This is the post-print of the following article: "Multicast Resource Allocation Enhanced by Channel State Feedbacks for Multiple Scalable Video Coding Streams in LTE Networks".
Article has been published in final form at: https://ieeexplore.ieee.org/abstract/document/7131571 DOI: 10.1109/TVT.2015.2449080

Multicast Resource Allocation Enhanced by Channel State Feedbacks for Multiple Scalable Video Coding Streams in LTE Networks

Massimo Condoluci, Giuseppe Araniti, Antonella Molinaro, Antonio Iera

***Abstract*—The growing demand of mobile multicast services such as IPTV and video streaming requires effective radio resource management (RRM) to handle traffic with strict Quality of Service constraints over Long Term Evolution (LTE) and beyond systems. Special care is needed to limit the system performance degradation when multiple multicast streams are simultaneously transmitted. To this aim, this paper proposes a RRM policy based on the subgrouping technique for the delivery of scalable multicast video flows in a cell. Our proposal enhances the legacy multicast transmission over LTE systems by exploiting the multi-user diversity and the users' channel quality feedbacks. Moreover, it is designed to take advantage from the frequency selectivity in the subgroup formation. Simulation results demonstrate the effectiveness of the proposed scheme, which outperforms existing approaches from the literature. It succeeds to achieve higher spectral efficiency and to guarantee adequate video quality to all the multicast receivers and improved quality to the ones with good channel conditions.**

***Index Terms*—LTE, LTE-A, RRM, Multicast.**

I. Introduction

LONG Term Evolution (LTE) [1] and beyond cellular systems represent the wireless technologies that will lead the growth of mobile broadband services in the years to come. LTE offers several benefits in terms of high data rates in both downlink and uplink directions, low latency, low cost per bit, high spectrum efficiency even for cell-edge users, and high system capacity. Such features are achieved through a *flat all-IP* network infrastructure and through transmissions that exploit Orthogonal Frequency Division Multiple Access (OFDMA) on the radio interface.

In a telecommunication scenario characterized by a fast growth of the mobile market, LTE is very appealing to network providers as a means to deliver high quality services. Especially *group-oriented* services, such as TV, be it managed IPTV or Over-The-Top (OTT), news feeds, weather forecast, video conferencing, Internet video streaming, are expected to be massively exchanged over LTE (4G) and future systems [2]. In this scenario, *multicast* transmissions are gaining in importance, in the view of simultaneously delivering data towards multiple destinations [3], and the Multimedia Broadcast Multicast Service (MBMS) as part of the 3GPP LTE standard [4] represents an attractive solution for their deployment in LTE systems [5].

Nevertheless, it is well known from the literature [6] that the resource allocation of multicast services raises several issues, which could affect the performance of wireless systems. Among them, a very critical issue turns to be the design of Radio Resource Management (RRM) policies that operate on a *per-group basis*, due to the interest of multiple destinations in receiving the same data traffic, which is conveyed through Point-to-Multipoint (PtM) transmissions. The group-based management limits the system spectral efficiency, mainly caused by cell-edge users, which force the group to be served with low data rate (robust) modulation and coding schemes (MCSs) due to their poor channel quality conditions. As a result, the high potential of OFDMA resource allocation is only partially exploited.

Moreover, multicast applications such as mobile TV are typically resource-hungry. This poses additional challenges to the effective utilization of the scarce available spectrum and may severely limit the overall capacity of the LTE system, especially when *multiple multicast services* are delivered in a single cell (as shown in Fig. 1). The presence of several multicast groups increases the system design criticalities, due to the high heterogeneity of channel conditions experienced by users belonging to different groups and to the dissimilar Quality of Service (QoS) requirements of the different video streams. Accordingly, serving preferably multicast groups whose members experience a high channel quality improves the system spectral efficiency at the expense of groups whose members are in bad channel conditions. In addition, starvation may occur for groups requiring videos with lower throughput constraints if, to increase the system capacity, preference is given to those asking for a higher throughput. Such issues are exacerbated by the potential differences in the size of the multicast groups; for example, giving preference to large groups may improve the system throughput at the expense of groups with a lower number of members.

This paper contributes to give an answer to the highlighted issues by proposing a novel RRM algorithm for the efficient resource allocation of multiple multicast video streams in LTE and beyond systems. The basic idea is to extend the LTE/MBMS capabilities by introducing a link-adaptation procedure, based on the channel quality feedback transmitted by multicast users. According to such a feedback, the proposed RRM exploits a *subgrouping* technique and splits each multicast group into different subgroups, with beneficial effects

Copyright ©2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

The authors are with the DIIES Department, University Mediterranea of Reggio Calabria, 89100, Italy (e-mail: massimo.condoluci@unirc.it; araniti@unirc.it; antonella.molinaro@unirc.it; antonio.iera@unirc.it).

DOI: 10.1109/TVT.2015.2449080

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

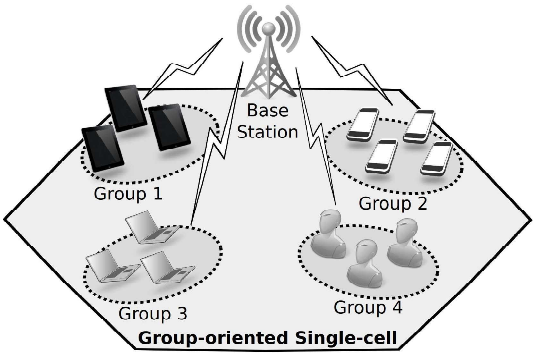


Fig. 1. Single-cell multi-group scenario.

on both the user and the network sides. Through the joint use of a *scalable video coding* (SVC) [7] [8] technique, the proposed subgroup formation leverages multi-user diversity and guarantees a "basic" quality to all the multicast receivers and an "improved" quality only to the ones with better channel conditions. *Frequency selectivity* is exploited by scheduling the assignment of each frequency resource to the subgroup that guarantees the highest spectral efficiency over such a resource.

The result of the presented research is the definition of the Multicast Subgrouping scheme for Multi-Layer video applications, which proves to be suitable for practical implementations thanks to its low computational cost. We present and analyze two different cost functions exploited by our schemes, both achieving high spectral efficiency and utilization, since they require a lower amount of radio resources for high quality video delivery compared to other policies in the literature.

The remainder of this paper is organized as follows. Section II provides an overview of multicast service provisioning techniques in current LTE systems. In Section III we briefly discuss the main literature related to our research work. The reference system model is described in Section IV, and our proposed policy with related cost functions in Section V. Simulation settings and results are illustrated in Sections VI and VII, whereas conclusive remarks are summarized in Section VIII.

II. The LTE System

Motivated by the increasing demand for high-quality mobile broadband services, the Third Generation Partnership Project (3GPP) carried out several activities, under the LTE and System Architecture Evolution (SAE) projects, finalized to define the radio access and the core network for the next generation of cellular systems [1]. Furthermore, being designed to natively support MBMS [9], the LTE system is one of the most promising wireless technologies to support the demand of high-quality group-oriented services.

The LTE/MBMS architecture [9] is shown in Fig. 2. The access network [1] is composed of the LTE base station (i.e., the eNodeB) and the MultiCell/Multicast Coordination Entity (MCE), which are responsible for transmission parameters configuration in single- and multi-cell mode, respectively. The core network [4] includes: Mobility Management Entity (MME) that is responsible for authentication, security, and mobility management procedures; MBMS Gateway (MBMS-GW), a logical entity whose principal function is data packets forwarding to eNodeBs; Broadcast Multicast-Service Center (BM-SC) that is the MBMS traffic source which also accomplishes service announcement and group membership functions. The MBMS traffic is delivered to interested users through two PtM downlink channels: the Multicast Traffic Channel (MTCH), designed for data delivery and the Multicast Control Channel (MCCH) that carries signalling information regarding one or several MTCHs (including the subframe allocation and MTCH transmission parameters).

LTE/MBMS is typically used in multicast-broadcast single-frequency network (MBSFN) mode; with the aim to enlarge the coverage and to improve the performance for users located at cell-edge, all cells in the MBSFN area use the same physical resources (where the cyclic prefix duration of OFDM symbols is properly set to reduce the interference between adjacent cells) at the same time with the same MCS. In our scenario, we consider that each multicast stream is transmitted separately within each cell (i.e., each base station performs the MCS adaptation according to the channel conditions measured by its own multicast receivers).

