IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 5, 2021, accepted April 12, 2021, date of publication April 20, 2021, date of current version April 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074360

A Framework for Radio Resource Allocation and SDMA Grouping in Massive MIMO Systems

WESKLEY V. F. MAURÍCIO[1,2], DANIEL C. ARAÚJO[1,2], TARCISIO FERREIRA MACIEL[1,2], AND FRANCISCO RAFAEL MARQUES LIMA[1,3], (Senior Member, IEEE)

[1]Wireless Telecommunications Research Group (GTEL), Federal University of Ceará, Fortaleza 60455-760, Brazil
[2]Department of Teleinformatics Engineering, Federal University of Ceará at Pici, Fortaleza 60440-900, Brazil
[3]Department of Computer and Electrical Engineering, Federal University of Ceará at Sobral, Sobral 62010-560, Brazil

Corresponding author: Weskley V. F. Maurício (weskley@gtel.ufc.br)

This work was supported in part by the Innovation Center Ericsson Telecomunicações S.A., Brazil, through Technical Cooperation Contract under Grant EDB/UFC.44, in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), in part by the Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP), and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)-Finance Code 001. The work of Tarcisio Ferreira Maciel was supported by the CNPq under Grant 308621/2018-2. The work of Francisco Rafael Marques Lima was supported by the FUNCAP (edital BPI) under Grant BP4-0172-00245.01.00/20.

ABSTRACT This work proposes a framework for multiuser massive Multiple Input Multiple Output (MIMO) systems which is composed of three parts – *clustering*, *grouping*, and *scheduling* – and aims at maximizing the total system data rate considering Quality of Service (QoS) constraints. We firstly use a clustering algorithm to create clusters of spatially correlated Mobile Stations (MSs). Secondly, in the grouping part, we select a set of Space-Division Multiple Access (SDMA) groups from each cluster. These groups are used as candidate groups to receive Scheduling Unit (SU) in the scheduling part. In order to compose a group, we employ a metric that takes into account the trade-off between the spatial channel correlation and channel gain of MSs. In this context, it is proposed a suboptimal solution to avoid the high complexity required by the optimal solution. Thirdly and finally, we use the candidate SDMA groups from the grouping part to solve the data rate maximization problem considering QoS requirements. The scheduling part can be solved by our proposed optimal solution based on Branch and Bound (BB). However, since it has high computational complexity, we propose a suboptimal scheduling algorithm that presents a reduced complexity. In the simulation results, we evaluate the performance of both optimal and suboptimal solutions, as well as an adaptation of the Joint Satisfaction Maximization (JSM) scheduler to a massive MIMO scenario. Although the suboptimal solution presents a performance loss compared to the optimal one, it is more suitable for practical settings as it is able to obtain a good performance-complexity trade-off. Furthermore, we show that the choice of a suitable trade-off between the spatial channel correlation and channel gain improves the system performance. Finally, for a low number of available SDMA groups, the suboptimal solution presents near optimal outage and a throughput loss of only 10% in comparison to the high-complexity optimal solution while it outperforms the JSM solution in terms of outage and system throughput.

INDEX TERMS Massive MIMO, channel hardening, SDMA grouping, radio resource allocation, quality of service.

I. INTRODUCTION

Nowadays, industry and academy intensified the research over Fifth Generation (5G) networks [1]. The main motivations for its development are the search for better QoS, higher transmit data rates, new services (multimedia) and evolution/massification of digital technology with new and increasingly powerful devices [2].

We highlight here massive MIMO as a technology capable of meeting the data rate requirements of 5G systems [3]. It allows the Base Stations (BSs) to be equipped with tens to hundreds of antennas, enabling them to create many narrow beams to serve simultaneously multiple MSs at the same SU.[1]

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

[1]The SU is the smallest unit of resources that can be allocated to a MS.

This is an attractive feature that especially fits to crowded areas as stadiums or concerts, where there is a very dense MS distribution over a typical area [4]. However, the use of classical digital beamforming in massive MIMO systems leads to low energy-efficiency solutions, as it is necessary that each antenna element has its own Radio Frequency (RF) chain [3]. As an alternative to the classical beamforming, the hybrid analog/digital architecture termed hybrid beamforming is the target of considerable attention nowadays for its reduced cost and power consumption [3]. In this architecture, the number of RF chains is reduced by splitting the signal precoding into analog and digital domains.

Massive MIMO systems need the Radio Resource Allocation (RRA) to manage the scarce available resources of the system (frequency, time, and beams associated to SUs) aiming to increase the system performance e.g., in terms of throughput, fairness, and QoS [5]. The large dimensionality of RRA problems in multi-user massive MIMO systems, on its own, drastically increases complexity and, when combined with the stringent target requirements of 5G, RRA problems become even more difficult to be solved. Therefore, the development of new RRA algorithms is particularly challenging into those scenarios [6].

Another difficulty of RRA in multi-user massive MIMO is that it should be designed jointly with SDMA. Since the scheduling at the same SU of different MSs with nearly orthogonal channels leads to high Spectral Efficiency (SE) [7]. We can classify the search for the best group of MSs to share a SU, i.e., SDMA grouping, into two categories: a solution that spatially splits the MSs into clusters followed by a scheduler that builds an SDMA group by choosing MSs from different clusters as shown in [8]; and a solution that places MSs into the SDMA groups iteratively based on some spatial compatibility metric [7]. However, in both cases the number of possible groups to schedule is high due to its combinatorial characteristics that increase with the minimum between the number of MSs and of transmit antennas [9].

One of major issues in Frequency Division Duplex (FDD) massive MIMO is the Channel State Information (CSI) feedback. The aforementioned work [8] proposes a solution to overcome this issue by employing the hybrid beamforming. Therefore, it is the base for several studies that investigate MS clustering and downlink SDMA group scheduling as [10]–[15]. In [10], the authors proposed an algorithm that performs joint dynamic clustering and CSI acquisition, and a scheduler that selects semi-orthogonal MSs. In [11], the authors used graph theory to propose a clustering and scheduling method. Their solution has polynomial computational complexity and deals with fairness among MSs. In [12], the authors propose a new hierarchical clustering method that builds the groups by merging clusters. In [13], the authors propose a method to jointly optimize the number of clusters, the clustering procedure, and the beamforming strategy. Furthermore, the authors in [12] and [13] propose a scheduler that utilizes a metric based in Signal to Leakage plus Noise Ratio (SLNR) to suppress inter-cluster and intra-cluster interference. In [14], the authors propose a greedy MS scheduling aiming to mitigate the inter-cluster and intra-cluster interference. However, the digital precoders of this method are calculated for each MS until the number of MSs per cluster is reached. Therefore, computational complexity becomes prohibitive as the number of MSs and cluster size increase. In [15] it is proposed a reinforcement learning based scheduling with the objective of maximizing the throughput while guaranteeing fairness among MSs. The main idea is to model the optimization problem as a Markov decision process, where the fairness is modeled in the transitions order. Note that, none of the aforementioned works [10]–[15] considers hybrid precoding.

Since hybrid beamforming gained a lot of attention by allowing the use of massive MIMO, [16]–[18] propose new scheduling methods for massive MIMO under hybrid beamforming architecture. The authors in [16] propose a scheduling based on statistical CSI with the objective of system throughput maximization. The scheduler main idea is to balance the channel gain and spatial channel correlation of the selected MSs leading to a high system throughput. In [17] it is proposed a scheduling based on matrix vectorization aiming at maximizing the system throughput. The scheduler main idea is to pre-select different sets of MSs based on the Pearson correlation coefficient, whose MSs are scheduled afterwards to maximize the throughput. In [18] the authors propose a scheduling based on contextual bandits to maximize the system throughput while satisfying the QoS. The main idea is that the scheduler learns over time the best scheduling strategy and adapts it to maximize the system performance. The authors in [19] propose a scheduling based on genetic algorithm and competitive learning to maximize the system throughput. The work in [20] propose a scheduler that jointly selects the beam and the MSs. The authors propose two schedulers, one based on stable matching and another based on greedy search. However, the works [16]–[18] do not consider the scheduling of multiple SUs.

As previously mentioned, the QoS satisfaction is of utmost relevance in mobile networks. Therefore, the resources need to be managed in a smart way to guarantee minimum QoS and serve more MSs [5], [9]. From the aforementioned works, only [10] and [11] consider QoS requirements. However, different traffic services can have distinct QoS requirements. Therefore, from the operator's point of view, different services should be provided with sustainable quality. Thus, another drawback of those works is that a multi-service scenario is not considered. In [21] the authors study the problem of throughput maximization guaranteeing the QoS requirements considering multiple services. The proposed scheduler, namely JSM, utilizes the derivatives of the sigmoidal function that are dynamically adapted to protect the most prioritized service satisfying the MSs QoS requirements. However, although the authors in [21] consider QoS requirements and multiple services they did not consider a MIMO scenario.

According to the presented literature review, none of the aforementioned works jointly considers hybrid beamforming,

QoS, multiple SUs and a multi-service scenario. Therefore, this work proposes a framework for multiuser massive MIMO systems considering all those issues, which is composed of three parts: *clustering*, *grouping*, and *scheduling*. In this paper, the proposed framework is used to solve the data rate maximization problem considering QoS requirements and a multi-service scenario, where it firstly clusterizes the MSs by means of a classification algorithm that exploits channel covariance matrices [8]. Although the clustering process helps to accomplish the task of finding the most suitable SDMA groups, an exhaustive search over all groups is still infeasible, especially in very dense urban scenarios. Therefore, we select a given number of MSs within each cluster to compose the SDMA groups by solving a quadratic problem that takes into account MS channel gains and spatial correlations [7]. The same set of SDMA groups is employed in all SUs (whole bandwidth) exploiting the channel hardening in massive MIMO systems [22]. Finally, those groups are the input to a scheduling problem aiming at maximizing the total system data rate and satisfying the QoS requirements of the MSs. We show in our results that this approach reduces the intra-cluster interference and, consequently, improves the overall system SE while satisfying a required number of MSs served in each service. Basically, the main contributions of this paper are:

- Proposal of a framework for RRA in massive MIMO systems that is divided into three parts: clustering, grouping and scheduling. The clustering and grouping steps aim at reducing the scheduling search space by creating low-correlated clusters of MSs and, afterwards, exploiting the channel hardening characteristic to generate a suitable set of SDMA groups that are going to be used in the scheduling step. The scheduling step assigns SUs to the SDMA groups generated in the previous step aiming at maximizing a given objective;
- Mathematical formulation of the grouping problem aiming at maximizing the throughput and satisfying QoS constraints considering multiple SUs and reducing the intra-cluster interference. Also, it uses only statistical CSI and exploits the channel hardening effect;
- Adaptation of the problem from data rate maximization considering QoS [9] to a massive MIMO scenario using hybrid beamforming;
- Proposal of an efficient and low complexity solution for the creation of SDMA groups considering the selection of more than one MS per cluster while reducing the intra-cluster interference;
- Proposal of an efficient and low complexity solution for the considered data rate maximization with QoS guarantees and a hybrid beamforming massive MIMO;
- Calculation of the computational complexity of the involved algorithms and their performance evaluation by means of computational simulations.

In Section II, we describe the system model and define the problem. Afterward, we propose a general framework to solve it. In Section III, the step 1 of the framework (clustering procedure) and hybrid beamforming design are shown. After that, Section IV presents the step 2 of the framework (grouping procedure) considering multiple SUs, fairness, and channel hardening, as well as its low complexity solution. Section VI describes the step 3 of the framework (scheduling procedure) and its low complexity solution. Section VII shows the performance evaluation step by step of the proposed framework in comparison with reference solutions. Finally, in Section VIII we present our conclusions.

II. SYSTEM MODEL

We consider the downlink of a cellular multi-user MIMO system based on Orthogonal Frequency Division Multiple Access (OFDMA) composed of a BS serving a set $\mathcal{J}$ of MSs randomly distributed within linearly spaced hotspots with a determined radius, where $|\mathcal{J}| = J$ and $|\cdot|$ denotes the set cardinality. As we are dealing with a multiservice scenario, we assume that the number of services provided by the system operator is S and that $\mathcal{S}$ is the set of all services. We consider that the set of MSs from service $s \in \mathcal{S}$ is $\mathcal{J}_s$ and that $J_s = |\mathcal{J}_s|$. Note that $\bigcup_{s\in\mathcal{S}} \mathcal{J}_s = \mathcal{J}$ and $\sum_{s\in\mathcal{S}} J_s = J$. Moreover, we consider that the BS is equipped with a Uniform Planar Array (UPA) composed of N_t antenna elements.

