in the network. When the network grows, all, the amount of UEs, APs, and CPUs grow with the same proportion and, hence, both the amount of UEs and the available resources grow with the same proportion. As a consequence, the system is scalable. However, this scalability claimed in [3] was questioned by [8], since [3] does not explicitly limit the number of UEs that can connect to one AP.

This raises the question of whether all the criteria in the scalability definition in [8] are necessary in practice. For instance, although one can argue the necessity of limiting the amount of UEs that can be served by one AP, the reality is that many works in the literature dealing with resource allocation problems do not consider this limitation. This is due to two main reasons. First, it is unlikely to saturate the APs by the amount of UEs, specially considering deployments with more APs than UEs like in cell-free massive MIMO systems. Second, real systems have access control mechanisms to avoid this situation. In this sense, we wonder whether it is reasonable to criticize the scalability of the solution in [3] because the amount of UEs served by an AP is not explicitly limited.

B. PROPOSED DEFINITION

From our point of view, the definition of scalability should take into account the system viability in terms of computation and resource requirements as the system increases in size. To do this, we have to discuss first when a system is viable. In this context, we can define viability using the system denial of service probability, i.e., the probability that a UE cannot be served by the system due to the lack of computational, front-haul or radio resources. Let D_k be the denial of service probability for the k-th UE. Then, the denial of service probability of the system $\mathcal{S}$ is defined as

$$P_{\mathrm{D}}(\mathcal{S}) = \lim_{K' \to K} \frac{1}{K'} \sum_{k=1}^{K'} D_k. \tag{3}$$

This definition is valid for both the case $K = \infty$, and the case $K < \infty$. In the latter, this definition can be rewritten as $P_{\mathrm{D}}(\mathcal{S}) = \frac{1}{K} \sum_{k=1}^{K} D_k$. We say that a system is viable if the denial of service probability is below a maximum threshold, i.e., $P_{\mathrm{D}}(\mathcal{S}) \leq P_{\max} < 1$.

Let $\mathcal{S}_m = \{\{\mathcal{L}_m\}, \{\mathcal{K}_m\}\}$ be a subsystem of the system $\mathcal{S}$ composed of the m-th cluster alone, whose denial of service probability is $P_{\mathrm{D}}(\mathcal{S}_m)$. Considering that the m-th cluster is viable, i.e., $P_{\mathrm{D}}(\mathcal{S}_m) \leq P_{\max}$, more clusters can be added without affecting the viability of the system if they are designed to meet the same criterion. This is explained by the fact that the availability of computational, front-haul, and radio resources of a cluster does not depend on the rest of the system. It is worth noticing that with the increase in UEs, these may experience a deterioration of the received signal but this is caused not by a reduction in the radio resources available, but by interference. Considering that for the k-th UE the denial of service probability is $D_k = P_{\mathrm{D}}(\mathcal{S}_{\tilde{m}(k)})$, we say that the system is viable since following (3), we get, $P_{\mathrm{D}}(\mathcal{S}) \leq P_{\max}$. In other words, if we design viable clusters, the overall system composed of those clusters will also be viable, and certainly, scalable.

We now look at the canonical cell-free massive MIMO case, where a unique CPU manages the reception and transmission to all UEs, and collects CSI information from all APs. Taking into account that the CPU computational capabilities, as well as the front-haul resources, are limited, the number

$$R_k^{\mathrm{DL}} = \log \frac{1 + \sum_{i=1}^{\tilde{k}(\tilde{m}(k)-1)} H_{k\tilde{m}(i)}^* Q_i H_{k\tilde{m}(i)} + \sum_{i=k}^{K} H_{k\tilde{m}(i)}^* Q_i H_{k\tilde{m}(i)}}{1 + \sum_{i=1}^{\tilde{k}(\tilde{m}(k)-1)} H_{k\tilde{m}(i)}^* Q_i H_{k\tilde{m}(i)} + \sum_{i=k+1}^{K} H_{k\tilde{m}(i)}^* Q_i H_{k\tilde{m}(i)}}. \tag{2}$$

of UEs with $D_k < 1$ is finite. This means that if $K = \infty$, the result of (3) is 1. In other words, the canonical cell-free massive MIMO system is not viable, and of course, not scalable. Note that this conclusion is valid even if we provide the CPU with infinite computational capabilities, i.e., the front-haul still makes the system not scalable.

Therefore, using our (less restrictive) definition of scalability, clustering ensures the scalability of the system, if cooperation between clusters is not allowed.

