Received 14 February 2022; revised 14 March 2022; accepted 18 March 2022. Date of publication 22 March 2022; date of current version 28 March 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3161312

Nonorthogonal Multiple Access and Subgrouping for Improved Resource Allocation in Multicast 5G NR

ENEKO IRADIER [1] (Member, IEEE), MAURO FADDA [2] (Member, IEEE), MAURIZIO MURRONI [2] (Senior Member, IEEE), PASQUALE SCOPELLITI [3] (Member, IEEE), GIUSEPPE ARANITI [3] (Senior Member, IEEE), AND JON MONTALBAN [1] (Senior Member, IEEE)

[1]Department of Communications Engineering, University of the Basque Country, 48013 Bilbao, Spain

[2]Department of Electrical and Electronic Engineering—DIEE/UdR CNIT, University of Cagliari, 09124 Cagliari, Italy

[3]Department of Information Engineering, Infrastructure, and Sustainable Energy—DIIES, Mediterranea University of Reggio Calabria, 89124 Reggio Calabria, Italy

CORRESPONDING AUTHOR: E. IRADIER (e-mail: eneko.iradier@ehu.eus)

This work was supported in part by the Italian Ministry of University and Research (MIUR), within the Smart Cities framework, Project Cagliari2020 ID: PON04a2_00381; in part by the Basque Government under Grant IT1234-19; and in part by the Spanish Government [Project PHANTOM under Grant RTI2018-099162-B-I00 (MCIU/AEI/FEDER, UE)].

ABSTRACT The ever-increasing demand for applications with stringent constraints in device density, latency, user mobility, or peak data rate has led to the appearance of the last generation of mobile networks (i.e., 5G). However, there is still room for improvement in the network spectral efficiency, not only at the waveform level but also at the Radio Resource Management (RRM). Up to now, solutions based on multicast transmissions have presented considerable efficiency increments by successfully implementing subgrouping strategies. These techniques enable more efficient exploitation of channel time and frequency resources by splitting users into subgroups and applying independent and adaptive modulation and coding schemes. However, at the RRM, traditional multiplexing techniques pose a hard limit in exploiting the available resources, especially when users' QoS requests are unbalanced. Under these circumstances, this paper proposes jointly applying the subgrouping and Non-Orthogonal Multiple Access (NOMA) techniques in 5G to increase the network data rate. This study shows that NOMA is highly spectrum-efficient and could improve the system throughput performance in certain conditions. In the first part of this paper, an in-depth analysis of the implications of introducing NOMA techniques in 5G subgrouping at RRM is carried out. Afterward, the validation is accomplished by applying the proposed approach to different 5G use cases based on vehicular communications. After a comprehensive analysis of the results, a theoretical approach combining NOMA and time division is presented, which improves considerably the data rate offered in each use case.

INDEX TERMS 5G, ADR, LDM, NOMA, P-NOMA, resource allocation, RRM, subgrouping, wireless communications.

I. INTRODUCTION

THE OBJECTIVE of the latest generation of cellular networks (i.e., 5G) is to provide groundbreaking connectivity to everyone and everything, everywhere and everytime [1]. Indeed, 5G networks are expected to cover a wide range of use cases with customized parameters: devices, type of users, requirements and mobility classes. The international research community, especially, 3^{rd} Generation Partnership Project (3GPP), is carrying out an important work designing a standard able to satisfy most of the current user demands. In particular, while the first 5G Releases (i.e., Rel-15 and Rel-16) focus on the design of the 5G New Radio (NR) and features related with the use cases defined in [1], Rel-17 and beyond is expected to include (but not limited to) enhanced support for wireless and wired convergence, multicast and broadcast architecture, proximity

services, multi-access edge computing, and network automation [2]. However, to enable all the potential use cases, obtaining high spectral efficiency becomes one of the main goals to be achieved. Spectral efficiency is closely related to radio resource management (RRM), and in 5G, as in the previous generation, orthogonal multiple access (OMA) techniques are the basis of the RRM module. Consequently, although 5G has increased the spectral efficiency compared to 4G adding flexibility to RRM in the definition of RB allocation, there appears to be still room for improvement.

An alternative to increase the overall network spectral efficiency is to use multicast transmissions. In these transmissions, the users access the same content, and the RRM module performs an efficient link adaptation procedure according to the channel conditions experienced by the multicast users. This adaptation has to be accomplished on a per-group basis, taking into account the Channel State Indicator (CSI) of all the users registered to a given multicast service. However, in single rate-transmissions the presence of cell-edge users, which experience poor channel conditions can strongly degrade the Quality of Service (QoS) that the cellular infrastructure could provide. Subgrouping is a solution that has been proposed in literature as an effective technique to overcome those limitations. Based on splitting the multicast users into subgroups and applying subgroup-based adaptive modulation and coding schemes (MCS), it enables more efficient exploitation of multi-user diversity by optimizing opportune cost functions [3]. This solution is very useful for vehicular communications, where the high and rapid variation of the receivers positions lead to unequal reception conditions [4], [5].

In general, vehicular communications imply unbalanced QoS requests from multicast services (e.g., groups with different throughput), and under this situation, OMA could limit the full exploitation of the resources. In such a scenario, Non-Orthogonal Multiple Access (NOMA) techniques have shown the power to allow increasing spectral efficiency [6]. With respect to OMA, where the available frequency/time resources in the network are orthogonally assigned to the different User Equipments (UEs), in NOMA, UEs share the available resources, both in frequency and time. On the one hand, in OMA at the receiver side, under perfect conditions, the desired data can be unequivocally separated from the rest of the information. On the other hand, in NOMA when decoding the desired content, the rest of the signals are considered an additional source of interference. In [7], NOMA techniques have shown higher efficiencies in comparison with OMA under specific conditions, such as when the throughput rate among different users is asymmetric or for the downlink transmission mode (i.e., from the transmitter to the users) [8].

Based on previous promising results, the combination of NOMA with subgrouping techniques can considerably enhance typical spectral efficiency values. In this line, authors in [9] explored by the first time the joint use of subgrouping multicast techniques and NOMA in an evolved multimedia broadcast multicast service (eMBMS)-like 5G scenario, where different quality video services were delivered to a group of users interested in the same contents. The potentiality of the proposed approach was preliminarily investigated in some envisaged 5G mobility environments, and the results in terms of QoS indicators such as maximum throughput (MT) and aggregated data rate (ADR) were promising. However, although [9] was oriented to 5G networks, the transceiver architecture used the numerology of 4G, and the cost functions of the resource allocation strategies were theoretical (i.e., Shannon's capacity approach). Doubtlessly, there is a relevant gap in the literature between [9] and realistic 5G networks, and our paper overcomes those limitations with a novel cost-function design and evaluation. Therefore, the objective of this paper is to provide an analysis of the requirements and benefits of subgrouping with NOMA techniques in 5G scenarios. Consequently, an in-depth analysis of the implications of introducing NOMA techniques in 5G subgrouping at the resource allocation level (i.e., RRM) is carried out, and the approach is formalized, validated and compared with existing solutions. In summary, the novel technical contribution of this paper is the comprehensive proposal and evaluation of subgrouping techniques combined with 5G and NOMA which includes:

1) Detection of the drawbacks of using NOMA for subgrouping and provide a solution based on the adaptation of the Signal-to-Noise Ratio (SNR) of the received signal.
2) Design of a new evaluation cost function for 5G subgrouping based on NOMA that maximizes the overall network throughput.
3) Identification of the network variable constraints that most affect the cost function performance and a comprehensive evaluation of the results.
4) Comparative study of the proposed solution performance with existing subgrouping methods.
5) Design of the system further evolution and theoretical demonstration for a joint T/FDMA+NOMA approach.

The rest of the paper is organized as follows. The next section describes the basic principles and related work published in literature for subgrouping based on both OMA and NOMA techniques. In Section III, the system model is introduced. Then, Section IV presents the problem formulation and the proposed solutions related with NOMA-based subgrouping. The simulation framework is presented in Section V, whereas Section VI is dedicated to the analysis of the results and a comparative study with previous works and other subgrouping techniques. Future works and system evolution are addressed in Section VII. Finally, Section VIII draws the conclusions.

II. BASIC PRINCIPLES AND RELATED WORK

Multicast multi-rate schemes allow each user to receive multimedia traffic based on the capabilities of the UE [10]. In particular, subgrouping is one of the multi-rate techniques

that exploits the group-splitting approach. Nevertheless, subgrouping techniques introduce new issues related to subgrouping formation: proper resource allocation, MCS selection, etc. Consequently, many schemes have been proposed by the research community to achieve the optimal solution. This section introduces the basic concepts of multicast subgrouping techniques (Section II-A) and summarizes the main contributions dividing them between the solutions that do not include NOMA (Section II-B.1) and the ones that take advantage of NOMA (Section II-B.2).

A. MULTICAST SUBGROUPING

Multi-rate multicast transmissions can be divided into two different categories, stream-splitting, and group-splitting. On the one hand, stream-splitting focuses on splitting high-rate multimedia contents into multiple low-rate substreams [11]. On the other hand, group-splitting techniques divide multicast members into different subgroups, each one formed by users experiencing similar channel conditions. In these cases, the resource manager selects the most suitable subgroup configuration dynamically based on the minimization (or maximization) of a given cost function. Moreover, the cost function takes into account the channel quality indicator (CQI) values of the users and the QoS constraints of the multicast session [12].