The MSs are equipped with a single omnidirectional antenna. Therefore, each Transmission Time Interval (TTI), K out of the J MSs are selected to receive data at the same SU. We also consider that the number of RF chains available at the BS is equal to the number of scheduled MSs K. Moreover, $\mathcal{N}$ is the set of available SUs and $|\mathcal{N}| = N_\mathrm{SU}$. Before transmission, for a given SU and TTI, the symbol x_k to be sent to MS k is prefiltered at the BS by the precoding vector $\mathbf{f}_k \in \mathbb{C}^{N_\mathrm{t}\times 1}$. The filtered symbols are then transmitted through the channel associated with the SU. Figure 1 illustrates the considered system model.

The downlink channel vector between the BS and the MS k is denoted by $\mathbf{h}_k \in \mathbb{C}^{N_\mathrm{t}\times 1}$, where this MS k belongs to a given SDMA group. The coefficients in the channel vector of a given SU refer to the middle subcarrier and the first Orthogonal Frequency Division Multiplexing (OFDM) symbol in a TTI, and these channel coefficients are assumed to remain constant within a TTI. Thus, the prior-filtering receive symbol y_k at the k^{th} selected MS is

$$y_k = \mathbf{h}_k^\mathrm{T}\mathbf{f}_k\sqrt{p_k}\,x_k + \sum_{i\neq k\ \ i\in\mathcal{M}} \mathbf{h}_k^\mathrm{T}\mathbf{f}_i\sqrt{p_i}\,x_i + z_k, \qquad (1)$$

where p_k is the power allocated to the k^{th} MS; the second term on the right-hand side of (1) represents the multi-user interference, also known as intra-cell interference, generated by the other $K-1$ MSs sharing the same SU; and z_k is the additive Gaussian noise, which is Independent and Identically Distributed (IID) as $\mathcal{CN}(0,\sigma^2)$, with standard deviation σ.

A. PROBLEM DEFINITION

Note that a smart design of precoders needs to be considered to avoid the degradation of system SE caused by the

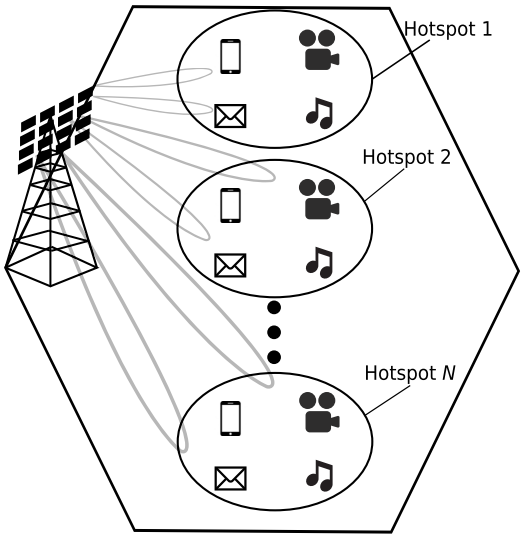


FIGURE 1. System model with massive MIMO system serving K MSs distributed in a set of linearly spaced hotspots, these MSs are using one of the different services provided by the system operator.

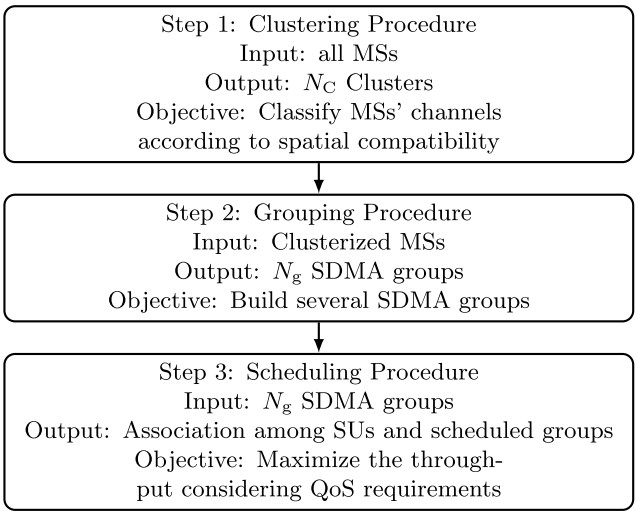


FIGURE 2. Proposed framework composed of three steps: clustering, grouping and scheduling.

intra-cell interference. The creation of SDMA groups containing spatially compatible MSs can avoid poor SE. In this sense, the total number of possible groups assuming J single-antenna MSs and R_f RF chains is given by [9]

$$\sum_{l=1}^{\min(R_f,J)} \binom{J}{l}. \tag{2}$$

The exhaustive search can be used to evaluate the best SDMA group among all possibilities. However, this brute force method leads to high computational costs. Therefore, low-complexity solutions to obtain the SDMA groups are required.

As previously mentioned in this work, we propose a general framework for RRA in massive MIMO systems. The propose framework is divided into three parts and summarized in Figure 2.

The first one, called *clustering procedure*, divides all MSs into N_C clusters, where each cluster contains MSs with similar spatial channel characteristics (correlated channels). This step can make the grouping process easier since, in general, MSs from different clusters will have spatially compatible channels. In this way, a grouping procedure is able to reduce the search space of SDMA groups. However, the resulting number of SDMA groups might still remain impracticable. In the second part, called *grouping procedure*, we select MSs from each cluster to form several SDMA groups. Intelligent strategies to build spectral efficient SDMA groups are employed here since the number of SDMA groups should be limited in order to not increase the complexity of the next step. Therefore, we build a total of N_g SDMA groups according to a metric that will be defined in details in Section IV, i.e., we drastically reduce the number of candidate SDMA groups through clusterization and grouping procedures. Finally, in the third part called *scheduling procedure*, we assign SUs to the SDMAs groups selected in the second step aiming at optimizing a predefined objective. Thus, we allocate SUs to those built SDMA groups aiming to satisfy the QoS requirements and maximizing the total system data rate.

III. CLUSTERING PROBLEM

The goal in this first step is to find MS clusters based on their channel characteristics. This classification is performed based on the knowledge of the long term CSI, which is represented by means of the covariance matrix

$$\Omega_j = \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbf{h}_{t,j}\mathbf{h}_{t,j}^{\mathrm{H}}, \tag{3}$$

where $\mathbf{h}_{t,j} \in \mathbb{C}^{N_t \times 1}$ is the channel vector of MS j at TTI t and τ indicates the number of channel samples considered to estimate the covariance matrix. The eigen decomposition of the covariance matrix is expressed as

$$\Omega_j = \mathbf{D}_j \Lambda_j \mathbf{D}_j^{\mathrm{H}}, \tag{4}$$

where $\mathbf{D}_j \in \mathbb{C}^{N_t \times N_t}$ and $\Lambda_j \in \mathbb{C}^{N_t \times N_t}$ contain the eigenvectors and the eigenvalues of Ω_j, respectively. The BS calculates the dominant eigenvector and eigenvalue of Ω_j and this information is used as input by a classification algorithm.[2] Each cluster has a central characteristic,[3] or centroid, defined by $\boldsymbol{\psi}_i \in \mathbb{C}^{N_t \times 1}$, where i is the cluster index. Specifically, we adopt the K-means algorithm [8]. The algorithm selects N_C out of J MSs to randomly initialize the centroids. Then, the cluster assignment is followed by a centroid update in each iteration. In the cluster assignment step, each MS $j \in \mathcal{J}$

[2]In the simulation presented in this article, we used (3) to estimate Ω_j. However, the estimation process and signaling schemes to make Ω_j available at the BS can be solved using prior art [8].

[3]The central characteristic of a cluster can be modeled as a matrix if the number of chosen eigenvectors is larger than one.

with dominant eigenvector $\mathbf{d}_j$ is assigned to the cluster i based on

$$\arg\min_i \|\mathbf{d}_j - \boldsymbol{\psi}_i\|_2^2, \tag{5}$$

where $\|\cdot\|_2$ is the Euclidean norm.

Then, in the centroid update step, the BS computes new values to $\boldsymbol{\psi}_i$ as the mean of the eigenvectors $\mathbf{d}_j$ of the MSs belonging to that cluster. The algorithm performs the assignment and centroid update steps until it reaches the convergence. Afterwards, the algorithm outputs the MSs clusters.

One specific drawback of K-means is the determination of the number of cluster, N_C. This is an important problem in clustering analysis since most clustering algorithms assume knowing a priori the number of clusters N_C. There is a variety of clustering validation measures and methods in the literature for evaluating clustering algorithms and determining the optimal number of clusters. In [23], the authors evaluate and compare 30 proposed methods to determine the optimal number of clusters. The silhouette index, proposed in [24], shows how good each item (MSs, in our case) is classified. However, although there are some works in the literature that focus on the clustering algorithm design and, therefore, on the optimum number of clusters, our work uses clustering as part of a more complex framework. More specifically, our focus lies on the scheduling and resource allocation, which is performed after the clustering stage. Therefore, for the sake of simplicity, we used the well-known K-means algorithm to solve the clustering problem, which needs (as many other clustering algorithms) as input the desired number of clusters in advance.

It is important to notice that our proposed framework is independent of the clustering algorithm since it only needs to know the clusters and the MSs belonging to them. Therefore, other clustering algorithms, besides K-means, can be used, such as the agglomerative clustering [12] that determines the number of clusters by either the average leakage level or the target number of clusters.

IV. GROUPING PROBLEM

The clustering procedure of Section III reduces the number of possible candidates SDMA groups, since selecting too many MSs from the same cluster does not make sense as they have correlated channels. However, the number of possible SDMA groups still remains unpractical. Therefore, a solution for further reducing the number of possible SDMA groups is necessary. In this section, we formulate a grouping problem that deals with such a challenge. The goal consists of generating N_g spatially compatible groups to maximize the sum-rate and meet QoS constraints. Therefore, instead of the scheduling step go through the whole search space to find the best SDMA group per SU, the grouping step will reduce the number of possible SDMA groups (N_g) that are pre-selected with the objective of maximizing the sum rate.

Assuming that the inter-cluster interference is negligible thanks to clustering and precoding, a grouping problem is applied to each cluster separately where a set of MSs is selected to form SDMA groups. In this selection process, we choose spatially compatible MSs to keep the intra-cluster interference at acceptable levels, if possible. One of the advantages of the grouping step is to evaluate spatial compatibility, without computing the precoding vectors as described in Section III for all possible MSs groups. Moreover, our grouping method exploits the channel hardening characteristic which avoids the computation of SDMA groups for each SU. Therefore, by exploiting channel hardening in the massive MIMO system, where the instantaneous channel gain of each MS can be approximated by its mean in the frequency domain, we can reduce the complexity of this task [22].

A. GROUPING SOLUTION

Let us define the matrix $\hat{\mathbf{D}} \in \mathbb{C}^{J_c \times N_t}$ containing all MSs dominant eigenmodes of a given cluster measured for the SU in the center of the bandwidth (middle SU):

$$\begin{aligned}\hat{\mathbf{D}} &= \left[(\mathbf{d}_1\lambda_1)\ (\mathbf{d}_2\lambda_2)\ \ldots\ (\mathbf{d}_{J_c}\lambda_{J_c})\right]^{\mathrm{T}} \\ &= \left[\hat{\mathbf{d}}_1\ \hat{\mathbf{d}}_2\ \ldots\ \hat{\mathbf{d}}_{J_c}\right].\end{aligned} \tag{6}$$

where λ_j is the highest eigenvalue obtained from Λ_j.

Consider $\mathbf{a} \in \mathbb{R}^{J_c \times 1}$ as the attenuation vector containing the inverse of the dominant eigenmode (channel gain) for the middle SU of all J_c MSs in a cluster. Then, we can express $\mathbf{a}$ using $\hat{\mathbf{D}}$ as:

$$\mathbf{a} = \left[\|\hat{\mathbf{d}}_1\|_2^{-2}\ \|\hat{\mathbf{d}}_2\|_2^{-2}\ \ldots\ \|\hat{\mathbf{d}}_{J_c}\|_2^{-2}\right]^{\mathrm{T}}. \tag{7}$$

Then using (6) and (7), we can write the spatial correlation matrix $\mathbf{C} \in \mathbb{R}^{J_c \times J_c}$ as

$$\mathbf{C} = |*|\sqrt{\mathrm{diag}(\mathbf{a})}\hat{\mathbf{D}}\hat{\mathbf{D}}^{\mathrm{H}}\sqrt{\mathrm{diag}(\mathbf{a})}. \tag{8}$$

Therefore, using (8) and considering N_g as the number of groups to be generated as output of the grouping step, we can define the following block diagonal spatial correlation matrix $\hat{\mathbf{C}} \in \mathbb{R}^{J_c N_g \times J_c N_g}$ as

$$\hat{\mathbf{C}} = \mathbf{I}_{N_g} \otimes \mathbf{C}, \tag{9}$$

where $\mathbf{I}_{J_c}$ is the $J_c \times J_c$ identity matrix and $\otimes$ is the kronecker product.

