IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 8, 2021, accepted October 30, 2021, date of publication November 2, 2021, date of current version November 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3124871

A Physical Layer Multicast Precoding and Grouping Scheme for Bandwidth Minimization

FRANCISCO J. MARTÍN-VEGA, FARSHAD ROSTAMI GHADI, F. JAVIER LÓPEZ-MARTÍNEZ, (Senior Member, IEEE), AND GERARDO GÓMEZ

Communications and Signal Processing Laboratory, Instituto Universitario de Investigación en Telecomunicación (TELMA), and CEI Andalucía TECH, ETSI Telecomunicación, Universidad de Málaga, 29010 Málaga, Spain

Corresponding author: Francisco J. Martín-Vega (fjmvega@ic.uma.es)

This work was supported in part by the European Fund for Regional Development (FEDER), Junta de Andalucía; and in part by the University of Málaga under Project P18-RT-3175, Project P18-TP-3587, Project UMA-CEIATECH-06, and Postdoctoral Grant DOC_00265 ("Selección de Personal Investigador Doctor Convocado Mediante Resolución de 21 de Mayo de 2020," PAIDI 2020).

ABSTRACT Physical layer multicasting exploits multiple antennas at the transmitter side to deliver a common message to a group of K users. To this end, two formulations have been well addressed in the literature: i) the max-min-fair criterion, which maximizes the signal-to-noise ratio (SNR) of the worst user for a fixed transmit power; and ii) the quality of service (QoS) formulation, which minimizes the transmit power subject to a target SNR. Nevertheless, it is known that the performance and complexity of these approaches is severely degraded as the group size grows. In this paper, we propose a different formulation that aims at minimizing the required bandwidth needed to provide the multicast service. This is achieved by dividing the users into smaller groups and assigning the bandwidth required to provide a target rate to each group. Contrary to the common belief, it is shown that dividing the users into different groups that use orthogonal bandwidth allocations can lead to a smaller aggregated bandwidth than the single-group with single bandwidth allocation counterpart, if an intelligent grouping scheme is used. An iterative algorithm to derive the optimal number of groups is presented with an stopping criterion to reduce the numerical complexity. It is shown through simulation that our proposed approach greatly reduces the required bandwidth compared to existing schemes that rely on single bandwidth allocation. Interestingly, results reveal that our proposed scheme also leads to a greater SNR for a randomly chosen user, and it reduces the variance of the required bandwidth, which eases the implementation in real networks.

INDEX TERMS 5G networks, multicast precoding, user grouping, clustering.

I. INTRODUCTION

The increasing request and popularity of on-demand video and broadcast-like applications for smartphones has fueled an intensive area of research and standardization activities. This wave was initiated with Long Term Evolution (LTE) under multimedia broadcast/multicast service (MBMS), multicast/broadcast single frequency network (MBSFN) and single cell-point to multipoint (SC-PTM) technologies; and it is now under discussion for the Fifth Generation (5G) New Radio (NR), which has a work item targeted for Release 17 on June 2022 [1], [2].

MBMS defines *MBMS service areas*, that are composed by a number of cells that announce a list of available broadcast services. user equipments (UEs) can subscribe to these services to receive broadcast/multicast data [3]. Since both channel state information (CSI) reports and hybrid automatic repeat request (HARQ) protocols are not supported by this technology, achievable rates provided by multicast services are limited. To overcome this limitation, MBSFN (also known as *enhanced* MBMS) was introduced in Release 9. With this scheme, *MBMS service areas* are divided into *MBSFN service areas*, which are composed by a group of cells that are time/frequency synchronized to mitigate inter-cell interference. While this technology efficiently increases the achievable rates, it lacks of flexibility since the service areas and the resources are allocated statically [4].

SC-PTM solves these issues, since it allows a flexible resource allocation on a per-cell basis. Although this technology does not support CSI reports nor HARQ on

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

current releases of the LTE standard, it has been shown that SC-PTM outperforms MBMS and MBSFN in some scenarios [4]. More importantly, SC-PTM might support CSI reports on 5G, which would open the door to great performance improvements mainly for two reasons. Firstly, broadcast/multicast precoding could be exploited at the transmitter side [2], [5] to determine a sub-optimal beamforming vector for the intended group of users. Secondly, multicast channel-aware user grouping and resource allocation can be used to maximize a given metric, i.e., system throughput or fairness among users.

A. RELATED WORK

There has been an extensive area of research on communication and signal theory communities to determine the optimal broadcast/multicast precoding, which is known as physical layer multicasting [6]. Attending to the specific scenario under consideration, existing studies focus either on single-group (i.e., broadcast) precoding [7]–[10] or multi-group (i.e., multicast) precoding [11]–[15]. The former involves that a single stream is delivered to a group of K UEs, using N transmit antennas, in the same time/frequency block of resources. The latter considers a type of group spatial division multiple access, where G groups are served in the same block of resources, but a different stream is transmitted to each group. Although multi-group multicast schemes have a greater flexibility, single-group multicast is normally preferred since it does not suffer from inter-group interference and it is simpler than the multi-group counterpart [6].

The determination of the optimal broadcast/multicast precoding admits two formulations [7]. The quality of service (QoS) formulation considers the minimization of the transmit power subject to a target signal-to-noise ratio (SNR) that must be fulfilled by each UE. An alternative formulation is the max-min-fair, which aims at maximizing the SNR of the worst UE, this being the key factor that limits the performance of the whole group, subject to a power constraint per UE.

Achieving the upper bound given by the multicast capacity requires precoding with high rank transmit covariance matrices, which is not feasible in practice [16]–[18]. For this reason, sub-optimal solutions that restrict to unit rank (i.e., transmit beamforming precoding) are widely adopted as single-group multicast schemes [7]–[10], [19]. Despite restricting to unit rank precoding, the aim of these works is to reduce the numerical complexity, since the computation of the multicast precoding requires treating the channel of a high number of users jointly.

A pioneering work is described in [7], where it is shown that max-min-fair and QoS formulations are equivalent NP-hard problems, which can be expressed as a non-convex quadratically constrained quadratic program (QCQP). The authors propose a semi-definite relaxation (SDR) programming, which relaxes the non-convex unit rank constraint to have a convex problem that can be solved by semi-definite programming (SDP) followed by relaxation and Gaussian randomization (SDR-G). It is shown that solving the SDP problem leads to the upper bound on the min SNR, whereas the SDR-G yields good sub-optimal results close to the upper bound.

The performance of SDR-G deteriorates as N and K grows, and this motivated a number of research works to propose better approximations to the multicast beamforming problem. One of the best solutions in terms of performance is the successive linear approximation (SLA) algorithm, which is proposed for QoS problem in [20]. This approach involves an iterative algorithm where the non-convex constraints are linearized at each iteration by using first-order Taylor series expansion. The resulting convex problem is solved and the obtained vector is used in the next iteration. As shown in [9] with simulations, SLA outperforms SDR-G, although at the expense of a higher computation time.

A different approximation for the QoS problem is presented in [21]. In this work, a low-complexity algorithm based on QR decomposition and channel orthogonalization is proposed. The proposed algorithm is shown to provide a better performance than SDR-G when $K \gg N$ with a smaller complexity.

On the other hand, the case of max-min-fair problem is addressed in [22], where the non-convex part of the problem is replaced with an equivalent non-convex bilinear trace constraint, that is solved with alternating maximization (AM). It is shown that AM leads to a greater min SNR than SDR-G, but at the cost of a higher computational complexity.

The computational complexity is highly reduced in [19], which presents an appealing algorithm named successive beamforming (SB). This algorithm exhibits a high performance with moderate and small number of users and it reaches the upper bound for the case of $K = 2$. The algorithm performs orthogonalizations of the subspace spanned by each user's channel vector in an iterative fashion until there are no more spatial degrees of freedom left, which results in a reduced number of iterations (i.e., $\min(M, K)$).

An interesting approach that achieves a good tradeoff between performance and complexity is proposed in [9] where class adaptive algorithms are developed. At each iteration, the beamforming vector is updated in the direction of an inverse SNR weighted linear combination of the SNR-gradient vectors of all the users. It is shown that the proposed algorithms feature guaranteed convergence and state-of-the-art performance at low complexity.

An algorithm to find the global solution is proposed in [23]. The algorithm is based on branch-and-bound strategy, combined with a new argument-cut technique that is used to design convex relaxations of non-convex constraints. Simulation results show that the proposal greatly outperforms state-of-the art techniques when N and K are high, but the computational complexity makes this approach unfeasible when the channel varies quickly.

All of the above works focus on multicast precoding given a fixed number of groups; nevertheless, user grouping and resource allocation offer additional degrees of freedom

that can be used to improve the system performance. The resource allocation and user grouping problems are often posed as maximization problems with an extremely large solution space [24], [25], which makes low-complexity solutions specially appealing. In [26] an opportunistic multicast scheduling (OMS) algorithm is proposed to exploit multi-user diversity and to increase aggregated throughput. Nevertheless, OMS does not guarantee fairness among multicast users. The fairness between multicast and unicast users is addressed in [27]–[29], which is important from the operators perspective, since it maintains a good balance between both services. Another strategy is to form groups that maximize the aggregated throughput [30]–[33]; however, these works restrict to single antenna case and do not consider multicast precoding. User grouping has been also studied for multi-antenna systems with hybrid beamforming. In this context, different user grouping strategies based on channel sparsity in the beam domain has been proposed in [34]–[37] for other applications such as physical layer security or wireless information and power transfer (SWIPT). These methods assume a spatial basis expansion model, e.g., beam-domain and angular-domain, which compress the dimension of the channel. This reduces the complexity of related tasks such as channel estimation, suppression of pilot contamination or user grouping. Here, users are grouped based on the similarity of the angle of arrivals [36] or the active beam domain sets to eliminate inter-group interference [35].

B. CONTRIBUTIONS

Despite of their relevance, none of the aforementioned works have addressed two paramount aspects: i) the problem of bandwidth minimization; and ii) the interactions between multicast precoding and user grouping.

The former aspect is closer to the problem that face operators and vendors in practical deployments, which aim to deliver a broadcast/multicast service with a given rate, using as less resources as possible to maximize their profits. Regarding multicast services, this bandwidth minimization problem has been only addressed for core networks [38]–[40]. These works propose different techniques to group service demands of the same content, thus saving bandwidth in the backbone. As for wireless access networks, the bandwidth minimization problem has been investigated only for unicast transmission, e.g., with non-orthogonal multiple access (NOMA), [41], [42] and wireless relaying scenarios [43]. Nevertheless, to the best of the authors' knowledge, this problem has not been investigated yet for physical layer multicast on wireless access networks.

The latter aspect is related to the fact that physical layer multicast has been mainly investigated under two different directions: i) multicast precoding; and ii) user grouping/resource allocations. On the one hand, the works that focus on multicast precoding normally assume fixed groups. On the other hand, the works that address user grouping and resource allocations either ignore the multi-antenna setups or assume simplified models that do not capture the dependence between the user group and the performance of multicast precoding.

These reasons motivated us to investigate the interactions between user grouping/resource allocation and multicast precoding. As a result, we propose an scheme that minimizes the required bandwidth to deliver a given broadcast/multicast service. The contributions of the present work can be summarized as follows:

- We propose a new problem formulation for multicast precoding and grouping. The proposed formulation aims at determining a user partitioning into groups and a multicast precoding for each group that minimizes the required bandwidth.
- We propose a novel low complexity algorithm for the above problem named adaptive multicast grouping (AMG). The proposed algorithm considers three grouping criteria that differ in terms of the required information and performance.
- The implementation aspects of the proposal are addressed and specialized for the context of 5G NR networks. It is discussed the availability of the information required by the AMG algorithm in current 5G networks. To this end, the signaling mechanisms, the available measurements and UL reports that might be used to implement the proposed algorithm are identified.
- The performance of the proposed scheme is assessed with extensive simulations. It has been investigated a plethora a performance indicators, which includes the average, the variance and the distribution of the minimum required bandwidth, the distribution of the optimal number of groups, the distribution of the SNR and the average computation time. Results reveal that the proposed approach greatly reduces the required bandwidth compared to existing schemes that rely on single bandwidth allocation. It is also shown that it leads to a greater SNR for a randomly chosen user, and it reduces the variance of the required bandwidth, which eases the implementation in real networks.