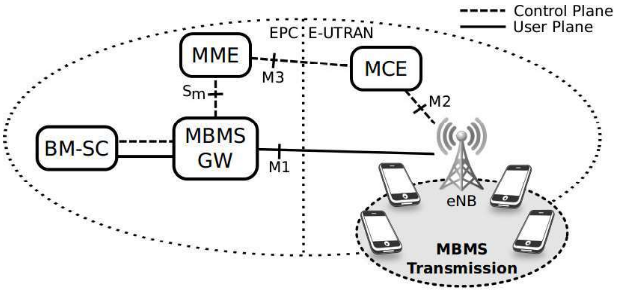


Fig. 2. The LTE/MBMS architecture.

The LTE downlink air interface is based on OFDMA. The spectrum is managed in terms of Resource Blocks (RBs), which is the smallest frequency resource unit that can be assigned to a User Equipment (UE). Each RB corresponds to 12 consecutive and equally spaced sub-carriers in the *frequency domain* and lasts 0.5 ms in the *time domain*. The overall number of available RBs depends on the channel bandwidth configuration; it can vary from 6 (1.4 MHz channel bandwidth) to 100 (20 MHz). In order to allow broadband wireless access, 3GPP defined the LTE-Advanced (LTE-A) system to support channel bandwidth up to 100 MHz through a carrier aggregation scheme that guarantees a higher spectrum utilization and backward compatibility with LTE devices.

The LTE resource allocation for unicast transmission is handled by the packet scheduler, whose detailed specifications are not defined by 3GPP, so it is up to implementation to define the preferred policy at the eNodeB. The packet scheduler can be decomposed into a time-domain and a frequency-domain scheduler [11]. In any scheduling frame, the Time Domain Packet Scheduler selects the flows to serve according to their QoS constraints. Every Transmission Time Interval (TTI, lasting 1 ms), the Frequency Domain Packet Scheduler assigns to each scheduled flow the adequate number of RBs (with

TABLE I
CQI-MCS MAPPING [10]

CQI index	Modulation Scheme	Code rate x 1024
1	QPSK	78
2	QPSK	120
3	QPSK	193
4	QPSK	308
5	QPSK	449
6	QPSK	602
7	16-QAM	378
8	16-QAM	490
9	16-QAM	616
10	64-QAM	466
11	64-QAM	567
12	64-QAM	677
13	64-QAM	772
14	64-QAM	873
15	64-QAM	948

relevant MCSs) on a RB-pair basis (i.e., two contiguous RBs in the time domain) by taking into account the status of the link. The assigned MCS is selected on the basis of a Channel Quality Indicator (CQI) feedback message transmitted by the UE to the eNodeB as an indication of the maximum supported MCS for a target Block Error Rate (BLER) value (as referred in Table I). The *frequency selectivity* can be exploited during the resource allocation procedures to improve the spectral efficiency. It consists in selecting, in the frequency domain, the most adequate portion of spectrum to assign to each served user.

III. RELATED WORK

In a single-cell scenario, group-oriented data services can be delivered towards multiple destinations in two modalities: Point-to-Point (PtP) and PtM. According to the former mode, data traffic is delivered to each group member by using a dedicated channel, thus, transmission parameters (i.e., MCS) are optimized on a *per-user* basis. On the contrary, the PtM mode feeds the whole multicast group with a single transmission. A performance analysis of PtP and PtM modes for group-oriented services in LTE systems is available from [12], where the authors clearly show that the PtP solution is unsuitable to handle multicast services, due to the large number of dedicated channels that shall be activated and that severely limits the number of group members, which can be served. PtM improves the resource utilization compared to PtP, and the achievable gain increases with the number of UEs. Nevertheless, the main disadvantage of PtM is that the MCS of the transmitted multicast flow should be selected to guarantee successful reception to all the multicast subscribers in the cell, and hence it is typically a low data rate MCS. This implies a session performance degradation, which affects the service quality perceived by the terminals.

To overcome this problem, the use of *channel aware* FDPSs has been considered in several works. For instance, the works [13] [14] extend the legacy LTE/MBMS baseline by introducing the CQI feedback transmission by the group members; this feature allows a MCS selection at the eNodeB that complies with the users' channel state variations. In general, we can say that the transmission of channel quality information by group members enables the design of enhanced RRM policies, specifically tailored to the multicast services. The numerous studies addressing RRM strategies in OFDMA-based systems [6] can be classified into three categories: *conservative*, *opportunistic*, and *subgrouping*. The *conservative* strategies select the single MCS for the group based on the multicast group member(s) with the worst channel condition (i.e., the lowest CQI among the collected ones) [6]. The performance of all group members will be bounded by the cell-edge multicast users that typically measure the poorest channel qualities, with consequent resource allocation inefficiencies [6] [15]. The conservative approach is at the basis of the proposal in [16], where the authors propose a single-rate policy for sub-channel allocation in multi-group scenarios.

The *opportunistic* strategies [17] follow the idea to dynamically change the MCSs (and, consequently, the portion of served users) within each scheduling frame, either by adopting threshold-based solutions [17] [18] [19] [20] or by optimizing a given objective cost function, such as spectral efficiency or throughput [21] [22]. For instance, author in [22] optimized the rate selections for all the system resources to maximize the throughput of the user with the worst quality, even in the general case that the channel qualities of terminals are non-identically distributed. As proposed in [23] [24], opportunistic-based schemes can support multi-rate applications by exploiting Multiple Description Coding (MDC). The data stream is fragmented into several substreams (or descriptors) and the received data quality depends on the number of successfully received descriptors. In [23] a weighted sum rate maximization method is proposed, whereas the work in [24] focuses on the fairness issue, although fairness is only considered as a constraint on the minimum number of sub-channels to assign to the groups. Validating the assumption that any combination of received sub-carriers can be decoded at the receiver is still an open issue. A coding algorithm is required to efficiently map the original data onto the assigned sub-channels, while avoiding high complexity on the receiver side [25]. Furthermore, as the portion of terminals served by the scheduler dynamically changes within the scheduling frame, opportunistic-based solutions need to work with rate-less coding schemes [21]; this adds further issues of computational burden, buffer size, decoding delay, and short-term fairness [20].

Finally, *subgrouping* [6] strategies, based on the multi-rate approach, have been proposed in the literature. To reduce the bottleneck effects of cell-edge users, these split the multicast members into different subgroups, each one including users with similar channel conditions, and serve the whole multicast group within every scheduling frame. For instance, in [26] the subgroup formation problem is outlined in a single-group scenario to the aim of maximizing the system throughput. For multicast video streaming applications, subgrouping could take advantage of SVC techniques that organize the original video stream into a base-layer and multiple enhancement layers. The goal of SVC is to improve the perceived video quality in scenarios where users experience heterogeneous channel conditions, at the cost of a bit-rate increase of at least 10% compared to a single-stream [7]. SVC can effectively work with subgroup-based scheduling strategies as shown in Fig.

3: the base layer (BL), which is essential for decoding the whole video frame, is received by *all* the multicast group members (e.g., users in both subgroups in Fig. 3), while each enhancement layer is delivered only to a subgroup of users (e.g., the enhancement layer E1 is transmitted to users in subgroup 2 only).

As shown for instance in [27], finding the optimal subgroup configuration, based on the maximization of a given objective function, is a NP-hard problem; in fact, the complexity of the subgroup formation exponentially depends on the available system resources and on the number of multicast members. Complexity increases in multi-group environments. To overcome this issue, RRM policies based on heuristic solutions, which run in polynomial time, as addressed in [27]-[28], are preferred in practical systems. More in detail, the works in [29], [28] focus on a policy that maximizes the total system throughput, but they do not account for any fairness issue. On the contrary, the work in [27] proposes a scheduling policy only based on proportional fairness.

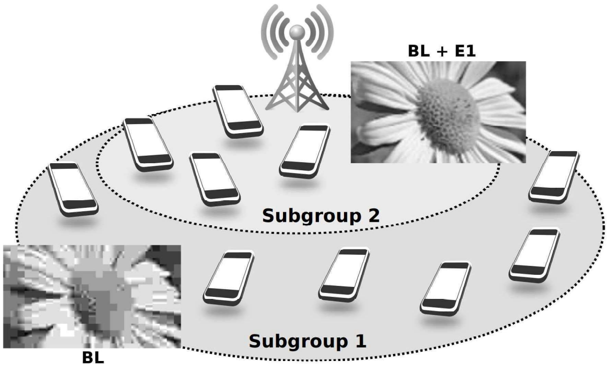


Fig. 3. Subgrouping with SVC application.