Analogously, let us define the following attenuation vector $\hat{\mathbf{a}} \in \mathbb{R}^{J_c N_g \times 1}$ as

$$\hat{\mathbf{a}} = \mathbf{1}_{N_g} \otimes \mathbf{a}, \tag{10}$$

where $\mathbf{1}_{N_g}$ is the $N_g \times 1$ vector composed of 1's and $\hat{\mathbf{a}}$ is a concatenation of attenuation vectors from all MSs of each SDMA group. Therefore, the block diagonal spatial correlation matrix $\hat{\mathbf{C}}$ and the stacked vector of channel gains $\hat{\mathbf{a}}$ refers to N_g independent SDMA groups.

Consider the binary selection vector

$$\mathbf{u} = \left[u_1\ u_2\ \ldots\ u_{J_c \cdot N_g}\right]^{\mathrm{T}}, \tag{11}$$

which selects MSs of each SDMA group, where u_i is equal to 1 when the k^{th} MS of a given cluster belongs to SDMA group g with $i = g \cdot N_g + k$. We can formulate a convex combination

TABLE 1. Description of main variables considered in the grouping part.

Variable	Definition
$\hat{\mathbf{D}}$	Concatenation of all MSs eingenmodes of a given cluster measured for the middle SU
$\mathbf{C}$	Spatial correlation matrix for a given cluster
$\hat{\mathbf{C}}$	Block diagonal spatial correlation matrix for a given cluster of each SDMA group
$\mathbf{T}$	Fairness constraint matrix of the SDMA groups
$\mathbf{V}_i$	Matrix to guarantee the diversity of MSs in the SDMA groups
$\mathbf{a}$	Attenuation vector for the middle SU of all MSs in a given cluster
$\hat{\mathbf{a}}$	Block diagonal attenuation vector for a given cluster of each SDMA group
$\mathbf{d}_j$	Dominant eigenvector of MS j
$\mathbf{u}$	Binary MS selection vector
β	Parameter to control the trade-off between spatial correlation and channel gain
λ_j	Highest eigenvalue of MS j
N_g	Number of generated SDMA groups
J_c	Number of MSs belonging to the cluster c
$\bar{J}_c$	Number of MSs selected in each SDMA group in cluster c
$\hat{\Pi}$	Set of MSs that belongs to a SDMA group g
$\Pi_{c,g}$	Set of MSs from cluster c that belongs to SDMA group g

that takes into account both the total spatial correlation and channel gains of all SDMA groups. Such a combination is defined as

$$m(\hat{\Pi}) = (1-\beta)\frac{\mathbf{u}^{\mathrm{T}}\hat{\mathbf{C}}\mathbf{u}}{\|\hat{\mathbf{C}}\|_{\mathrm{F}}} + \beta\frac{\hat{\mathbf{a}}^{\mathrm{T}}\mathbf{u}}{\|\hat{\mathbf{a}}\|_{\mathrm{F}}}, \tag{12}$$

where $\hat{\Pi}$ is the set of MSs belonging to a SDMA group, and $0 \leq \beta \leq 1$ is a parameter to control the trade-off between spatial correlation and channel gain. We also introduced the normalization factors $\frac{1}{\|\hat{\mathbf{C}}\|_{\mathrm{F}}}$ and $\frac{1}{\|\hat{\mathbf{a}}\|_{\mathrm{F}}}$ to balance $\hat{\mathbf{C}}$ and $\hat{\mathbf{a}}$, i.e., to make the β parameter unbiased [7].

Let us define the matrix $\mathbf{T} \in \{0,1\}^{J_c \times J_c N_{\mathrm{g}}}$ as

$$\mathbf{T} = \mathbf{1}_{N_{\mathrm{g}}}^{\mathrm{T}} \otimes \mathbf{I}_{J_c}, \tag{13}$$

where $\mathbf{T}$ is a matrix that we introduce to cope with fairness constraints of the SDMA groups. Indeed each row of the matrix $\mathbf{T}$ is associated to MS j and all SDMA groups, and it will be used to guarantee that each MS will be present in at least one of the N_{g} generated SDMA groups.

Let us define the following matrix $\mathbf{V}_i \in \{0,1\}^{J_c N_{\mathrm{g}} \times J_c N_{\mathrm{g}}}$

$$\mathbf{V}_i = \mathbf{G}_i \otimes \mathbf{I}_{J_c}, \quad \forall i \in \{1, N_{\mathrm{g}}\}, \tag{14}$$

where $\mathbf{G}_i \in \{0,1\}^{N_{\mathrm{g}} \times N_{\mathrm{g}}}$ is a diagonal matrix whose unique non-zero element corresponds to the i^{th} element in the main diagonal and its value is 1; and $\mathbf{V}_i$ is a matrix introduced to guarantee the MS diversity in the SDMA groups as it will be explained later, which helps to satisfy the QoS requirements. The description of variables considered in the grouping problem is shown in Table 1.

Using the definitions above, the multiple groups and fairness optimization problem, which should be solved separately for each cluster, can be formulated as

$$\mathbf{u}^{\star} = \arg\min_{\mathbf{u}} \left\{ (1-\beta)\frac{\mathbf{u}^{\mathrm{T}}\hat{\mathbf{C}}\mathbf{u}}{\|\hat{\mathbf{C}}\|_{\mathrm{F}}} + \beta\frac{\hat{\mathbf{a}}^{\mathrm{T}}\mathbf{u}}{\|\hat{\mathbf{a}}\|_{\mathrm{F}}} \right\} \tag{15a}$$

$$\text{subject to: } (\mathbf{I}_{N_{\mathrm{g}}} \otimes \mathbf{1}_{J_c}^{\mathrm{T}})\mathbf{u} = \mathbf{1}_{N_{\mathrm{g}}}\bar{J}_c, \tag{15b}$$

$$\mathbf{T}\,\mathbf{u} \geq \mathbf{1}_{N_{\mathrm{g}}} \lfloor * \rfloor \frac{N_{\mathrm{g}}}{J_c}, \tag{15c}$$

$$\mathbf{1}_{N_{\mathrm{g}}J_c}(\mathbf{V}_i - \mathbf{V}_g)\mathbf{u} \neq 0, \quad \forall i, g \in \{1, N_{\mathrm{g}}\}, \text{ and } \forall i \neq g, \tag{15d}$$

$$\mathbf{u} \in \{0,1\}^{J_c N_{\mathrm{g}}}, \tag{15e}$$

where $\mathbf{u}^{\star}$ is the solution containing the best N_{g} SDMA groups containing $\bar{J}_c$ MSs of a given cluster that have low total spatial correlation and low total channel attenuation, depending of the chosen β. Constraint (15b) ensures that only $\bar{J}_c$ MSs per group are selected, totalizing a number of $N_{\mathrm{g}}\bar{J}_c$ MSs selected per cluster assuming all SDMA groups. Constraint (15c) ensures that every MS is present in at least $\lfloor * \rfloor \frac{N_{\mathrm{g}}}{J_c}$ SDMA groups, i.e., we impose a fairness constraint among MSs. Constraint (15d) ensures that we do not form groups containing the same MSs, i.e., we impose a variability constraint among SDMA groups. The last constraint assures that $\mathbf{u}$ is binary. Problem (15) can be solved by exhaustive search, which consists by enumerating all the possible SDMA group compositions and choosing the best one. However, this method has impractical computational complexity. Therefore, efficient suboptimal algorithms are required, as presented in the following.

B. PROPOSED ALGORITHM FOR GROUPING

The proposed low complexity solution for the grouping part is presented in Algorithm 1. The main idea here is to select MSs from each cluster to compose all N_{g} SDMA groups that will be used in the scheduling part. In lines 7 to 13, the proposed algorithm firstly selects an initial MS in order to calculate the correlation metric $m(g)$. The chosen MS is the one with highest eigenmode gain. This selected MS will have a low priority to be chosen as the initial MS of other SDMA groups to fulfill constraint (15c). After that, in lines 14 to 18, the algorithm employs a greedy search to find the MS which can form a possible and different combination (constraint (15d)) that minimizes the metric $m(g)$ when the MS is added to the group. Thus, the metric is calculated for each MS that does not belong to the SDMA group formed by the already selected MSs of the cluster. Then, the same procedure is repeated until $\bar{J}_c$ MSs associated with each cluster are selected, as to respect constraint (15b). In line 19, we remove the selected combination from the set of all possible combinations. In lines 20 to 24, if a given MS has already formed all of its possible SDMA groups, then this MS cannot be chosen to compose another SDMA group. These steps are repeated until N_{g} groups are formed for each cluster. The pseudo-code of the proposed algorithm is presented in Algorithm 1.

At this point, it is also important to present a computational complexity analysis for the proposed algorithms. Therefore,

Algorithm 1 Proposed Algorithm for Grouping

Define $\Pi_{c,g}$ as the set of MSs from cluster c that belongs to SDMA group g
for $c = 1$ to $c = N_{\mathrm{C}}$ **do**
  Define $\mathcal{J}_c$ as the set of MSs belonging to cluster c
  Sort $\mathcal{J}_c$ in descending order of dominant eigenvalue
  Define $\mathcal{A}$ as the set with all possible combinations of $\bar{J}_c$ MSs taken from $\mathcal{J}_c$ ▷ Constraint (15d)
  for $g = 1$ until $g = N_{\mathrm{g}}$ **do**
    $\hat{\Pi} = \emptyset$
    Define $\mathbf{u}$ as the binary selection vector for group g
    $\mathbf{u} \leftarrow \mathbf{0}_{J_c}$
    $j^\star \leftarrow$ first element of $\mathcal{J}_c$ ▷ Constraint (15c)
    Put $j^\star$ in the last position in the set $\mathcal{J}_c$ ▷ Constraint (15c)
    $\hat{u}_{j^\star} \leftarrow 1$
    $\hat{\Pi} \leftarrow \{j^\star\}$
    while $\mathbf{1}_{J_c}^{\mathrm{T}}\mathbf{u} \neq \bar{J}_c$ **do** ▷ Constraints (15b)
      $j^\star \leftarrow \arg\min_{j} \left\{ m(\hat{\Pi}) \right\}, \forall \hat{\Pi} \subseteq \mathcal{A}$
      $\hat{u}_{j^\star} \leftarrow 1$
      $\hat{\Pi} \leftarrow \hat{\Pi} \cup \{j^\star\}$
    end while
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{\hat{\Pi}\}$
    for each MS j in $\hat{\Pi}$ **do**
      if $j \notin \mathcal{A}$ **then**
        $\mathcal{J}_c \leftarrow \mathcal{J}_c \setminus \{j\}$
      end if
    end for
    $\Pi_{c,g} \leftarrow \hat{\Pi}$
  end for
end for
return Π

as in [9], we consider summations, multiplications, and comparisons as the most relevant and time-consuming operations. We use the asymptotic notation $\mathcal{O}(\cdot)$ to represent the worst case computational complexity. Thus, in the following we calculate the worst-case computational complexity of Algorithm 1. It sorts the MSs within a cluster based on the dominant eigenvalue. Note that, in the worst case a simple sorting of the J_c MSs can have a complexity of $\mathcal{O}(J_c \log(J_c))$ per cluster. Then, Algorithm 1 selects the first MS, i.e., the MS with highest dominant eigenvalue. Note that, $N_{\mathrm{C}}N_{\mathrm{g}}$ selections are done. Then, the selected MS is excluded from the search for the next MS to compose the SDMA group. If we need to select more than one MS per cluster, i.e., if the number of streams per cluster is greater than one, the algorithm chooses another MS that minimizes the compatibility metric in (12). The process is repeated until the number of MSs selected per cluster is equal to the number of streams per cluster. Based on this, the total number of comparisons considering all the clusters and SDMA groups are $N_{\mathrm{C}}N_{\mathrm{g}}\sum_{i=0}^{\bar{J}_c-1}(J_c - i)$. Therefore, the algorithm worst case complexity is $\mathcal{O}(N_{\mathrm{C}}N_{\mathrm{g}}(1+\sum_{i=0}^{\bar{J}_c-1}(J_c-i)))$, which is polynomial. Note that this solution has a very low complexity compared to solvers based on the well-known BB methods that have exponential complexity [25].