The remainder of this manuscript is structured as follows. The system model and problem formulation is depicted in Section II whereas the proposed multicast scheme is described in Section III. Section IV illustrates the benefits of our proposal with extensive simulation results. Finally, some conclusions are drawn in Section V.

Notation: The following notation is used throughout the text. Matrices and vectors are represented with boldface uppercase and lowercase letters, respectively. If Σ is a matrix, $[\Sigma]_{n,m}$ is used to identify its (n, m)-th element. A matrix with N_1 rows and N_2 columns where all its elements are equal to 0 is written as $\mathbf{0}_{N_1 \times N_2}$. $(\bullet)^*$ denotes conjugate of a complex number whereas $\mathrm{Re}\{\bullet\}$ and $\mathrm{Im}\{\bullet\}$ denote the real and imaginary parts, respectively; $\mathrm{j} = \sqrt{-1}$ stands for the imaginary unit. $(\bullet)^T$ denotes transpose operation whereas $(\bullet)^H$ denotes Hermitian transpose. If $\mathbf{X}$ is an Hermitian matrix, then $\mathbf{X} \succeq 0$ indicates that such a matrix is positive

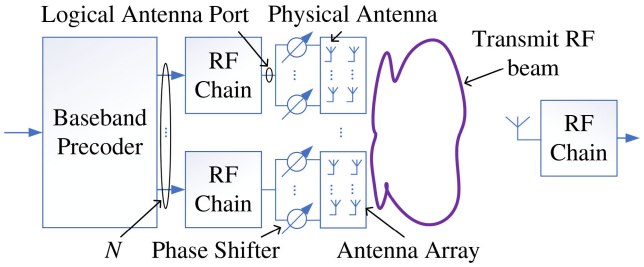


FIGURE 1. Simplified block diagram of hybrid beamforming. The relevant blocks of the BS are drawn at the left hand side of the figure whereas the blocks related to the UE are at the right hand side.

semi-definite. $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$ stand for the sets of natural, real and complex numbers respectively. If $\mathcal{A}$ is a set, then $|\mathcal{A}|$ is the cardinality of that set; however, if a is a complex number, $|a|$ represents its modulus. $\mathcal{CN}(\mathbf{0}_{K\times 1}, \Sigma)$ denotes the circular symmetric complex Gaussian distribution with zero mean and covariance Σ.

II. SYSTEM MODEL

We consider a hybrid beamforming scheme, as considered in 5G for millimeter-wave (mmW) bands, where the precoding is composed of two stages: i) a digital base-band beamforming; and ii) an analog/radio frequency (RF) beamforming [44]. On the one hand, base-band digital precoding is implemented by multiplying the complex IQ constellation symbols by a complex beamforming vector, and thus different precoding vectors can be applied to different blocks of time/frequency resources. By contrast, RF beamforming is implemented with phase shifters after digital-to-analog converters (DACs) and frequency up-conversion in the transmitter side, and thus RF beams can be multiplexed in time domain only, and not in the frequency domain [44], [45]. It is considered a hierarchical and modular RF and digital beam management as in 5G networks [46], [47]. This implies that RF beam management and digital precoding computation are performed independently.

Fig. 1 illustrates a simplified block diagram of the hybrid beamforming scheme considered in this work. On the base station (BS) side, it is observed that the base-band digital beamforming block receives a single stream of IQ constellation symbols and outputs N streams which are delivered to N RF chains for digital-to-analog and frequency up-conversion. In the context of 5G, the output of each RF chain is named logical antenna port [48], and it is connected to a different antenna array. The number of physical antennas that form the different antenna arrays determine the beamwidth and gain of the transmit RF beams that can be synthesized [44], [49]. In this work it is assumed that the UE has a single RF chain with a single physical antenna to reduce the cost of mobile handsets.

The transmit RF beams are modeled using the widely adopted sectored-pattern model [50]–[53], where it is considered that the main lobe has a constant gain of G_m with a beamwidth of θ radians, and it is centered at the steering angle φ. Angles that do not fall within the main lobe have a constant back lobe gain G_b. It is assumed that the RF gain of the main lobe is related to the beamwidth as $G_s \approx 2\pi/\theta$ [52], [54]. It is considered that UE receiver has a single RF chain connected to a single physical antenna for the sake of cost saving. For this reason it is assumed that the gain of the receive beam is just 1. Thus, with this system model, the relevant parameters are the beamwidth of the transmit RF beam, which also determines its beam gain, and the number of RF chains, N, which determines the gains that can be achieved by digital multicast beamforming.

Finally, it is considered that RF beams are time-domain multiplexed so there is no inter-beam interference. In addition, it is considered that mmW systems are noise-limited rather than interference-limited and thus inter-cell interference can be neglected, e.g., due to the high path-loss exhibited at mmW bands, the directivity of the RF beam patterns and the use of frequency planning strategies [55], [56].

A. SPATIAL MODELING

The probe BS is assumed to be placed at the origin, giving service to a cell of radius d_c meters and it is equipped with N RF chains, whereas the UEs have a single RF chain. We focus the analysis on a probe RF beam, and thus we consider the set of K UEs that are served by that RF beam. The K UEs can be divided into $G \in [1, K] \subset \mathbb{N}$ groups, and a different digital beamforming vector, $\mathbf{w}_g \in \mathbb{C}^{N\times 1}$, can be used for each group $g \in [1, G] \subset \mathbb{N}$.

Each UE is identified by a given index, $k \in \mathcal{K} = [1, K] \subset \mathbb{N}$, where $\mathcal{K}$ represents the set of UEs. The set of UEs that belong to the group $g \in [1, G] \subset \mathbb{N}$ is expressed as $\mathcal{K}_g \subseteq \mathcal{K}$. Each UE is assigned to a single group, i.e., $\bigcup_{g=1}^{G} \mathcal{K}_g = \mathcal{K}$ and $\bigcap_{g=1}^{G} \mathcal{K}_g = \emptyset$. Besides, we define the function $\mathcal{K}_g(u)$, as the u-th ordered element of the set $\mathcal{K}_g$. For instance, if we have the following UE set, $\mathcal{K} = \{1, 2, 3, 4, 5\}$, with $\mathcal{K}_1 = \{1, 3\}$ and $\mathcal{K}_2 = \{2, 4, 5\}$ for $G = 2$; then $\mathcal{K}_1(1) = 1$, $\mathcal{K}_1(2) = 3$, $\mathcal{K}_2(1) = 2$, $\mathcal{K}_2(2) = 4$ and $\mathcal{K}_2(3) = 5$.

The UEs are associated with the RF beam that provides the highest received power [47], [57], [58] and thus the UE locations fall within the region $\mathcal{R} \in \mathbb{R}^2$, which is defined by its main lobe as $\mathcal{R} = \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \leq d_c, \angle\mathbf{x} \in [\varphi - \frac{\theta}{2}, \varphi + \frac{\theta}{2}]\}$, being $\|\mathbf{x}\|$ the Euclidean norm of $\mathbf{x}$, and $\angle\mathbf{x} \in (-\pi, \pi]$ its angle.

In addition, the UE locations are drawn randomly according to a point process (PP), $\Phi = \{\mathbf{x}_1, ., \mathbf{x}_K\} \subset \mathbb{R}^2$. In this work we have considered two spatial distributions: i) the uniform Binomial point process (BPP), that models zero interaction between node locations; and ii) a clustered point process (CPP) that models spatial correlation between nodes [59]. The BPP places randomly K points within the region $\mathcal{R}$ with uniform distribution. The CPP is expressed

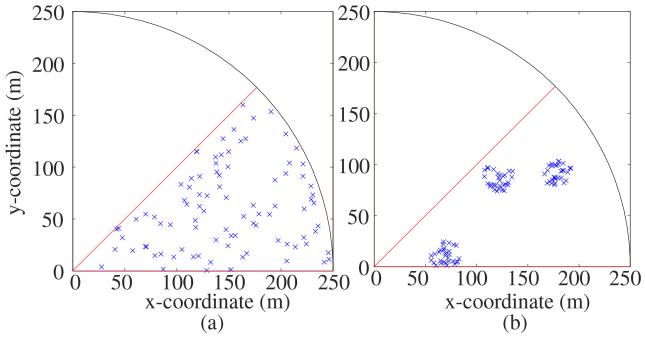


FIGURE 2. Spatial realizations of $K = 30$ UE locations for a cell of radius $d_c = 250$ m with a RF beamwidth of $\theta = \frac{2\pi}{8}$ for: (a) the BPP; and (b) the CPP with $C = 3$ clusters and $d_d = 15$ m.

as $\Phi = \bigcup_{i=1}^{C} \Phi_i + \mathbf{x}_i$, where C is the number of clusters and $\mathbf{x}_i$ represents the center of the i-th cluster. The cluster centers are drawn randomly within the region $\mathcal{R}$ with uniform distribution. The PP Φ_i represents the distribution of daughter points and it is modeled as a BPP that places C_i points within a disk centered at the origin with radius d_d. The sum of all daughter points placed by each of the clusters is K, i.e., $\sum_{i=1}^{C} C_i = K$. Fig. 2 shows a given spatial realization for the two PPs considered in this work.

B. CHANNEL AND SIGNAL MODELING

For a given allocation block of time/frequency resources, the channel is assumed to be flat in the time and frequency domains. A base-band equivalent channel model is considered, and thus, the channel between the BS and the k-th UE, $\ddot{\mathbf{h}}_k \in \mathbb{C}^{N\times 1}$, is modeled as a zero-mean complex random vector that accounts for the RF beam patterns gains, the path-loss and fast fading. Thus, the received signal by the k-th UE, which belongs to user group g, is given by

$$r_k = \mathbf{w_g}^H \ddot{\mathbf{h}}_k s_g \sqrt{p_t} + z_k \tag{1}$$

where $\mathbf{w_g} \in \mathbb{C}^{N\times 1}$ is the multicast beamforming vector and s_g represents the IQ constellation symbols intended for group g with zero mean and unit power, i.e., $\mathbb{E}[|s_g|^2] = 1$ and $\mathbb{E}[|s_g|] = 0$. p_t represents the transmit power per constellation symbol which spans a bandwidth of Δf Hz. Wide-sense stationary additive Gaussian noise is represented as z_k, and its power is p_n on Δf Hz. It is assumed a constant power spectral density (PSD) in the frequency domain, and thus $\varrho_t = p_t/\Delta f$ W/Hz and $N_0 = p_n/\Delta f$ W/Hz represent the transmit and noise power spectral densities, respectively. The channel vector $\ddot{\mathbf{h}}_k \in \mathbb{C}^{N\times 1}$ is expressed as

$$\ddot{\mathbf{h}}_k = \breve{\mathbf{h}}_k \|\mathbf{x}_k\|^{-\alpha/2} \sqrt{G_s} \tag{2}$$

where $\|\mathbf{x}_k\|$ is the distance between the k-th UE and the probe BS, and $\breve{\mathbf{h}}_k = [\breve{h}_{1,k}, ., \breve{h}_{N,k}]^T \in \mathbb{C}^{N\times 1}$ models the fast fading. The path-loss model is based on the classical power law, $\|\mathbf{x}_k\|^{-\alpha/2}$, where $\alpha > 0$ is an environmental dependent path-loss exponent. The complex gain due to fast fading between the i-th transmit logical antenna and the k-th user has a marginal distribution according to a zero mean unit power complex Gaussian distribution $\breve{h}_{k,i} \sim \mathcal{CN}(0, 1)$.