A. A step forward

In this paper we propose a novel RRM scheme that exploits *frequency selectivity* for the resource allocation of *multiple* multicast groups in LTE and beyond systems. The proposed solution extends our previous work in [26], which focused on subgrouping techniques applied to a single-group scenario without accounting for frequency selectivity. Likewise in [26], our scheme enhances the current LTE/MBMS baseline by considering the transmission of CQI feedback by MBMS subscribers, as addressed, for instance, in other researches such as [13] and [14]. We also advance our work presented in [30], in which five proposed policies to manage multiple SVC streams in a LTE cell are compared. These differ in the subgroup formation approach implemented and in the logic followed to select the multicast stream to serve and were representatives of the cited conservative, opportunistic, and subgrouping strategies. These techniques will be considered for performance comparison in this paper, and will be briefly detailed in Section VII.

From [30], it emerged that there is not any single solution that can satisfy both the system and the user requirements; specifically, those solutions that guarantee the multicast members with a higher session quality require a great amount of radio resources; conversely, the policies that offer a higher spectral efficiency and a lower resource consumption cannot always guarantee an adequate user quality. In the present research work we advance the study presented in [30] by designing a fresh new RRM policy that is able to offer high video quality to the multicast users while also guaranteeing high spectral efficiency. The proposed Multicast Subgrouping for Multi-Layer video applications (MSML) scheme outperforms the previous approaches by improving both the subgroup formation and group selection policies. As for the former issue, MSML adopts a novel subgrouping technique that creates subgroups for the purpose of guaranteeing intra-group spectral efficiency, i.e., the subgroup which offers the highest spectral efficiency improvement is enabled. Spectral efficiency is also taken into account for group selection, where we propose two different cost functions designed to guarantee inter-group fairness by considering the ratio of received data (i.e., previously scheduled layers) and the overall amount of data relevant to a given group (i.e., all layers of the video stream). Thanks to the above mentioned features, we demonstrate that MSML (exploiting both proposed cost functions) is able to outperform the schemes in [30] in terms of spectrum utilization and service quality. As a consequence, the proposed solution is suitable for implementation in practical systems wherein multicast streams share the available bandwidth with unicast services.

TABLE II
NOTATIONS USED IN THE PAPER

$\mathcal{G}$	Multicast group set
$\mathcal{K}_g$	Set of users in multicast group g
$\mathcal{N}$	Set of available resources in the frame
$\mathcal{C}$	Set of admissible CQI levels
$c_{g,k,n} \in \mathcal{C}$	CQI of user k in group g on the RB n
$\overline{c}_{g,k} \in \mathcal{C}$	Mean CQI of user k in the group g
L_g	Number of layers related to the video of group g
$d_{g,l}$	Number of bits of the l-th layer for the group g
$\mathcal{K}_{g,l} \subseteq \mathcal{K}_g$	Users in group g receiving the l-th layer
$\mathcal{N}_{g,l} \subseteq \mathcal{N}$	RBs for delivering the l-th layer of group g

IV. SYSTEM MODEL

In this work we refer to a *single-cell* scenario, like the one illustrated in Fig. 1, where the eNodeB exploits PtM transmissions to serve multiple multicast groups in the cell.

Let us denote by $\mathcal{G}$ the *multicast group set*, which includes all the groups served by the eNodeB. Let $\mathcal{K}_g$ be the *user set* which collects the users that joined the multicast group $g \in \mathcal{G}$.

The set of available resources in a frame, i.e., the *RB set*, is denoted with $\mathcal{N}$. The channel quality perceived over each RB is represented by an integer value that indicates the maximum supported MCS [10] (refer to Table I). Let us denote by $\mathcal{C}$ the *CQI set* and by $c_{g,k,n} \in \mathcal{C}$ the CQI value, relevant to the RB $n \in \mathcal{N}$ experienced by the user k belonging to the group g (i.e., $k \in \mathcal{K}_g$). Finally, $\overline{c}_{g,k} \in \mathcal{C}$ is the mean CQI achieved by such a user over the whole available spectrum.[1]

Each multicast group is served with a video flow encoded through SVC techniques. Let L_g be the number of layers of the

[1] In LTE systems, the CQI experienced by a user on the available spectrum is referred to as wideband CQI.

video flow delivered to the group g. We indicate by $\mathcal{K}_{g,l} \subseteq \mathcal{K}_g$ the subset of users that joined the multicast group g and that receive the l-th layer (with $l = 0, 1, \ldots, L_g - 1$), where $l = 0$ indicates the base layer, $l = 1$ the first enhancement layer, and so on. Let $d_{g,l}$ denote the number of bits related to the l-th layer relevant to the multicast flow g. Finally, $\mathcal{N}_{g,l} \subseteq \mathcal{N}$ represents the set of RBs selected for the transmission of such a layer.

A. System constraints

The proposed RRM scheme must meet a number of constraints in order to suitably perform resource allocation in a multicast scenario and to successfully exploit SVC techniques. These constraints are briefly discussed.

1) Resource Constraints: The RBs allocated in a scheduling-frame shall not exceed the number of those available:

$$\sum_{g \in \mathcal{G}} \sum_{l=0}^{L_g - 1} |\mathcal{N}_{g,l}| \leq |\mathcal{N}| \tag{1}$$

Each scheduled resource shall be assigned for the transmission of *one* layer towards *one* multicast group:

$$\mathcal{N}_{g,l} \cap \mathcal{N}_{g^*,l^*} = \{\emptyset\}, \forall g, g^* \in \mathcal{G} | g \neq g^*, \forall l, l^* | l \neq l^* \tag{2}$$

The MCSs related to the RBs assigned to the group g for the transmission of the l-th layer can be supported by all users selected to receive such a layer:

$$m_n = \min_{k \in \mathcal{K}_{g,l}} c_{g,k,n}, \ \forall n \in \mathcal{N}_{g,l} \tag{3}$$

where $m_n \in \mathcal{C}$ is the index of the selected MCS for the transmission over the RB n.

2) Layer Constraints: The base layer shall be delivered to *all* the multicast receivers of a given group:

$$\mathcal{K}_{g,0} = \mathcal{K}_g, \ \forall g \in \mathcal{G} \tag{4}$$

Finally, the users selected for the reception of a given layer shall be already scheduled for the reception of previous layers:

$$\mathcal{K}_{g,l} \subseteq \mathcal{K}_{g,l-1}, \ \text{with } l = 1, 2, \ldots, L_g - 1, \ \forall g \in \mathcal{G} \tag{5}$$

V. The MSML Algorithm

The proposed MSML scheme is designed to guarantee high spectral efficiency, high video quality, and intra- and inter-group fairness. Similarly to [27], we assume that video layers are synchronized and that data are grouped on a per-layer basis, i.e., bits relevant to a given video layer are managed by the packet scheduler as a single data unit. According to this model, the data unit corresponding to a given layer is scheduled only if the units associated to the preceding layers have been already scheduled. We also consider that MBMS members update their CQI values every scheduling frame, to allow MSML to select the most suitable subgroup configuration for video delivery according to the channel quality variations.[2]

The proposed MSML is designed to assign resources to the subgroups by considering the frequency selectivity. An example is shown in Fig. 4, where the channel quality experienced on each RB is drawn for four sample users; one can observe that, according to our algorithm, the scheduled users have assigned the RBs in which they experience the highest channel quality. Users in subgroup 1 are assigned the RB1-RB4 resources that allow to adopt a high-rate MCS according to the experienced channel conditions. This way, the available resources are more efficiently exploited.

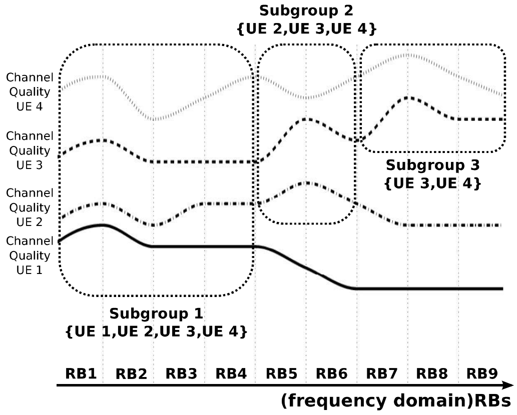


Fig. 4. Frequency selectivity in multicast subgrouping environments.