V. HYBRID BEAMFORMING AND DATA RATE

To perform the scheduling part, we need to know how the data rate are calculated and, in consequence, the hybrid beamforming scheme. Therefore, in this section, we present the hybrid precoding scheme and the data rate calculation. We assume that the BS already performed the clustering step and built the SDMA groups. Consider a given SDMA group g that is composed of G_g MSs and consider that $\bar{J}_c$ is the number of MSs from SDMA group g belonging to cluster c. Then, let us define the matrix $\mathbf{E}_c \in \mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{t}}}$ as the average of the eigenvector matrices $\mathbf{D}_j$ (defined in (4)) belonging to the cluster c, i.e.

$$\mathbf{E}_c = \frac{1}{J_{\mathrm{c}}}\sum_{j\in\mathcal{J}_c}\mathbf{D}_j \tag{16}$$

where $\mathcal{J}_c$ is the set of MSs belonging to cluster c and $|\mathcal{J}_c| = J_{\mathrm{c}}$. Thus, let us define $\mathbf{K}_c \in \mathbb{C}^{N_{\mathrm{t}}\times\bar{J}_c}$ as the matrix containing the $\bar{J}_c$ strongest eigenvectors of matrix $\mathbf{E}_c$ for each cluster c as

$$\mathbf{K}_c = \left[\mathbf{e}_{c,1}\ \mathbf{e}_{c,2}\ \ldots\ \mathbf{e}_{c,\bar{J}_c}\right], \tag{17}$$

where the vector $\mathbf{e}_{c,b}$ is the b^{th} strongest eigenvector from matrix $\mathbf{E}_c$ of cluster c. Therefore, $\mathbf{K}_c$ contains the $\bar{J}_c$ best beams of the cluster c.

In the following, we present the computational complexity analysis to obtain the $\bar{J}_c$ strongest eigenvectors of the covariance matrix. We are using the Singular Value Decomposition (SVD) to decompose (4) and, according to [26], the computational complexity to compute the SVD of a $m \times n$ matrix is $O(m^2n + mn^2 + n^3)$. After that, we need to employ a sorting algorithm in the eigenvalue matrix. In general, according to [27], the worst-case computation complexity to sort a vector of size m is $O(m^2)$. Substituting m and n by N_{t}, the computational complexity to obtain the strongest eigenvectors of a MS is $O(N_{\mathrm{t}}^3)$. Therefore, we perform those operations $\bar{J}_c$ times to obtain the $\bar{J}_c$ strongest eigenvectors, which give us a worst case computational complexity of $O(\bar{J}_cN_{\mathrm{t}}^3)$.

Herein, the analog precoder $\mathbf{F}_g^{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{t}}\times G_g}$ from SDMA group g is obtained using (17) for each cluster, and can be written as

$$\mathbf{F}_g^{\mathrm{RF}} = \left[\mathbf{K}_1\ \mathbf{K}_2\ \ldots\ \mathbf{K}_{N_{\mathrm{C}}}\right]. \tag{18}$$

The precoder (18) does not fulfill the constant amplitude constraint of analog beamformers. However, it is possible to implement such a precoder by combining two RF chains for each MS stream, as discussed in [28]. According to [28], this approach achieves the same performance of digital precoding with the requirement that the number of RF chains should be twice the number of spatial streams. There are other methods,

such as those presented in [28] that enable the use of one RF chain per stream up to a negligible performance loss. We choose the simplest approach that consists in constructing the analog beamforming using the phases of each entry in the matrix defined in (18).

Let us define the group channel matrix $\mathbf{H}_g \in \mathbb{C}^{G_g \times N_t}$ of the MSs belonging to the SDMA group g as

$$\mathbf{H}_g = \left[\mathbf{h}_{\zeta_{1,g}}\ \mathbf{h}_{\zeta_{2,g}}\ \ldots\ \mathbf{h}_{\zeta_{G_g,g}}\right]^{\mathrm{T}}, \tag{19}$$

where $\zeta_{k,g}$ is the MS k of SDMA group g. The group channel matrix $\mathbf{H}_g$ and the group analog precoder $\mathbf{F}_g^{\mathrm{RF}}$ form the equivalent channel matrix

$$\bar{\mathbf{H}}_g = \mathbf{H}_g \mathbf{F}_g^{\mathrm{RF}} \ \in \mathbb{C}^{G_g \times G_g}. \tag{20}$$

To suppress the residual inter-cluster interference, we exploit the digital beamformer, that is part of hybrid precoding, using the Zero-Forcing (ZF) digital filter defined as [5]

$$\mathbf{F}_g^{\mathrm{BB}} = \frac{\bar{\mathbf{H}}_g^H (\bar{\mathbf{H}}_g \bar{\mathbf{H}}_g^H)^{-1}}{\|\bar{\mathbf{H}}_g^H (\bar{\mathbf{H}}_g \bar{\mathbf{H}}_g^H)^{-1}\|_{\mathrm{F}}}, \tag{21}$$

where $\|\cdot\|_{\mathrm{F}}$ represents the Frobenius norm.

The total power constraint is enforced by normalizing the digital and analog filters, such that $\|\mathbf{F}_g^{\mathrm{RF}} \mathbf{F}_g^{\mathrm{BB}} \sqrt{\mathbf{P}_g}\|_{\mathrm{F}}^2 = p_{\mathrm{SU}}$, where $\mathbf{P}_g \in \mathbb{R}_+^{G_g \times G_g}$ is a diagonal power matrix with the power allocated to each MS belonging to the SDMA group g and p_{SU} is the transmit power for a given SU. We consider that the number of MSs in the SDMA group is equal to the number of streams. Finally, $\mathbf{F}_g^{\mathrm{RF}}$ and $\mathbf{F}_g^{\mathrm{BB}}$ can be combined to compose the hybrid precoding matrix $\mathbf{F}_g = \mathbf{F}_g^{\mathrm{RF}} \mathbf{F}_g^{\mathrm{BB}} \in \mathbb{C}^{N_t \times G_g}$.

The receive information vector $\hat{\mathbf{y}}_g \in \mathbb{C}^{G_g \times 1}$ of the SDMA group is given by

$$\hat{\mathbf{y}}_g = \mathbf{H}_g \mathbf{F}_g \sqrt{\mathbf{P}_g}\, \mathbf{x}_g + \mathbf{z}_g, \tag{22}$$

where $\mathbf{x}_g \in \mathbb{C}^{G_g \times 1}$ is the group symbol vector and $\mathbf{z}_g \in \mathbb{C}^{G_g \times 1}$ is the group noise vector. The average Signal to Interference-plus-Noise Ratio (SINR) perceived by a selected MS i from group g can be calculated as

$$\Gamma_i = \frac{|q_{i,i}|^2}{\sum\limits_{j \neq i} |q_{i,j}|^2 + \sigma^2}, \tag{23}$$

where $q_{i,j}$ is the element at the i^{th} row and j^{th} column of $\mathbf{Q}_g = \mathbf{H}_g \mathbf{F} \sqrt{\mathbf{P}_g} \in \mathbb{C}^{G_g \times G_g}$ and σ^2 is the noise power. The data rate of MS i is calculated according to Shannon capacity formula and is given by

$$R_i = B \log_2(1 + \Gamma_i), \tag{24}$$

where B is the bandwidth of the SU.

Note that, when RRA is concerned, our presented scenario has similar challenges to the conventional MIMO scheduling, which is combinatorial. However, there are additional issues in our scenario, such as the higher number of antennas and, therefore, the number of multiplexed MSs, as well as the assumption of hybrid beamforming. In this scenario, the data rate of each MS thus depends on the employed analog and digital precoders, which, in their turn, depends on the chosen SDMA group and clusterization. So, there is a hard inter-dependence of scheduling and hybrid beamforming. Consequently, in order to find an optimal solution, it would be necessary to use brute force enumeration to estimate the data rate of each MS at each possible SDMA group, which is impracticable for massive MIMO systems.

VI. SCHEDULING PROBLEM

This section presents the maximization of total data rate considering QoS and a multiservice scenario. This problem has been already studied in [9] for a conventional MIMO system. However, in [9], the authors optimize the system performance by evaluating the possible transmit data rates considering all possible combinations of SDMA groups, as shown in (2), which is impracticable in real systems, especially with massive MIMO.

A. OPTIMAL SOLUTION

Let us define some relevant variables. Assume that $\mathbf{O} \in \{0, 1\}^{N_g \times N_{SU}}$ is an assignment matrix whose element $o_{g,n}$ assumes the value 1 if the SU n is assigned to the SDMA group g and 0 otherwise. Let $\mathbf{R} \in \mathbb{R}^{N_g \times J \times N_{SU}}$ be a tensor whose element $r_{g,j,n}$ is the system data rate of the MS j in SU n if MS j belongs to SDMA group g and 0 otherwise. Let us define the vector $\boldsymbol{\rho} \in \{0, 1\}^{J \times 1}$ as a binary selection vector whose element ρ_j assumes the value 1 if MS j is selected to be satisfied and 0 otherwise. The vector $\mathbf{l} \in \mathbb{R}^{J \times 1}$ is defined as a vector whose element l_j is the required data rate necessary to satisfy MS j. Note that, as in [9], we map the long-term data rate requirements as instantaneous data rate requirements. The minimum satisfaction constraint for each service is defined as a vector $\mathbf{w} \in \mathbb{Z}^{S \times 1}$ whose element w_s is the minimum number of MSs from service s that should be satisfied. Note that, we sequentially dispose the index of MSs in $r_{g,j,n}$ and in l_j according to the service, i.e, the MSs $j = J_{s-1} + 1$ to $j = J_s$ are from service s, where J_s is the number of MSs from service s. The description of variables considered in the scheduling problem is shown in Table 2.

According to the previous considerations, the resource assignment problem can be formulated as the following optimization problem:

$$\max_{\mathbf{x}, \boldsymbol{\rho}} \left(\sum_{g \in \mathcal{G}} \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{J}} o_{g,n} r_{g,j,n} \right), \tag{25a}$$

$$\text{subject to: } \sum_{g \in \mathcal{G}} o_{g,n} = 1, \forall n \in \mathcal{N}, \tag{25b}$$

$$\sum_{g \in \mathcal{G}} \sum_{n \in \mathcal{N}} o_{g,n}\, r_{g,j,n} \geq \rho_j\, l_j, \forall j \in \mathcal{J}, \tag{25c}$$

$$\sum_{j \in \mathcal{J}_s} \rho_j \geq w_s, \forall s \in \mathcal{S}, \tag{25d}$$

$$o_{g,n} \in \{0, 1\}, \forall g \in \mathcal{G} \text{ and } \forall n \in \mathcal{N}, \tag{25e}$$

$$\rho_j \in \{0, 1\}, \forall j \in \mathcal{J}. \tag{25f}$$

TABLE 2. Description of main variables considered in the scheduling part.

Variable	Definition
$\mathcal{J}$	Set containing MSs of the system
$\mathcal{J}_s$	Set containing the MSs using the service s
$\mathcal{N}$	Set containing the available SUs
$\mathcal{S}$	Set containing the available services
$\mathcal{G}$	Set containing the available SDMA groups
$\mathbf{O}$	Binary assignment matrix of SUs and SDMA groups
$\mathbf{R}$	Tensor containing the system data rate of MSs, SU, and SDMA groups
$\boldsymbol{\rho}$	Binary selection vector of the MSs selected to be satisfied
$\mathbf{l}$	Vector containing the MSs required data rate
$\mathbf{w}$	Vector containing the minimum number of MSs necessary to be satisfied per service
$o_{g,n}$	Element of $\mathbf{O}$ that assumes the value 1 if the SU n is assigned to the SDMA group g and 0 otherwise
$r_{g,j,n}$	System data rate of the MS j in SU n belonging to SDMA group g and 0 otherwise
ρ_j	Element of $\boldsymbol{\rho}$ that assumes the value 1 if MS j is selected to be satisfied and 0 otherwise
l_j	Required data rate of MS j
w_s	Required minimum number of MSs of service s

The objective function shown in (25a) is the maximization of the total downlink data rate transmitted by the BS. The first constraint (25b) assures that an SU will not be shared by different SDMA groups. Constraints (25c) and (25d) state that a minimum number of MSs should be satisfied for each service. Problem (25) is a combinatorial optimization problem with linear constraints (25d). Hence depending on the problem dimension, its optimal solution has prohibitive computational complexity [9].