In general, subgrouping techniques consist of three stages:

- *CQI collection:* the resource manager collects the CQI feedbacks from each UE belonging to the multicast group. In particular, each UE estimates its CQI value based on the Signal to Interference and Noise Ratio (SINR) and the target Block Error Rate (BLER). Afterward, each UE forwards the CQI back (associated with the highest supported MCS) to the resource manager.
- *Subgroup creation:* the multicast members are split into subgroups. The subgrouping algorithm determines the subgroup configuration that: *(i)* allows to maximize the system capacity; *(ii)* guarantees that each served UE can successfully demodulate the received signal; *(iii)* optimizes a given cost function.
- *Resource allocation:* the resource manager shares the time, and frequency channel resources (i.e., Resource Blocks (RBs)) among all the received feedback (i.e., CQI), depending on the configuration defined in the previous step. Particularly, all the users reporting a CQI value equal to the minimum CQI selected for each subgroup will be served with the closest supported MCS.

The optimization of the resource allocation stage can follow different strategies. In particular, the resource manager bases the RB allocation among users on different communication and network aspects, such as maximizing the system throughput, spectral efficiency, energy efficiency, or minimizing the resource utilization. The optimal configuration is selected by evaluating a cost function, which is a mathematical expression made by network or user parameters. In addition, the cost function is continuously evaluated until the maximization/minimization of the target parameter is achieved. Moreover, according to [3], when the allocation strategy is the network data rate maximization, the optimal subgroup configuration is obtained using no more than two subgroup configurations.

B. RELATED WORK

This subsection shows the most relevant and recent works in subgrouping techniques. The subsection is divided into two parts: the contributions based on OMA techniques and NOMA works.

1) OMA TECHNIQUES

One of the most relevant OMA-based subgrouping works is presented in [13]. Casetti *et al.* presented a two-step procedure that merges significantly overlapped areas in space considering similar users' contents. Their algorithm returns multi-content groups that include cells with similar content interests and broadcasts the most demanded items, maximizing the system throughput. Another paper showing the basis of the subgrouping theory is [14], where the authors proposed to introduce expert users in each subgroup to facilitate the subgroup configuration of large-scale groups.

Then, a particular scenario where subgrouping techniques have had a very positive reception is eMBMS. Particularly, authors in [15] presented a heuristic algorithm for eMBMS environments over single frequency networks (MBSFN). In this case, the work studied the area formation in MBSFN to increase the ADR exploiting the multi-rate transmissions. In addition, the algorithm evaluates simultaneously scalable video coding (SVC) techniques and radio resource allocation for efficient spectrum utilization in the configured MBSFN areas.

Apart from mobile networks, OMA-based subgrouping techniques have also been proposed for satellite communications [16]. In this case, the authors presented a multicast subgroup formation method and an Application-Layer Joint Coding (ALJC) technique combination to improve the performance of multicast satellite transmission in terms of throughput and perceived quality of the video streams.

Eventually, in [17], a common message with a precoding structure is proposed to enhance the fairness among users in group multicasting. This technique turned into a rate-splitting algorithm that combines private and shared information in each group. Results indicated that the method performance is optimized for overloaded systems. On the other hand, in [18], the authors combined game theory developments with multicast grouping algorithms. In this case, the request nodes (RNs) are grouped, and the broadcast content is transmitted by the head node (HN), which assigns different reputation values depending on the delivery.

According to the presented literature review, due to their performance and low complexity, OMA multiplexing schemes are widely accepted for subgrouping techniques.

However, these techniques can still improve their network spectral efficiency using a NOMA approach.

2) NOMA TECHNIQUES

One of the first works was presented in [19], where the authors analyzed the benefits of Layered Division Multiplexing (LDM) technology with subgrouping algorithms into the resource allocation mechanism of LTE-A standard for multicast vehicular communications.

Then, [20]–[22] consider NOMA for resource allocation. First, [20] investigates the possibility of a more spectrum efficient solution in mixed unicast-broadcast service delivery in 5G-MBMS using power-based non-orthogonal multiplexing (P-NOM) (i.e., LDM). Particularly, the authors examined the potential capacity gains of using NOMA over OMA in broadband systems (i.e., LTE and 5G-NR). On the other hand, [21] describes a new method to gather users in P-NOMA and NOMA-2000 [22] over different Rayleigh fading channels. NOMA-2000 combines OFDMA signal and multi carrier-code division multiple access (MC-CDMA) signal, spread and superposed to OFDMA waveform.

During the last few years, the use of NOMA for multicasting has been applied in different use cases apart from multimedia content distribution. For example, in [23], a cooperative transmission scheme is proposed for unicast/multicast transmissions in a downlink Cognitive Radio (CR)-NOMA system. In this case, the authors evaluated the performance improvement when the number of secondary users is increased. Then, the work in [24] presents a multiple access method called power domain sparse code multiple access (PSMA) for heterogeneous networks (HetNets). This method uses codebooks that can be reused more than once in the coverage area of a base station to improve the system spectral efficiency. The signal model and the PSMA detection techniques with power domain (PD)-NOMA and sparse code multiple access (SCMA) were compared. In addition, in [25], a simple user grouping and pairing scheme for NOMA in a downlink visible light communication system is presented. The proposed scheme is a mix of conventional NOMA and OMA schemes where every two users are paired using NOMA. Then, all pairs are allocated with conventional OMA. The performance of the proposed scheme is compared to conventional OMA in terms of maximum sum rate.

As shown in the previous references, NOMA is a promising alternative to increase the network spectral efficiency. However, the already published works do not cover the implications of implementing NOMA under the 5G stringent requirements. Therefore, there is a gap in current literature considering the evaluation of NOMA techniques in combination with subgrouping techniques to increase the network spectral efficiency of 5G.

III. SYSTEM MODEL

In this paper, we refer to a single-cell 5G coverage area represented by a next-generation base station (gNB), providing services to users with different mobility types. The network is composed of K users (i.e., $e_1, e_2, \ldots, e_K$), where each user informs the gNB of the instantaneous CQI. Based on the CQIs, the transmission from the gNB to the users is set with a given MCS value. Let M be the number of available MCS levels (i.e., $m = 1, 2, \ldots, M$). For each MCS, a certain spectral efficiency is achieved in the transmission. We identify with t_m the spectral efficiency of each MCS value in m. The greater is the spectral efficiency (t_m), the lower is the number of required resources to achieve a given data rate.

This work is based on the 5G NR allocation system, which is carried out in the RB basis for the resource allocation issue. Mainly, one RB is the smallest frequency resource that can be assigned to a UE; each RB corresponds to 12 consecutive and equally spaced subcarriers. NR supports multiple numerologies with different Subcarrier Spacing (SCS) values, according to the following equation:

$$\Delta f = 15 * 2^{\mu}(kHz), \tag{1}$$

where μ is the numerology. In this work, $\mu = 0$ is used, which is oriented to enhanced Mobile Broadband (eMBB) applications. Let Nrb be the number of available RBs in the channel bandwidth and r the portion of RBs assigned to a particular multicast group, where $r = 1, 2, \ldots, Nrb$. In this work, each RB is a sub-channel of 180 kHz formed by 12 consecutive and equally spaced sub-carriers.

Then, the gNB enables the subgrouping configuration that maximizes the network data rate (i.e., ADR). In this case, subgrouping configurations imply the number of users in each subgroup using a particular MCS value and the number of RBs allocated for each group. The gNB evaluates different configurations and uses the one maximizing the ADR. Therefore, the optimization problem can be formulated as:

$$\max \sum_{k=1}^{K} t_{m,k} \cdot r_k, \tag{2}$$

where $t_{m,k}$ is the spectral efficiency received by the k^{th} user using the m^{th} MCS value, and r_k is the number of RBs dedicated to the k^{th} user. In this case, the received spectral efficiency ($t_{m,k}$) and the number of resource blocks dedicated to the user (r_k) are defined according to the group to which they belong.

According to the proposed system model, all the multicast users decide their CQI level (m) based on the instantaneous SNR. However, as shown in Fig. 1, the instantaneous SNR is different in NOMA and TDMA. Remarkably, while in TDMA, the SNR is calculated as the difference between the signal power and the noise power, in NOMA, the signal dedicated for Group 2 interferes with the Group 1 signal. Therefore, the straightforward relation between CQI and MCS is not possible in NOMA, and an SNR adaptation is required.

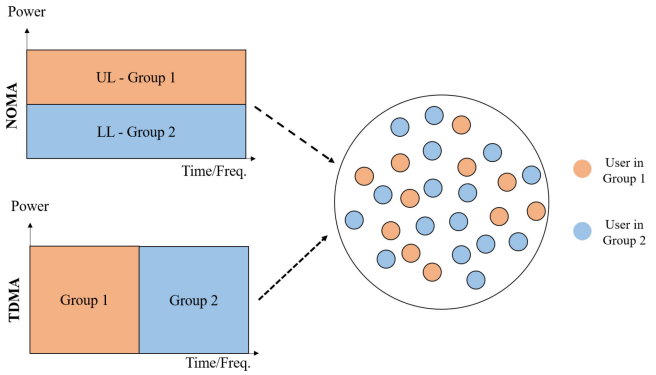

FIGURE 1. Graphical representation of the channel resource sharing.