MSML is carried out in two phases, described in details in the subsequent subsections. First, the algorithm provides each scheduled multicast group with the resources needed to deliver the base layer. Subsequently, the perceived quality is increased by allocating resources for the enhancement layer(s) depending on channel conditions and available resources.

A. Base Layer Allocation

Table III summarizes the algorithm for base layer allocation. To the purpose of improving spectral efficiency and increasing the number of supported groups, eNodeB exploits frequency selectivity and tries to minimize the amount of RBs used for base layer transmissions.

For each multicast group g, based on the channel conditions of *all* the multicast destinations (*lines 1-7*), the algorithm computes the sustainable MCS of each available RB, i.e., $m_{g,n}$. This meets the constraints (3) and (4). In detail, $m_{g,n}$ is the minimum MCS among those supported by the users of multicast group g over the RB n. The selection of $m_{g,n}$ according to *line 5* guarantees that the MCS adopted for the considered RB is supported by all multicast members according to the experienced CQI. Once $m_{g,n}$ is computed for each group and for each RB, then MSML starts the iterations for the base layer assignment (*lines 8-18*).

At every iteration, MSML computes the set $\mathcal{N}_{g,0} \subseteq \mathcal{N}$ for each group (*line 10*) which still needs to be served with the base layer (condition in the for loop at *line 9*). Such a set collects the "best" resources to convey the base layer towards the generic group g. In detail, $\mathcal{N}_{g,0}$ is the minimum set of RBs that guarantees the base layer delivery, i.e., the RBs associated to the highest MCSs among those supported by the group

[2] More sophisticated CQI reporting schemes [14] could be conceived that foresee, for instance, the update of CQI values only when the UE observes a channel variation for more than a pre-defined number of scheduling frames. This would help to reduce the signaling load and the frequent switching ON/OFF of the enhancements layers for terminals with fast channel variations. The analysis of this aspect is however left for future work.

TABLE III
BASE LAYER ALLOCATION

```
1: for all $g \in \mathcal{G}$ do
2:     $l_g = 0$
3:     $\mathcal{K}_{g,0} = \mathcal{K}_g$
4:     for all $n \in \mathcal{N}$ do
5:         Compute $m_{g,n} = \min_{k \in \mathcal{K}_{g,0}} c_{g,k,n}$
6:     end for
7: end for
8: repeat
9:     for all $g \in \mathcal{G} | l_g = 0$ do
10:        $\begin{cases} \mathcal{N}_{g,0} = \arg\min_{\tilde{\mathcal{N}}_{g,0} \subseteq \mathcal{N}} |\tilde{\mathcal{N}}_{g,0}| \\ s.t. \\ \sum_{n \in \tilde{\mathcal{N}}_{g,0}} f(m_{g,n}) = d_{g,0} \end{cases}$   (6)
11:            if (6) can not be accomplished then
12:                Update $\mathcal{G} = \mathcal{G} - \{g\}$
13:            end if
14:    end for
15:    Select $g^* = \arg\min_{g \in \mathcal{G} | l_g = 0} |\mathcal{N}_{g,0}|$
16:    $l_{g^*} = 1$
17:    Update $\mathcal{N} = \mathcal{N} - \mathcal{N}_{g^*,0}$
18: until $|\{g \in \mathcal{G} | l_g = 0\}| = 0 \vee |\mathcal{N}| = 0$
```

members, as indicated in the constraint (6). The value $f(\cdot)$ in (6) indicates the number of achievable bits over the considered RB [10]. This varies according to the selected MCS, i.e., $m_{g,n}$. In case the available resources cannot guarantee the base layer transmission, such a group is deleted from the $\mathcal{G}$ set.

Line 15 in Table III indicates that the group g^*, which requires the lowest amount of resources, is selected.[3] If several groups require the same amount of resources, then the algorithm selects the one with the highest number of served users. This aims at improving the system capacity.

The approach described aims at minimizing the resource consumption and maximizing the system capacity, since it aims at serving the highest possible number of multicast groups. Once the group g^* is selected, the set $\mathcal{N}$ of available resources is updated and the parameter l_g^* (i.e., the index value of the next layer to be delivered to the multicast group) is set to 1.

Iterations stop either when all groups are served, or when no more resources are available.

The complexity of the code in *lines 1-7* is $\mathcal{O}(GNK)$, where K is the number of UEs in the most populated group, whereas the complexity of the code in *lines 8-18* is $\mathcal{O}(G^2N)$. Without loss of generality, we can assume that $K > G$, i.e., the number of users in the most populated group is higher than the number of served multicast flows, hence the overall complexity for the base layer allocation is equal to $\mathcal{O}(GNK)$.

B. Enhancement Layer Allocation

The algorithm for the enhancement layer allocation is summarized in Table IV.

TABLE IV
ENHANCEMENT LAYER ALLOCATION

```
1: for all $g \in \mathcal{G}$ do
2:     if $L_g > 1$ then
3:         $l_g = 1$
4:         for all $k \in \mathcal{K}_{g,0}$ do
5:             Compute $\mathcal{U}_{g,k} = \{u \in \mathcal{K}_{g,0} : \bar{c}_{g,u} \geq \bar{c}_{g,k}\}$
6:             for all $n \in \mathcal{N}$ do
7:                 Compute $m_{g,k,n} = \min_{u \in \mathcal{U}_{g,k}} c_{g,u,n}$
8:             end for
9:         end for
10:    else
11:        $\mathcal{G} = \mathcal{G} - \{g\}$
12:    end if
13: end for
14: repeat
15:    for all $g \in \mathcal{G}$ do
16:        for all $k \in \mathcal{K}_{g,l_g - 1}$ do
17:            $\begin{cases} \mathcal{R}_{g,k} = \arg\min_{\tilde{\mathcal{R}}_{g,k} \subseteq \mathcal{N}} |\tilde{\mathcal{R}}_{g,k}| \\ s.t. \\ \sum_{n \in \tilde{\mathcal{R}}_{g,k}} f(m_{g,k,n}) = d_{g,l_g} \end{cases}$   (7)
18:        end for
19:        if (7) can not be accomplished then
20:            Update $\mathcal{G} = \mathcal{G} - \{g\}$
21:        else
22:            $\bar{k}_g = \arg\max_{k \in \mathcal{K}_{g,l_g - 1}} \frac{d_{g,l_g} \cdot (|\mathcal{U}_{g,k}|/|\mathcal{K}_g|)}{|\mathcal{R}_{g,k}|}$
23:            Define $\tilde{\mathcal{K}}_{g,l_g} = \mathcal{U}_{g,\bar{k}_g}$
24:            Define $\tilde{\mathcal{N}}_{g,l_g} = \mathcal{R}_{g,\bar{k}_g}$
25:        end if
26:    end for
27:    Perform group selection according to (8) or (9)
28:    $l_{g^*} = l_{g^*} + 1$
29:    $\mathcal{K}_{g^*,l_{g^*}} = \tilde{\mathcal{K}}_{g^*,l_{g^*}}$
30:    $\mathcal{N}_{g^*,l_{g^*}} = \tilde{\mathcal{N}}_{g^*,l_{g^*}}$
31:    if $l_{g^*} > L_{g^*} - 1$ then
32:        Update $\mathcal{G} = \mathcal{G} - \{g^*\}$
33:    end if
34:    Update $\mathcal{N} = \mathcal{N} - \mathcal{N}_{g^*,l_{g^*}}$
35: until $|\mathcal{G}| = 0 \vee |\mathcal{N}| = 0$
```

We recall that, from the previous phase, the $\mathcal{G}$ set includes the groups served with the base layer, and $\mathcal{N}$ indicates the resources still available after the base layer assignment.