In order to write this problem in a compact form we will represent the problem variables and inputs in vector and matrix forms. Thus, let us define the matrix $\tilde{\mathbf{R}} \in \mathbb{Z}^{J \times N_g N_{SU}}$ as follows

$$\tilde{\mathbf{R}} = \begin{bmatrix} r_{1,1,1} & r_{2,1,1} & \cdots & r_{N_g,1,1} & r_{1,1,2} & \cdots & r_{N_g,1,N_{SU}} \\ r_{1,2,1} & r_{2,2,1} & \cdots & r_{N_g,2,1} & r_{1,2,2} & \cdots & r_{N_g,2,N_{SU}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ r_{1,J,1} & r_{2,J,1} & \cdots & r_{N_g,J,1} & r_{1,J,2} & \cdots & r_{N_g,J,N_{SU}} \end{bmatrix}.$$

Therefore, we can rewrite problem (25) as

$$\max_{\mathbf{O},\boldsymbol{\rho}} \left(\mathbf{1}_J^{\mathrm{T}} \tilde{\mathbf{R}} \operatorname{vec}(\mathbf{O}) \right), \tag{26a}$$

$$\text{subject to: } (\mathbf{1}_{N_g}^{\mathrm{T}} \otimes \mathbf{I}_{N_{SU}}) \operatorname{vec}(\mathbf{O}) = \mathbf{1}_{N_{SU}}, \tag{26b}$$

$$\tilde{\mathbf{R}} \operatorname{vec}(\mathbf{O}) \geq \operatorname{diag}(\mathbf{l})\, \boldsymbol{\rho}, \tag{26c}$$

$$(\mathbf{1}_{J_s} \otimes \mathbf{I}_S)^{\mathrm{T}} \boldsymbol{\rho} \geq \mathbf{w}, \tag{26d}$$

$$\mathbf{O} \in \{0,1\}^{N_g \times N_{SU}}, \tag{26e}$$

$$\boldsymbol{\rho} \in \{0,1\}^{J}, \tag{26f}$$

where the operator $\operatorname{vec}(\cdot)$ maps a matrix to a vector by stacking its columns on top of each other and returns a column vector. Therefore, the original problem is recasted as a standard Integer Linear Problem (ILP) and can be solved using BB methods.

In the following, we calculate the worst-case computational complexity to obtain the optimal solution of problem (26). For an arbitrary number of integer variables υ, the number of linear programming subproblems to be solved is at least $(\sqrt{2})^{\upsilon}$ [25]. Since in problem (26) there are $N_g N_{SU} + J$ integer variables and $N_{SU} + J + S$ constraints, and by retaining only the high order operations, the worst-case computational complexity for problem (26) is $O\left(\sqrt{2}^{(N_g N_{SU}+J)}\right)$. Motivated by this exponential computational complexity, we present in the next section a low-complexity suboptimal solution.

B. LOW COMPLEXITY SOLUTION

We propose a low complexity heuristic algorithm for the scheduling, which is divided into two parts: unconstrained maximization and reallocation. The unconstrained maximization part is responsible for allocating the SUs into groups to maximize the throughput without taking into account the QoS constraints. The reallocation part is responsible for distributing the SUs that have been assigned in the previous part to another group to satisfy the QoS constraints. Flowcharts describing unconstrained maximization and reallocation parts are shown in Figures 3 and 4, respectively.

Before initializing our proposed algorithm, we consider that the achievable data rates of all MSs on all resources when belonging to any SDMA group formed by the grouping problem (15) are known. One way to do this is by calculating the precoders (as explained in Section V), then the SINR and the capacity according to (24). In the unconstrained maximization part, the basic idea is to have a good initial solution that gives us a capacity upper bound. Firstly, in step 1, we define the set of available MSs composed of all MSs that is used along the algorithm. In step 2, we assign the SUs to the SDMA groups with the highest data rate (maximum rate allocation). After that, we define a set with the MSs that have fulfilled their data rate requirements and another set with the ones that are still unsatisfied. If the minimum number of satisfied MSs for all services is fulfilled (according to the constraint (25d)), we have found the optimum solution to problem (25). However, in general, only a few groups get assigned most of the SUs due to the unfairness of the employed assignment.

In case the satisfaction constraint for any service is not fulfilled, a MS of the available MS set will be disregarded. By disregarding a MS, we mean it will not contribute to the reallocation metric (28) at the current TTI, i.e., the algorithm will not try to satisfy this MS. The criterion to select the MS $j^\star$ to be disregarded is given by

$$j^\star = \arg\min_{j \in \mathcal{J}} \frac{\left(\sum_{g \in \mathcal{G}} \sum_{n \in \mathcal{N}} \frac{\Gamma_{g,j,n}}{\kappa_j \cdot N_{SU}} \right)}{l_j}, \tag{27}$$

where $\Gamma_{g,j,n}$ is the SINR of the MS j in SU n belonging to the SDMA group g, κ_j is the number of SDMA groups that MS j belongs to and N_{SU} is the number of available SUs. The adopted criterion to disregard a MS is quite reasonable: we disregard the MS that requires, on average, more SUs to be satisfied. The selected MS is taken out of the available MS set. After that, if the SDMA groups that contain the MS $j^\star$ have only disregarded MSs, then these groups are also disregarded.

The next step is to check whether the service of the MS $j^\star$, chosen using (27), can have another MS disregarded without infringing the minimum number of MSs necessary to satisfy the QoS constraint (25d). If so, we perform the maximum rate allocation considering the remaining SDMA groups. Otherwise, no MSs from this service will be disregarded anymore and all the MSs from this service are taken out of the available MS set. This procedure is repeated until we find a feasible solution, or no MS can be disregarded anymore.

Finally, we check if at least one MS is satisfied. If so, we define the receiver set $\mathcal{R}$ and available resource set $\mathcal{D}$. The receiver set is composed of the unsatisfied MSs, which have to receive SUs from the donors to satisfy their data rate requirements, where the donors are the satisfied MSs, which can donate/share SUs to/with unsatisfied MSs. Finally, the available resource set is composed of all SUs that are assigned to MSs of the donors, and which can be donated/shared to/with MSs of the receiver set. In case there is no satisfied MS after executing the first part, the proposed algorithm is not able to find a feasible solution, i.e., the algorithm is not able to satisfy the constraints of problem (25).

As the unconstrained maximization part does not deal with QoS guarantees, it is necessary to reallocate the SUs that were previously allocated in order to satisfy the QoS requirements. Therefore, in the reallocation part, we exchange SUs among SDMA groups, changing the initial allocation provided by the unconstrained maximization part, to satisfy the MSs from the receiver set.

We start by creating the set $\mathcal{G}$ that is composed of all SDMA groups that contain the MS $j^\star$, which is the most difficult MS to be satisfied from receiver set $\mathcal{R}$. This MS can be found according to (27). The main motivation for choosing the most difficult MS to be served firstly is to assign the minimum number of SUs to satisfy the MSs in an unfavorable situation and assign the remaining SUs to MSs with better channel conditions. After that, we must identify the SDMA groups and SU pairs that are candidate to be chosen in the reallocation procedure. Therefore, the next step is to calculate the number of MSs that each pair of SDMA group from $\mathcal{G}$ and available SU from $\mathcal{D}$ can satisfy. Then, we can compose the set $\mathcal{F}$ containing the pairs of $\mathcal{G}$ and $\mathcal{D}$ that maximize the number of satisfied MSs.

The next step is to define a metric to reallocate an SU to the SDMA group that leads the receivers ($\mathcal{R}$) to satisfaction, while not causing a high SE loss. This can be achieved by the following metric

$$\varphi_{g,n} = \left(\frac{\sum\limits_{j \in \mathcal{R}_{j^\star}} |l_j - (\hat{l}_j + r_{g,j,n} - r_{g',j,n})|}{\sum\limits_{j \in \mathcal{R}_{j^\star}} |l_j - \hat{l}_j|} \right) \frac{\Phi^{cur}}{\Phi^{new}_{g,n} \cdot \pi_g}, \quad (28)$$

where $\mathcal{R}_{j^\star}$ is the set of receivers that belong to the SDMA group g of the chosen MS $j^\star$, $\hat{l}_j$ consists in the required data rate of MS j according to the current resource assignment, Φ^{cur} is the sum of the data rate achieved by all User Equipments (UEs) in all Resource Block (RB) according to the current resource assignment, $\Phi^{new}_{g,n}$ is the sum of the data rate when the SDMA group g receives via reallocation the SU n without modifying the assignment on the other SUs, and π_g is the number of receivers in SDMA group g. This is an adaptation of the reallocation metric used in [9].

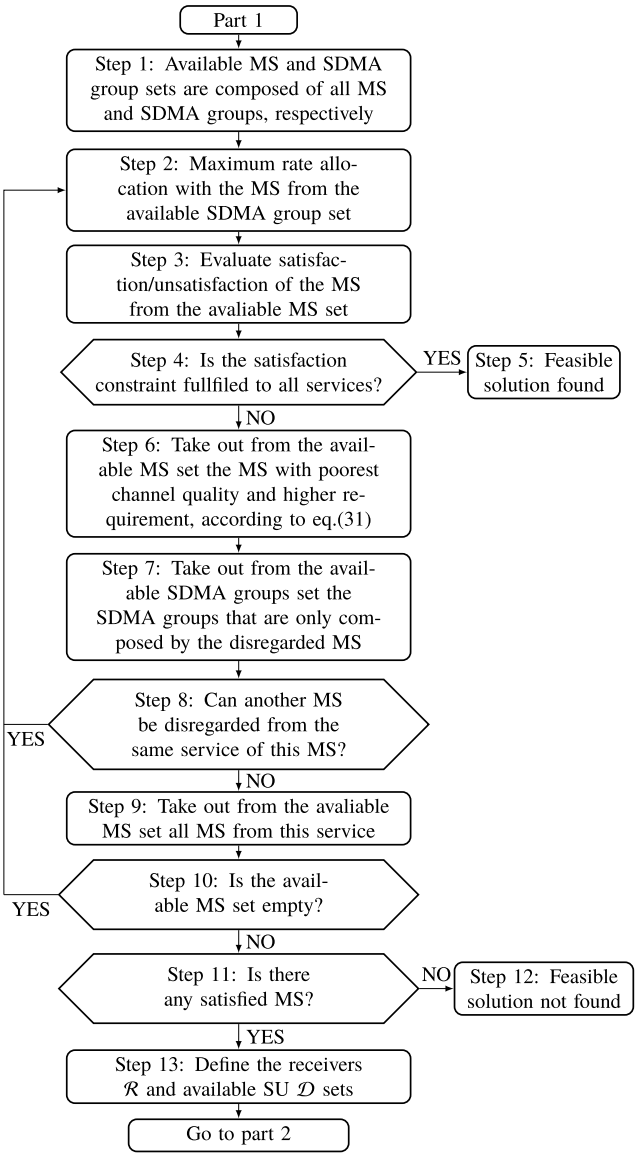


FIGURE 3. Unconstrained maximization part.

We consider that g' is the SDMA group that was chosen to SU n in the first part of the proposed solution (unconstrained maximization). Also, $r_{g',j,n}$ is the data rate of MS j on SU n when present in SDMA group g'. Note that $\Phi^{new}_{g,n} \leq \Phi^{cur}$ since we begin with the maximum rate solution in the unconstrained maximization part of our solution and to satisfy MSs we lose spectral efficiency. The SDMA group and SU chosen in the reallocation part are those that minimize the reallocation metric (28).

The next step is to check whether the reallocation would lead any MS from the donor SDMA group to become unsatisfied. If so, the reallocation is not performed, and the chosen

TABLE 3. Description of main variables considered in the low complexity solution.

Variable	Definition
$\mathcal{R}$	Set with receivers MSs (unsatisfied)
$\mathcal{D}$	Set with available SUs to be donated/shared
$\mathcal{G}$	Set with all SDMA groups that contain the MS j^*
$\mathcal{F}$	Set with the pairs $\mathcal{G}/\mathcal{D}$ that maximize the number of satisfied MSs
$\mathcal{R}_{j^\star}$	Set of receivers that belong to a given SDMA group of the chosen MS $j^\star$
$j^\star$	MS to be disregarded
$\varphi_{g,n}$	Reallocation metric
$\Gamma_{g,j,n}$	SINR of the MS j in SU n in the SDMA group g
κ_j	Number of SDMA groups that the MS j belongs to
N_{SU}	Number of available SUs
$\hat{l}_j$	Required data rate of MS j according to the current resource assignment
g'	SDMA group chosen to a given SU in the unconstrained maximization part
$r_{g',j,n}$	Data rate of MS j on SU n in SDMA group g'
Φ^{cur}	System data rate of the current resource assignment
$\Phi^{new}_{g,n}$	System data rate of the new SU allocation
π_g	Number of receivers in SDMA group g

SU is removed from the available SU set and cannot be chosen anymore. Otherwise, the reallocation is performed, and the MSs' data rates are updated. Then, the algorithm checks whether any receiver has become satisfied after reallocation. If so, these MSs are taken out from the receiver set. After that, the SU that was assigned to a new SDMA group is removed from the available SU set. Then, it is checked whether the MSs of all services became satisfied. If so, the algorithm ends, and a feasible solution is found. Otherwise, the reallocation process should continue. An outage event happens when still exist MSs in the receiver set, and there is no SU for reallocation. The variables used in the algorithms are those already presented in Table 2 and the remaining variables are defined in Table 3.