IV. CLUSTERING APPROACHES

As shown in Section III, the solution to guarantee scalability is self-presenting: divide and conquer. However, the division of the network into clusters and the posterior allocation of resources should be done in a smart way, trying to take full advantage of the flexibility of cell-free deployments and user-centric solutions. In this sense, our work focuses on the study of the strengths and weaknesses of current clustering solutions with the aim of proposing more advanced solutions to achieve better performance.

In what follows, two clustering approaches are analyzed. The first approach focuses on ensuring that the UEs are closed to the center of the cluster, thus avoiding edge effects. In the extreme case of this approach, the UEs are placed at the center creating a cluster of APs for each individual UE. Due to this, each cluster of APs serves only one UE as shown in Figure 1. This approach requires the highest number of CPUs in a particular area. However, in this case, each CPU can have lower computational capabilities than in other cases where a CPU has to serve several UEs. These clusters can be overlapping. The second approach includes several UEs in each cluster of APs and focuses on using interference cancellation techniques within each cluster to reduce the effect of interference. In the extreme case of this approach, the AP clusters are enlarged as much as possible including all the UEs in their coverage area in order to maximize the amount of canceled interference in the clusters. This produces disjoint clusters as shown in Figure 2. For the same average amount of APs in each cluster, the disjoint cluster approach is the one that requires the lowest number of clusters in a particular area, hence, the lowest number of CPUs. However, in this case, each CPU needs to serve the highest amount of UEs. Therefore, this solution demands more powerful CPUs but, at the same time, a smaller number of them.

It is worth noticing that the channel estimation and signaling overheads do not depend directly on the clustering approach but on the number of APs in each cluster and is, at least, linearly proportional to the total number of UEs served. Synchronization is also a critical aspect. With this respect, in [3], a synchronization mechanism is proposed, which could be used in any of the clustering approaches. With this synchronization mechanism, the overhead is proportional to the number of APs each AP is collaborating with and to the number of clusters that contain each pair of AP.

In the following sections, these two approaches are compared in terms of sum data rate. For this purpose, first, different clustering techniques are analyzed for both uplink and downlink. Then, conclusions are reached on the suitability of using one clustering solution or another, depending on the characteristics of the scenario and the inherent needs of each of the transmission modes.

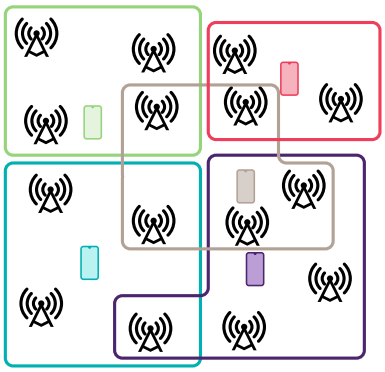

FIGURE 1. Representation of overlapping clusters.

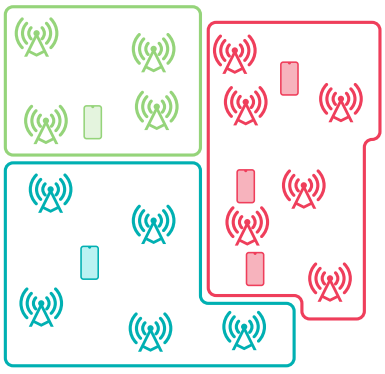

FIGURE 2. Representation of disjoint clusters clusters.

V. ANALYSIS OF THE CLUSTERING SOLUTIONS FOR THE UPLINK

In this section, we present a clustering solution for each of the two clustering approaches described in Section IV for the uplink. Subsequently, we study the performance of the two solutions in a particular scenario to gain insights into the performance of the two clustering approaches.

A. USER-AT-CENTER

This solution is part of the first clustering approach in Section IV. Therefore, each UE is at the center of a custom-designed cluster and therefore does not experience edge effects. The clusters of different UEs can overlap. In [8], the clusters are formed according to the following procedure. The UE appoints the AP with the strongest large-scale fading channel coefficient as its master AP. The maximum amount of UEs that one AP can serve is limited. So, when the UE requests the service, the AP will assign the channel it considers to be the least affected by interference from the other UEs it serves. The master AP informs a limited set of neighboring APs that it is going to serve this UE. In particular,

the informed APs are those whose channel with the UE is at most a "threshold" weaker than the channel of the master AP. Then, the informed APs decide whether they serve the UE or not. To make this decision, the APs have to take into account other UEs they are serving. Using this clustering solution, as many clusters as UEs are created, with the UEs being relatively centered in their respective clusters. This implies that the last UE of the m-th cluster is m, i.e., $\tilde{k}(m) = m$ and the cluster where the k-th UE belongs to is k, i.e., $\tilde{m}(k) = k$. Considering this, the achievable data rate of the k-th UE for uplink and downlink results in, respectively,

$$R_k^{\text{UL-UaC}} = \log \frac{|I + P\sum_{i=1}^{K} H_{ik}H_{ik}^*|}{|I + P\sum_{i=1,i\neq k}^{K} H_{ik}H_{ik}^*|}, \tag{4}$$

$$R_k^{\text{DL-UaC}} = \log \frac{1 + \sum_{i=1}^{K} H_{ik}^* Q_i H_{ik}}{1 + \sum_{i=1,i\neq k}^{K} H_{ik}^* Q_i H_{ik}}. \tag{5}$$