As shown in *lines 1-13*, MSML computes all the admissible subgroups that could be formed for each multicast group. Each candidate solution is indicated by $\mathcal{U}_{g,k} \in \mathcal{K}_g$, with $k \in \mathcal{K}_g$. The subgroup $\mathcal{U}_{g,k}$ contains the users belonging to $\mathcal{K}_g$ that experience a mean channel quality greater or equal to the one of the member k in such a group, i.e., $\bar{c}_{g,k}$. Hence, the overall number of considered subgroup configurations is equal to $|\mathcal{K}_g|$. Once the admissible subgroup configurations are defined, the algorithm evaluates the sustainable MCS for available RBs to select the most performing portion of spectrum to assign the transport block relevant to each subgroup configuration. Let $m_{g,k,n}$ be the MCS for the transmission over the RB n according to the number of users belonging to the candidate subgroup configuration $\mathcal{U}_{g,k}$.

At *line 14*, MSML phase 2 begins its iterations. Since each subgroup of a given multicast group collects users that experience different channel qualities, MSML must evaluate (*lines 16-20*) the most adequate portion of RBs for the delivery of the required layer, i.e., l_g, to each candidate subgroup. At every iteration, the candidate subgroups are those which

[3] It is worth noting that, in a scenario where the available resources cannot guarantee the base layer reception to all the scheduled flows, the proposed resource allocation minimizes the number of "not served" flows.

contain the users scheduled for the reception of the previous layer in order to fulfill the constraint (5).

We indicate by $\mathcal{R}_{g,k} \subseteq \mathcal{N}$ the RB set relevant to $\mathcal{U}_{g,k}$, i.e., the set which contains the lowest number of resources to convey the l_g-th layer according to the channel conditions of the users in $\mathcal{U}_{g,k}$. If the available resources cannot guarantee the transmission of the considered layer for any of the subgroup configurations, then the given multicast group is deleted from the $\mathcal{G}$ set. Once all $\mathcal{R}_{g,k}$ sets are created, the best subgroup configuration, denoted by $\tilde{\mathcal{K}}_{g,l_g}$ and $\tilde{\mathcal{N}}_{g,l_g}$, is selected (*line 22*), which is able to convey the l_g-th layer for the group g in the current iteration.[4] According to *line 22*, the selected subgroup is the one that guarantees the highest intra-group spectral efficiency. At the end of this phase, the algorithm has selected the best subgroup (with the associated resources) for each group.

Finally (*line 27*), the scheduled multicast group is selected for the current iteration. For this step, we propose the use of two different cost functions tailored to guarantee the highest spectral efficiency while assuring inter-group fairness, selected based on the ratio between the number of users in the subgroup, $|\tilde{\mathcal{K}}_{g,l}|$, and the number of resources requested by the subgroup, $|\tilde{\mathcal{N}}_{g,l}|$. The proposed cost functions vary according to the approach used to take into account the inter-group satisfaction fairness. The first cost function is defined as:

$$g^* = \arg\max_{g\in\mathcal{G}} \frac{\log\left(\frac{\sum_{l=1}^{l_g} d_{g,l}}{\sum_{l=1}^{L_g} d_{g,l}}\right)\cdot|\tilde{\mathcal{K}}_{g,l_g}|}{|\tilde{\mathcal{N}}_{g,l_g}|} \tag{8}$$

i.e., fairness is considered through the logarithmic ratio between the obtained and the maximum data rate values. The second cost function is defined as:

$$g^* = \arg\max_{g\in\mathcal{G}} \frac{\log\left(\sum_{l=1}^{l_g} d_{g,l}\right)\cdot\frac{L_g}{l_g}\cdot|\tilde{\mathcal{K}}_{g,l_g}|}{|\tilde{\mathcal{N}}_{g,l_g}|} \tag{9}$$

In this case, the fairness requirement is met by accounting for the ratio (L_g/l_g), i.e., the ratio between the total number of video layers and the index of the next video layer to schedule for a group. Such a value gives higher priority to groups that still miss a greater number of layers compared to others.

Once the group g^* is selected[5], the RBs belonging to the set $\mathcal{N}_{g^*,l_{g^*}}$ are marked as *not available* and the layer l_{g^*} is marked as *scheduled*. Finally, the group g^* is deleted by the $\mathcal{G}$ set if all its enhancement layers have been assigned. The iterations stop either when no more resources are available or when all layers have been transmitted towards all multicast groups.

The complexity of the code in *lines 1-15* is $\mathcal{O}(GK^2N)$. The maximum number of iterations of the code in *lines 16-38* can be expressed as $\mathcal{O}(GL)$, where L is the maximum number of layers to be transmitted, whereas the complexity of the code in *lines 18-28* is $\mathcal{O}(GKN)$. Hence the complexity of the enhancement layer allocation is equal to $\mathcal{O}(GK^2N+G^2LKN)$. By assuming $K > N > G > L$, the overall complexity of the proposed MSML algorithm is $\mathcal{O}(GK^2N)$.

VI. Simulation Assumptions

The performance analysis is conducted in accordance with the guidelines defined in [31]. Channel quality is evaluated in terms of Signal to Noise and Interference Ratio (SINR) experienced over each sub-carrier [32]:

$$SINR_i = \frac{P_0 \times PL_0 \times h_0}{\sum_{j=1}^{N_{BS}}(P_j \times PL_j \times h_j) + N_o} \tag{10}$$

where P_0, PL_0, and h_0 are, respectively, the transmission power, the path loss, and the small scale fast fading of the link between the UE and the serving base station; whereas, P_j, PL_j and h_j are the transmission power, the path loss, and the small scale fast fading of the link between the UE and the j-th interfering base station; N_o is the noise power. The Exponential Effective SIR Mapping (EESM) [33] is used to map the channel state into the effective SINR. Finally, the effective SINR is mapped onto the CQI level ensuring a BLER value lower than 1% [32][34]. More details on the LTE system settings are listed in Table V.

The members of each multicast group are randomly distributed in a concentrated area within the macrocell, so to represent a typical on-campus scenario. We consider that MBMS users are distributed in the area covered by one serving cell, which is placed in the center of a larger cell deployment scenario (i.e., an hexagonal grid with 19 cell sites, 3 sectors per site [31]). Each adjacent cell acts as an interference source and serves a set of 50 best effort with infinite buffer users. A proportional fairness scheduler is implemented at the interfering cells. Various multicast video sessions are activated by different multicast groups in the simulated cell; the source data rate settings of the base layer (BL) and enhancement layers (E1, E2, and E3) are generated according to [35]. Table VI shows the average source bit rate relevant to different layers for the video flows considered in our analyses.

We simulated a video delivery period of 1s, i.e., 1000 TTIs. Each simulation run has been repeated several times to get 95% confidence intervals for the most relevant results.

A. Performance metrics

The described RRM techniques are compared in terms of the performance metrics listed below:

- *Spectral Efficiency* is the ratio between the number of bits received by the multicast users and the channel bandwidth exploited for the multicast transmission; this metric indicates how efficiently the system resources are exploited during the multicast service provisioning.
- *Resource Consumption* indicates the amount of resources consumed to support the multicast traffic delivery; it is computed as the percentage of used RBs, during a scheduling frame, with respect to the whole set of

[4] For a given group that is not scheduled in a iteration of the loop in *lines 14-35*, this implies that the portion of users selected for receiving an enhancement layer may change in the successive iterations. Indeed, the subgroup selection for each group is influenced by the still available resources. This procedure allows to adapt in every iteration the subgroup configuration for enhancement layer delivery according to the available resources.

[5] In both lines 22 and 27, if more solutions achieve the same maximum cost function value, then the algorithm selects the one with the highest number of served users. In case of same number for several admissible solutions, then the one requiring the lowest amount of RBs is chosen.

TABLE V
MAIN SIMULATION ASSUMPTIONS

Parameter	Value
Cell layout	3GPP Macro-cell case #1, Hexagonal grid, 19 cell sites, 3 sectors per site [31]
Inter Site Distance	500 m
Distance attenuation	128.1+37.6*log(d), d [km]
Shadow fading	Log-normal,0 mean, $\sigma = 8$ [dB]
Shadowing Correlation distance	50 m [31]
Fast Fading	ITU-R PedB (extended for OFDM)
Carrier frequency	2 GHz
Scheduling frame	10 ms
RB size	12 sub-carriers, 0.5 ms
Sub-carrier spacing	15 kHz
Data/Control OFDM symbols	11/3
BLER target	1%
TTI	1 ms
CQI scheme	eNodeB-configured subband feedback
EUTRA UE	Antenna gain 0 dBi, Noise Figure 9 dB [31]
EUTRA Node-B	Antenna gain 14 dBi, Noise Figure 5 dB [31]
eNodeB transmit power	43 dBm [31]
MIMO Configuration	1 Tx, 2 Rx
Thermal Noise	-174 dBm/Hz

TABLE VI
DATA RATE [KBPS] PER LAYER [35]

Name	BL	E1	E2	E3
CREW	306	578	814	1184
FOOTBALL	442	827	1114	1621
MOBILE	189	322	442	649
CITY	448	923	1288	1943
FOREMAN	170	407	589	890
BUS	185	390	567	857
HARDBOUR	577	1025	1379	1929
NEWS	121	259	372	564
SOCCER	385	795	1095	1651
ICE	277	548	767	1123

available RBs. Please note that the resource consumption is not simply the reciprocal of the spectral efficiency. In fact, it takes into account only the number of RBs used for traffic delivery; differently, the spectral efficiency considers how such consumed resources are used (i.e., it accounts also for the number of bits transmitted over the RBs).