In the following, we present the computational complexity analysis of the proposed algorithm. Part 1 of the proposed solution in Figure 3 is clearly dominated by the maximum rate allocation which needs $N_{\mathrm{SU}}N_{\mathrm{g}}$ comparisons. Part 2 is dominated by the calculation of the reallocation metric in (28), which is calculated for every available SU and group. This is repeated until the SU set becomes empty. Therefore, as the reallocation metric needs to be calculated for each available SU and MS, the worst-case computational complexity is $\mathcal{O}(N_{\mathrm{SU}}N_{\mathrm{g}} + N_{\mathrm{SU}}^2\, N_{\mathrm{g}})$. Thus, by retaining only the high order operations, the worst-case computational complexity is $\mathcal{O}(N_{\mathrm{SU}}^2\, N_{\mathrm{g}})$.

VII. PERFORMANCE EVALUATION

In this section, we evaluate the framework proposed step by step. In Section VII-A we evaluate the step 1 using the K-means algorithm. In Section VII-B we evaluate the step 2 comparing the Proposed Grouping (G-PROP) algorithm against the Optimal Grouping (G-OPT) solution, considering that the K-means algorithm was utilized in step 1. In Section VII-C we evaluate the step 3 comparing the Proposed Scheduling (S-PROP) algorithm proposed in

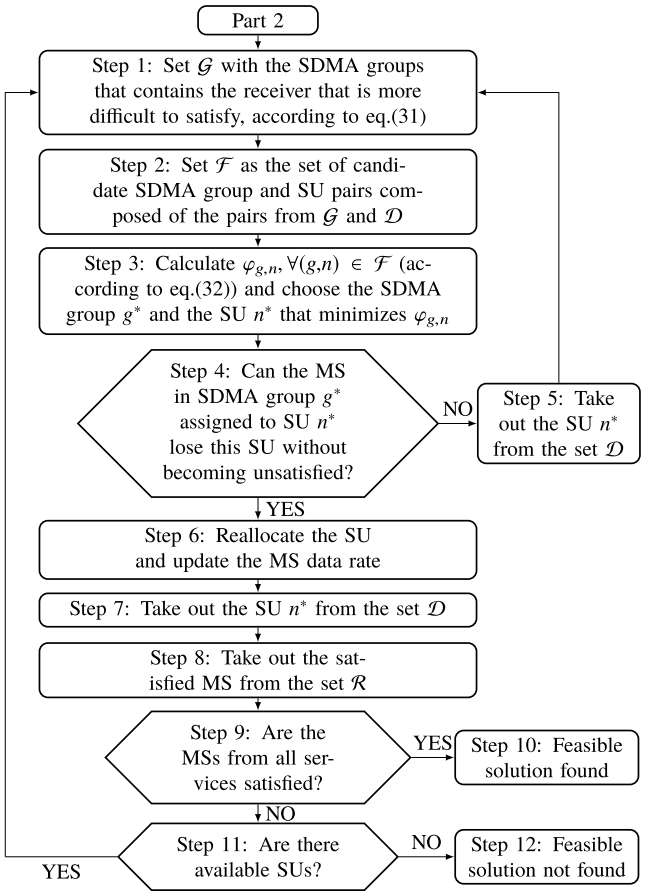


FIGURE 4. Reallocation part.

Section VI-B against the Optimal Scheduling (S-OPT) solution obtained using the CPLEX solver [29] and an adaptation of the JSM algorithm [21]. The JSM algorithm [21] determines the MSs' priority using the derivative of a sigmoidal function. Since the JSM solution needs to estimate the instantaneous data rate, we estimate it by using the dominant eigenvalue and eigenvector that are the CSI available for the others algorithms. Furthermore, in order to deal with the interference among clusters and keep fairness when comparing with our proposed framework, the JSM algorithm is employed after step 1.

The simulation scenario consists in a Urban Micro (UMi) Line Of Sight (LOS) [30] single cell system with an 8×8 UPA ($N_{\mathrm{t}} = 64$). We also assume that the system works with a bandwidth of 100 MHz, a frequency of 28 GHz and that the MSs are equipped with a single-antenna. Based on [31, Table 2], we generate a set of 125 SUs, each composed of 12 equally spaced subcarriers of 60 kHz. Moreover, each frame has 10 subframes carrying 14 symbols each and the TTI duration is 0.25 ms. The considered channel model is the 3-dimensional Quasi Deterministic Radio Channel Generator (QuaDRiGa) [30]. We consider that a set of 40 MSs is equally divided into two groups (forming two circular hotspots) with 15 m of radius. The MSs are uniformly disposed inside

TABLE 4. Simulation parameters.

Parameter	Value	Unit
Simulation time	25	ms
Channel Samples	25	ms
Number of simulation rounds	2000	–
Cell radius	200	m
Total transmit power	35	dBm
Noise figure	9	dB
Noise spectral density	-174	dBm/Hz
Shadowing standard deviation	3.1	dB
Number of MSs	40	-
Number of Clusters	2	-
MSs speed	3	km/h
Total Number of SUs	125	-
Used Number of SUs	25	-
Number of services	2	-
Number of MSs per service	20	-

hotspots, which are linearly distributed in a 60° cell sector.[4] The total power is fixed and Equal Power Allocation (EPA) among SUs and among spatial subchannels is employed. Since we are using hybrid precoding, as many other works in literature [16]–[18], [20], [32], we are considering the number of available RF chains as at most 10% of the number of BS antennas. Also, we are considering that the BS serves 2 or 3 MSs per SU and cluster. Therefore, the BS can simultaneously serve more than 100 MSs since we are considering in our simulations 2 clusters and 25 SUs. Furthermore, the number of simulated MSs is limited by the complexity to obtain the optimal solution that has an exponential computational complexity. Other relevant simulation parameters are listed in Table 4.

A. K-MEANS ALGORITHM EVALUATION (STEP 1)

In Figure 5 we evaluate the effectiveness of the employed clustering (step 1 of the proposed framework) in the proposed scenario with two hotspots. Let us clarify the difference between the terms hotspot and group of MSs, a hotspot is a group of MS in a confined area, and the cluster is the grouping done by the clustering step, which can even select MSs of different hotspots. For this analysis, basically, we increase the angle between the line segments from the BS to the center of the two hotspots. With this, the channel among MSs of different hotspots tends to be more uncorrelated. As the angle increases, we expect that the probability of the clustering algorithm to group MSs of different hotspots decreases. We define the expected clustering difference as the probability of clustering together MSs that do not belong to the same hotspot. Focusing on performance, we can see that as the angle between hotspots increases, the formed clusters get more and more close to the given physical clusters (hotspots).

In Figure 6 we evaluate the step 1 (clustering step) of our algorithm by means of the mean-squared error between the centroids formed in each iteration and those formed when the stop criteria is met (clusters do not change or a maximum

[4]Note that, a UPA with 64 antenna elements radiating with the 3rd Generation Partnership Project (3GPP) antenna model has an effective coverage of a 60° sector.

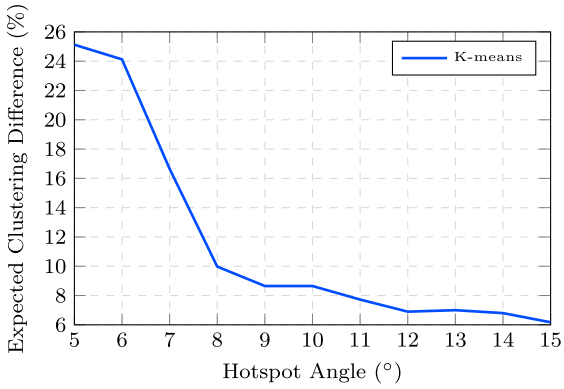


FIGURE 5. Expected clustering difference of K-means clustering for different clusters dispositions.

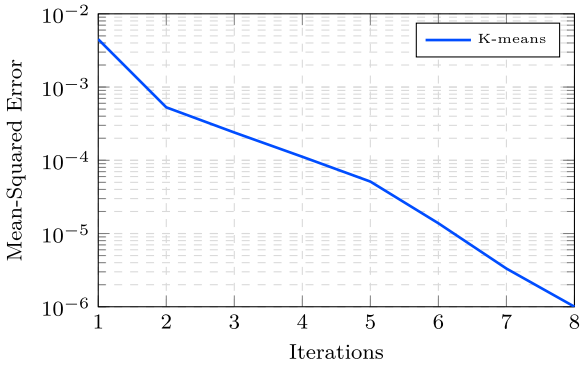


FIGURE 6. Convergence of K-means clustering.

number of iterations). We can see that the mean-squared error decays very fast over the iterations and converge in 8 iterations. This happens because the MSs are already disposed in a defined number of hotspots and an angle of 15° was assumed between clusters, which helps the algorithm to converge. Therefore, from the analyses provided in this subsection, the K-means algorithm can reach a good clusterization with a small number of iterations in the considered scenarios.

B. GROUPING ALGORITHM EVALUATION (STEP 2)

In this section we evaluate the step 2 (grouping step) in terms of the system capacity. The capacity is the cell capacity at the 50th percentile of the Cumulative Distribution Function (CDF). After this step, the SUs are allocated to the SDMA groups aiming at maximizing the system capacity. For the sake of comparison, we implemented the optimal SDMA grouping solution (G-OPT), that is obtained by enumerating all the possible SDMA group compositions and choosing the best one for each SU. Note that, due to the complexity to obtain the optimum solution, we had to reduce the number of SDMA groups of the problem. For this reason, we decided to reduce the number of MSs per cluster. For example, considering 20 MSs in each cluster, 2 clusters and 2 MS served per cluster, the number of possible SDMA groups is 36, 100, which is impracticable. Motivated by this, the performance analysis of the G-PROP against the G-OPT considers a reduced scenario with 2 clusters each one containing 10 MSs.

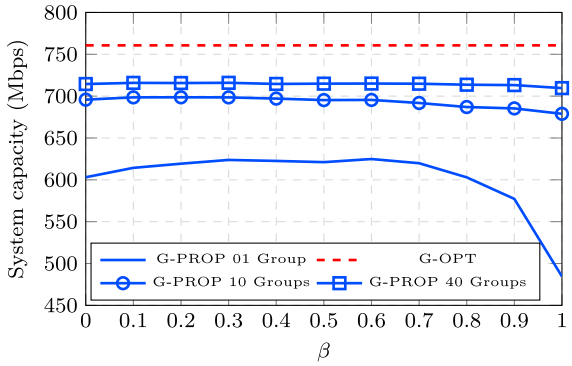


FIGURE 7. System capacity of G-PROP and OPT solutions for a scenario considering 2 MSs selected per cluster and different number of SDMA groups.

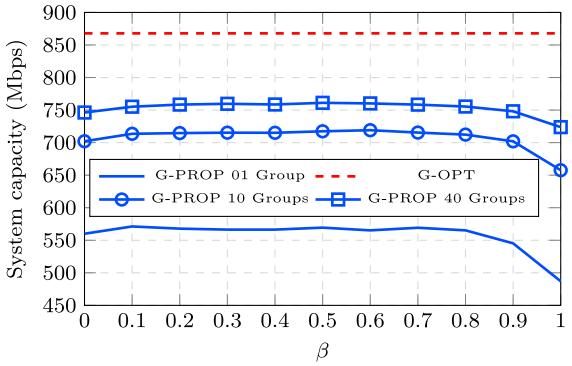


FIGURE 8. System capacity of G-PROP and G-OPT solutions for a scenario considering 3 MSs selected per cluster and different number of SDMA groups.

In Figure 7, we evaluate the step 2 (grouping step) in terms of total system capacity for the G-OPT and G-OPT solutions for a scenario considering 2 MSs selected per cluster when the number of SDMA groups N_g varies. As we can see, the impact of β on the system performance decreases as the number of SDMA groups N_g increases. This behavior happens due to the fairness constraint 25d, which avoids aiming only at maximizing the system capacity, i.e., the algorithm tries to select SDMA groups that include in a balanced way all MSs and not only groups that maximize the SE. Focusing on performance, selecting the best β of each curve, the G-PROP algorithm compared with the optimal solution has a loss of 21%, 8% and 6% for $N_g = 1$, $N_g = 10$ and $N_g = 40$, respectively.