In addition, in NOMA systems, the data transmitted in each layer is separated by an Injection Level (IL). In particular, IL is a power difference applied to the layers that directly affects the SNR of each layer. Therefore, the optimization problem stated in Eq. (2) is only valid for OMA subgrouping systems, and, so, a specific solution is required for NOMA.

The following section details the basis of each problem and presents a solution for each one.

IV. PROBLEM FORMULATION AND SOLUTION PROPOSAL

After a deep analysis of the system model and its implications, we have identified two problems for implementing NOMA in 5G-based subgrouping techniques: 1) the SNR variation compared to OMA subgrouping, and 2) not possible to apply the OMA-based data rate maximization techniques straightforwardly in NOMA. This section focuses on presenting the characteristics of the identified problems and proposing solutions.

A. SNR ADAPTATION

OMA techniques organize their services on single-layer mode, where the services do not share the time/frequency resources and, therefore, in the resource grid, each RB can be allocated for a single layer. Therefore, the SNR ratio associated with single-layer systems (i.e., γ_{sl}) can be calculated as:

$$\gamma_{sl} = \frac{P_0|h_0|^2}{\sum_{z=1}^{Z} P_z|h_z|^2 + n_0}, \tag{3}$$

where, $z = 1, \ldots, Z$ is the set of interfering signals, P_0 is the power of the multicast signal, P_z is the power of the interference signal, n_0 is the gaussian noise, and h_0 and h_z are the channel coefficients of the multicast signal and the z^{th} interfering signal, respectively.

As different services can share the same time/frequency resources using NOMA, each RB is organized in two layers, one for each subgroup. When configuring the overall content distribution, the Injection Level (IL) is one of the most relevant parameters of NOMA systems. In particular, IL is the power allocation difference between the Upper Layer (UL) and the Lower Layer (LL), which indicates how many dB separate the UL from the LL. Therefore, variations in the IL have direct implications on the available SNR values for each layer. This effect is translated in an increase on the required SNR to decode both layers compared with the case where OMA is used. In particular, the SNR required for each layer in linear units can be calculated as:

$$\gamma_i = \frac{1}{\frac{\sigma_i}{\gamma_{SL}} - \sum_{j=i+1}^{N} \sigma_j}, \tag{4}$$

where, N is the total number of layer (i.e., $N = 2$ in this work), i is the layer identifier (i.e., from 1 to N), γ_{SL} is the equivalent SNR value of the corresponding layer configuration in single-layer mode and σ_i is power allocation ratio for layer i.

Following the analysis carried out in [26], σ_i can be calculated as:

$$\sigma_i = \frac{10^{\frac{\Delta_{i-1}}{10}}}{\sum_{i=1}^{N} 10^{\frac{\Delta_{i-1}}{10}}}, \tag{5}$$

where, Δ is the IL expressed in dB.

Then, considering Eq. (4) and Eq. (5), the SNR adaptation for a two-layer system in dB units can be mathematically expressed based on the single-layer case, as defined in [27]. First, the required SNR to decode the UL can be calculated as follows:

$$\gamma_{ul} = \gamma_{sl} + 10 * \log_{10}\left(\frac{1 + 10^{(\Delta/10)}}{1 - 10^{((\gamma_{sl}+\Delta)/10)}}\right), \tag{6}$$

where γ_{ul} and γ_{sl} are the required SNR to decode the UL and the equivalent single-layer signal, respectively. Regarding Eq. (6), the argument of the log function could be negative and, consequently, γ_{ul} would not be defined. To avoid this, a validity range of γ_{ul} is defined with following condition: $\gamma_{sl} + \Delta < 0$. Therefore, it is not possible to apply any Δ value. The election of the IL is conditioned by the required SNR to the decode the signal in single-layer mode. Moreover, the more robust γ_{sl} is (i.e., low SNR values), the greater is the range of applicable Δ values. In addition, as expected, the higher the IL (Δ) value, the more similar are the γ_{ul} and γ_{sl}.

Then, the required SNR to decode the LL can be also calculated based on the single-layer SNR [27]:

$$\gamma_{ll} = \gamma_{sl} - \Delta + 10\log_{10}\left(1 + 10^{(\Delta/10)}\right), \tag{7}$$

where γ_{ll} is the required SNR to decoded the LL.

In Fig. 2, the performance degradation for both layers is shown for different ILs and MCS schemes. The x-axis represents each of the available MCS values in 5G NR according to [28, Table 5.1.3.1-2], while in the y-axis the required SNR (in dB) value to decode each MCS is shown. The solid black line represents the single-layer case. Regarding UL, it is probed that the higher the IL is, the lower is the difference between single-layer and NOMA case. On the other hand, LL presents the opposite behavior.

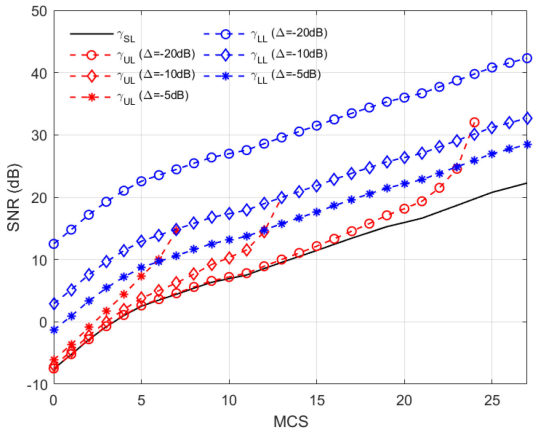

FIGURE 2. SNR variation for different MCS cases, layers, and IL values.

B. USER REORGANIZATION

The users' feedback is delivered by the CQI values, which are related to the maximum MCS that the UE is able to decode correctly. The CQI calculation is performed assuming a single-layer transmission mode, and therefore, before applying any NOMA-based subgrouping technique, the SNR adaptation formulas (6) and (7) have to be implemented. However, for the gNB, the CQI feedback represents a range of possible SNRs.. In fact, when a user requests a specific MCS, the current SNR value that the user has is inside a range of possible SNR values limited by the boundary SNR values of the MCSs. The following expression represents the SNR range: $\gamma_N \leq \gamma_{real} < \gamma_{N+1}$, where N is the MCS requested by the user and γ_N and γ_{N+1} are the minimum required SNR values to decode the MCS N and $N+1$, respectively. Consequently, when the gNB receives the CQI, it cannot guess the exact SNR value that the user has in order to carry out the SNR adaptation in the NOMA case. This uncertainty could downgrade the delivered capacity or, in the worst case, let some users out of the service because they have a real SNR value below the assumed value. The uncertainty cannot be completely solved, but in this paper it is proposed to reorganize the CQIs of users in the subgrouping process to minimize its effect. In particular, three different user reorganization algorithms have been developed and included in the SNR adaptation process together with the cost function evaluation.

1) WORST CASE ELECTION

The first technique assumes that, within the range of possible SNR values (i.e., $\gamma_N \leq \gamma_{real} < \gamma_{N+1}$) that a user can have before requesting a specific MCS values, the user will have the worst possible:

$$\gamma_i' = \gamma_N, \tag{8}$$

where γ_i' is the SNR value assumed by this method for the i^{th} user. This implies that the capacity that NOMA is going to offer after the subgrouping configuration will be downgraded since not all users will require such strict SNR values within the range of possible SNRs. On the other hand, none of the users will receive a signal with a required SNR higher than their current SNR. In short, it is not a realistic assumption and it is very pessimistic.

2) UNIFORM SNR DISTRIBUTION

This algorithm takes for granted that users requesting the same MCS value will have different SNR values. Consequently, the UEs are uniformly distributed among the possible SNR values:

$$\gamma_i' = U(\gamma_N, \gamma_{N+1}). \tag{9}$$

In this case, each of the i^{th} users have a different γ_i' because they belong to different realizations of the uniform function. It represents a more realistic case and the final capacity will not be severely downgraded. However, it might be the case where a particular user is assigned to a group with higher requirements than its actual SNR value.

3) UNIFORM DISTANCE DISTRIBUTION

Another alternative is to assume that the position of the users within the coverage ring related to a particular MCS value are uniformly distributed:

$$d_i' = U(d_N, d_{N+1}), \tag{10}$$

where d_N and d_{N+1} are the distances between user and gNB associated to γ_N and γ_{N+1}, respectively. Then, the obtained distance for the i^{th} user (i.e., d_i') is converted to a SNR value (i.e., γ_i') following the corresponding distance attenuation relation (see Section V-B). As in the previous case, it represents a more realistic scenario and, on the other hand, the throughput will not be downgraded. Nevertheless, a particular mismatch could also happen between the real and the assumed SNR after the reorganization.

C. SUBGROUPING ALGORITHM

This section shows the resource allocation algorithms for NOMA and TDMA. Both define different subgroup configurations by sharing the available channel time and frequency resources (i.e., RB), evaluating the CQI-based feedback received from the users. As mentioned in Section II-A, the resource allocation stage evaluates a cost function following a particular strategy. In this case, the followed strategy is to maximize the network ADR.