- *Mean Throughput* is the average data rate experienced by the multicast group members; the greater the throughput the higher the service quality and the "satisfaction" level of the multicast users.
- *Network Coverage*, computed as the empirical cumulative distribution function of the throughput of multicast members; this metric measures the throughput-fairness trade-off.
- *Standard deviation*, σ_T, of the throughput of multicast members normalized to the maximum allowable one (i.e., the rate associated to the highest video quality perceived when all enhancement layers have been received) [26]; this metric indicates how "fair" the resource allocation is in terms of user "satisfaction". Indeed, the higher the σ_T value the greater the difference in terms of "satisfaction" among multicast members; i.e., a portion of users achieves a higher satisfaction level compared to other users in the same group.[6]

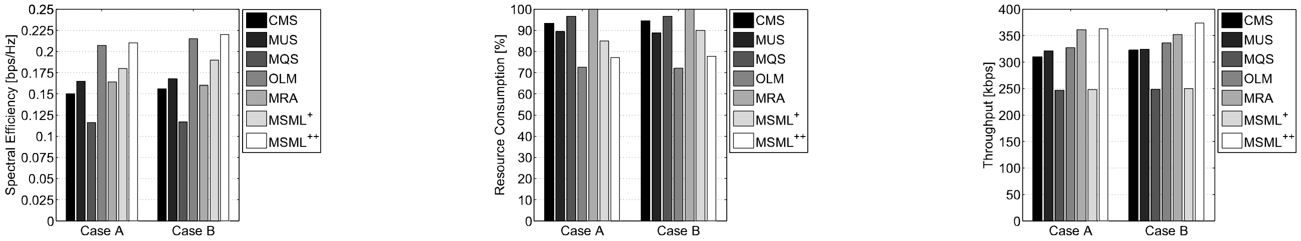
VII. PERFORMANCE ANALYSIS

MSML is compared with other strategies tailored for the resource allocation of multi-group environments; such strategies have been adapted to our considered SVC scenario in our previous work in [30]. We consider two different versions of our proposed MSML: *MSML*$^+$ indicates MSML exploiting the cost function (8) for enhancement layer group selection; *MSML*$^{++}$ indicates MSML exploiting the cost function (9). The compared schemes are briefly described here for the sake of completeness.

The *Conservative Multicast Scheme (CMS)*, based on the idea presented in [16], aims to maximize the intra-group fairness by delivering each enhancement layer following a conservative strategy, i.e., according to the user with the worst channel quality. As a consequence, all terminals belonging to the same group will experience the same video quality. By considering our scenario, for each video layer, the purpose of CMS is to serve the multicast streams through a round-robin approach by starting from the group that requires the lowest number of resources to deliver the video layer [30]. The *Median User Scheme (MUS)* is based on the class of opportunistic techniques such as [23] [24] [17] [20]. In SVC environments, the MUS dynamically adapts the portion of scheduled users by delivering a given enhancement layer only to 50% of the users which received the previous layer [30]. In doing so, the system throughput can be improved and the overall resources saved. Indeed, among the best subgroups of each multicast group, MUS select the one that guarantees the highest spectral efficiency increase. The *Median Quality Scheme (MQS)* is based on strategies like [23] [24] [18] [19]; the aim is to use the system resources in an efficient way by scheduling the users according to a threshold CQI value: in SVC scenarios, the MQS is tailored to convey each enhancement layer to the terminals which experience a CQI value higher than a "mean" CQI [30]. Similarly to MUS, the served subgroups are those which guarantee the highest spectral efficiency increase. The *Opportunistic Layered Multicasting (OLM)* policy performs the resource allocation (i.e., the group selection) and the subgroup formation so as to minimize the amount of RBs necessary for the delivery of the enhancement layer [21] [30]. Finally, the *Multicast Resource Allocation (MRA)* extends the idea in [27] by implementing a proportional fair resource allocation.

In [30], we showed that CMS and MRA guarantee high service quality at the expenses of a great amount of allocated radio resources; whereas OLM, MUS, and MQS achieve higher spectral efficiency at the expenses of a lower quality. A comparison of MSML with the more traditional PtP-based policy is not considered as fair in this context, since, as outlined for instance in [12], it is well known that the PtP performance drastically decreases when the number of

[6] Fairness is not measured through the well-known *Jain's fairness index (JFI)* since, in the evaluated scenario, each multicast group is subject to different data rate constraints. Indeed, JFI indicates how close to each other the throughput values of the multicast users are, without considering how much the achieved throughput is close to the requested one.



(a) Spectral Efficiency

(b) Resource Consumption

(c) Throughput

Fig. 5. Performance results in the Scenario with four groups and fixed bandwidth in section VII-A.

multicast users increases, and this aspect makes PtP unsuitable to MBMS delivery in the reference context.

Performance analyses are carried out in different simulation scenarios, as outlined in the following subsections.

A. Four groups with fixed bandwidth

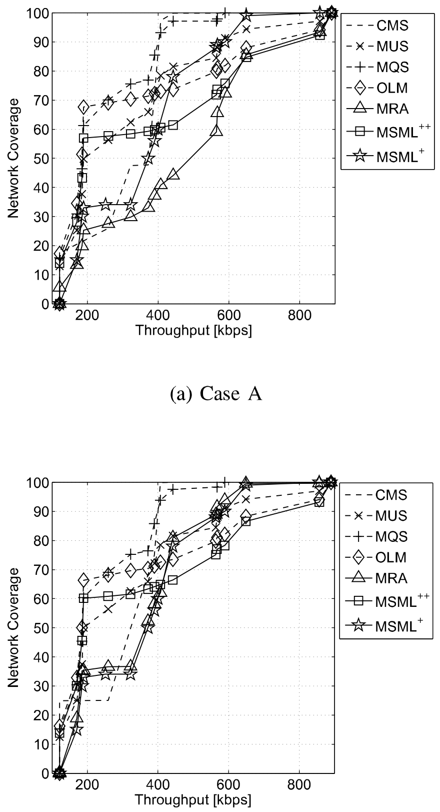
In the first analysis four video flows (MOBILE, FOREMAN, BUS, and NEWS) are transmitted by the eNodeB towards 400 multicast members in its cell over a channel bandwidth equal to 10 MHz (i.e., 50 RBs). We analyze two different simulation cases: *Case A*, where the 400 multicast destinations are uniformly distributed among the four multicast services (i.e, each multicast group interested in a given video stream is composed of 100 members); *Case B*, where the destinations are unequally distributed among the four multicast services (i.e., the multicast groups are composed of a different number of users). This task is performed through the *randfixedsum* function, provided by the Matlab software. For a given simulation of the considered scenario, such a function generates an array with four positions, each one containing a random value from the interval $[1,x]$, under the constraint that the sum of all values is equal to x, with $x = 400$ in our case.

Purpose of this simulation campaign is to explore to what extent the user distribution among the multicast groups influences the performance.

Let's first focus on the spectral efficiency results in Fig. 5(a). It clearly emerges that the CMS and the MQS policies suffer from poor spectral efficiency (0.153 and 0.116 bps/Hz, respectively), and their behavior does not meaningfully change in the two addressed scenarios. The MUS and the MRA policies achieve similar results in Case A (0.16 bps/Hz), whereas the MUS technique outperforms MRA in Case B (0.17 and 0.16 bps/Hz, respectively). The OLM technique reaches a spectral efficiency equal to 0.21 and 0.215 bps/Hz in Cases A and B, respectively. Finally, the proposed MSML^{++} achieves the highest spectral efficiency in both cases, 0.217 and 0. 222 bps/Hz in Cases A and B, respectively, while MSML^{+} obtains a performance equal to 0.18, on average. The mean gain of MSML^{++} compared to the OLM and the MRA policies is equal to 3.5% and 37%, respectively.