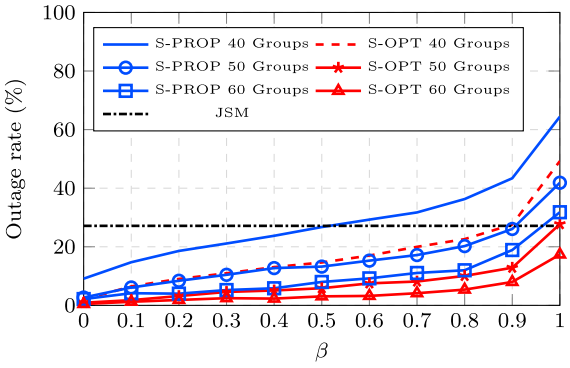
In Figure 8, we evaluate the step 2 (grouping step) in terms of total system capacity for the G-PROP and G-OPT for a scenario considering 3 MSs selected per cluster when the number of SDMA groups N_g varies. Note that, as we increased the number of MSs selected per cluster, the setting of β parameter should be performed more carefully than in the previous scenario (considering 2 MSs selected per cluster). Therefore, differently of Figure 7, the impact of β on the system performance can be seen even considering 40 SDMA groups. Focusing on performance, selecting the best β of each curve, the G-PROP algorithm compared with the G-OPT solution has a loss of 51%, 20% and 14% for $N_g = 1$, $N_g = 10$ and $N_g = 40$, respectively. The reason for the increase in the performance gap between solutions is the increase in the complexity (search space) of the problem, since the search space grows combinatorially with the number of selected MSs per cluster.

The total number of possible SDMA groups evaluated by G-OPT solution are $2,025$ and $14,400$ for the scenarios considering 10 MSs in each cluster serving 2 and 3 MSs per cluster, respectively. Therefore, from the analyses of the results, the G-PROP algorithm achieves good performance even when a very small percentage of the possible SDMA compositions is considered. Furthermore, as shown in Section IV-B, the computational complexity of the G-PROP suboptimal algorithm is polynomial and much lower than that of the G-OPT solution, thus offering a good performance-complexity trade-off.

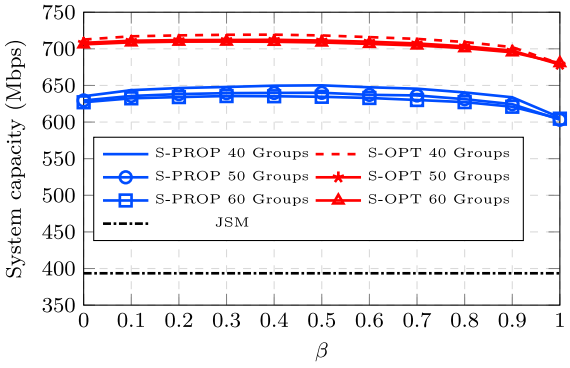
C. SCHEDULING ALGORITHM EVALUATION (STEP 3)

In this section we evaluate the proposed step 3 (scheduling step) considering 40 MSs in the system, 2 clusters, 2 MSs selected per cluster, and 25 available SUs which makes possible to serve all 40 MSs simultaneously. Note that, the total number of scheduled MSs in an SU (4 in this section) is limited by the number of available RF chains. In all figures of this section, the β parameter in (12) varies from 0 to 1. We consider three performance metrics: the total system capacity, the outage rate and the average number of satisfied MSs. Note that, only solutions that are feasible for all β are utilized. An outage event happens when the problem constraints cannot be fulfilled by the algorithm. Note that, the problem itself can be infeasible, depending on the MSs' positions, channel gains, and data rate requirements. Thus, we can define the outage rate as the ratio between the number of outage events and the total number of simulation rounds. Therefore, this performance metric represents if the algorithms are capable of finding a feasible solution to the studied problem. The third and final metric is the ratio between the total number of satisfied MSs and the total number of MSs in the system.

In Figure 9, we evaluate the step 3 (scheduling step) in terms of outage rate and the total system capacity for the S-PROP, JSM and S-OPT solutions when the number of groups N_g varies. As we can see, the selection of β impacts on the system performance achieving different system capacity and outage values as β varies. For example, we can see that the lowest outage rate can be achieved with $\beta = 0$ (selection of MSs with lowest channel correlation within a cluster). This behavior occurs for the rest of the figures in this section. According to this, the spatial correlation cannot be neglected, e.g., $\beta = 1$, since this increases the intra-cluster interference. However, this figure also shows that the best system outage and capacity are achieved for values of $\beta = 0$ and $\beta = 0.5$, respectively, showing that both channel correlation and channel gain should be carefully taken into account, depending on the performance objectives



(a) Outage.



(b) System capacity at 50th percentile.

FIGURE 9. System outage and capacity of our proposed and OPT solutions for a scenario considering 2 MSs selected per cluster, requirement of 5 Mbps per MS, requirements of 100% of satisfied MSs and different number of SDMA groups.

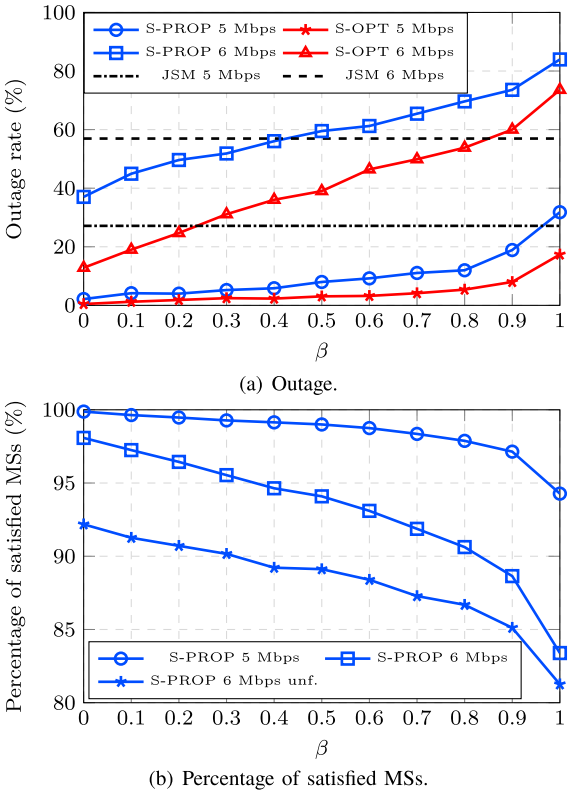
of the system. Moreover, the performance loss of the JSM algorithm in comparison with other solutions is due to the fact that this solution does not take into account the intra-cluster interference. This behavior occurs for the rest of the figures in this section.

As we can see in Figure 9(a), it is possible to reduce the outage rate by increasing the number of candidate SDMA groups. Focusing on the relative performance among algorithms we can see that, when the β parameter is close to 0, the S-PROP algorithm performs near optimally for 50 and 60 groups, while a higher performance loss can be seen for 40 SDMA groups. Therefore, depending on the number N_g of SDMA groups, the grouping procedure can return a solution where the constraint (25d) might be unfeasible or hard to be solved, i.e., as the number of SDMA groups increases the QoS constraints become easier to be fulfilled. However, to increase the number of SDMA groups, it is necessary to build more SDMA groups in the grouping procedure, that in its turn leads to an increase in the search space for the scheduling step and, consequently, the complexity of both procedures increases. Thus, as the number of SDMA groups increases, a trade-off between complexity and performance takes place. Another observation is that we consider only a small fraction of the total number of possible SDMA groups, which for the considered scenario is 91, 390, according to (2). Therefore, it is unpractical to solve the problem considering all possible SDMA groups.

Analyzing Figure 9(b), we can see that the capacity is almost unchanged and does not depend on the number of SDMA groups, i.e., the increase of N_g has more impact on the outage than on the capacity. This can be justified due to the grouping metric (15a), which tries to create SDMA groups that maximize capacity. Due to that, the capacity for the feasible solutions has a similar behavior. Focusing on the relative performance among algorithms, the S-PROP and JSM algorithms have a loss of 10% and 20%, respectively, in comparison to the S-OPT solution.

In Figure 10, we evaluate the step 3 (scheduling step) in terms of outage rate for the S-PROP, JSM, and S-OPT solutions and the average number of satisfied MSs for the S-PROP solution when the required data rate (l_j) varies from 5 to 6 Mbps. As we can see in Figure 10(a), the outage rate increases when the required data rate per MSs increases. Focusing on the relative performance among algorithms, the S-PROP algorithm performs near optimally for the requirement of 5 Mbps, and a performance loss is noted for a requirement of 6 Mbps.

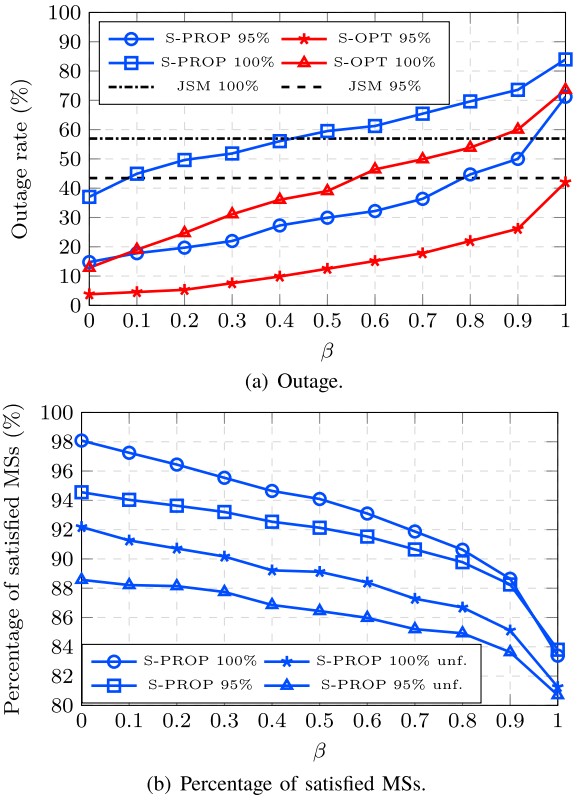
In the next analyses, we evaluate the performance of the S-PROP algorithm in scenarios which do not have a feasible solution, or it is hard to obtain a feasible solution. We denote this case as "unf.". An unfeasible solution happens when the analyzed algorithm is not able to find a solution that satisfies all the constraints of problem (25). Note that, it is interesting to analyze this scenario since an important feature that a QoS constrained RRA algorithm should seek is to provide a good result within the presented circumstances. This "unf." case is compared against the scenario considering all the simulation rounds. Therefore, as we can see in Figure 10(b), even when the S-PROP algorithm is not able to find a solution, it provides a result that satisfies a good number of MSs. Focusing on performance, the average percentage of satisfied MSs is almost 100% for a requirement of 5 Mbps and 98% for a requirement of 6 Mbps considering all simulation rounds. When only unfeasible simulation rounds are considered, the S-PROP algorithm satisfies in average 92% of the MSs for $\beta = 0$, which is reasonable considering the hard nature of the scenario. This happens due to step 3 of part 2 of the proposed RRA algorithm in Figure 4, wherein at each iteration the algorithm tries to satisfy with only one SU reallocation a high number of unsatisfied MSs. Therefore, although in Figure 10(a) we can see a high outage rate when the required data rate is 6 Mbps, most of the MSs are satisfied. Thus, a way to deal with these unsatisfied MSs is to satisfy them in the upcoming TTIs, i.e., the MSs can receive a priority inversely proportional to the data rate obtained until now. Note that, we do not present results for the S-OPT solution regarding the percentage of satisfied MS since the solver only returns a valid solution when the problem is feasible. Therefore, the average percentage of satisfied MSs cannot be analyzed to the S-OPT algorithm. In Figure 10(a) we can see an outage of approximately 0% when we consider the requirements of 5 Mbps, therefore, the number of simulations rounds containing only unfeasible solutions is very low (close to 0).



(a) Outage.

(b) Percentage of satisfied MSs.

FIGURE 10. System outage and satisfaction of our proposed and OPT solutions for a scenario considering 2 MSs selected per cluster, $N_g = 60$, requirement of 100% of satisfied MSs and different data rate requirements.



(a) Outage.

(b) Percentage of satisfied MSs.

FIGURE 11. System outage and satisfaction of our proposed and OPT solutions for a scenario considering 2 MSs selected per cluster, $N_g = 60$, requirement of 6 Mbps per MS and different number of satisfied MSs per service.

Thus an analysis containing only unfeasible solutions cannot be performed for those requirements. Moreover, we skip the analysis of the average percentage of satisfied MSs to the JSM algorithm since it has a higher loss in outage rate in comparison to our proposed algorithm.