Besides the uncorrelated case, it is considered the case where the fading of different users can be correlated based on their distance. This latter distance dependent correlation model is based on the observation made in some works, (e.g., [60] and references therein), where it is shown that there exists a correlation between nearby locations that tend to decrease as the distance increases. Therefore, it is proposed to model the correlation between the k-th and q-th UEs as $\rho_{k,q} = \mathbb{E}[\breve{h}^*_{i,k}\breve{h}_{i,q}] = \exp(-\beta\|\mathbf{x}_k - \mathbf{x}_q\|)$, being $\beta \geq 0$ a factor that models how strong the correlation is and $\|\mathbf{x}_k - \mathbf{x}_q\|$ the distance between the two UEs. It has been chosen a decreasing exponential to capture the fact that the correlation is stronger between nearby locations and decreases as the distance increases. Besides of this, the proposed model considers the case of maximal correlation, i.e., $\rho_{k,q} = 1$, and the independent case as particular cases that are modeled with $\beta = 0$ and $\beta \to \infty$ respectively. Finally, the vector of complex gains for the i-th transmit logical antenna w.r.t. the K UEs is generated according to a multivariate Gaussian distribution as $\breve{\mathbf{h}}^{(i)} = [\breve{h}_{i,1}, ., \breve{h}_{i,K}] \in \mathbb{C}^{1\times K} \sim \mathcal{CN}(\mathbf{0}_{1\times K}, \Sigma)$, where $[\Sigma]_{k,q} = \rho_{k,q}$ and $\Sigma \in \mathbb{R}^{K\times K}$.

C. BANDWIDTH ALLOCATION AND RATE ADAPTATION

It is assumed that a broadcast/multicast service is intended to be delivered to K UEs associated with the probe RF beam. The target binary rate of such a service is R_T bps. The K UEs are divided into G groups that use a different beamforming vector, $\mathbf{w}_g$, on a different bandwidth allocation, $\mathcal{B}_g$, of $|\mathcal{B}_g|$ Hz. The bandwidth allocations to different groups are orthogonal to avoid inter-group interference and thus $\bigcap_{g=1}^{G} \mathcal{B}_g = \emptyset$. The overall bandwidth allocated to the broadcast/multicast service is

$$\mathcal{B} = \bigcup_{g=1}^{G} \mathcal{B}_g \tag{3}$$

The SNR of UE k that belongs to group g, $\gamma_k(\mathbf{w}_g)$, is given by

$$\gamma_k(\mathbf{w}_g) = \frac{|\mathbf{w}_g \ddot{\mathbf{h}}_k|^2 p_t}{p_n} = |\mathbf{w}_g \mathbf{h}_k|^2 \tag{4}$$

where $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$ is the scaled channel gain, which is expressed as

$$\mathbf{h}_k = \ddot{\mathbf{h}}_k \sqrt{\frac{\varrho_t}{N_0}} = \breve{\mathbf{h}}_k \sqrt{\bar{\gamma}_k} \tag{5}$$

being $\bar{\gamma}_k = \|x_k\|^{-\alpha} G_s \frac{\varrho_t}{N_0}$ the average SNR when $N = 1$.

The transmission rate of each group is adapted to the min SNR of the group, and thus the spectral efficiency (SE) (i.e., rate) of group g is expressed as $\log_2(1 + \min_{k\in\mathcal{K}_g} \gamma_k(\mathbf{w}_g))$ bps/Hz.

III. PROPOSED USER GROUPING, BANDWIDTH ALLOCATION, AND PRECODING SCHEME

A. PROBLEM FORMULATION AND PROPOSED ALGORITHM

We consider a precoding and grouping scheme that aims at minimizing the bandwidth $|\mathcal{B}|$ needed to provide a broadcast/multicast service that requires R_T bps. Therefore, the problem can be posed as obtaining the user division into G groups, $\mathcal{K}_g$, with $g \in [1, G]$, and multicast beamforming vector for each group, $\mathbf{w}_g \in \mathbb{C}^{N\times 1}$, that minimizes the overall required bandwidth. More formally, this problem can be formulated as

$$\arg \min_{G,\{\mathcal{K}_1,\cdots,\mathcal{K}_G\},\{\mathbf{w}_1,\cdots,\mathbf{w}_G\}} |\mathcal{B}| \tag{6}$$

$$\text{with } |\mathcal{B}| = R_T \sum_{g=1}^{G} \log_2^{-1}(1 + \min_{k\in\mathcal{K}_g} \gamma_k(\mathbf{w}_g)) \tag{7}$$

$$\text{s.t. } \|\mathbf{w}_g\| = 1, \quad \forall g \in [1, G] \subset \mathbb{N} \tag{8}$$

where (7) comes after (3) and the fact that the achievable binary rate of every group $g \in [1, G] \subset \mathbb{N}$ must be equal to R_T. To seek for a sub-optimal solution, we split this problem in two parts: user grouping and multicast precoding. The user grouping assumes a number of groups and outputs the partition of users into G groups, $\mathcal{K}_g$ $\forall g \in [1, G] \subset \mathbb{N}$.

1) USER GROUPING

The proposed user grouping algorithm relies on the observation that highly correlated channels increases the multicast beamforming gain [9], [18]. Thus, the proposed algorithm aims at assigning the same group to users whose channel is similar. To this end, a K-means++ (KM) algorithm has been selected due to its reduced complexity and quality of final solution [61]. This clustering algorithm partitions a data set of K points into G groups using an iterative algorithm to minimize the sum of data sample-to-centroid distances, summed over all G clusters. The centroid of each group is the mean of the points that belong to the group and it is also an output of the algorithm. We define three different data set types, which are used as input to KM, to devise three grouping algorithms that differ in performance and complexity.

i) Scaled channel matrix: This option considers that the data set is the matrix $\mathbf{H} \in \mathbb{C}^{N\times K}$, which is built by stacking the scaled channel gain of each user, i.e., $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$ $\forall k \in \mathcal{K}$ as column vectors. Since KM algorithm restricts to real data samples, the input data set for KM algorithm is formed by stacking the real and imaginary parts of $\mathbf{H}$ as $[\text{Re}\{\mathbf{H}\}^T, \text{Im}\{\mathbf{H}\}^T]^T \in \mathbb{R}^{2N\times K}$. Hence, each data sample is a real point in $\mathbb{R}^{2N}$. This grouping algorithm is labeled as KM-CSI.

ii) Location information: Here the input to the KM algorithm is the position $\mathbf{x}_k \in \mathbb{R}^2$ of every user $k \in \mathcal{K}$. The data set is the PP of all UEs locations, Φ, which can be arranged as a matrix $\Phi \in \mathbb{R}^{2\times K}$. Compared to KM-CSI, this type of data set lacks of information about the instantaneous channel gain of every user; nevertheless, the size of the data set is smaller as well as its complexity. This grouping algorithm is labeled as KM-loc.

iii) Reference signal received power (RSRP): The reference signal received power (RSRP) represents the average received power by a given user k and it is expressed as: $\mu_k = \|\mathbf{x_k}\|^{-\alpha} G_s \varrho_t$. The data set is then a vector of K real elements $\boldsymbol{\mu} = [\mu_1, ., \mu_K] \in \mathbb{R}^{1\times K}$. Contrarily to KM-loc and KM-CSI, this method lacks of information about distances between the different users and it only has information about distances towards the BS. However, it has a smaller data size and complexity than the other two options. It is labeled as KM-RSRP.

The proposed algorithm, AMG, requires any of these three types of information. The availability of such information in 5G networks is discussed in Section III-B whereas the impact of the type of information used on the performance is assessed in Section IV-B.

2) MULTICAST BEAMFORMING

The multicast precoding considers a partition of the users into G groups and then it computes a beamforming vector for each group that maximizes the min SNR. The following state-of-the art multicast precoding vectors have been considered in this paper:

i) SDR-G: This algorithm was proposed in [7]. The algorithm approximates the max-min-fair problem into the following convex problem that can be solved via SDP. Hence, with SDR-G, the multicast beamforming vector for group g can be computed as follows:

1) Solve the relaxed SDP problem to obtain the positive semi-definite matrix, $\mathbf{X} \in \mathbb{C}^{N\times N}$:

$$\max_{\mathbf{X}\in\mathbb{C}^{N\times N}, t\in\mathbb{R}} t \tag{9}$$

$$\text{s.t. } \text{trace}(\mathbf{X}\mathbf{Q}_k) \geq t, \ \forall k \in \mathcal{K}_g \tag{10}$$

$$\text{trace}(\mathbf{X}) = 1, \ \mathbf{X} \succeq 0 \tag{11}$$

where $\mathbf{Q}_k \in \mathbb{C}^{N\times N}$ and $\mathbf{Q}_k = \mathbf{h}_k \mathbf{h}_k^H \succeq 0$.

2) Perform Gaussian randomization. This involves that M triplets of candidate beamforming vectors are randomly generated and the best one after M realizations is selected. Each triplet consists on the following vectors, $\mathbf{w}_g^{(a)}, \mathbf{w}_g^{(b)}, \mathbf{w}_g^{(c)} \in \mathbb{C}^{N\times 1}$, that are generated as

$$\mathbf{w}_g^{(a)} = \mathbf{U}\Lambda^{\frac{1}{2}}\mathbf{e}_g^{(a)} \tag{12}$$

where $\mathbf{U} \in \mathbb{C}^{N\times 1}$ is obtained after eigen-decomposition of $\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^H$ and $[\mathbf{e}_g^{(a)}]_i = \exp(\text{j}\theta_i)$ being θ_i uniformly distributed on $[0, 2\pi)$, being

$$[\mathbf{w}_g^{(b)}]_i = \sqrt{[\mathbf{X}]_{i,i}}\mathbf{e}_g^{(b)} \tag{13}$$

where $\mathbf{e}_g^{(b)} \in \mathbb{C}^{N\times 1}$ are generated using the same procedure as for $\mathbf{e}_g^{(a)} \in \mathbb{C}^{N\times 1}$, and finally

$$\mathbf{w}_g^{(c)} = \mathbf{U}\Lambda^{\frac{1}{2}}\mathbf{e}_g^{(c)} \tag{14}$$

where $\mathbf{e}_g^{(c)} \in \mathbb{C}^{N\times 1}$ is a vector of zero-mean, unit-variance complex circularly symmetric uncorrelated Gaussian random variables.

ii) Adaptive update (AU): This is an iterative algorithm which is proposed in [9]. With this algorithm, each update takes a step in the direction of an inverse SNR weighted linear combination of the SNR-gradient vectors of all $|\mathcal{K}_g|$ users. At iteration m, the beamforming vector of group g is updated as follows

$$\tilde{\mathbf{w}}_g^{(m+1)} = \mathbf{w}_g^{(m)} + \xi \left(\sum_{k=1}^{|\mathcal{K}_g|} \frac{\mathbf{Q}_{\mathcal{K}_g(k)}}{\left(\mathbf{w}_g^{(m)}\right)^H \mathbf{Q}_{\mathcal{K}_g(k)} \mathbf{w}_g^{(m)} + \epsilon} \right) \mathbf{w}_g^{(m)}$$
$$\mathbf{w}_g^{(m+1)} = = \frac{\tilde{\mathbf{w}}_g^{(m+1)}}{\|\tilde{\mathbf{w}}_g^{(m+1)}\|} \tag{15}$$

where ξ is the fixed positive step-size for every iteration and ϵ is a positive constant that is introduced for numerical stability.

iii) SB: this algorithm involves a maximum of $\min(N, K)$ iterations. With iteration k the algorithm equalizes the SNRs of the k UEs with smallest SNR, and it stops if the SNRs of the rest of users is greater than that value or if there are no more degrees of freedom, i.e., $k+1 > \min(N, K)$. The details are described in [19].

iv) Random beamforming (RBF): this scheme considers that the beamforming vector is randomly generated according to a complex Gaussian distribution where $[\mathbf{w}_g]_i \sim \mathcal{CN}(0, 1/N)$.