First, a TDMA-based cost function has been developed as shown in Algorithm 1. The input is the CQI feedback (i.e., CQI_k) perceived by each k^{th} UE connected to the gNB in each Transmission Time Interval (TTI). Using these collected feedback values, the CQI distribution vector $U = \{u_1, u_2, \ldots, u_C\}$ is generated (line 6), where u_c indicates the number of UEs perceiving a CQI equal to c, c varying from 0 to 27, according to [28, Table 5.1.3.1-2].

Afterwards, the MCS values duple $s_m = [s_{1m}, s_{2m}]$, which provides the maximum cost function value has to be found. The first item, s_{1m}, represents the MCS value assigned to the

Algorithm 1 TDMA Algorithm

1: **Define:** N_{rb}. The number of RB available in the channel bandwidth;
2: **Define:** $m = 1, ..., M$. Set of possible MCS configuration couples;
3: **Let** t_m be the rate achieved using a single resource block using a MCS value = m;
4: **Define:** $r = 1, ..., N_{rb}$. Set of possible RB assigned to the first subgroup;
5: **INITIALIZATION:** The CQI requested by each user has been previously obtained as CQI_k;
6: **Create the CQI distribution vector:**
$$U = \{u_1, u_2, ..., u_C\};$$
7: **ADR calculation of the possible MCS couples and RBs distribution**:
8: **for** $m \leq M$ **do**
9: **for** $r \leq N_{rb}$ **do**
10: $ADR_{r,m} = r * u_{1_m} * t_{1_m} + (N_{rb} - r) * u_{2_m} * t_{2_m}$
11: **end for**
12: **end for**
13: $[s_{1_{r,m}}, s_{2_{r,m}}]$ = Arg max $\{ADR_{r,m}\}$;

Algorithm 2 NOMA Algorithm

1: **Define:** N_{rb}. The number of RB available in the channel bandwidth;
2: **Define:** $m = 1, ..., M$. Set of possible configurations;
3: **Define:** $l = \Delta_1, ..., \Delta_L$. Set of possible injection levels;
4: **INITIALIZATION:** The CQI requested by each user has been previously obtained as CQI_k;
5: **Create the CQI distribution vector:**
$$U = \{u_1, u_2, ..., u_C\};$$
6: **Calculate the SNR related to each CQI according to:**
$$\gamma_{min} = (2^{eff} - 1) * (-log(5 * BER)/1.5);$$
7: **Evaluate different IL values:**
8: **for** $l \leq \Delta_L$ **do**
9: **Apply SNR adaptation and calculate** γ_{min-UL} **and** γ_{min-LL}
10: **Apply the corresponding SNR reorganization technique:**
$$U_{UL} = Reorganize(U, \gamma_{min-UL});$$
$$U_{LL} = Reorganize(U, \gamma_{min-LL});$$
11: **ADR calculation of the possible MCS couples and** Δ **values:**
12: **for** $m \leq M$ **do**
$ADR_{m,l} = N_{rb} * (s_{1,m} * U_{UL,m} + s_{2,m} * U_{LL,m});$
13: **end for**
14: **end for**
15: $[s_{1,m}, s_{2,m}]$ = Arg max $\{ADR_{m,l}\}$;

first group and s_{2m} represents the minimum supported MCS for the second group. Users with lower channel gain are grouped into the s_{1m} subgroup, whereas users under better channel condition are served through higher MCS in the s_{2m} subgroup. For each duple $[s_{1m}, s_{2m}]$, the cost function is iteratively calculated varying the number of resources N_{rb} assigned to each subgroup. Eventually, the configuration that achieves the maximum ADR is the selected (lines 8-12). The ADR is computed as the sum of the datarates of users belonging to both subgroups, according to the MCS values ($[s_{1m}, s_{2m}]$) of the selected configuration and to the number of RB assigned to each subgroup:

$$ADR = r * u_{1_m} * t_{1_m} + (N_{rb} - r) * u_{2_m} * t_{2_m}, \qquad (11)$$

where r is the amount of RBs assigned to first subgroup, t_{1_m} is the rate achieved with a single resource unit (i.e., 1 RB) by users belonging the first subgroup (s_{1m}), t_{2_m} is the rate achieved with a single resource unit by users belonging the second subgroup (s_{2m}), and u_{1_m} and u_{2_m} are the number of users connected to groups one and two, respectively, using the m^{th} MCS value. Finally, the algorithm evaluates the different combinations of the cost function shown in Eq. (11). The optimal solution is the one providing the maximum ADR, as shown in line 13.

In [9], although NOMA was evaluated, the cost function was based on Shannon capacity formula. Therefore, in this work, several modifications have been included to take into account the numerology that imposes 5G NR [28]. The cost function implemented for carrying out 5G subgrouping using NOMA is presented in Algorithm 2. In this case, it is possible to choose one of the three available user reorganization techniques described in Section IV-B (i.e., NOMA (I), NOMA (II), and NOMA (III), respectively) as an additional input. The first step, in line 5, is the same as in Algorithm 1. In the second step (line 6), the minimum SNR (γ_{min}) required to decode the MCS related to each CQI in single-layer mode is evaluated:

$$\gamma_{min} = \left(2^{eff} - 1\right) * (-log(5 * BER)/1.5). \qquad (12)$$

The formula is obtained from [29] and according to the authors, the Bit Error Rate (BER) has been set to $5 \cdot 10^{-5}$. Then, a loop is defined for different Δ values (line 8), evaluating the rest of the steps iteratively. In order to set the range of possible injection levels (Δ), the same range as in ATSC 3.0 has been used [30]. It is composed by 31 values, where steps of 1 dB are used for Δ between -25 dB and -5 dB and steps of 0.5 dB between -5 dB and 0 dB.

The first step inside the loop is to calculate the minimum SNR required to decode the MCS related to each CQI in the two-layer mode (γ_{min-UL} and γ_{min-LL}). This evaluation process is carried out according to the formulas (6) and (7). Once the limit SNR values are calculated, one of the three available user reorganization techniques described before is applied and new CQI distribution vectors are obtained, for UL and LL, respectively (line 10). The results obtained for each of the reorganization techniques will be compared in Section VI. Then, similar to the second step in Algorithm 1, different configurations are searched to find the one with the highest ADR. As shown in line 12, in this case the whole bandwidth is used for configuring both groups, that is, each of the NOMA layers uses N_{rb} resources. This search is iteratively carried out by varying Δ. The combination of ($s_m = [s_{1,m}, s_{2,m}]$) and Δ finally provides the output of the algorithm. Therefore, the evaluated cost function is shown

in line 12 and the optimal solution is the one providing the maximum ADR, as shown in line 15.

1) COMPLEXITY

In addition to the functionality of each cost function, computational complexity is also considered another KPI in RRM algorithms. In this case, the number of operations that define the computational burden of each algorithm is based on a double iterative process that evaluates the cost function for different conditions. Regarding the iterative processes, while in the TDMA-based cost function, the parameters that define the iterative process are m and r, in the case of NOMA are l and m. Therefore, the computational complexity of TDMA presents a linear trend that depends on the number of evaluated MCS and the number of available resource blocks: $O(m \cdot r)$. In the same way, the complexity of NOMA depends on the evaluated MCS values and the possible IL values: $O(m \cdot l)$. In this case, the NOMA-based cost function has fewer iterations since the number of IL values is lower than the number of available RBs for typical 10 MHz and 20 MHz channels (31 vs. 50 and 31 vs. 100, respectively).

Furthermore, it should be underlined that in further implementations, the channel bandwidth size might be broadened (e.g., in the mmWave frequency bands) while the number of evaluated ILs will remain static. The main reason is that increasing the number of evaluated IL values would not significantly affect the performance. On the one hand, ILs below -25 dB would only increase unnecessarily the required SNR of the LL without benefiting the UL, and, on the other hand, including more granularity would not provide high difference with the current IL selection. Consequently, the NOMA-based approach presents better computational complexity characteristics than TDMA, especially when the channel bandwidth increases.

V. SIMULATION ENVIRONMENT

In this section, the simulation methodology and the use cases defined to evaluate the proposed algorithms are shortly described.

A. METHODOLOGY

The simulation framework is based on a dual simulation tool. On the one hand, a network level simulator is used to emulate a specific network configuration and to obtain the periodical CQI reports from the users. And on the other hand, a mathematical simulator is used to implement the algorithms presented in Section IV-C.

Regarding the network level simulator, OMNeT++ has been used [31]. Using this simulation tool, each user in the cell periodically sends its CQI to the gNB and these feedback values are saved and used as input for the second part of the simulation framework. In particular, the CQI election of the users is conditioned by the propagation channel and by the mobility model that each user follows. In particular, the propagation channel model is emulated using the Tapped Delay Line (TDL) model in [32], specifically, TDL-D and TDL-E. These models represent the propagation channel of a current urban area assuming different multipath sources. Furthermore, following the recommendations on [32], different delay spread values are applied to replicate short-, medium- and large-transmission distances. Moreover, both uplink (i.e., feedback transmission from the users to the gNB) and downlink (i.e., multicasting from the gNB to the users) propagation channels are simulated with the above-presented channel models.

Finally, concerning the mathematical simulation tool, MATLAB has been used. In this part, the subgrouping algorithms described in Section IV-C are studied. Two different subgrouping algorithms have been evaluated, based on TDMA and NOMA, Algorithm 1 and Algorithm 2, respectively. Both use the CQI values obtained from the mobility model generator module as input to perform their assessment.