Fig. 5(b) analyzes the resource consumption. The performance of all the considered policies does not vary substantially in the two evaluated cases. CMS, MUS and MQS require the 94%, 89% and 96% of available RBs, respectively, whereas this value is close to 100% for MRA. This result testifies to the unsuitability of these schemes whenever multicast services coexist in the same cell with additional services, such as unicast flows. Results from 85% to 90% are obtained by MSML^{+}. As expected, OLM uses the lowest percentage of RBs, equal to 72%. The proposed MSML^{++} achieves a performance close to the OLM technique, i.e., uses 78% of available RBs. The results obtained by MSML demonstrate that the proposed scheme offers a reasonable low resource consumption for the multicast service delivery. These results are highlighted by the performance of MSML^{++} which allows to preserve the system resources while guaranteeing a high spectral efficiency performance. CMS, MUS and MRA achieve poor spectral efficiency since they exploit all the available resources even the ones with low channel quality for to many users.

Fig. 5(c) depicts the mean throughput experienced by the multicast members. MQS is the worst performing policy, showing a performance equal to 248 kbps, on average; this value is lower than the one achieved by CMS (310 and 323 kbps in Cases A and B, respectively) and MUS (322 kbps, on average). The performance of MSML^{+} is of about 249 kbps. MUS can guarantee a throughput performance close to CMS by exploiting a lower amount of resources. OLM has a performance varying from 327 to 336 kbps, while the proposed MSML^{++} guarantees a throughput equal to 363 and 374 kbps, respectively, since these two approaches are able to efficiently exploit the multi-user diversity in selecting the portion of users to serve. Finally, it is worth noting that the MRA policy is the only one influenced by the user distribution. Indeed, MRA is designed to guarantee intra-group fairness thanks to a proportional fairness allocation, but it cannot guarantee adequate inter-group fairness, since it does not account for the amount of free resources and the number of conveyed video layers. In Case A, the throughput for MRA is close to the one of MSML^{++} technique (i.e, 361 kbps) while the performance decreases down to 352 kbps in Case B. The reason is that, in Case B, MRA schedules large multicast subgroups (as demonstrated by the increase in the spectral efficiency) although this does not correspond to a higher throughput (the throughput depends on the data rate requirements of the



(a) Case A

(b) Case B

Fig. 6. Network Coverage in the Scenario with four groups and fixed bandwidth in section VII-A.

served video layer). The achieved results highlight that the proposed MSML^{++}, i.e., the exploitation of cost function (9) for enhancement layer group selection, guarantees a high throughput performance, and the heterogeneity in the number of users among the served multicast groups does not influence its behavior (it is worth noting that also the performance of MSML^{+} is not influenced by the multicast group size). Indeed, the proposed MSML approach assures inter-group fairness, in terms of spectral efficiency and number of layers to convey for a given group, since it avoids that the most populated groups have more chances to be scheduled compared to smaller groups.

From the Network Coverage depicted in Fig. 6, we can note that CMS is the fairest[7] solution in terms of throughput values experienced by multicast members, although it does not reach throughput value as high as the other techniques. Among those, MRA achieves the fairest behavior since it reaches the maximum throughput value faster (i.e., the network coverage curve with the lowest slope). As expected, MQS, MUS, and OLM offer low fairness, as they guarantee high data rates only to a small percentage of users. From Fig. 6(a), one can observe that MSML-based approaches and MRA are characterized by a similar behavior in homogeneous conditions, whereas different behaviors are revealed in Case B, as depicted in Fig. 6(b). More specifically, by changing from Case A to Case B, the MRA performance varies. In Case A, 37% of users achieves a throughput equal to or lower than 390 kbps (Fig. 6(a)), whereas this percentage becomes 58% in Case B and, as a consequence, the mean throughput is lower (as shown in Fig. 5(c) and analyzed above). Differently, the performance of MSML^{+} and MSML^{++} do not vary meaningfully in the two considered cases.

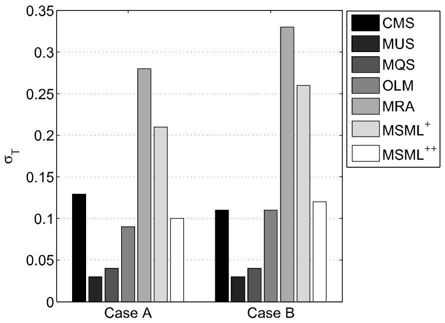
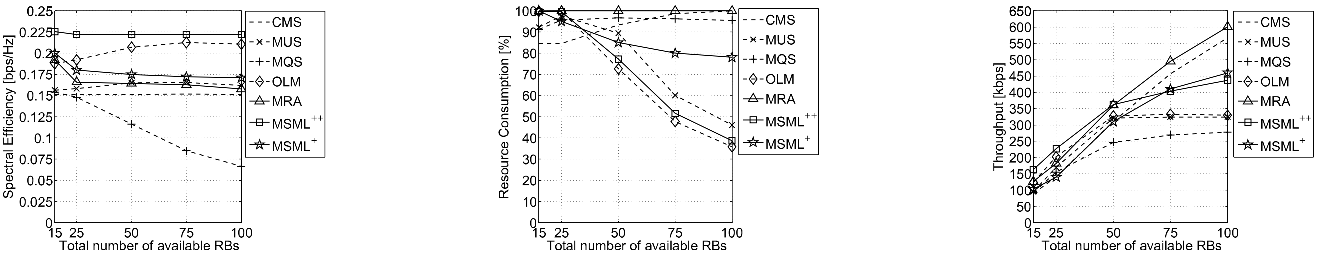


Fig. 7. Throughput standard deviation σ_T in the Scenario with four groups and fixed bandwidth in section VII-A.

This aspect is further explored in Fig. 7, which shows the throughput standard deviation. The proposed MSML^{+} and MSML^{++} have a σ_T performance that does not differ meaningfully between both the evaluated cases. A similar behavior is also observed for CMS, MUS, MQS, and OLM. In detail, CMS, OLM, MSML^{+} and MSML^{++} have a σ_T equal to 0.11, whereas this value is equal to 0.02 and 0.04 for MUS and MQS, respectively. The results achieved by MRA varies from 0.31 (Case A) to 0.33 (Case B). This demonstrates that not only the throughput but also the "satisfaction fairness" for MRA is influenced by the user distribution within the groups. In the Case B scenario, MRA schedules a higher number of resources for the most populated subgroups and, as a consequence, the difference among the "satisfaction" of the considered subgroups increases. This behavior is underlined by considering the portion of users served per layer for each video flow (not depicted in the paper due to the lack of space). MRA is influenced by the multicast group size. Indeed, in Case B, the percentage of users served with enhancement layers is higher for MOBILE and NEWS video services (i.e., the most populated groups), while in Case A (when all groups have the same size) the percentage of users per layer is almost equal for all served video streams. On the contrary, the results of CMS, MUS, MQS, OLM, and MSML approach do not meaningfully vary in the two considered cases. In particular, MSML^{+} serves enhancement layers to a portion of about 25% of users, while this portion increases up to 30% for MSML^{++} (these percentages and those of each enhancement layer vary by varying the video settings); this underlines that the use of

[7] It is worth noting that the perfect fairness is observed on the Network Coverage through a vertical line indicating that all users get the same throughput performance. In our scenarios, due to different data rate constraints of video flows, the fairness can be measured according to the slope of the Network Coverage.



(a) Spectral Efficiency

(b) Resource Consumption

(c) Throughput

Fig. 8. Performance results in the Scenario with four groups and variable bandwidth in section VII-B.

cost function (9) allows to enhance the performance compared to the function (8).

The analysis in this subsection demonstrated that the proposed MSML approach, in particular the MSML^{++} scheme, substantially improves the spectral efficiency compared to other approaches from the literature, while requiring a low amount of system resources. As for the throughput, MSML^{++} is the policy achieving the best performance. At the same time, MSML^{++} allows for a significant reduction (about 22% with respect to MRA that has throughput performance close to MSML^{++}) in terms of exploited resources. Moreover, the resource allocation performed by MSML^{++} is not affected by the user distribution within the multicast groups whereas MRA is meaningfully influenced by the number of multicast users per groups. Besides the multicast subgroup size, MSML^{++} also accounts for the number of layers already scheduled for each group; this allows accomplishing a fairer resource allocation among all multicast destinations.