Both stages, user grouping and multicast precoding, assume a number of groups, G. Hence, the proposed algorithm goes through an increasing number of groups in an iterative fashion, starting with a single group. The proposed algorithm for user grouping, bandwidth allocation and beamforming, referred to as AMG, is summarized in Algorithm 1, and it is explained as follows. Firstly, the algorithm sets the initial required bandwidth with the highest value, and the variable *stop*, which is used as stopping criterion (lines 1 to 3). Then, a loop is executed to search for the best number of groups, ℓ, within the range $[1, K-1]$ (lines 4 - 11). At each iteration, the algorithm partitions the users in ℓ groups according to either KM-CSI, KM-loc or KM-RSRP criteria (line 6). Afterwards, it computes the multicast precoding for each group, $\mathbf{w}_g \in \mathbb{C}^{N\times 1}$, and the required bandwidth, $|\mathcal{B}^{(\ell)}|$, (lines 7 - 10). To save computation time, we can limit the maximum number of iterations (i.e., groups) to a given value, G_{max}. The algorithm iterates until the maximum number of groups to explore, G_{max}, is reached, or the stopping criterion is fulfilled. We propose as stopping criterion whether the required bandwidth of current iteration, ℓ, is greater than the bandwidth of the previous iteration (line 11). The effect of such an stopping criterion will be assessed in Section IV with simulation results. The obtained solution, which can explore a number of groups up to $K-1$, is finally compared with the results of unicast transmission, where the number of groups is K (lines 12 - 16). This unicast transmission uses MRT beamforming which achieves the capacity of the MISO channel. If the bandwidth of multicast transmission with ℓ groups is smaller than the one required by unicast transmission, then $G^\star = \ell$, whereas unicast transmission is selected otherwise.

Algorithm 1 AMG

Input: $\mathcal{K}$, $\mathbf{h}_k$ $\forall k \in \mathcal{K}$
Output: $G^\star$, $\gamma^\star_{\min,g}$, $|\mathcal{B}^\star_g|$, $\mathbf{w}^\star_g \in \mathbb{C}^{N\times 1}$, $\mathcal{K}^\star_g$ $\forall g \in [1, G^\star]$
Data: Data for user grouping:
KM-CSI: $[\mathrm{Re}\{\mathbf{H}\}^T, \mathrm{Im}\{\mathbf{H}\}^T]^T \in \mathbb{R}^{2N\times K}$
KM-loc: $\Phi \in \mathbb{R}^{2\times K}$
KM-RSRP: $\boldsymbol{\mu} \in \mathbb{R}^{1\times K}$

1: Set $|\mathcal{B}^{(0)}| \to \infty$
2: $\ell = 0$
3: $stop = false$
4: **while** $\ell \leq \min(G_{max}, K-1)$ & $(\sim stop)$ **do**
5: $\quad \ell = \ell + 1$
6: $\quad$ Partition the UE set in ℓ groups, $\mathcal{K}_g^{(\ell)}$ $\forall g \in [1, \ell]$, according to the chosen *user grouping* algorithm: KM-CSI, KM-loc or KM-RSRP
7: $\quad$ Compute the *multicast beamforming* vector, $\mathbf{w}_g^{(\ell)} \in \mathbb{C}^{N\times 1}$, for each group $g \in [1, \ell]$ according to the chosen method: SDR-G, adaptive update (AU), SB or random beamforming (RBF)
8: $\quad$ Compute the min SNR per group, $\gamma^{(\ell)}_{\min,g} = \min_{k\in\mathcal{K}_g} \gamma_k\left(\mathbf{w}_g^{(\ell)}\right)$ using (4)
9: $\quad$ Compute the required bandwidth per group as $|\mathcal{B}_g^{(\ell)}| = \frac{R_T}{\log_2(1+\gamma^{(\ell)}_{\min,g})}$
10: $\quad |\mathcal{B}^{(\ell)}| = \sum_{g=1}^{\ell} |\mathcal{B}_g^{(\ell)}|$
11: $\quad stop = |\mathcal{B}^{(\ell)}| > |\mathcal{B}^{(\ell-1)}|$
12: **end while**
13: Compute the required bandwidth for unicast transmission as $|\mathcal{B}^{(K)}| = \sum_{k=1}^{K} \frac{R_T}{\log_2(1+\|\mathbf{h}_k\|^2)}$, which uses maximum ratio transmission (MRT) beamforming $\mathbf{w}_k^{(K)} = \mathbf{h}_k^H$ $\forall k \in [1, K]$.
14: **if** $|\mathcal{B}^{(\ell)}| < |\mathcal{B}^{(K)}|$ **then**
15: $\quad G^\star = \ell$
16: **else**
17: $\quad G^\star = K$
18: **end if**
19: $|\mathcal{B}^\star_g| = |\mathcal{B}^{(G^\star)}_g|$ $\forall g \in [1, G^\star]$
20: $\mathbf{w}^\star_g = \mathbf{w}^{(G^\star)}_g \in \mathbb{C}^{N\times 1}$ $\forall g \in [1, G^\star]$
21: $\gamma^\star_{\min,g} = \gamma^{(G^\star)}_{\min,g}$ $\forall g \in [1, G^\star]$
22: $\mathcal{K}^\star_g = \mathcal{K}^{(G^\star)}_g$ $\forall g \in [1, G^\star]$

B. IMPLEMENTATION ASPECTS

In this subsection the implementation aspects of the proposed AMG scheme in real systems and its application to 5G NR

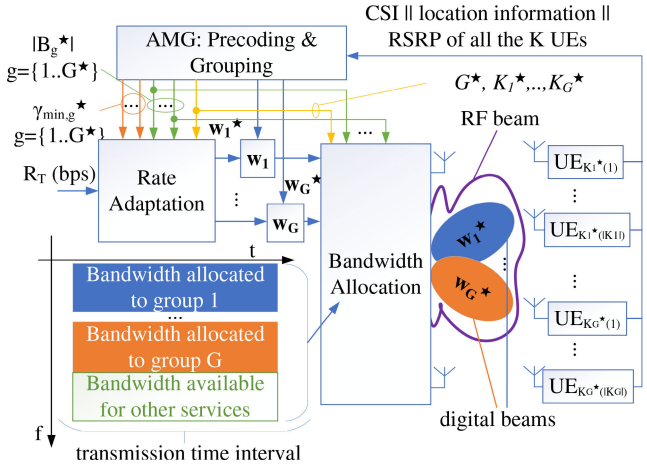


FIGURE 3. Block diagram of AMG implementation.

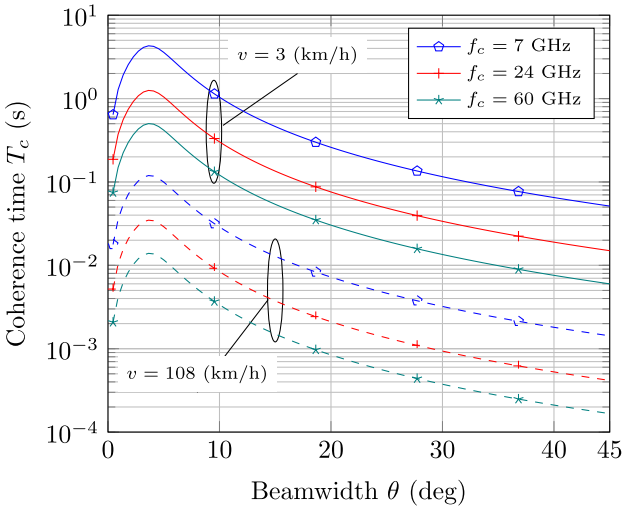


FIGURE 4. Coherence time of the channel versus the beamwidth, θ, for a non-line-of-sight (NLOS) scenario. The figure evaluates equation (36) of [68], assuming a scattering radius of 1000 wavelengths, an original pointing direction of 5 degrees and a target correlation of 0.5. The expression is evaluated for frequency bands centered at 7, 24 and 60 GHz, and user speeds of 3 and 108 km/h (i.e., 0.83 and 30 m/s).

are discussed. The block diagram of the proposed scheme is illustrated in Fig. 3. The time is divided into transmission time intervals (TTIs), and it is considered that the channel is time invariant within the TTI although it varies between different TTIs. It is observed that the AMG block computes the number of groups, $G^\star$, the min SNR, $\gamma^\star_{\min,g}$, and the required bandwidth, $|\mathcal{B}^\star_g|$, per group. All these metrics are forwarded to the *Rate Adaptation* block, which adapts the symbol rate intended to each group to the link conditions of the worst UE, given by $\gamma^\star_{\min,g}$. In real systems, e.g., 5G NR, this is achieved by selecting an appropriate modulation and coding scheme (MCS).

The constellation symbols for each group, s_g, are delivered to the precoding stage, where they are multiplied by the beamforming vector per group, $\mathbf{w}^\star_g \in \mathbb{C}^{N\times 1}$, which is obtained by the AMG block. The precoded symbols for each group are delivered to the *Bandwidth Allocation* block which stacks blocks of symbols for each group and maps them into the portion of bandwidth allocated to each group. As seen in the figure, the bandwidth that is not used by the broadcast/multicast service is available to other services. Since the required overall bandwidth, $|\mathcal{B}^\star_g|$, depends on the channel of every UE, it varies on a TTI basis. Finally, the digitally beamformed stream to each group is transmitted using the same RF beam, which is received by the K UEs.

The 5G standard supports several reporting quantities that can be used to develop the proposed scheme. Firstly, the scaled channel matrix per UE, $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$, is used by the *multicast beamforming* stage as well as user grouping in case of KM-CSI. This metric can be obtained with the precoding matrix indicator (PMI) and channel quality indicator (CQI) report quantities [47], [57]. The former report quantity leads to a precoding vector that belongs to a given codebook, whose type is specified by higher layer configuration [62]. This precoding vector can be understood as a quantized version of the Hermitian channel vector, i.e., $\mathbf{h}^H_k \in \mathbb{C}^{1\times N}$, since this is the optimal precoder of the single user MISO channel [63]. The CQI, on the other hand, indicates the appropriate MCS to achieve a block error rate (BLER) below a given target value, which can be configured to either 10^{-1} or 10^{-5} in 5G networks [64]. This CQI is computed by the UE based on the estimated SNR [65], and thus, it can be used in real systems in combination with the PMI to get the scaled channel matrix given by (5).

Another required metric is the RSRP, which is needed at the user grouping stage if KM-RSRP algorithm is used. This metric actually corresponds to an existing report quantity which is named L1-RSRP in the 5G standard [66].

Lastly, the location information is used by the user grouping stage in case of KM-loc. This can be achieved thanks to 5G NR positioning protocol of current releases 15 and 16. Some methods that are part of the standard and can be used to this end are uplink time difference of arrival (UTDOA), enhanced cell ID (E-CID), multicell round trip time (Multi-RTT) and uplink angle of arrival (UL-AoA) [67].

As it was justified in the introduction, a paramount challenge of multicast precoding algorithms is the computational time due to the complexity of the underlying optimization problems that needs to be solved. The solution of the user grouping and multicast precoding determined by the AMG algorithm is valid for a time period where the joint channel of all the users can be considered as roughly constant. Thus, this period of time, which is named coherence time, acts as a system requisite for the chosen multicast precoding and user grouping algorithm. As shown in [68], the coherence time highly depends on different factors such as the velocity of the users, the frequency band and the beamwidth. From Fig. 4 it is observed that the coherence time ranges from thousands of ms down to a few ms or even fractions of ms. It is seen that reducing the beamwidth increases the coherence time, whereas it can be reduced by increasing the frequency band

TABLE 1. Default simulation parameters.