B. SCENARIO

An urban mobile network environment has been emulated as close as possible to the real environments. Two types of users have been defined: low-speed pedestrian and car-mounted receivers. Due to the characteristics of urban environments and the user types, the Random Way Point (RWP) mobility model [33] has been implemented. In addition, the speed is also specified for each receiver type. In this case, pedestrian receivers walk inside the cell with a mean speed of 3 km/h and the car receivers randomly vary their speed between 30 and 50 km/h.

Taking into account that mobile receivers will use small size screens to display the video content, the minimum data rate for delivering High Definition (HD) or Ultra High Definition (UHD) content has been established between 1 and 4 Mbps [34]. The minimum data rate that has to be guaranteed for both groups of users will be discussed and analyzed in detail in the following section.

The CQI feedback is sent from each user to the gNB in each time slot. According to the implemented numerology (i.e., $\mu = 0$, SCS $= 15$ kHz), this TTI is equal to 1 ms. The rest of the network simulation parameters are summarized in Table 1.

VI. RESULTS

In this section, the most relevant results are presented and analyzed considering two approaches. In the first case, neither capacity constraint, nor user constraints have been applied. In the other cases, minimum capacity constraints and minimum percentages of served users have been defined and applied. Then, the obtained results are compared with other existing subgrouping methods.

A. CASE 0: WITHOUT CONSTRAINTS

Firstly, in Fig. 3, ADR results are presented for car-mounted users moving with a speed between 30 km/h and 50 km/h. The vertical and the x-axis represent the ADR in Mbps and the multiplexing configurations, respectively. In the

TABLE 1. Simulation parameters.

Parameter	Value
Center Frequency	2 GHz
Tx Power	44 dBm
Distance attenuation	128.1 + 37.6*log(d)
Type of nodes	Pedestrians, car-mounted
Speed	Pedestrians: 3 km/h Car-mounted: 30-50 km/h
Number of Receivers	100
Mobility type	RWP
ISD	500 m
Delay Spread	90, 360, 1100 ns
Noise Power	-90 dBm
Noise Figure	9 dB
Time slot size	1 ms
Evaluated time slots	10000
SCS	15 kHz
Bandwidth	10, 20 MHz
Injection levels	As defined for ATSC 3.0 [30]
Channel model	TDL-D, TDL-E

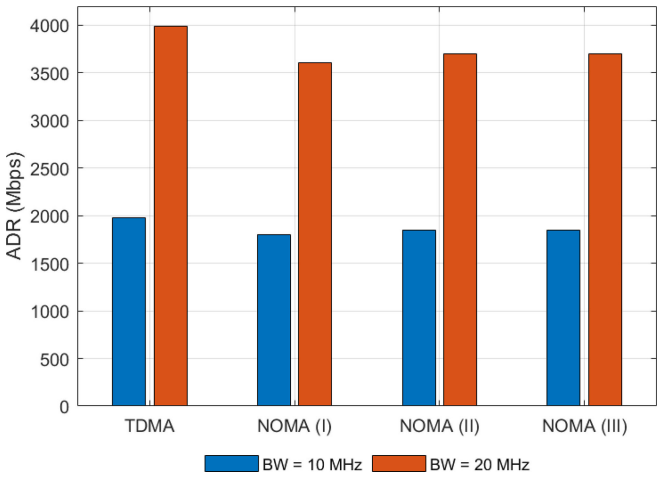

FIGURE 3. Case 0 - ADR results.

NOMA case, the three reorganization techniques presented in Section IV-B are depicted. The main conclusion is that similar results are obtained for both multiplexing technologies (i.e., TDMA and NOMA). Considering the 10 MHz channel bandwidth, results are almost equal, while in the case of 20 MHz, TDMA offers slightly better results. This is due to the limitation of the IL values, which maximum is set to -25 dB. If more separation between layers is applied, the difference between TDMA and NOMA is reduced.

However, when the mean data rate delivered by each group and the percentage of users served by each group are analyzed, as shown in Table 2, both TDMA and NOMA present unfair configurations. On the one hand, in TDMA, the algorithm assigns almost all the resources to one group and the other group receives just one RB. With this configuration the group with the lowest datarate offers 80 kbps, which is not enough to guarantee mobile video reception. On the other hand, in the NOMA-based solutions, the drawback is that due to the SNR penalization suffered in both layers, the algorithms present groups where users are not served (i.e., 38.03-38.56%) because of the high SNR requirements.

Simulations have also been carried out for pedestrians moving with a mean speed of 3 km/h. The results have not been included in this work because, without any constraint, they show very similar behavior to the one shown in Fig. 3 and Table 2.

B. CASE 1: CAPACITY AND USERS CONSTRAINTS

In order to guarantee that each group is fairly configured, a minimum capacity and percentage of served users constraints have been defined. For the capacity constraint, four different values have been defined: 1, 2, 3, and 4 Mbps. This constraint has to be guaranteed simultaneously in each group and it has been selected according to the potential users requirements described in Section V-B. If it is not possible to guarantee the constraint, a single-group configuration is applied. For the users percentage constraint, two different restrictions have been applied: 90% and 95%. This constraint represents the total number of users of the configuration, which is calculated by adding the number of users served in each group.

Fig. 4 depicts the results obtained for pedestrian and car-mounted users using 10 MHz (light blue box) and 20 MHz (light orange box) channel bandwidths. Overall, NOMA-based configurations outperform the results obtained with TDMA. Nonetheless, the difference between the technologies varies considerably depending on the constraints. If the minimum capacity constraint is used (i.e., 1 Mbps), similar results are obtained for both technologies with the 10 MHz bandwidth (i.e., results inside light blue boxes), while TDMA outperforms NOMA in the 20 MHz case (i.e., results inside light orange boxes). However, when the highest constraint is used (i.e., 4 Mbps), the gain of NOMA is maximized for all the cases. The maximum gain appears for the 10 MHz bandwidth, where the ADR that NOMA offers is more than two times the ADR of TDMA configuration: 120.7% for pedestrian users and 155.9% for car-mounted users. The main reason for these results is that, as stated in literature [6], [9], NOMA performs better than TDMA when the received CQI feedbacks generate a high unbalance between both subgroups.

Also, the influence of the channel bandwidth can be analyzed thanks to graphical results summarized in Fig. 4. On the one hand, more extreme values are obtained using a 10 MHz bandwidth and the minimum and the maximum gains of NOMA are maximized in this case. Consequently, the gain that NOMA offers in relative terms is maximized. On the other hand, in the 20 MHz bandwidth case, results are more linear compared with the ones obtained for the 10 MHz case. In fact, results are constant for all the capacity constraints in the NOMA cases. This effect is because while each layer (subgroup) can exploit 100% of the bandwidth resources (i.e., RB) in NOMA, in TDMA, the available RBs have to be shared between both subgroups. Therefore, if the transmission is delivered using the lowest MCS (i.e.,

TABLE 2. Served users and offered capacity analysis in Case 0 results.

BW	Metric	TDMA		NOMA (I)		NOMA (II)		NOMA (III)	
		G1	G2	G1	G2	G1	G2	G1	G2
10 MHz	Served users (%)	42.50	57.50	0.4	61.51	0.97	61.00	0.2	61.24
	Non-served users (%)	0		38.09		38.03		38.56	
	Data rate (Mbps)	0.08	34.88	7.83	29.59	7.75	30.66	7.86	30.42
20 MHz	Served users (%)	42.48	57.52	0.4	61.51	0.97	61.00	0.2	61.24
	Non-served users (%)	0		38.09		38.03		38.56	
	Data rate (Mbps)	0.08	70.45	15.66	59.17	15.50	61.32	15.72	60.84

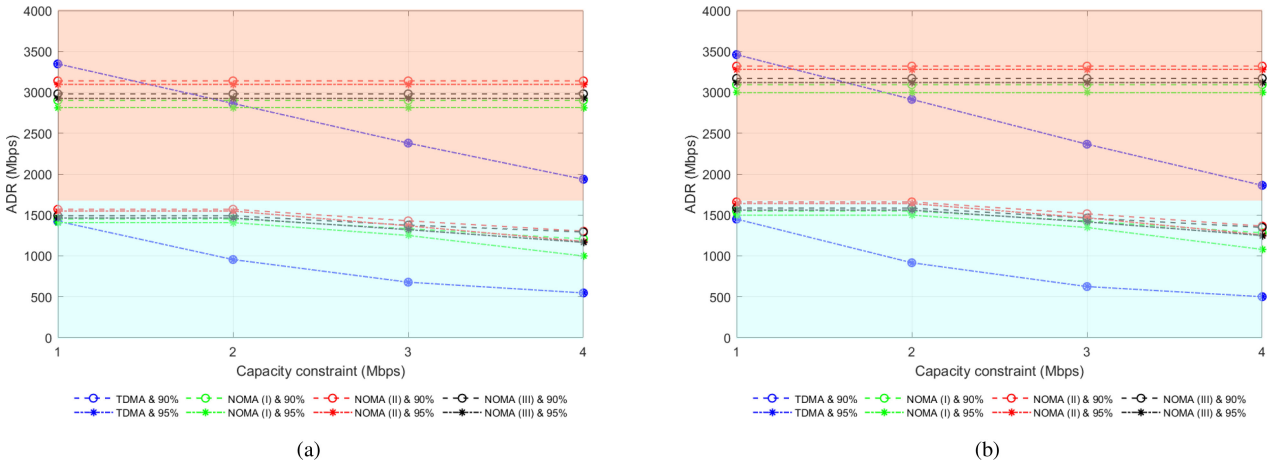

FIGURE 4. ADR results with minimum capacity and served users constraints for different configurations and mobility types: (a) Pedestrian users, (b) Car-mounted users.