In Figure 11, we evaluate the step 3 (scheduling step) in terms of outage rate for the S-OPT, JSM, and S-OPT solutions and the percentage of satisfied MSs for the S-PROP solution when the number of satisfied MSs per service varies. As we can see in Figure 11(a), the outage rate increases with the number of MSs that need to be satisfied. Note that the gap among S-PROP, JSM, and S-OPT solutions is reduced when the number of required satisfied MSs decreases. This behavior is similar to that one seen in Figure 10(a) when the required data rate decreases. Focusing on the relative performance among algorithms, when β is close to 0, the S-PROP and JSM solutions present a higher performance loss. However, despite this higher loss in Figure 11(a), we will see in the sequel that the S-PROP solution is capable of satisfying, on average, almost the target satisfaction percentage.

In Figure 11(b) we present the percentage of satisfied MSs for the scenarios in Figure 11(a) that presented higher performance losses. As in Figure 10, we consider both the case where all simulation rounds are considered and the case where only unfeasible rounds are taken into account, i.e., unf. In both cases, we can see that the S-PROP solution can satisfy almost the required satisfaction target (95% and 100%). In Figure 11(b), when we aim to satisfy 100% and 95% of the MSs, the S-PROP solution is capable of satisfying 98% and 94% of them considering the optimal β value and all simulation rounds. First of all, it is important to clarify the meaning of the outage rate. The outage rate, as already defined previously, consists in the percentage of the simulation rounds in which an algorithm is not able to satisfy the problem constraints, namely, data rate guarantees for MSs and minimum number of satisfied MSs. Consider for example that in a given simulation round, our proposed solution satisfied 95% of MSs but the requirement was to satisfy 100% of them. This simulation round is considered as an outage. In this sense, an outage rate of 15% for a solution means that this solution was not able to satisfy the problem constraints in 15% of the simulation rounds. However, the outage rate does not provide information on how close the algorithm was to meet the required number of satisfied MSs in the simulation rounds that were in outage. This is our idea with the plots with the percentage of satisfied MSs. We plot the percentage of satisfied MSs assuming two cases: considering the average over all simulation rounds (with and without outage) and assuming only the simulation rounds in outage (unf. in the plots). Even if the outage rate is relatively high, e.g., 15%, the percentage of satisfied MSs can be close to the target. This result indicates that even in outage events, our

proposed solution manages to satisfy a high number of MSs that is just below the target. Considering only the unfeasible simulation rounds, we can see that the gap between the target and obtained satisfaction is reduced when the number of required satisfied MSs decreases. As explained before, a way to deal with these unsatisfied MSs is to satisfy them in the next TTIs. We did not simulate this priority mechanism, letting it for future works.

In summary, from the analyses of the results, the proposed S-PROP algorithm achieves good performance compared to the S-OPT solution considering the problem objective and constraints. As shown in Section VI, the computational complexity of the S-PROP suboptimal algorithm is polynomial and much lower than that of the S-OPT solution, thus offering a good performance-complexity trade-off.

VIII. CONCLUSION

In this study, we proposed and evaluated a framework for resource allocation for hybrid precoding massive MIMO communication systems that consists of three parts. First, the MSs were partitioned into clusters containing spatially correlated MSs using the K-means clustering algorithm. The analog part of the hybrid precoder is obtained from the cluster centroids. In this case, it was necessary to consider a low complexity metric to find suitable MSs from each cluster to allocate them to SUs while avoiding the computation of the digital precoders for every possible candidate group of MSs. Secondly, an optimization problem was formulated using a spatial compatibility metric to build SDMA groups. The solution to this problem generates a set of SDMA groups suitable for all SUs exploiting the channel hardening characteristic. Finally, it was necessary to allocate the SUs to SDMA groups to meet QoS requirements while maximizing the data rate. Moreover, a suboptimal algorithm was proposed to solve the scheduling part, and it was compared against the optimal solution.

Simulation results showed that our proposed framework presented a good performance especially in low and moderated system loads. In high loads, even when the proposed algorithm was not able to find a feasible solution, it provided good results in terms of MS satisfaction. We also show that a suitable trade-off between the spatial channel correlation and channel gain should be chosen to improve the system performance. Also, there may have an optimum trade-off to the outage rate and another to capacity, i.e., this choice depends on the system objective. Moreover, the spatial compatibility and channel hardening characteristics can be exploited to reduce drastically the possible number of SDMA groups that need to be built in a system. Furthermore, the proposed suboptimal solution presented a good performance-complexity trade-off.

REFERENCES

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[2] *Requirements Related to Technical Performance for IMT-Advanced Radio Interface(s)*, document ITU-R M.2134, International Telecommunication Union, 2008.

[3] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

[4] H. Yu, H. Lee, and H. Jeon, "What is 5G? Emerging 5G mobile services and network requirements," *Sustainability*, vol. 9, no. 10, p. 1848, Oct. 2017.

[5] E. Castaneda, A. Silva, A. Gameiro, and M. Kountouris, "An overview on resource allocation techniques for multi-user MIMO systems," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 239–284, 1st Quart., 2017.

[6] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo, and P. Soldati, "Learning radio resource management in RANs: Framework, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 138–145, Sep. 2018.

[7] T. F. Maciel and A. Klein, "On the performance, complexity, and fairness of suboptimal resource allocation for multiuser MIMO–OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 59, no. 1, pp. 406–419, Jan. 2010.

[8] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming, user grouping and simplified downlink scheduling," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 876–890, Oct. 2014.

[9] F. R. M. Lima, T. F. Maciel, W. C. Freitas, Jr., and F. R. P. Cavalcanti, "Improved spectral efficiency with acceptable service provision in multiuser MIMO scenarios," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2697–2711, Jul. 2014.

[10] A. Destounis and M. Maso, "Adaptive clustering and CSI acquisition for FDD massive MIMO systems with two-level precoding," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2016, pp. 1–6.

[11] A. Maatouk, S. E. Hajri, M. Assaad, H. Sari, and S. Sezginer, "Graph theory based approach to users grouping and downlink scheduling in FDD massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[12] X. Sun, X. Gao, G. Y. Li, and W. Han, "Agglomerative user clustering and cluster scheduling for FDD massive MIMO systems," *IEEE Access*, vol. 7, pp. 86522–86533, 2019.

[13] J. Chen and D. Gesbert, "Joint user grouping and beamforming for low complexity massive MIMO systems," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2016, pp. 1–6.

[14] G. Wang, J. Zhao, X. Bi, Y. Lu, and F. Hou, "User grouping and scheduling for joint spatial division and multiplexing in FDD massive MIMO system," *Int. J. Commun., Netw. Syst. Sci.*, vol. 10, no. 08, pp. 176–185, 2017.

[15] G. Bu and J. Jiang, "Reinforcement learning-based user scheduling and resource allocation for massive MU-MIMO system," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2019, pp. 641–646.

[16] W. V. F. Mauricio, D. C. Araujo, F. C. H. Neto, F. M. R. Lima, and T. F. Maciel, "A low complexity solution for resource allocation and SDMA grouping in massive MIMO systems," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018, pp. 1–6.

[17] H. Xu, T. Zhao, S. Zhu, D. Lv, and J. Zhao, "Agglomerative group scheduling for mmWave massive MIMO under hybrid beamforming architecture," in *Proc. IEEE 18th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2018, pp. 347–351.

[18] W. V. F. Mauricio, T. F. Maciel, A. Klein, and F. R. M. Lima, "Learning-based scheduling: Contextual bandits for massive MIMO systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[19] J. Jiang, J. Chen, Y. Xie, H. Lei, and L. Zheng, "Modified-PBIL based user selection for multi-user massive MIMO systems with massive connectivity," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Jul. 2020, pp. 1260–1265.

[20] Z. Cheng, Z. Wei, and H. Yang, "Low-complexity joint user and beam selection for beamspace mmWave MIMO systems," *IEEE Commun. Lett.*, vol. 24, no. 9, pp. 2065–2069, Sep. 2020.

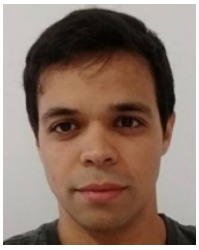
[21] R. P. Antonioli, E. B. Rodrigues, T. F. Maciel, D. A. Sousa, and F. R. P. Cavalcanti, "Adaptive resource allocation framework for user satisfaction maximization in multi-service wireless networks," *Telecommun. Syst.*, vol. 68, no. 2, pp. 259–275, Jun. 2018, doi: 10.1007/s11235-017-0391-3.

[22] V. F. Monteiro, I. L. da Silva, and F. R. P. Cavalcanti, "5G measurement adaptation based on channel hardening occurrence," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1598–1602, Sep. 2019.

[23] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, Jun. 1985, doi: 10.1007/BF02294245.

[24] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 1990.
[25] G. Sierksma, *Linear and Integer Programming: Theory and Practice* (Advances in Applied Mathematics), 2nd ed. New York, NY, USA: Taylor & Francis, 2001.
[26] G. H. Golub and C. F. van Loan, *Matrix Computations*, 4th ed. Baltimore, MD, USA: JHU Press, 2013. [Online]. Available: http://www.cs.cornell.edu/cv/GVL4/golubandvanloan.htm
[27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2009.
[28] D. C. Araujo, E. Karipidis, A. L. F. de Almeida, and J. C. M. Mota, "Hybrid beamforming design with finite-resolution phase-shifters for frequency selective massive MIMO channels," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 6498–6502.
[29] IBM. *IBM ILOG CPLEX Optimizer*. [Online]. Available: https://www.ibm.com/analytics/cplex-optimizer
[30] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.
[31] A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, and I. Da Silva, "Waveform and numerology to support 5G services and requirements," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 90–98, Nov. 2016.
[32] Z. Wang, M. Li, Q. Liu, and A. L. Swindlehurst, "Hybrid precoder and combiner design with low-resolution phase shifters in mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 256–269, May 2018.

WESKLEY V. F. MAURÍCIO received the B.Sc. degree in computer engineering and the M.Sc. degree in electrical and computer engineering with emphasis in telecommunications from the Federal University of Ceará (UFC), Sobral, Brazil, in 2015 and 2017, respectively. He is currently pursuing the D.Sc. degree in telecommunications engineering with UFC, Fortaleza. Since 2014, he has been a Collaborator and a Researcher with the Wireless Telecom Research Group (GTEL), UFC, where he works on projects in cooperation with Ericsson Research. His research interest includes radio resource allocation algorithms for QoS guarantees in scenarios with multiple services, resources, antennas, and users.



DANIEL C. ARAÚJO received the bachelor's degree in telecommunications engineering from the University of Fortaleza (UNIFOR), in 2010, and the M.S. and Ph.D. degrees in teleinformatics engineering from the Federal University of Ceará, Brazil, in 2012 and 2016, respectively. During his studies, he was supported by the Brazilian Agency CAPES and Ericsson. During his Ph.D., he was a Visiting Researcher with Ericsson, in 2013 and from 2015 to 2016. He is currently a Postdoctoral Researcher with GTEL. His research interests include concern to channel estimation, sparse signal processing, massive MIMO, and hybrid beamforming.

TARCISIO FERREIRA MACIEL received the B.Sc. and M.Sc. degrees from the Federal University of Ceará (UFC), in 2002 and 2004, respectively, and the Dr.-Ing. degree from the Technische Universität Darmstadt (TUD), Germany, in 2008, all in electrical engineering. Since 2001, he has been actively participated in several projects in a technical and scientific cooperation between the Wireless Telecom Research Group (GTEL), UFC, and Ericsson Research. From 2005 to 2008, he was a Research Assistant with the Communications Engineering Laboratory, TUD. Since 2008, he has been a member of the Postgraduation Program in teleinformatics enginnering, UFC. In 2009, he was a Professor of computer engineering with UFC-Sobral, and has also been a Professor with the Center of Technology, UFC, since 2010. His research interests include radio resource management, numerical optimization, and multiuser/multiantenna communications.

FRANCISCO RAFAEL MARQUES LIMA (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering, and the M.Sc. and D.Sc. degrees in telecommunications engineering from the Federal University of Ceará, Fortaleza, Brazil, in 2005, 2008, and 2012, respectively. In 2008, he has been in an internship at Ericsson Research, Luleå, Sweden, where he studied scheduling algorithms for LTE system. Since 2010, he has been a Professor with the Department of Computer Engineering, Federal University of Ceará, Sobral, Brazil. He is also a Researcher with the Wireless Telecom Research Group (GTEL), Fortaleza, where he works at projects in cooperation with Ericsson Research. He has published several conference papers and journal articles, and patents in the wireless telecommunications field. His research interests include radio resource allocation algorithms for QoS guarantees in scenarios with multiple services, resources, antennas, and users.