Since this clustering solution creates a different AP cluster for each UE, it is not possible to cancel interference between any of the UEs. Therefore, all the UEs in this scenario are considered interferers by the k-th UE.

B. DISJOINT CLUSTERS

This solution is part of the second clustering approach in Section IV. For this solution, the APs are grouped into non-overlapping clusters of a specific size, i.e., $|\mathcal{L}_m| = N$ for all m. Regarding the UE clusters, the clustering process is as follows. First, the UEs select their master APs as in the previous solution, i.e., the APs with the strongest large-scale fading channel coefficient. Then, the m-th UE cluster is composed of the UEs whose master AP is in the m-th AP cluster. It is clear that, following this clustering process, the UE clusters are composed of several UEs in general. Therefore, since the signals of the UEs in the same cluster are processed by the same CPU, some interference cancellation technique could be used for the decoding/encoding. However, UEs are not prevented from suffering cluster border effects.

C. COMPARISON OF SOLUTIONS FOR UPLINK

In this section, we study the performance of the two previous clustering solutions in a particular scenario. The scenario is a square of 32 by 32 ceiling-mounted APs in a squared grid with an inter-AP distance of 10 m. In order to avoid scenario border effects, a wrap-around technique is implemented. The UEs are randomly distributed in the scenario. The heights of APs and UEs are 6 m and 1.5 m respectively. The channel model used is the "Industrial indoor scenario" presented in [9]. The UEs have 10 mW of available power, and the noise power at the receivers is $1.5887 \cdot 10^{-10}$ mW, which corresponds to the noise power at 15 °C and a transmission bandwidth of 20 MHz. In this scenario, the AP clusters of the "disjoint clusters" (DC) solution are squared groups of APs located in a squared grid over the scenario.

First, the results for several configurations of the "user-at-center" (UaC) solution are obtained. From one configuration to the other the only parameter that changes is the maximum number of UEs that can be served by one AP, denoted by α.

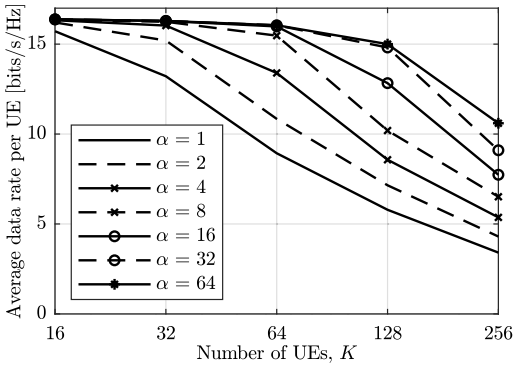

FIGURE 3. Achievable data rate in the uplink for the UaC solution.

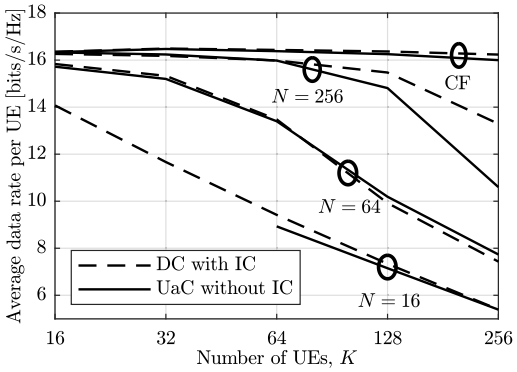

FIGURE 4. Achievable data rate in the uplink fixing the average cluster size. The lines for canonical cell-free massive MIMO (CF) are obtained with only one cluster with and without IC.

As shown in Figure 3, the average data rate per UE decreases as the number of UEs in the scenario increases. This decrement can be softened by allowing each AP to serve more UEs, i.e. by increasing α.

In the case of the UaC solution, the size of the AP clusters is variable. Assuming that the master AP of each UE informs all APs in the scenario to create the AP cluster of this UE, then all APs serve $\min(K, \alpha)$ UEs. Using this assumption, we can compute the average AP cluster size as $L\min(K, \alpha)/K$. For instance, in the case of 256 UEs and $\alpha = 16$, the average cluster size is $1024 \times 16/256 = 64$. In order to facilitate the comparison between the UaC solution and the DC solution for different cluster sizes, we will compare the data rate of certain average cluster size of the UaC solution with the data rate of the same cluster size of the DC solution.