QPSK 120/1024 and 0.2344 bps/Hz), the obtained capacity is 4.22 Mbps, which is always above the capacity constraints.

Table 3 shows the data rate that is offered by each group for some cases shown in Fig. 4. As in the case of the ADR, the gain obtained by using NOMA is maximized when the most challenging conditions are imposed. In fact, for the 10 MHz channel, the high capacity group (i.e., G2) offers up to three times the capacity offered in TDMA when 4 Mbps are required, while in the case of 1 Mbps, the gain is considerably reduced.

On the other hand, the user percentage constraint does not affect TDMA cases since as neither SNR adaptation, nor user reorganization are required, all the user in the network are always served. However, in NOMA cases when the percentage of users is increased, the obtained ADR is reduced (see Fig. 4). Although the reduction varies depending on the case, it is always below 18%. The worst situation is obtained in the case of 10 MHz bandwidth with the 4 Mbps capacity constraint, where a deterioration of 17.3% is obtained for pedestrians and 15.8% for car-mounted users when increasing from 90% to 95% served users.

Moreover, the difference between the proposed NOMA-based algorithms in terms of ADR should be remarked. The algorithm that assumes uniform SNR distribution (i.e., NOMA (II)) is the one that performs better. However, the difference between the NOMA-based solutions is around 10% with respect to the worst case election algorithm (i.e., NOMA (I)) and around 5% with respect to the uniform distance distribution algorithm (i.e., NOMA (III)). These results indicate that a trade-off has to be assumed between the maximum capacity that is going to be offered and the potential mismatch between the requested and the offered MCS.

C. COMPARISON WITH EXISTING SOLUTIONS

To test the relevance of the proposed methods, a comparison with the previous works is required. First, the work in [9] has been used as the baseline to calibrate the impact of the results since very similar use cases, boundary conditions, and mobility models are implemented. Nevertheless, in [9], a theoretical analysis based on Shannon's capacity is presented and the results are obtained for a 10 MHz channel bandwidth. In our paper, a simulation set-up closer to the real implemention is proposed (see Table 1). The comparison aims at testing if the outcomes of the theoretical approach are still valid for a more realistic use case. Table 4 shows a comparison between the previous and current work in terms of mean data rate per user.

The first relevant difference with the previous work is the lack of flexibility of NOMA systems. There is no numerology in [9] when the multicast group MCS is selected, whereas, in this work, the possible MCS values are the ones in [28, Table 5.1.3.1-2], limiting the possible offered capacity. This effect can be seen in Table 4, where NOMA cannot adjust the mean data rate of G1 to the minimum capacity constraint. Then, it should be highlighted that the mean data rate values per user are generally higher in this work than in the previous due to two reasons. The first reason is that although [9] was oriented to cover 5G use cases, the baseline technology was LTE, and so, the target data rate values

TABLE 3. Example of mean data rate offered to each group for car-mounted users and 95% of served broadcast users.

	TDMA		NOMA (I)		NOMA (II)		NOMA (III)	
	G1	G2	G1	G2	G1	G2	G1	G2
10 MHz & 1 Mbps	1.0	25.6	2.1	32.8	2.1	32.4	2.1	32.9
10 MHz & 4 Mbps	4.0	5.6	4.2	9.9	4.2	18.9	4.2	20.0
20 MHz & 1 Mbps	1.0	61.0	4.2	65.5	4.2	64.8	4.2	65.8
20 MHz & 4 Mbps	4.1	32.4	4.2	65.5	4.2	64.8	4.2	65.8

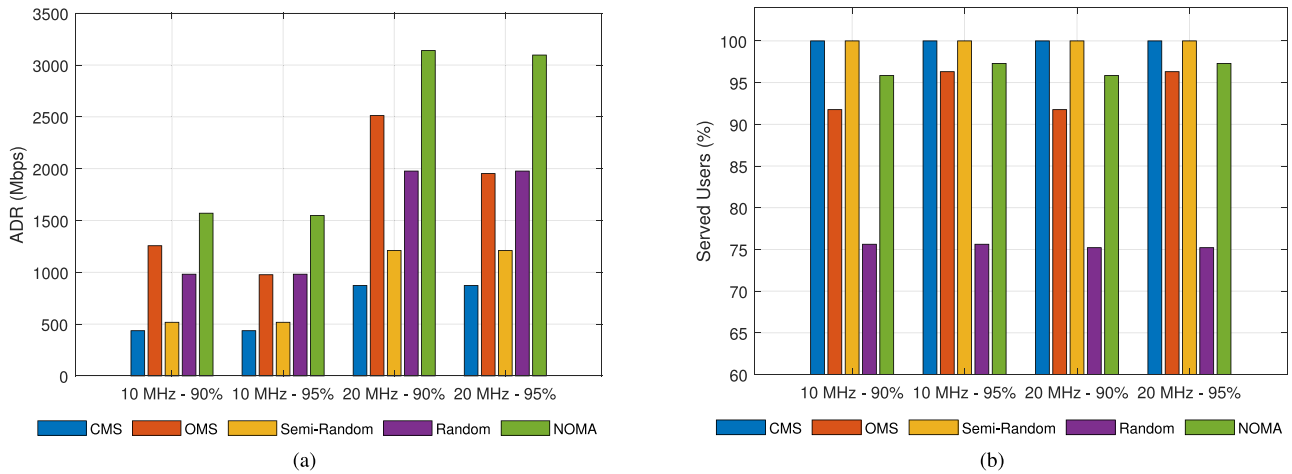

FIGURE 5. NOMA-based subgrouping algorithm comparison with other techniques in terms of: (a) ADR, (b) Served users.

TABLE 4. Mean data rate per user comparison with previous work.

Reference	Scheme	G1 (Mbps)	G2 (Mbps)	G1 Gain	G2 Gain
[9]	TDMA	0.1	3.0	0%	83%
	NOMA	0.1	5.5		
This work (1 Mbps)	TDMA	1.0	25.6	100%	29%
	NOMA	2.0	32.0		
This work (4 Mbps)	TDMA	4.0	5.6	5%	257%
	NOMA	4.2	20.0		

were lower than in 5G numerology. On the other hand, the second reason is the mobility model. In this case, the simulation parameters in Table 1 are not the same as in the previous work and, therefore, the mobility model generated is not the same, which makes the results not directly comparable. Consequently, the gain of both groups (G1 and G2) is presented in percentages in Table 4. Those values show the gain of NOMA versus TDMA in terms of mean data rate per user. As it might be seen, this work provides in both cases (1 and 4 Mbps of capacity constraint) higher gain than in [9]. In the first case, most of the gain is concentrated in G1, while in the second case, the gain focuses on G2.

Then, the performance of the NOMA-based subgrouping algorithm has been compared with other more realistic solutions: Conventional Multicast Scheme (CMS), Opportunistic Multicast Scheduling (OMS), semi-random and random [35], [36]. First, CMS is a conservative approach, which decides the system data rate according to the user with the worst channel conditions. OMS is a variation of CMS that allows not offering service to the worst channel condition users to maximize the offered capacity. Random algorithms do not analyze the received feedbacks and configure the MCS values of each group randomly. Finally, in the semi-random case, only the high capacity group is randomly organized and the low capacity group is configured based on the worst channel condition users. Fig. 5 presents the comparative study in terms of ADR (Fig. 5(a)) and percentage of served users (Fig. 5(b)). Two bandwidth sizes (10 MHz and 20 MHz) and two minimum percentages of served users (90% and 95%) have been assumed. In addition, in the case of NOMA, a minimum data rate per service of 1 Mbps has been fixed. On the one hand, concerning ADR results, NOMA clearly outperforms the rest of the techniques for each evaluated bandwidth and served user percentage. On the other hand, the final served user values show that NOMA is only outperformed by CMS and semi-random methods. However, NOMA is always guaranteeing the minimum required percentage. Therefore, considering both graphics (i.e., Fig. 5(a) and (b)), it could be concluded that NOMA provides a better performance in comparison with other existing methods.

VII. FUTURE WORK AND SYSTEM EVOLUTION

Although generally, in downlink communications, NOMA-based subgrouping techniques have better performance than TDMA-based techniques, they cannot be considered the optimum solution. In Fig. 4(a) and Fig. 4(b), the ADR results are almost constant for different capacity constraints. Consequently, a solution based on the combination of NOMA and T/FDMA techniques could achieve an optimum exploitation of the resource management. In this section, a theoretical approach is provided in order to prove that future T/FDMA+NOMA solutions are viable.

TABLE 5. Analysis of the theoretical offered capacity using a combined T/FDMA + NOMA approach.