Parameter	Description	Value
PP	PP type	BPP
β	Correlation parameter	∞ (uncorrelated)
Multicast beamforming	-	AU
Grouping algorithm	-	KM-loc
α	path-loss exponent	4
R_T	Target rate	1 Mbps
ϱ_t	Transmit PSD	-100 dBm/Hz
θ	RF beamwidth	$2\pi/8$ radians
N_0	Noise PSD	-167 dBm/Hz
K	Number of UEs	30
N	Number of RF chains	8

and/or the user velocity. More specifically, for a beamwidth of 45 degrees a pedestrian that moves at 3 km/h and receives a multicast transmission at 7 GHz has a coherence time around 50 ms. If the set of users has this coherence time, the AMG algorithm should take less than 50 ms to compute the user partition into groups and the related multicast beamforming vectors. Nevertheless, as explained in Section III, the AMG algorithm can use different precoding (e.g., SDR-G, AU, RBF) and grouping algorithms (i.e., KM-CSI, KM-loc, KM-RSRP) that lead to different computational time as well as different performance. So an appropriate choice can be selected based on the expected coherence time to have an appropriate balance between performance and computational time.

IV. SIMULATION RESULTS AND DISCUSSIONS

The performance of the proposed scheme is assessed with simulations. It is considered a path-loss exponent of $\alpha = 4$, and a beamwidth of $\theta = 2\pi/8$ radians with a main lobe gain of $G_s = 8$. The thermal noise is assumed to be $N_{th} = -174$ dBm/Hz, with a noise figure, $N_F = 7$ dB, and thus the noise PSD is $N_0 = N_{th} + N_F = -167$ dBm/Hz. The results have been obtained through Monte Carlo simulations with 10^3 realizations.

The default parameters considered in this section are summarized in Table 1. These parameters have been used to obtain each figure unless otherwise stated in the caption.

The performance of the proposed algorithm, AMG, is compared to two extreme alternatives: i) *unicast* transmission, which considers MRT beamforming and orthogonal bandwidth allocation to each UE; and ii) *broadcast*, which involves a single group with single bandwidth allocation and a single multicast beamforming vector.

Three multicast beamforming algorithms are considered as described in Section III. The AU algorithm uses a step size $\xi = 0.1$, with a factor of $\epsilon = 10^{-3}$ to avoid numerical instability and 100 iterations. The SDR-G algorithm uses $M = 30NK$ randomizations as recommended in [7], which involves 7200 random trials for the default parameters.

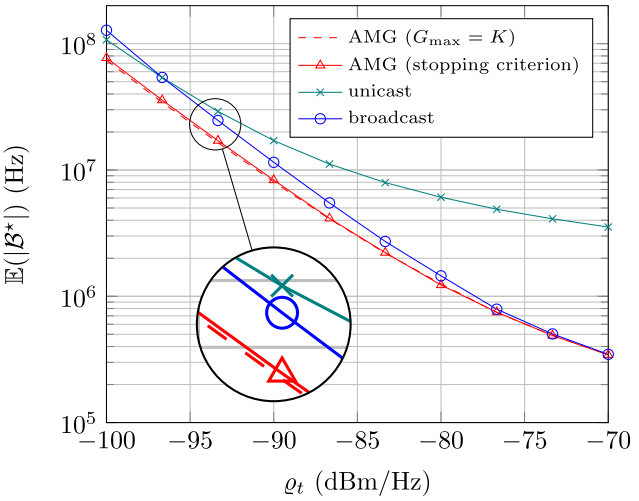


FIGURE 5. Average required bandwidth for BPP spatial distribution with uncorrelated fading and $N = 8$ RF chains.

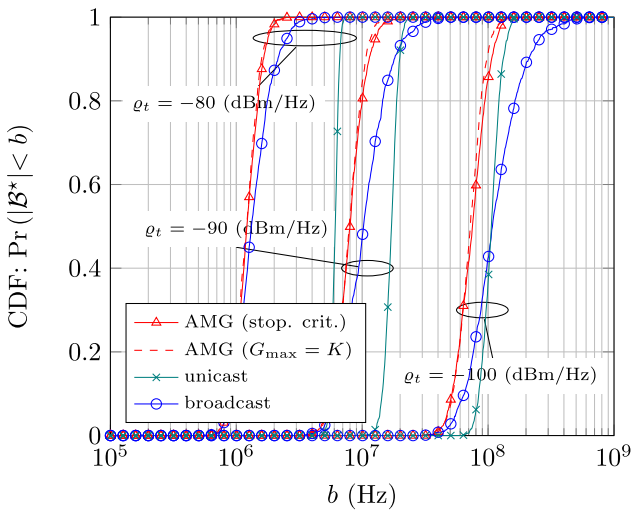


FIGURE 6. Cumulative distribution function (CDF) of required bandwidth for BPP spatial distribution with uncorrelated fading and $N = 8$ RF chains.

The proposed algorithm described in Algorithm 1 considers an stopping criterion as described in line 11. To assess the impact of such a stopping criterion it is also considered the case where the algorithm executes G_{max} iterations (line 11 is removed). This allows us to evaluate the increment in the required bandwidth due to the early stopping but also the savings in computation time.

Next subsections illustrate the performance of the proposed algorithm under diverse scenarios to get insights about the performance trends and interplay between different parameters.

A. PERFORMANCE WITH INDEPENDENT FADING AND BPP SPATIAL DISTRIBUTION

Firstly, the performance of AMG is assessed for a BPP spatial distribution. This type of distribution models independent

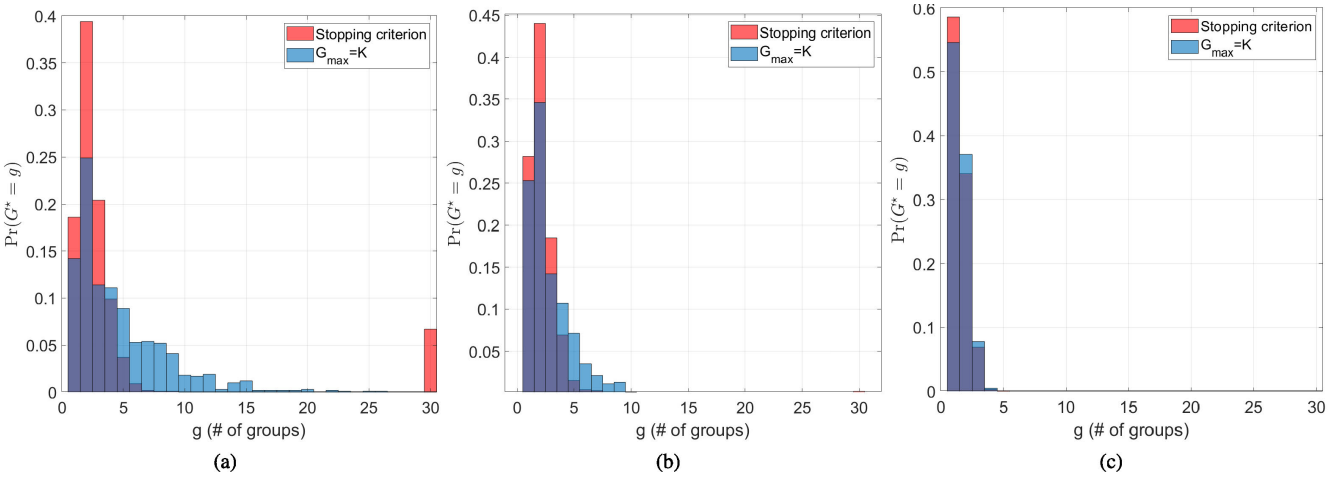


FIGURE 7. Probability mass function (PMF) of the chosen number of groups $G^\star$ with (*red*) and without (*light blue*) a stopping criterion $|\mathcal{B}^{(\ell)}| < |\mathcal{B}^{(\ell-1)}|$ for: (a) $\varrho_t = -100$ dBm/Hz; (b) $\varrho_t = -90$ dBm/Hz; and (c) $\varrho_t = -80$ dBm/Hz. The UE locations follow a BPP with $K = 30$ and it is considered uncorrelated fading with $N = 8$ RF chains. The PMF is represented as a semi-transparent colored bar diagram. The intersection of probability values with and without stopping criteria is represented in *dark blue*.

locations of the UEs, which are drawn randomly with uniform distribution. The fading is also assumed to be independent.

Figures 5 and 6 show the mean and cumulative distribution function (CDF) of the required bandwidth versus the transmit PSD for the proposed AMG algorithm, with and without an stopping criterion. Results for the case of unicast and broadcast transmissions are also presented.

It is observed that our proposal greatly outperforms broadcast and unicast transmissions for a wide range of transmit PSD values. This demonstrates that it is beneficial to divide the users into different groups, since the required bandwidth can be potentially reduced. On the one hand, the performance of multicast precoding algorithms is deteriorated as the group size increases [6]. This involves that dividing the users into smaller groups will lead to a higher min SNR, $\min_{k\in\mathcal{K}_g} \gamma_k\left(\mathbf{w}_g\right)$, of the groups than the single group counterpart. If the increase in the min SNR is high enough, then it leads to an smaller required bandwidth than using a single group. On the other hand, grouping users with similar channel realizations increases the gain of the multicast precoding. This statement is based on the observation that correlated scaled channels (including fast fading and average SNR) increase the performance gain of multicast precoding [9], [18]. Hence, our proposed AMG scheme relies on these two ideas to find a sub-optimal number of groups, $G^\star$, user partition, $\mathcal{K}_g^\star$ $\forall g \in [1, G^\star]$, and multicast beamforming that minimizes the required bandwidth.

The performance loss due to early stopping is negligible as it can be observed from the average required bandwidth in Fig. 5 as well as from its distribution in Fig. 6. Interestingly, the average required bandwidth is smaller for unicast than for broadcast transmission in the low transmit power regime. This means that, on average, the impact of the increase in min SNR, which yields an increase on the SE, leads to an smaller aggregated bandwidth than the single group counterpart. However, as the transmit power is increased, the required bandwidth of broadcast transmission tends to be much smaller than the unicast alternative.

Fig. 7 illustrates the probability mass function (PMF) of the sub-optimal number of groups, $G^\star$, with (red) and without (blue) an stopping criterion, for different transmit PSD values. It can be observed that the distribution of $G^\star$ takes a wider set of values when the transmit power is small. Nevertheless, as the transmit power is increased, the range of values is narrower and more concentrated around small values. This might be expected in view of Fig. 5 and 6 since the performance of broadcast gets closer to the performance of AMG as the transmit power is increased.

If we compare the PMF of the number of groups with and without stopping criterion, we observe that the effect of the stopping criterion is to concentrate the distribution around smaller values of $G^\star$. This is related to the fact that the early stop prevents from searching solutions related to greater number of groups.

As it is seen in the pseudo-code of Algorithm 1, the algorithm always searches a solution in the unicast case, since computing the unicast beamforming vector is trivial in terms of computational complexity. When an stopping criterion is used and $\varrho_t = -100$ dBm/Hz, there is a probability of 0.067 to find the unicast case as sub-optimal solution (i.e., $G^\star = K$).

Nevertheless, for the same transmit PSD of -100 dBm/Hz, the unicast case is not selected if no stopping criterion is considered, which is labeled in the legend as $G_{max} = K$. This involves that the optimal solution is never the unicast solution, i.e., $G^\star = K$, if an exhaustive search is considered. Yet, the unicast solution is better than the solution found by the loop of lines 4-11 in Algorithm 1, if an stopping criterion is used, with a probability of 0.067.