Figure 4 shows the average data rate per UE versus the total number of UEs in the scenario for UaC and DC solutions for cluster sizes of 16, 64 and 256 APs. The figure also shows the average data rate for canonical cell-free massive MIMO with and without interference cancellation as a benchmark. Note that canonical cell-free massive MIMO without interference cancellation is a particularization of the UaC solution, whereas canonical cell-free massive MIMO with interference cancellation is a particularization of the DC solution, both for $N = 1024$. In Figure 4, when cluster sizes of 16 APs and 64 APs or less than 64 UEs are considered, the compared solutions have similar behaviour. However, when considering

clusters of 256 APs and more than 64 UEs, the DC solution clearly outperforms the UaC solution. Objectively speaking, according to what it is shown, there is no solution whose performance prevails over the other for any number of UEs. This motivates the definition of a mixed solution in the following section.

D. MIXED SOLUTION FOR THE UPLINK

This solution is inspired by the idea of merging the best of the previous solutions: the avoidance of border effects and the possibility of applying interference cancellation. The UE and AP clustering is performed as follows. First, the APs are grouped into non-overlapping cluster cores as for the disjoint clusters approach. Then, more APs are added to those clusters. The criterion to follow for such cluster addition could be, e.g., proximity. As a result, the cores will be surrounded by other APs of their clusters. Second, each UE is linked to its best-serving AP, as previously. The cluster of UEs served by each AP cluster is the set of UEs linked to the APs in the corresponding core. The latter implies that the cluster of UEs is centered with respect to the cluster of APs preventing them from experimenting edge effects. Figure 5 illustrates this solution. In the figure, four clusters are delimited by colored solid lines. The core of each cluster is formed by the APs that have the same color as the mentioned lines. The UEs are represented in the color of the cluster they are served by. For the sake of simplicity, for the results in this section and for the same scenario described in the previous section, the AP clusters are assumed to be squares of $\sqrt{|\mathcal{L}_m|} \times \sqrt{|\mathcal{L}_m|}$ APs, whereas the cluster cores are squares of $\sqrt{|\mathcal{L}_m|}/2 \times \sqrt{|\mathcal{L}_m|}/2$ APs at the center of the corresponding cluster, for $m = 1, \ldots, M$.

In Figure 6 the achievable data rate per UE for a cluster average size of 256 APs is shown for the UaC, the DC and the mixed solutions. The curve of the UaC solution was obtained following the procedure explained in Figure 4. As can be noticed, the performance with the mixed solution is better than the UaC and the DC solutions. Therefore, in this case, the best performance is not obtained by only (i) avoiding the cluster border effects or (ii) canceling all the interference of UEs with the master AP in the cluster. In this case, we need to use a solution that mixes the two clustering approaches. In any case, the DC solution is the most similar to the mixed solution. This highlights the importance of canceling interference.

VI. ANALYSIS OF THE CLUSTERING SOLUTIONS FOR THE DOWNLINK

In the uplink, the APs can be considered passive receivers, since they just need to send the signal received to the corresponding CPUs. However, in the downlink, the APs are actively transmitting so the available power needs to be allocated to each of those transmissions. This power allocation becomes more challenging if the clusters are overlapping. In that case, as each AP could be connected to more than one CPU, all the CPUs involved would have to cooperate and this could lead to a non-scalable system. In order to preserve the scalability of the system, suboptimal strategies should be considered. One of the options is to allocate power equally as will be detailed later in Section VI-A for the UaC solution. Another option would be to design a scheme to make a fixed pre-allocation of power and then optimize the power allocation in each cluster. However, there is no clear criterion for such pre-allocation. Due to this and in order to simplify the subsequent analysis, we will not consider a mixed solution, which requires the APs to pre-allocate power to all the clusters they belong to.

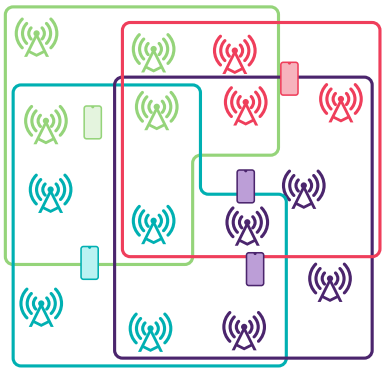

FIGURE 5. Representation of the mixed solution.