	NOMA-only		T/FDMA+NOMA (1 Mbps)		T/FDMA+NOMA (2 Mbps)		T/FDMA+NOMA (3 Mbps)		T/FDMA+NOMA (4 Mbps)	
	# RB	Mbps	# RB	Mbps	# RB	Mbps	# RB	Mbps	# RB	Mbps
G1 - UL MCS 0	100	4.22	24	1.01	48	2.03	72	3.04	95	4.01
G2 - LL MCS 4	100	21.16	24	5.08	48	10.16	72	15.24	95	20.11
G2 - SL MCS 10	0	0	76	35.16	52	24.06	28	12.95	5	2.31
Total G2	100	21.16	100	40.24	100	34.22	100	28.19	100	22.42

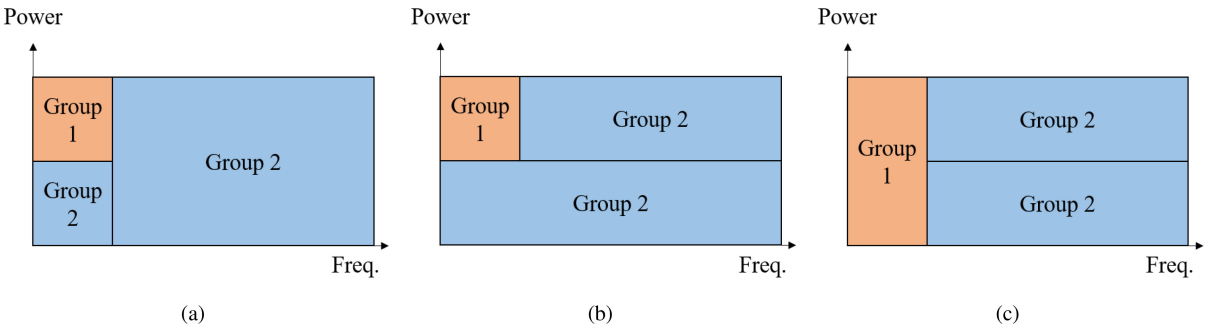

FIGURE 6. Different resource allocation representations of the TDMA+NOMA solution: (a) Group 1 in UL and Group 2 in LL and single-layer, (b) Group 1 in UL and Group 2 in UL and LL, (c) Group 1 in single-layer and Group 2 in UL and LL.

Taking into account the algorithm developed for NOMA (see Algorithm 2), both subgroups can manage all the RBs of the channel bandwidth independently, which in comparison with TDMA, allows doubling the number of available RBs. Therefore, as the low capacity group of NOMA (i.e., UL - Group 1) uses all the available RBs, the minimum capacity offered can be calculated as follows:

$$C_{G1} = N_{RB} \cdot BW_{RB} \cdot eff, \tag{13}$$

where, BW_{RB} is the bandwidth size of a unique RB, in this case 180 kHz (i.e., 12 subcarriers with 15 kHz of subcarrier spacing). Assuming that the lowest MCS value is used for the transmission, the channel bandwidth defines the offered capacity. In the case of 10 MHz bandwidth (i.e., 50 RB), the minimum capacity that is going to be offered using NOMA is 2.11 Mbps, while if a 20 MHz bandwidth (i.e., 100 RB) is used, 4.22 Mbps can be offered. Therefore, in the case of NOMA the capacity constraint is not as relevant as in the case of T/FDMA, the capacity is always above the constraints established.

The solution is to offer a combination of both techniques. In Fig. 6, different alternatives for the combined NOMA + T/FDMA resource allocation are presented. Fig. 6(c) provides the most robust Group 1 services among the three options since single layer mode is used. On the other hand, Fig. 6(b) presents the highest spectral efficiency improvement since all the channel resources are shared in both layers. However, this alternative is the less robust solution since all the services are affected by the injection level. Then, Fig. 6(a) presents an intermediate solution, where services of the Group 2 are provided in single-layer mode and in the LL.

Doubtlessly, the RRM algorithm for each solution is different. For example, concerning Fig. 6(a), first, some channel resources are assigned to the NOMA based part to configure the low capacity group (i.e., Group 1) and will vary depending on the capacity constraint that is applied. Then, the high capacity group (i.e., Group 2) is configured in the NOMA part with the same number of RBs as in the case of Group 1. Finally, the remaining RBs are used to maximize the capacity offered to Group 2 and the transmission is carried out in single-layer mode. It is important to highlight that the portion of data delivered in the single-layer mode has to be configured to be decoded with the same SNR value (or lower) than the LL. It is quite evident that these approaches provide a more accurate RRM, in exchange for an increase in the complexity of the algorithm since more SNR and capacity values will have to be handled and, therefore, the possible solutions will considerably increase. Consequently, this paper includes a theoretical analysis of the potential capacity gain that combined NOMA+T/FDMA solutions will have in future implementations.

The preliminary evaluation results are summarized in Table 5. For the calculations, a 20 MHz (i.e., 100 RB) channel bandwidth is assumed following the approach presented in Fig. 6(a). The reference NOMA configuration is a configuration obtained from the simulations carried out in Section VI-B, where Group 1 uses the lowest MCS, Group 2 is delivered with MCS 4 and the IL is -5 dB. Then, four different approaches of the T/FDMA+NOMA solutions are presented for four different capacity constraints: 1, 2, 3, and 4 Mbps. The gain obtained in the capacity of Group 2 should

be highlighted. In fact, the lower the capacity constraint, the higher the gain obtained with the combined solution. For a minimum capacity equal to 1 Mbps, the overall capacity offered for Group 2 is around the double of the capacity offered in the NOMA-only mode.

VIII. CONCLUSION

NOMA-based 5G NR subgrouping techniques have been designed and tested. Besides the design and implementation of the required algorithms, the impact on the SNR deterioration due to the use of NOMA has been detected and opportune solutions based on user reorganization techniques have been proposed.

The first relevant conclusion of this study is that depending on the configurations both proposals, NOMA, and TDMA, can present unfair configurations from the served users or minimum capacity point of view, respectively. Then, different decisions have been taken to avoid unfairness, such as minimum capacity and number of user constraints. Consequently, the second main conclusion is that, under the evaluated use case conditions, NOMA appears as a better candidate since NOMA outperforms TDMA in terms of ADR in almost all cases. In this line, we have concluded that some configuration parameters (capacity and users constraints and channel bandwidth) are critical to determining the gain of NOMA. In particular, the gain of NOMA is maximized in the most challenging situations (i.e., less number of RB, higher required throughput). Finally, the third major conclusion is that both multiplexing technologies, NOMA, and TDMA, should be considered suboptimal solutions since higher spectral efficiency (and so, throughput) can be obtained by combining TDMA and NOMA solutions in the same subgrouping algorithm. Moreover, this conclusion has been confirmed following the theoretical analysis shown in Section VII, where the novel T/FDMA+NOMA solution outperforms each of the proposed single multiplexing technology solutions.

There are still some future works on the horizon. On the one hand, the performance of these subgrouping algorithms should be analyzed from the network layer perspective by using a network simulator. This approach could lead to a novel PHY/MAC analysis of the subgrouping techniques. In addition to the KPIs used in this work (ADR or data rate per group), KPIs related with the PHY and the MAC layers could be used, such as latency, effective throughput or Packet Error Rate (PER). On the other hand, the combination of TDMA and NOMA solutions should be analyzed more in detail and the influence of the rest of the affecting parameters should be studied. Then a simulation platform that integrates a new cost function for TDMA+NOMA case should be developed.

REFERENCES

[1] "Digital cellular telecommunications system (phase 2+) (GSM); universal mobile telecommunications system (UMTS); LTE; 5G; release description; release 15, V15.0.0," 3GPP, Sophia Antipolis technology park in France, Rep. ETSI TR 121 915, Oct. 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/121900_121999/121915/15.00.00_60/tr_121915v150000p.pdf

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

[3] 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.

[4] L. Militano, D. Niyato, M. Condoluci, G. Araniti, A. Iera, and G. M. Bisci, "Radio resource management for group-oriented services in LTE-A," *IEEE Trans. Veh. Technol.*, vol. 64, no. 8, pp. 3725–3739, Aug. 2015.

[5] M. Condoluci, G. Araniti, A. Molinaro, and A. Iera, "Multicast resource allocation enhanced by channel state feedbacks for multiple scalable video coding streams in lte networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 2907–2921, May 2016.

[6] E. Iradier *et al.*, "Using NOMA for enabling broadcast/unicast convergence in 5G networks," *IEEE Trans. Broadcast.*, vol. 66, no. 2, pp. 503–514, Jun. 2020.

[7] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G Systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[8] Q. Wu, W. Chen, D. W. K. Ng, and R. Schober, "Spectral and energy-efficient wireless powered IoT networks: NOMA or TDMA?" *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6663–6667, Jul. 2018.

[9] J. Montalban *et al.*, "Multimedia multicast services in 5G networks: Subgrouping and non-orthogonal multiple access techniques," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 91–95, Mar. 2018.

[10] 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.

[11] C. Suh and J. Mo, "Resource allocation for multicast services in multicarrier wireless communications," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 27–31, Jan. 2008.

[12] G. Araniti, V. Scordamaglia, M. Condoluci, A. Molinaro, and A. Iera, "Efficient frequency domain packet scheduler for point-to-multipoint transmissions in LTE networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 4405–4409.