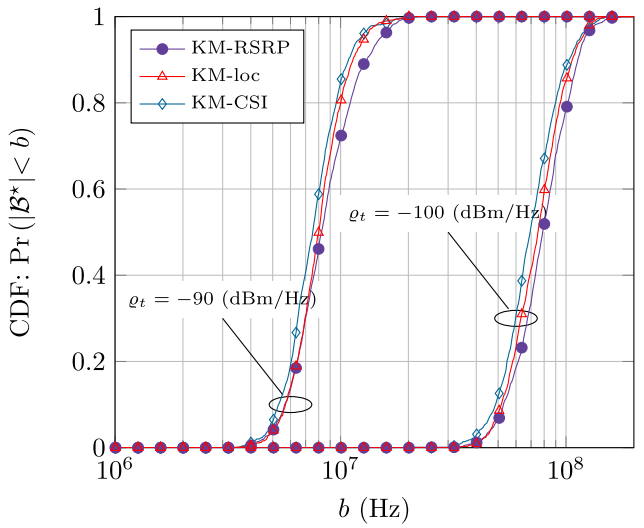


FIGURE 8. CDF of required bandwidth for the 3 proposed grouping algorithms. The UE locations follow a BPP with $K = 30$ and it is considered uncorrelated fading with $N = 8$ RF chains.

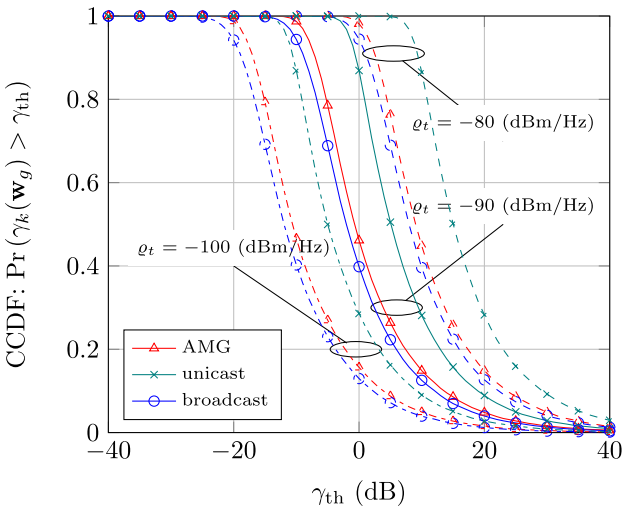


FIGURE 9. Complementary cumulative distribution function (CCDF) of SNR of a randomly chosen UE for different transmit PSDs values, $\varrho_t = \{-100, -90, -80\}$ (dBm/Hz). The UE locations follow a BPP with $K = 30$ and it is considered uncorrelated fading with $N = 8$ RF chains.

B. EFFECT OF GROUPING ALGORITHMS

The effect of the three proposed grouping algorithms is illustrated in Fig. 8. As it is mentioned in Section III, among the 3 grouping algorithms, KM-CSI is the one that requires a higher amount of data, i.e., $2NK$ real numbers. This algorithm is followed by the KM-loc, which requires $2K$ real numbers and KM-RSRP, which requires K real numbers. As it can be seen in Fig. 8, the performance of the different grouping algorithms follows the amount of data used. Thus, the smallest bandwidth in statistical terms is obtained by KM-CSI, whereas KM-RSRP leads to the highest bandwidth and KM-loc obtains intermediate results.

C. SNR DISTRIBUTION AND VARIANCE OF REQUIRED BANDWIDTH

The complementary cumulative distribution function (CCDF) of the SNR of a randomly chosen UE is shown in Fig. 9 for AMG, unicast and broadcast schemes with different transmit PSD values. It can be observed that unicast transmission, which is based on MRT, achieves the highest SNR in statistical terms. This is expected since multicast precoding uses the same beamforming vector for a group of users, and thus its performance deteriorates as the group size increases. Nevertheless, broadcast transmission, which uses single group multicast precoding, achieves a much smaller required bandwidth than unicast transmission for a broader set of transmission power values, as it was discussed in Fig. 5 and 6. Interestingly, AMG achieves a smaller required bandwidth than broadcast and unicast approaches, but at the same time, it leads to a higher SNR than broadcast transmission.

Fig. 10, on the other hand, shows the variance of the required bandwidth for AMG, unicast and broadcast schemes versus the transmit PSD, ϱ_t. The variance of the required

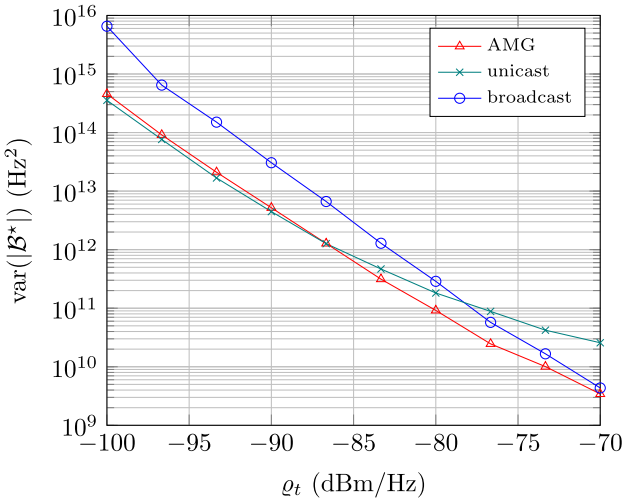


FIGURE 10. Variance of the required bandwidth versus ϱ_t. The UE locations follow a BPP with $K = 30$ and it is considered uncorrelated fading with $N = 8$ RF chains.

bandwidth is an important metric for frequency planning. As it was mentioned in Section III-B, the bandwidth that is not used by the multicast/broadcast service can be used by other services. Therefore, having an small variance of the required bandwidth is highly appealing, since it eases the frequency planning of the other services that can be accommodated in the available bandwidth. It can be observed that AMG achieves an smaller variance than broadcast transmission in the considered range of transmit power and it also achieves an smaller variance than unicast transmission for $\varrho_t > -85$ dBm/Hz, while for $\varrho_t < -85$ dBm/Hz the difference in terms of performance between AMG and unicast is negligible.

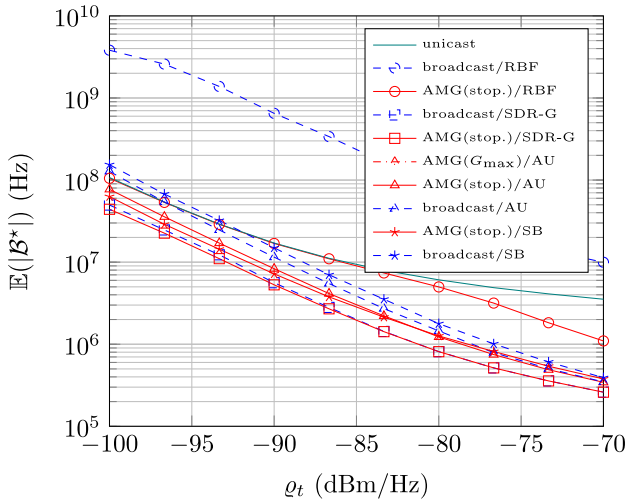


FIGURE 11. Average required bandwidth versus ϱ_t for different multicast precoding algorithms. The UE locations follow a BPP with $K = 30$ and it is considered uncorrelated fading with $N = 8$ RF chains.

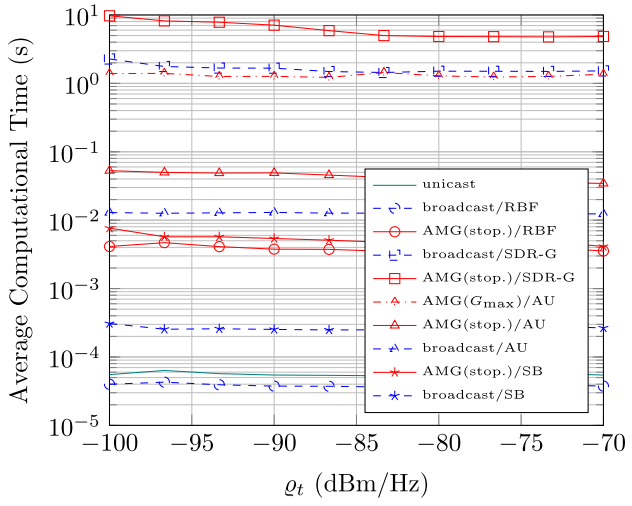


FIGURE 12. Average computation time for different multicast precoding algorithms. The UE locations follow a BPP with $K = 30$ and it is considered uncorrelated fading with $N = 8$ RF chains.

D. PERFORMANCE VERSUS COMPLEXITY

The trade-off between performance and complexity for different multicast precoding algorithms is shown in Fig. 11 and 12. Results reveal that unicast transmission and broadcast with RBF lead to the smallest average computation time. Nevertheless, the required bandwidth for broadcast with RBF is the highest, and its performance is clearly inferior to other alternatives. These results highlight the importance of multicast precoding, which leads to great performance improvements compared to the RBF case, thanks to the use of channel information of all the users and different RF chains. It is observed that AMG greatly improves the performance compared to broadcast and unicast alternatives. Even with RBF, AMG leads to an smaller required bandwidth than the unicast case and its broadcast counterpart. This is specially relevant at high transmit powers, where the required bandwidth of AMG with RBF tends to be also greatly smaller than the unicast case.

As seen from Fig. 11, the smallest required bandwidth is obtained with SDR-G precoding. However, as it is observed in Fig. 12, this precoding leads to the highest computation time with both (the broadcast and AMG) alternatives. Hence, SDR-G precoding does not seem appropriate for real implementations due to its high numerical complexity. After SDR-G, AU is the precoding technique that leads to the smallest required bandwidth for broadcast transmission as it is illustrated in Fig. 11. In case of broadcast, SB requires a greater bandwidth than AU as it is expected, since AU leads to a higher minimum SNR for a high number of users [9]. Nevertheless, it is observed that AMG/SB leads to a smaller bandwidth than AMG/AU, which suggests that SB performs better with AMG. The reason behind this is that SB exhibits a high performance for a small number of users, and it reaches the upper bound for the case of $K = 2$ as shown in [19]. Since AMG partitions the users into smaller groups, this can explain why SB leads to an smaller bandwidth when it is used by the AMG algorithm.

The greater differences observed in terms of computation time are due to the chosen multicast precoding option. For instance, with broadcast transmission under the simulated parameters listed in Table 1 and $\varrho = -100$ dBm/Hz, SDR-G precoding requires around 176 times more computation time than AU; AU consumes roughly 42 times more time than SB; whereas SB requires around 8 times more time than RBF.

The increment in terms of computation time due to the use of AMG is clearly smaller than the differences observed between different multicast precoding algorithms. More specifically, the increment of computation time with AMG/AU with respect to broadcast/AU is around 4.10 times; whereas the increment of AMG/SDR-G with respect to broadcast/SDR-G is roughly 4.28 times.

As it was discussed in Section III-B, the coherence time of the channel imposes a system requisite for the chosen multicast precoding and user grouping algorithm. It is observed from Fig. 12 that AMG/AU leads to around 50 ms of average computational time, whereas AMG/SB leads to around 8 ms and AMG/RBF leads to 4 ms. AMG/SDR-G requires the highest computation time, which is around 5 s due to its high complexity. As it is observed, SB requires a small computational time compared with SDR-G and AU. The reason behind this is that the number of iterations performed by SB is limited to $\min(N, K)$, which is greatly smaller than the number of iterations and randomizations required by AU and SDR-R (i.e., 100 iterations and 7200 random trials respectively). These simulation results have been obtained with an Intel i7 processor and MATLAB R2020b. Nevertheless, such computation times can be greatly reduced implementing and optimizing the code, e.g., for C++ and using more powerful processor as used in commercial 5G base stations.