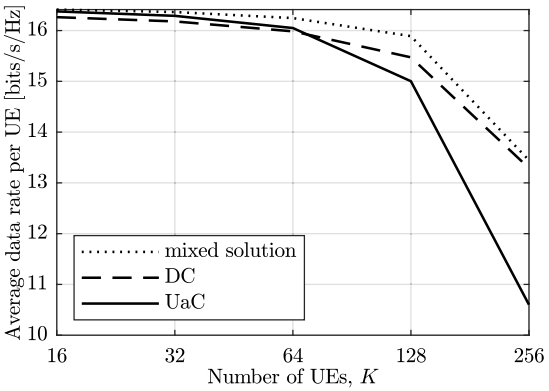

FIGURE 6. Achievable data rate comparison in the uplink for $N = 256$.

A. USER-AT-CENTER

According to [8], in the downlink, the AP clusters are created as explained in Section V-A. We recall that this clustering solution produces overlapping AP clusters that serve only one UE each. Moreover, since the APs serve a maximum of α UEs (or belong to a maximum of α clusters), they split their available power into this amount of UEs. This implies that the APs pre-allocate a certain amount of power for the different clusters they belong to. Letting P be the available power at the APs, they use a maximum of P/α for each UE or cluster [8]. The actual power used to transmit to the UEs is obtained by multiplying P/α by a normalized precoding vector, which, in the case of the k-th UE, is computed as

$$w_k = \frac{H_{kk}}{|H_{kk}|}. \tag{6}$$

Then, the power used by the l-th AP of the k-th cluster to transmit to the k-th UE is

$$\frac{P}{\alpha}|w_{kl}|^2 \leq \frac{P}{\alpha}, \tag{7}$$

which indicates that the total power used to transmit to the UEs by all APs is, indeed, P/α. This power allocation, which is proposed in [8], ensures the APs do not use more power than P, although in a very conservative way. In particular, on average, the APs use significantly less power than P to transmit to all the UEs. In order to illustrate this fact, we assume that each AP serves exactly α UEs. Then, the average amount of APs serving each UE is

$$L_{\text{serving}} = \frac{L\alpha}{K}. \tag{8}$$

As mentioned before, the total power used to transmit to each UE is P/α, which is split into the APs serving this UE. Therefore, the average power used per AP and UE is $P/\alpha L_{\text{serving}}$. Since the APs serve α UEs, the average total power used per AP is

$$\frac{P}{\alpha L_{\text{serving}}}\alpha = \frac{PK}{L\alpha}. \tag{9}$$

We consider here two numerical examples. If $\alpha = 4$, $K = 256$, and $L = 1024$, the average total power used per AP would be $P/16$, and if $\alpha = 8$, $K = 64$, and $L = 1024$, the average drops to $P/128$. This conclusion is quite striking, as it shows that a large part of the available power is unused.

B. DISJOINT CLUSTERS

As in the case of the uplink, we can also consider the creation of disjoint clusters in the downlink with size $|\mathcal{L}_m| = N$ for all m. As discussed before, the resource allocation in each cluster should not be conditioned to the particular allocation performed in neighboring clusters. If not, the effect of each cluster would propagate throughout the scenario, making the resource allocation not scalable. In order to avoid this scalability problem, we propose to perform the resource allocation in each cluster as if they were isolated from the rest, i.e., assuming that the interference from the other clusters do not affect their UEs. Without loss of generality, we are going to particularize this resource allocation for the first cluster, i.e., $m = 1$. In this case, and taking into account that we neglect the interference from other clusters, the achievable rate of the k-th UE is, using (2),

$$R_k^{\text{DL-2}}(\{Q_i\}_{i=1}^{\tilde{k}(1)}) = \log \frac{1 + H_{k1}^* \left(\sum_{i=k}^{\tilde{k}(1)} Q_i\right) H_{k1}}{1 + H_{k1}^* \left(\sum_{i=k+1}^{\tilde{k}(1)} Q_i\right) H_{k1}}, \tag{10}$$

for $k = 1, \ldots \tilde{k}(1)$. The rate expression in (10) is the one used for the resource allocation, i.e., the computation of the transmit covariances $\{Q_i\}_{i=1}^{\tilde{k}(1)}$. However, the actual achievable rate is shown in (2), which takes all the interference into account. The expression in (10) is the one used in the following section to compute the achievable rate in the figures.