[13] C. Casetti, C.-F. Chiasserini, F. Malandrino, and C. Borgiattino, "Area formation and content assignment for LTE broadcasting," *Comput. Netw.*, vol. 126, pp. 174–186, Oct. 2017.

[14] A. Labella, R. M. Rodriguez, G. De Tre, and L. Martinez, "A cohesion measure for improving the weighting of experts' subgroups in large-scale group decision making clustering methods," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jun. 2019, pp. 1–6.

[15] G. Araniti, F. Rinaldi, P. Scopelliti, A. Molinaro, and A. Iera, "A dynamic MBSFN area formation algorithm for multicast service delivery in 5G NR networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 808–821, Feb. 2020.

[16] G. Araniti, I. Bisio, M. De Sanctis, F. Rinaldi, and A. Sciarrone, "Joint coding and multicast subgrouping over satellite-eMBMS networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 1004–1016, May 2018.

[17] A. Z. Yalcin, M. Yuksel, and B. Clerckx, "Rate splitting for multi-group multicasting with a common message," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12281–12285, Oct. 2020.

[18] B. Zhang, Y. Chen, J. Yu, and Z. Han, "An indirect-reciprocity-based incentive framework for multimedia service through device-to-device multicast," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 10969–10980, Nov. 2019.

[19] E. Iradier, J. Montalban, G. Araniti, M. Fadda, and M. Murroni, "Adaptive resource allocation in LTE vehicular services using LDM," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2016, pp. 1–6.

[20] Y. Xue, A. Alsohaily, E. Sousa, W. Li, L. Zhang, and Y. Wu, "Using layered division multiplexing for mixed unicast-broadcast service delivery in 5G," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2019, pp. 1–6.

[21] Y. Yin, Y. Peng, M. Liu, J. Yang, and G. Gui, "Dynamic user grouping-based NOMA over Rayleigh fading channels," *IEEE Access*, vol. 7, pp. 110964–110971, 2019.

[22] H. Sari, A. Maatouk, E. Caliskan, M. Assaad, M. Koca, and G. Gui, "On the foundation of NOMA and its application to 5G cellular networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Barcelona, Spain, Apr. 2018, pp. 1–6.

[23] L. Lv, J. Chen, and Q. Ni, "Cooperative non-orthogonal multiple access in cognitive radio," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2059–2062, Oct. 2016.
[24] M. Moltafet, N. Mokari, M. R. Javan, H. Saeedi, and H. Pishro-Nik, "A new multiple access technique for 5G: Power domain sparse code multiple access (PSMA)," *IEEE Access*, vol. 6, pp. 747–759, 2017.
[25] E. M. Almohimmah, M. T. Alresheedi, A. F. Abas, and J. Elmirghani, "A simple user grouping and pairing scheme for non-orthogonal multiple access in VLC system," in *Proc. 20th Int. Conf. Transp. Opt. Netw. (ICTON)*, Bucharest, Romania, Jul. 2018, pp. 1–4.
[26] L. Zhang *et al.*, "Layered-division-multiplexing: Theory and practice," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 216–232, Mar. 2016.
[27] J. Montalbán *et al.*, "Cloud transmission: System performance and application scenarios," *IEEE Trans. Broadcast.*, vol. 60, no. 2, pp. 170–184, Jun. 2014.
[28] "Technical specification group services and system aspects; NR; physical layer procedures for data (release 15), v15.3.0," 3GPP, Sophia Antipolis, France, Rep. TS 38.214, Sep. 2018.
[29] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.
[30] "ATSC standard: Physical layer protocol," ATSC, Washington, DC, USA, Rep. A/322, 2016.
[31] A. Varga and R. Hornig, "An overview of the OMNeT++ simulation environment," in *Proc. 1st Int. Conf. Simulat. Tools Techn. Commun. Netw. Syst. Workshops*, 2008, p. 60.
[32] "Radio access network working group; study on channel model for frequencies from 0.5 to 100 GHz (release 15)," 3GPP, Sophia Antipolis, France, Rep. TR 38.901, 2018.
[33] E. Hyytiä, H. Koskinen, P. E. Lassila, A. Penttinen, J. T. Virtamo, and J. Roszik, *Random Waypoint Model in Wireless Networks*, Helsinki Univ. Technol., Espoo, Finland, 2005.
[34] X. Xu, J. Liu, and X. Tao, "Mobile edge computing enhanced adaptive bitrate video delivery with joint cache and radio resource allocation," *IEEE Access*, vol. 5, pp. 16406–16415, 2017.
[35] C. K. Tan, T. C. Chuah, and S. Tan, "Adaptive multicast scheme for OFDMA-based multicast wireless systems," *Electron. Lett.*, vol. 47, no. 9, pp. 570–572, 2011.
[36] G. Araniti, M. Condoluci, L. Militano, and A. Iera, "Adaptive resource allocation to multicast services in LTE systems," *IEEE Trans. Broadcast.*, vol. 59, no. 4, pp. 658–664, Dec. 2013.

ENEKO IRADIER (Member, IEEE) received the M.S. and Ph.D. degrees in telecommunications engineering from the University of the Basque Country in 2018 and 2021, respectively. Starting in 2017 and for a year and a half, he worked as a student researcher with IK4-Ikerlan Technology Center, where he developed URLLC communications and ultra-low consumption systems. Since 2015, he has been a part of the TSR Research Group, UPV/EHU, where he is currently a Postdoctoral Researcher. During his doctoral studies, he did an internship with the Communications Research Centre Canada, Ottawa. His current research interests include designing and developing new technologies for the future physical layer of communication systems and broadcasting in 5G environments.

MAURO FADDA (Member, IEEE) received the M.Sc. degree in telecommunication engineering from the University of Bologna, Italy, in 2006, and the Ph.D. degree from the University of Cagliari, Italy, in 2013, where he is an Assistant Professor with the Department of Electronic and Information Engineering. In 2007, he spent one year as a Researcher with the National Research Center, Bologna, developing UMTS mobile-network simulation software. From 2008 to 2009, he was a Researcher with the Research Center of Sardinia, Pula, implementing different Web communication applications. From 2014 to 2015, he was a Researcher for the research unit of the Italian University Consortium for Telecommunications (CNIT), University of Cagliari. His main research interests are telecommunications, cognitive radio systems, mobile technologies, and broadcasting systems.

MAURIZIO MURRONI (Senior Member, IEEE) received the M.Sc. degree in electronic engineering and the Ph.D. degree in electronic engineering and computers from the University of Cagliari in 1998 and 2001, respectively. He is an Associate Professor with the Department of Electrical and Electronic Engineering (DIEE), University of Cagliari and a member of the CNIT-National Inter-University Consortium for Telecommunications. He has coauthored the 1900.6-2011-IEEE Standard for Spectrum Sensing Interfaces and Data Structures for Dynamic Spectrum Access and other Advanced Radio Communication Systems. He has coauthored an extensive list of journal articles and peer-reviewed conference papers and received several best paper awards. His research focuses on quality of experience, multimedia data transmission and processing, broadcasting, cognitive radio system, and signal processing for radio communications. He served as a chair for various international conferences and workshops. He was a guest editor for several journals. He is a Distinguished Lecturer for the IEEE Broadcast Technology Society and an Associate Editor for the IEEE TRANSACTIONS ON BROADCASTING. He is a Senior Member of the IEEE Communications Society, IEEE Broadcast Technology Society, IEEE Vehicular Technology Society, and IEEE Signal Processing Society.

PASQUALE SCOPELLITI (Member, IEEE) received the M.Sc. degree in telecommunications engineering and the Ph.D. degree in information engineering from the University Mediterranea of Reggio Calabria, Italy, in 2013 and 2018, respectively. He is currently a Postdoctoral Researcher with the University Mediterranea of Reggio Calabria, Italy. His current research interests include cellular systems, radio resource management, multicast and broadcast services over 5G cellular networks, and heterogeneous networks. He is an IEEE BTS Member.

GIUSEPPE ARANITI (Senior Member, IEEE) received the Laurea degree and the Ph.D. degree in electronic engineering from the University Mediterranea of Reggio Calabria, Italy, in 2000 and 2004, respectively, where he is currently an Assistant Professor of Telecommunications. He is also with CNIT, Italy. His major area of research is on 5G/6G networks and it includes personal communications, enhanced wireless and satellite systems, traffic and radio resource management, multicast and broadcast services, device-to-device, and machine-type communications.

JON MONTALBAN (Senior Member, IEEE) received the M.S. and Ph.D. degrees in telecommunications engineering from the University of the Basque Country, Spain, in 2009 and 2014, respectively. He is part of the TSR (Radiocommunications and Signal Processing) Research Group, University of the Basque Country, where he is an Assistant Professor with the Department of Electronic Technology. He has held visiting research appointments with Communication Research Centre, Canada, and Dublin City University, Ireland. He is currently involved in research activities related to broadcasting in 5G environments and wireless systems for reliable industrial communications. His main research interest is in the area of network architectures for wireless communications. He is a co-recipient of the Scott Helt Memorial Award to recognize the Best Paper published in the IEEE TRANSACTIONS ON BROADCASTING in 2019. He has served as a reviewer for several renowned international journals and conferences in the area of wireless communications and currently serves as an Associate Editor for the IEEE ACCESS and IEEE TRANSACTIONS ON BROADCASTING.