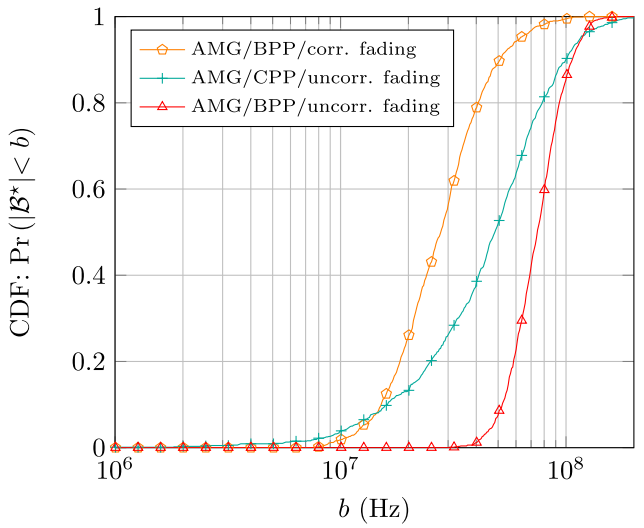


FIGURE 13. Effect of spatial and fading correlations in the distribution of the required bandwidth. It is considered $K = 30$ and $\varrho = -100$ dBm/Hz. The case of spatial correlation considers a CPP with 3 and a cluster radius, $d_d = 15$ m, but it considers that the fading of different UEs is independent. The case of fading correlation considers the distance dependent correlation model with $\beta = 0.01$, but the UE locations follow a BPP.

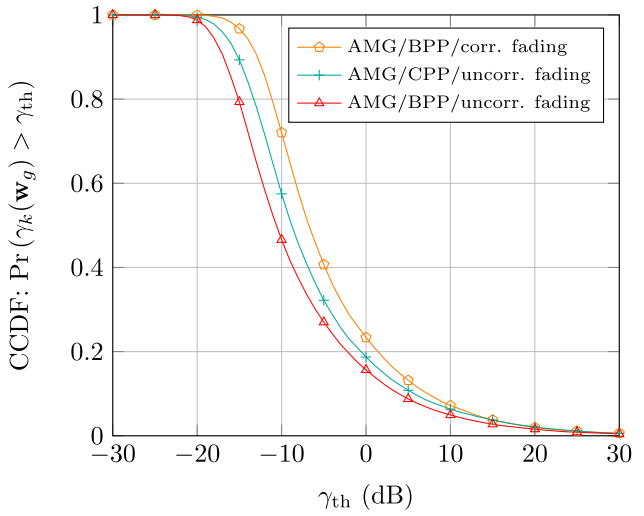


FIGURE 14. Effect of spatial and fading correlations in the complementary distribution of the SNR of a randomly chosen UE.

Finally, the saving in computation time due to the use of an early stopping is also assessed for the case of AMG/AU in Fig. 12. It is observed that such a saving is around 26 times, which is considerable bearing in mind that the early stopping leads to minor reduction in required bandwidth.

E. EFFECT OF SPATIAL AND FADING CORRELATIONS

The effect of the spatial and fading correlations is assessed in terms of the CDF of the required bandwidth, in Fig. 13, and CCDF of the SNR of the typical UE, in Fig. 14. To model the fading correlation, it has been selected a factor of $\beta = 0.01$ for the distance dependent correlation model.

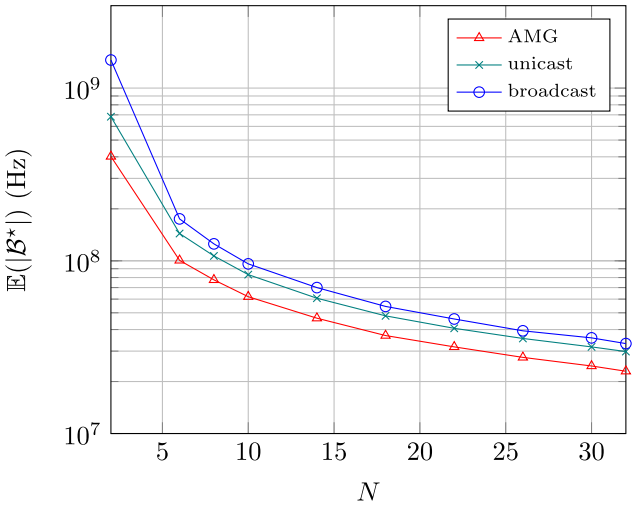


FIGURE 15. Average required bandwidth versus number of RF chains, N. The UE locations follow a BPP with $K = 30$ and it is considered uncorrelated fading with $\varrho_t = -100$ dBm/Hz.

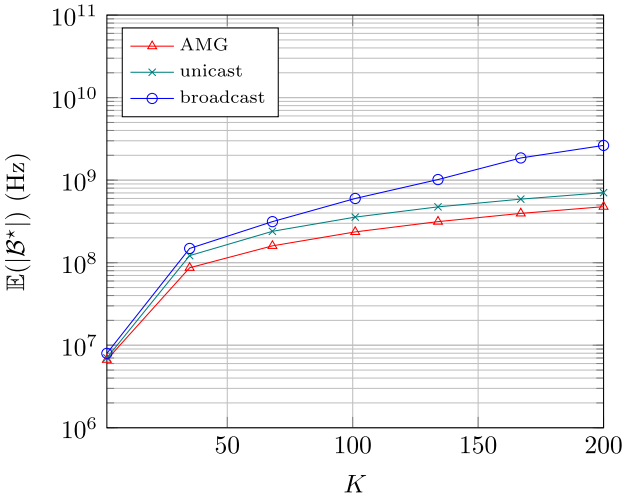


FIGURE 16. Average required bandwidth versus number of users, K. The UE locations follow a BPP and it is considered uncorrelated fading with $\varrho_t = -100$ dBm/Hz.

The spatial correlation has been modeled with a CPP, where 3 clusters of UEs are randomly placed over the sector of the probe RF beam. Each cluster is modeled as a disk of radius $d_d = 15$ m, with 10 UEs each randomly placed.

It is confirmed that both the spatial and fading correlations are beneficial in terms of required bandwidth and SNR of the UEs since they increase the SNR and reduce the bandwidth in statistical terms.

F. EFFECT OF THE NUMBER OF UEs AND RF CHAINS

To conclude this section, the effect of the number of served users and available RF chains is shown in Fig. 15 and 16 respectively. Fig. 15 shows the average required bandwidth ranging from 2 up to 32 RF chains. It is shown the high impact of the number of RF chains. For instance, with AMG, the

decrease of average required bandwidth between $N = 2$ and $N = 32$ is around 17.5 times. In all the simulated range, the improvement of AMG with respect to the other alternatives is notorious.

The average bandwidth versus the number of users, K, is illustrated in Fig. 16. As it is seen, the required bandwidth of all the considered techniques increases as the number of users increases. Nevertheless, the performance of broadcast transmission greatly worsens as the number of users increases, compared to the AMG and unicast approaches. More specifically, the bandwidth reduction of AMG with respect to broadcast transmission is 42% for $K = 35$, 60% for $K = 100$, and 82% for $K = 200$ users. This confirms the observation made in other papers (e.g., [6], [7], [9]) that the performance of multicast precoding deteriorates as the group size grows. This exacerbation of the increase of required bandwidth that happens with broadcast transmission does not happen with AMG, as observed in Fig. 16. This is due to the fact that AMG smartly divides the UE set into smaller groups and thus it benefits from higher multicast beamforming gains, even when the number of users increases. Since AMG searches for a sub-optimal number of groups, the growth in the required bandwidth as K increases is less notorious than in the broadcast case, and thus AMG is even more appealing as K increases.

V. CONCLUSION

In this paper a novel grouping and precoding scheme, named AMG, has been proposed. This algorithm relies on a new formulation of the multicast problem that aims at minimizing the required bandwidth, since it is a key metric to increase the operator profits. The implementation aspects of the proposal have been addressed. To this end, its suitability to be integrated in the context 5G NR, using the signaling mechanisms and the available measurements has been discussed. Extensive simulation results have been provided to demonstrate the benefits of the proposal. Hence, the proposal has been assessed in terms of the mean and distribution of the required bandwidth, the complementary distribution of the SNR, the PMF of the optimal number of groups, and the average computation time. Different grouping and multicast precoding algorithms have been compared under different fading and spatial correlation models. Results reveal that the proposed approach reduces the required bandwidth up to 82% for 200 users compared to existing schemes. It is also shown that AMG leads to a greater SNR for a randomly chosen user, and it reduces the variance of the required bandwidth, which eases the implementation in real networks.

REFERENCES

[1] J. J. Gimenez, J. L. Carcel, M. Fuentes, E. Garro, S. Elliott, D. Vargas, C. Menzel, and D. Gomez-Barquero, "5G new radio for terrestrial broadcast: A forward-looking approach for NR-MBMS," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 356–368, Jun. 2019.

[2] E. Garro, M. Fuentes, J. L. Carcel, H. Chen, D. Mi, F. Tesema, J. J. Gimenez, and D. Gomez-Barquero, "5G mixed mode: NR multicast-broadcast services," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 390–403, Jun. 2020.

[3] *Multimedia Broadcast/Multicast Service (MBMS); Architecture and Functional Description*, document TS 23.246, Rev. 16.1.0, 3GPP, Nov. 2020.

[4] D. Striccoli, G. Piro, and G. Boggia, "Multicast and broadcast services over mobile networks: A survey on standardized approaches and scientific outcomes," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1020–1063, 2nd Quart., 2019.

[5] *Study on Single-Cell Point-to-Multipoint Transmission for E-UTRA*, document TR 36.890, 3GPP, Jul. 2015.

[6] M. Alodeh, D. Spano, A. Kalantari, C. G. Tsinos, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Symbol-level and multicast precoding for multiuser multiantenna downlink: A state-of-the-art, classification, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1733–1757, 3rd Quart., 2018.

[7] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[8] X. Xu, B. Du, and C. Wang, "On the bottleneck users for multiple-antenna physical-layer multicasting," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2977–2982, Jul. 2014.

[9] B. Gopalakrishnan and N. D. Sidiropoulos, "High performance adaptive algorithms for single-group multicast beamforming," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4373–4384, Aug. 2015.

[10] K.-X. Li, L. You, J. Wang, and X. Gao, "Physical layer multicasting in massive MIMO systems with statistical CSIT," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1651–1665, Feb. 2020.

[11] E. Karipidis, N. D. Sidiropoulos, and Z. Q. Luo, "Far-field multicast beamforming for uniform linear antenna arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 10, pp. 4916–4927, Oct. 2007.

[12] E. Karipidis, N. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[13] D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Weighted fair multicast multigroup beamforming under per-antenna power constraints," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5132–5142, Oct. 2014.

[14] S. Gautam, E. Lagunas, A. Bandi, S. Chatzinotas, S. K. Sharma, T. X. Vu, S. Kisseleff, and B. Ottersten, "Multigroup multicast precoding for energy optimization in SWIPT systems with heterogeneous users," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 92–108, 2020.

[15] M. Dong and Q. Wang, "Multi-group multicast beamforming: Optimal structure and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 3738–3753, May 2020, doi: 10.1109/TSP.2020.2994753.

[16] N. Jindal and Z.-Q. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 1841–1845.

[17] S. Y. Park and D. J. Love, "Capacity limits of multiple antenna multicasting using antenna subset selection," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2524–2534, Jun. 2008.

[18] S. Y. Park, D. J. Love, and D. H. Kim, "Capacity limits of multi-antenna multicasting under correlated fading channels," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 2002–2013, Jul. 2010.

[19] H. Kim, D. J. Love, and S. Y. Park, "Optimal and successive approaches to signal design for multiple antenna physical layer multicasting," *IEEE Trans. Commun.*, vol. 59, no. 8, pp. 2316–2327, Aug. 2011.

[20] L.-N. Tran, M. F. Hanif, and M. Juntti, "A conic quadratic programming approach to physical layer multicasting for large-scale antenna arrays," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 114–117, Jan. 2014.

[21] A. Abdelkader, A. B. Gershman, and N. D. Sidiropoulos, "Multiple-antenna multicasting using channel orthogonalization and local refinement," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3922–3927, Jul. 2010.

[22] O. T. Demir and T. E. Tuncer, "Alternating maximization algorithm for the broadcast beamforming," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 1915–1919.