To obtain the transmit covariances, we propose to maximize the sum data rate in the cluster taking into account the limited available power in the APs. Mathematically, for the cluster $m = 1$, we want to solve

$$\begin{aligned} \max_{\{Q_k\}_{k=1}^{\tilde{k}(1)}} & \sum_{k=1}^{\tilde{k}(1)} R_k^{\text{DL-2}}(\{Q_i\}_{i=1}^{\tilde{k}(1)}), \\ \text{s.t. } & Q_k \succeq 0, \ k = 1, \ldots, \tilde{k}(1), \\ & \sum_{k=1}^{\tilde{k}(1)} q_{kl} \leq P, \ l = 1, \ldots, N, \end{aligned} \tag{11}$$

where q_{kl} is the l-th element of the main diagonal of Q_k. To find a solution for (11), we can use the algorithm proposed in [10], which is based on [11]. With the optimum transmit covariance matrices in (11), the APs use all their available power. It is worth recalling that this solution is optimum locally, i.e., for one isolated cluster. The use of all the available power can be dangerous in terms of the interference caused to other clusters. Due to this, we also consider a different resource allocation.

For the alternative resource allocation, we consider that the available power can be shared among all the APs in the cluster. By doing this, the optimum covariance matrices contain, in the eigenvectors, the optimum precoding matrices without the limitation of the power constraints. These precoding matrices define the optimum power distribution to maximize the data rate. Due to this, we propose to use a scaled version of these covariance matrices. The scaling is necessary to ensure that the power constraints are met. More specifically, we solve

$$\begin{aligned} \max_{\{Q_k\}_{k=1}^{\tilde{k}(1)}} & \sum_{k=1}^{\tilde{k}(1)} R_k^{\text{DL-2}}(\{Q_i\}_{i=1}^{\tilde{k}(1)}), \\ \text{s.t. } & Q_k \succeq 0, \ k = 1, \ldots, \tilde{k}(1), \\ & \sum_{k=1}^{\tilde{k}(1)} Q_k \leq NP. \end{aligned} \tag{12}$$

Let $\{\bar{Q}_k\}_{k=1}^{\tilde{k}(1)}$ be the optimum solution of (12). In general, these covariance matrices do not satisfy the individual power constraints in each AP. Due to that, we compute

$$Q_k = \frac{P}{\max_l \left(\sum_{k=1}^{\tilde{k}(1)} \bar{q}_{kl}\right)} \bar{Q}_k, \ k = 1, \ldots, \tilde{k}(1), \tag{13}$$

where $\bar{q}_{kl}$ is the l-th element of the main diagonal of $\bar{Q}_k$. The covariance matrices $\{Q_k\}_{k=1}^{\tilde{k}(1)}$ in (13) satisfy the power constraints in each AP and, since they are scaled versions of $\{\bar{Q}_k\}_{k=1}^{\tilde{k}(1)}$, they have the same eigenvectors and, hence, they define the same precoding matrices. The optimization problem in (12) can be solved with the algorithm presented in [10], or with other techniques that make use of the uplink-downlink duality like in [12] and [7].

C. COMPARISON OF SOLUTIONS FOR DOWNLINK

In this section, we study the performance of the two previous clustering solutions in the same scenario described in

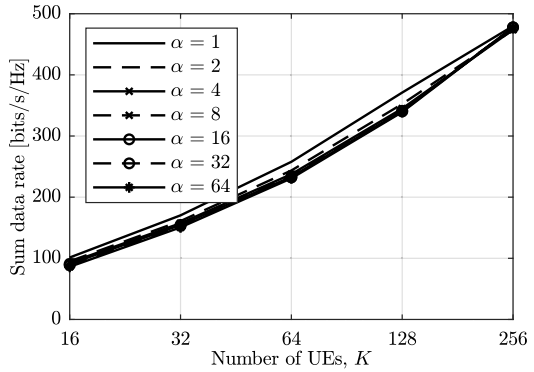

FIGURE 7. Achievable data rate in the downlink for the UaC solution.

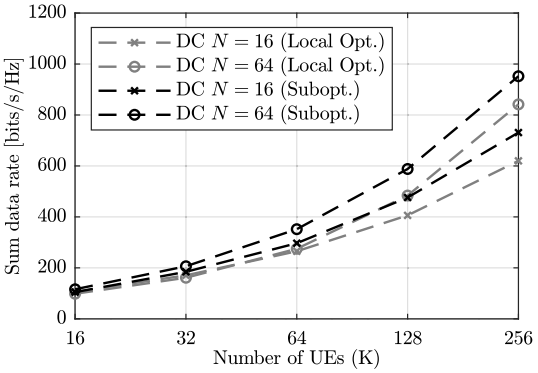

FIGURE 8. Achievable data rate in the downlink for the DC solution.

Section V-C, with the same available power at the APs and the same noise power at the receivers. We start showing the achievable sum data rate with different configurations of the UaC solution. Figure 7 shows the sum data rate versus the number of UEs for the UaC solution for different values of α, i.e., the maximum number of UEs that each AP can serve. As it can be observed, the difference between the curves is not significant, even considering cases where the average total power used per AP is very low (see Section VI-A). This suggests that the system is limited by interference.