[23] C. Lu and Y.-F. Liu, "An efficient global algorithm for single-group multicast beamforming," *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3761–3774, Jul. 2017.

[24] A. De La Fuente, J. J. Escudero-Garzas, and A. Garcia-Armada, "Radio resource allocation for multicast services based on multiple video layers," *IEEE Trans. Broadcast.*, vol. 64, no. 3, pp. 695–708, Sep. 2018.

[25] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast scheduling and resource allocation algorithms for OFDMA-based systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 240–254, 1st Quart., 2013.

[26] T.-P. Low, M.-O. Pun, Y.-W. P. Hong, and C.-C. J. Kuo, "Optimized opportunistic multicast scheduling (OMS) over wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 791–801, Feb. 2010.

[27] S. Pizzi, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G.-M. Muntean, "A unified approach for efficient delivery of unicast and multicast wireless video services," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8063–8076, Dec. 2016.

[28] L. Christodoulou, O. Abdul-Hameed, A. M. Kondoz, and J. Calic, "Adaptive subframe allocation for next generation multimedia delivery over hybrid LTE unicast broadcast," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 540–551, Sep. 2016.

[29] J. Chen, M. Chiang, J. Erman, G. Li, K. K. Ramakrishnan, and R. K. Sinha, "Fair and optimal resource allocation for LTE multicast (eMBMS): Group partitioning and dynamics," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 1266–1274.

[30] G. Araniti, M. Condoluci, M. Cotronei, A. Iera, and A. Molinaro, "A solution to the multicast subgroup formation problem in LTE systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 149–152, Apr. 2015.

[31] C. Tan, T. Chuah, and S. Tan, "Adaptive multicast scheme for OFDMA-based multicast wireless systems," *Electron. Lett.*, vol. 47, no. 9, pp. 570–572, Apr. 2011. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/el.2011.0481

[32] G. Araniti, M. Condoluci, A. Iera, A. Molinaro, J. Cosmas, and M. Behjati, "A low-complexity resource allocation algorithm for multicast service delivery in OFDMA networks," *IEEE Trans. Broadcast.*, vol. 60, no. 2, pp. 358–369, Jun. 2014.

[33] C. Tan, T. Chuah, S. Tan, and M. Sim, "Efficient clustering scheme for OFDMA-based multicast wireless systems using grouping genetic algorithm," *Electron. Lett.*, vol. 48, no. 3, pp. 184–186, Feb. 2012. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/el.2011.3429

[34] L. You, X. Gao, G. Y. Li, X.-G. Xia, and N. Ma, "BDMA for millimeter-wave/terahertz massive MIMO transmission with per-beam synchronization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1550–1563, Jul. 2017.

[35] K. Xu, Z. Shen, Y. Wang, X. Xia, and D. Zhang, "Hybrid time-switching and power splitting SWIPT for full-duplex massive MIMO systems: A beam-domain approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7257–7274, Aug. 2018.

[36] Z. Shen, K. Xu, X. Xia, W. Xie, and D. Zhang, "Spatial sparsity based secure transmission strategy for massive MIMO systems against simultaneous jamming and eavesdropping," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3760–3774, Jun. 2020, doi: 10.1109/TIFS.2020.3002386.

[37] Z. Shen, K. Xu, and X. Xia, "Beam-domain anti-jamming transmission for downlink massive MIMO systems: A stackelberg game perspective," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2727–2742, Mar. 2021, doi: 10.1109/TIFS.2021.3063632.

[38] D. Eager, M. Vernon, and J. Zahorjan, "Minimizing bandwidth requirements for on-demand data delivery," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 5, pp. 742–757, Sep. 2001.

[39] Y. Cui, B. Li, and K. Nahrstedt, "OStream: Asynchronous streaming multicast in application-layer overlay networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 1, pp. 91–106, Jan. 2004.

[40] J. Choi, A. Reaz, and B. Mukherjee, "A survey of user behavior in VoD service and bandwidth-saving multicast streaming schemes," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 1, pp. 156–169, Feb. 2012.

[41] K. Chitti, F. Rusek, and C. Tumula, "Multiuser bandwidth minimization with individual rate requirements for non-orthogonal multiple access," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Jun. 2017, pp. 1–7.

[42] K. Chitti, F. Rusek, and C. Tumula, "Bandwidth minimization under probabilistic constraints and statistical CSI for NOMA," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.

[43] N. Krishnan, R. D. Yates, N. B. Mandayam, and J. S. Panchal, "Bandwidth sharing for relaying in cellular systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 117–129, Jan. 2012.

[44] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[45] H. Ju, Y. Long, X. Fang, R. He, and L. Jiao, "Systematic beam management in mmWave networks: Tradeoff among beam coverage, link budget, and interference control," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15325–15334, Dec. 2020.

[46] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, Sep. 2018.

[47] E. Onggosanusi, S. Rahman, L. Guo, Y. Kwak, H. Noh, Y. Kim, S. Faxer, M. Harrison, M. Frenne, S. Grant, R. Chen, R. Tamrakar, and Q. Gao, "Modular and high-resolution channel state information and beam management for 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 48–55, Mar. 2018.

[48] M. Enescu, *5G New Radio: A Beam-based Air Interface*, M. Enescu, Ed. Hoboken, NJ, USA: Wiley, 2020.

[49] K. Roth and J. A. Nossek, "Arbitrary beam synthesis of hybrid beamforming systems for beam training," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 714–717, Dec. 2017.

[50] T. Bai and R. W. Heath, Jr., "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb. 2014.

[51] M. Di Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5038–5057, Sep. 2015.

[52] J. Fan, L. Han, X. Luo, Y. Zhang, and J. Joung, "Beamwidth design for beam scanning in millimeter-wave cellular networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1111–1116, Jan. 2020.

[53] M. Cheng, J.-B. Wang, Y. Wu, X.-G. Xia, K.-K. Wong, and M. Lin, "Coverage analysis for millimeter wave cellular networks with imperfect beam alignment," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8302–8314, Sep. 2018.

[54] Y. Li, J. G. Andrews, F. Baccelli, T. D. Novlan, and J. C. Zhang, "Design and analysis of initial access in millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6409–6425, Oct. 2017.

[55] T. Novlan, R. Ganti, A. Ghosh, and J. Andrews, "Analytical evaluation of fractional frequency reuse for OFDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4294–4305, Dec. 2011.

[56] H.-B. Chang and I. Rubin, "Optimal downlink and uplink fractional frequency reuse in cellular wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2295–2308, Apr. 2015.

[57] J. Liu, K. Au, A. Maaref, J. Luo, H. Baligh, H. Tong, A. Chassaigne, and J. Lorca, "Initial access, mobility, and user-centric multi-beam operation in 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 35–41, Mar. 2018.

[58] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "Standalone and non-standalone beam management for 3GPP NR at mmWaves," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 123–129, Apr. 2019.

[59] M. Haenggi, *Stochastic Geometry for Wireless Netw.*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[60] A. Hughes *et al.*, "Measuring the impact of beamwidth on the correlation distance of 60 GHz indoor and outdoor channels," *IEEE Open J. Veh. Technol.*, vol. 2, pp. 180–193, Mar. 2021, doi: 10.1109/OJVT.2021.3067673.

[61] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[62] *Technical Specification Group Radio Access Network; NR; Radio Resource Control (RRC) Protocol Specification (Release 16)*, Standard TS 38.331, Rev. 16.3.1, 3GPP, Jan. 2021.

[63] H. Huang, C. B. Papadias, and S. Venkatesan, *MIMO Communication for Cellular Networks*. New York, NY, USA: Springer, 2012.

[64] *Technical Specification Group Radio Access Network; NR; Multiplexing Channel Coding (Release 16)*, Standard TS 38.212, Rev. 16.4.0, 3GPP, Jan. 2021.

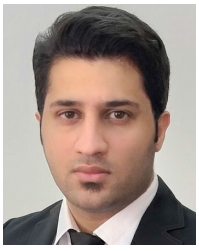
[65] V. Kumar and N. B. Mehta, "Exploiting correlation with wideband CQI and making differential feedback overhead flexible in 4G/5G OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2579–2591, Apr. 2021.

[66] *Tech. Specification Group Radio Access Network; NR; Phys. Layer Procedures for Data (Release 16)*, Standard TS 38.214, Rev. 16.4.0, 3GPP, Jan. 2021.

[67] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, "5G evolution: A view on 5G cellular technology beyond 3GPP release 15," *IEEE Access*, vol. 7, pp. 127639–127651, 2019.

[68] V. Va, J. Choi, and R. W. Heath, Jr., "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5014–5029, Jun. 2017.

FRANCISCO J. MARTÍN-VEGA received the B.Sc. and Ph.D. degrees in telecommunication engineering from the University of Málaga, Spain, in 2011 and 2017, respectively, and the M.Sc. degree in signal processing from the University of Vigo, Spain, in 2014. From 2011 to 2017, he was an Associate Researcher with the University of Málaga, where he participated in contracts with several industry partners related to cellular and satellite communications. He was with Keysight Technologies as a SW Research and Development Engineer, from 2017 to 2021, developing 5G cutting-edge technology for real-time communication systems. He was a Visiting Researcher with Paris-Saclay University, in 2014, and the Queen Mary University of London, in 2016. He has been a Postdoctoral Researcher with the University of Málaga, since February 2021. His main research interest includes mathematical analysis and optimization of 5G and 6G networks. He was awarded the best master's and Ph.D. thesis by the Official College of Telecommunication Engineers of Spain, in 2012 and 2018, respectively.



FARSHAD ROSTAMI GHADI received the B.Sc. degree in electrical communication engineering from the Babol Noshirvani University of Technology, Babol, Iran, in 2014, the M.Sc. degree in electrical communication systems engineering from the Shahrood University of Technology, Shahroud, Iran, in 2017, and the Ph.D. degree (Hons.) in electrical communication systems engineering from the Ferdowsi University of Mashhad, Mashhad, Iran, in 2021. He is currently a Postdoctoral Research Fellow with the Communication Engineering Department, Universidad de Málaga, Málaga, Spain. His research interests include analyzing wireless communication networks, network information theory, and copula theory, with an emphasis on wireless channel modeling and physical layer security.

F. JAVIER LÓPEZ-MARTÍNEZ (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunication engineering from the University of Málaga, Spain, in 2005 and 2010, respectively. He was an Associate Researcher with the Communication Engineering Department, University of Málaga, between 2005 and 2012. He was a Marie Curie Postdoctoral Fellow with the Wireless Systems Laboratory, Stanford University, from 2012 to 2014, and the University of Málaga, from 2014 to 2015. Since 2015, he has been a Faculty Member of the Communication Engineering Department, University of Málaga, where he is currently an Associate Professor. He has been a Visiting Researcher with University College London, in 2010, and Queen's University Belfast, in 2018. His research interests include a diverse set of topics in the wide areas of communication theory and wireless communications, including stochastic processes, wireless channel modeling, physical layer security, and wireless powered communications. He has received several research awards, including the Best Paper Award from the Communication Theory Symposium at the IEEE GLOBECOM 2013, the IEEE COMMUNICATIONS LETTERS Exemplary Reviewer Certificate, in 2014 and 2019, and the IEEE TRANSACTIONS ON COMMUNICATIONS Exemplary Reviewer Certificate, in 2014, 2016, and 2019. He is an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, in the area of wireless communications.

GERARDO GÓMEZ received the B.Sc. and Ph.D. degrees in telecommunication engineering from the University of Málaga, Spain, in 1999 and 2009, respectively. From 2000 to 2005, he was with Nokia Networks and Optimi Corporation (acquired by Ericsson), leading the area of quality of service (QoS) for 2G and 3G cellular networks. In 2005, he joined the University of Málaga, where he is currently an Associate Professor with the Communications Engineering Department. His research interests include the field of mobile communications, especially physical layer security, stochastic geometry, interference management, and QoS/QoE evaluation for multimedia services.