Regarding the DC solution, the two power allocation approaches presented are compared for $N = 16$ and $N = 64$ in Figure 8. In the legend, the results corresponding to the cluster-optimal resource allocation are labeled as "Local Opt.", whereas the results of the suboptimal resource allocation are labeled as "Subopt.". The results highlight that, although the optimal resource allocation per cluster (neglecting inter-cluster interference) allows for achieving higher per-cluster rate values, the effect of interference makes the total data rates worse than those obtained with the suboptimal resource allocation algorithm. In other words, by using all the available power in the APs, the sum data rate is more affected by the interference than benefiting from the transmitted power.

Figure 9 shows the sum data rate versus the number of UEs for the best result obtained for UaC, i.e. $\alpha = 1$, the DC solution with suboptimal resource allocation for cluster sizes, N, equal to 4, 16, 64, 256, and the canonical cell-free massive MIMO with DPC. The latter corresponds to the DC solution

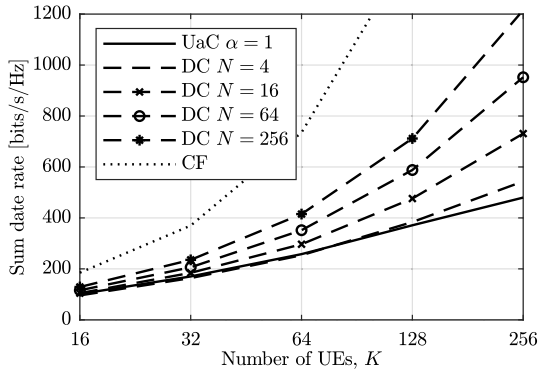

FIGURE 9. Achievable data rate comparison in the downlink.

with $N = 1024$. As can be observed, the DC solution shows a better performance than the UaC solution. Specifically, the performance could improve by 2.5 times for $N = 256$ and 256 UEs in the scenario. Note also that the rate improvement from one cluster size to another is almost constant, e.g., the absolute rate improvement from the $N = 4$ case to the $N = 16$ case is similar to that from the $N = 16$ case to the $N = 64$ case. An exception is the rate improvement from the $N = 256$ case to the $N = 1024$ case. As mentioned before, the latter case corresponds to canonical cell-free massive MIMO. Since, in this case, there is only one cluster, this cluster does not suffer from external interference, and this is the cause of the extra rate gain between the $N = 256$ and the $N = 1024$ cases.

VII. CONCLUSION

In this work, we studied scalability issues of cell-free massive MIMO. After considering a scalability definition presented in the literature, we proposed a new definition that is less restrictive and includes more systems that are intuitively scalable.

We also studied two clustering approaches to design scalable cell-free massive MIMO systems. The first approach is to design clusters for each UE in such a way that the UEs are at the center of their clusters, hence avoiding cluster-edge effects. The second approach is to include several UEs in each cluster to be able to use interference cancellation techniques inside each cluster. In the case of the uplink, none of the two approaches seem to be the best for all cluster sizes and amounts of UEs. This motivated the development of a clustering solution that mixes the two approaches. This solution showed the best performances in all tested configurations. In the case of the downlink, whether or not the clustering solution allows overlapping AP clusters has a major impact on the performance. In particular, in order to avoid a global optimization to make the system scalable, if an AP belongs to several clusters, pre-allocating certain amount of power to each cluster is needed. This is an additional resource management procedure not present in the uplink. This power pre-allocation makes APs underutilize their available power. Due to this, a disjoint clustering solution that follows the second approach, i.e., interference cancellation, provided the best performance.

REFERENCES

[1] 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.

[2] 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.

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

[4] S. Verdu, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[5] D. Calabuig, R. H. Gohary, and H. Yanikomeroglu, "Optimum transmission through the multiple-antenna Gaussian multiple access channel," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 230–243, Jan. 2016.

[6] M. H. M. Costa, "Writing on dirty paper (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 439–441, May 1983.

[7] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

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

[9] G. Interdonato, E. Björnson, H. Quoc Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–3, Aug. 2019.

[10] H. Huh, H. Papadopoulos, and G. Caire, "MIMO broadcast channel optimization under general linear constraints," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2009, pp. 2664–2668.

[11] L. Zhang, R. Zhang, Y.-C. Liang, Y. Xin, and H. V. Poor, "On Gaussian MIMO BC-MAC duality with multiple transmit covariance constraints," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 2064–2078, Apr. 2012.

[12] W. Yu, "Uplink-downlink duality via minimax duality," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 361–374, Feb. 2006.

DANAISY PRADO-ALVAREZ graduated in electronics and telecommunication engineering from the Technological University of Havana "José Antonio Echeverría," Cuba, in 2015. She received the M.Sc. degree in telecommunication technologies, systems, and networks and the Ph.D. degree in telecommunications from Universitat Politècnica de València (UPV), Valencia, Spain, in 2017 and 2022, respectively. In 2018, she joined Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), UPV. Her current research interests include 5G&6G wireless technologies design, channel modeling, network planning, and resource allocation.

DANIEL CALABUIG (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications from Universitat Politècnica de València (UPV), Valencia, Spain, in 2005 and 2010, respectively. In 2005, he joined Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), UPV. During the Ph.D. degree, he participated in some European projects and activities, such as NEWCOM, COST2100, and ICARUS, where he was working on radio resource management in heterogeneous wireless systems and Hopfield neural network optimization. In 2010, he received the Marie Curie Fellowship from the European Commission to research cooperative multipoint transmissions. He visited the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, from 2010 to 2012. In 2012, he returned to the iTEAM and worked with the European projects Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS), METIS-II, and 5G for Connected and Automated Road Mobility in the European Union (5G-CARMEN).

JOSE F. MONSERRAT (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees (Hons.) in telecommunications engineering from Universitat Politècnica de València (UPV), in 2003 and 2007, respectively. He is currently a Full Professor with the Communications Department, UPV, and the Vice-President of UPV. He has been involved in several European projects, especially his significant participation in NEWCOM, PROSIMOS, WINNER+, and METIS/METIS-II, where he led the simulation activities, or 5G-CARMEN on autonomous driving and 5G-SMART on industry 4.0. He also participated in one External Evaluation Group, ITU-R, on the performance assessment of the candidates for the future family of standards IMT-Advanced, in 2010. He co-edited two Special Issues in *IEEE Communications Magazine* on IMT-advanced and 5G technologies and is a co-editor of the book *Mobile and Wireless Communications for IMT-Advanced and Beyond* (Wiley) and *5G Mobile and Wireless Communications Technology* (Cambridge). He manages around €0.5 million yearly budget, holds nine patents, and has published more than 100 journal articles. He is a group head of five postdoctoral fellows, eight Ph.D. students, and two master's students. He has been an advisor to the European Parliament and the World Bank. His current research interests include the design of future 5G wireless systems and their performance assessment. He was a recipient of the First Regional Prize of Engineering Studies for his outstanding student record, in 2003, and the Best Thesis Prize from UPV, in 2008. In 2009, he was a recipient of the Best Young Researcher Prize of Valencia. In 2016, he received the Merit Medal from the Spanish Royal Academy of Engineering, in the young researcher category.

SAMER BAZZI received the B.E. degree in computer and communications engineering from the American University of Beirut, in 2008, and the M.Sc. and Dr.-Ing. degrees in communications engineering from Technische Universität München (TUM), in 2010 and 2016, respectively. From 2010 to 2014, he was a member of DOCOMO Euro-Labs, Munich, Germany, and an External Dr.-Ing. Candidate with the Chair of Signal Processing Methods, TUM, where he worked on coordinated multi-point techniques and precoding for interference channels. From 2015 to 2022, he was a member of the research staff at the European Research Center, Huawei Technologies Duesseldorf GmbH, Munich. His research interests include multiple-input-multiple-output systems, parameter estimation, interference management, and general signal processing techniques for wireless communications.

WEN XU (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Dalian University of Technology (DUT), China, in 1982 and 1985, respectively, and the Dr.-Ing. degree in electrical engineering from Technische Universität München (TUM), Germany, in 1996. From 1995 to 2006, he was with Siemens Mobile (later BenQ Mobile), Munich, where he was the Head of the Algorithms and Standardization Laboratory. As a competence center, the laboratory was responsible for physical layer and multimedia signal processing, and partly protocol stack aspects of 2G, 3G, and 4G mobile terminals, and was actively involved in standardization activities of ETSI, 3GPP, DVB, and ITU. From 2007 to 2014, he was with Infineon Technologies AG (later Intel Mobile Communications GmbH), Neubiberg, focusing on wireline and wireless system concepts/architectures and software/hardware implementations. In 2014, he joined the European Research Center, Huawei Technologies Duesseldorf GmbH, Munich, where he is currently the Head of the Radio Access Technologies Department. His research interests include signal processing, source/channel coding, and wireless communications systems in general. He is a member of the Verband der Elektrotechnik, Elektronik, Informationstechnik (VDE), Germany.