B. Four groups with variable channel bandwidth

The second simulation analysis considers the transmission of the same multicast streams (MOBILE, FOREMAN, BUS, NEWS) towards the 400 multicast members uniformly distributed among the groups. The focus is on the behavior of the considered policies in different system deployment scenarios with a channel bandwidth that varies from 15 RBs (i.e., 3 MHz) to 100 RBs (i.e., 20 MHz).

The spectral efficiency is shown in Fig. 8(a). The performance of MUS, and OLM increases with the number of RBs, whereas the one of the MRA and MQS policies decreases. The spectral efficiency of both MSML^{+} and MSML^{++} is not significantly influenced by the channel bandwidth and a similar trend can be observed for CMS.

Hence, the gain introduced by the proposed MSML^{+} and MSML^{++} schemes, compared to MRA and MQS, increases as the bandwidth becomes larger. In details, the most performing scheme is MSML^{++}, whose efficiency ranges from 0.222 up to 0.225 bps/Hz. The performance of MRA decreases from 0.19 down to 0.16 bps/Hz, with a reduction of about 16%. Moreover, the spectral efficiency of MRA becomes close to one of CMS when the bandwidth increases. This emphasizes that, in large bandwidth scenarios, MRA preferably schedules large multicast subgroup; a consequence is the inefficient exploitation of multi-user diversity (similar to CMS).

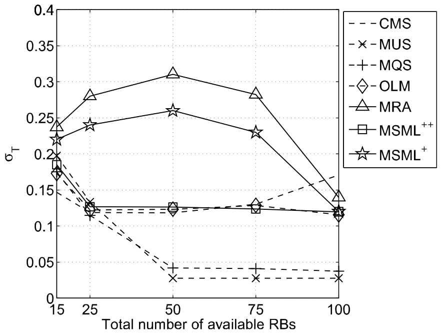
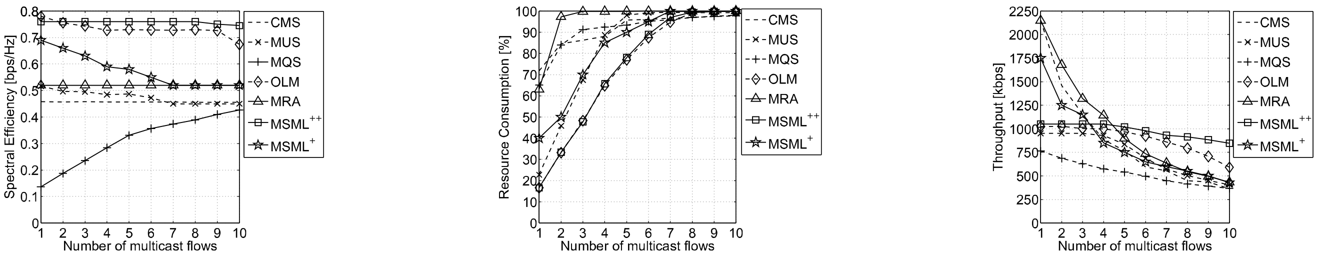


Fig. 9. Throughput standard deviation σ_T in the Scenario with four groups and variable bandwidth in section VII-B.

Plots in Fig. 8(b) show the performance in terms of resource consumption. The CMS policy requires an amount of resources higher than 85% to convey the multicast services in all the evaluated deployment cases, whereas this percentage varies from 91% to 91% for MQS. MRA has a performance equal to 100% in all evaluated scenarios, while the resource consumption of our proposed MSML^{+} decreases down to 80% in case of large bandwidth values. The OLM policy, designed to minimize the resource consumption, exploits 36% of the RBs in the best case (i.e., 100 RBs), whereas this percentage for both MUS and MSML^{++} is equal to 38% and 46%, respectively. Hence, also in this analysis, MSML^{++} outperforms MSML^{+} and achieves a performance close to the one of OLM. In details, compared to MRA, MSML^{++} reduces the percentage of the required RBs by a factor equal to about 60%.

Fig. 8(c) shows the analysis in terms of mean throughput. Obviously, all the policies provide an increased throughput when the bandwidth becomes higher. It is worth noting that the proposed MSML^{++} achieves the greatest throughput value in case of low system bandwidth, i.e., 163 and 226.6 kbps in the 15 and 25 RBs cases, respectively. MSML^{++} is outperformed by the CMS, the MRA and the proposed MSML^{+} policies when the bandwidth increases, because MSML^{++} aims at preserving the system resources. When focusing on a com-



(a) Spectral Efficiency

(b) Resource Consumption

(c) Throughput

Fig. 10. Performance results in the Scenario with fixed bandwidth and variable multicast groups in section VII-C.

parison of MSML^{++} versus CMS and MRA in case of 100 RBs available, on the one hand the throughput is reduced by a factor equal to 23% and 27% compared to CMS and MRA, respectively; on the other hand, MSML^{++} allows a reduction in terms of needed RBs almost equal to 60% compared to these policies. MSML^{++} achieves a mean throughput equal to 437 kbps in this case; and this means that a large portion of users receives the layers up to the second enhancement layer (refer to Table VI), on the average. The consequence is a high video session quality for the multicast members. Please note that the high throughput performance for the CMS and MRA policies does not correspond to a high spectrum utilization (Fig. 8(a)).

Fig. 9 shows the results in terms of throughput standard deviation σ_T. It is observed that MSML^{++} is more stable in terms of σ_T compared to CMS, MRA and MSML^{+}. MSML^{+} and MRA have a performance close to MSML^{++} only in case of large bandwidth, i.e., 100 RBs, otherwise MRA is the worst performing policy in terms of σ_T.

C. Carrier aggregation scenario with variable number of multicast groups

In this scenario, the number of available RBs is increased through the aggregation of three component carriers, each one composed of 50 RBs. The number of multicast flows served in the cell varies from 1 to 10; this allows to analyze the impact of the number of video sessions on the system performance. In each simulation, all multicast services are served with the CREW video flow. A uniform distribution of users among the considered groups is assumed, with 100 users, on average, per multicast group.

Fig. 10(a) shows the performance in terms of spectral efficiency. MQS is the worst performing policy, as its efficiency varies from 0.13 up to 0.42 bps/Hz. As the number of multicast groups grows, the efficiency of OLM varies between 0.78 and 0.67 bps/Hz while the MRA performance is equal to 0.52 bps/Hz and does not vary when the number of groups increases. Also in this scenario, MSML^{++} achieves the highest performance, with a spectral efficiency varying from 0.76 to 0.74 bps/Hz. Hence, the proposed MSML^{++} scheme can guarantee high spectral efficiency also in LTE-A with carrier aggregation and when one or several streams are supported. The mean spectral efficiency gain compared to OLM and MRA is equal to 4% and 44%, respectively. It is worth noting that MSML^{+} reaches high values of spectral efficiency until six groups, then its performance becomes equal to that of CMS.

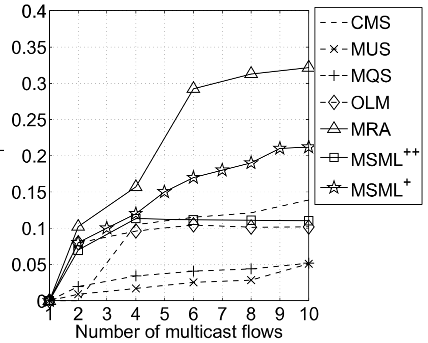


Fig. 11. Throughput standard deviation σ_T in the Scenario with fixed bandwidth and variable multicast groups in section VII-C.

Fig. 10(b) shows the percentage of assigned RBs and points out that the MRA quickly wastes the available system resources. In fact, from 1 to 3 groups, the percentage of RBs exploited by the MRA policy varies from 63% to 99%, whereas for MSML^{++} this percentage varies from 16% to 47% and from 23% to 70% for MSML^{+}.

In Fig. 10(c) the mean throughput analysis is depicted. As expected, the throughput decreases for all the policies as the number of multicast groups becomes larger. The proposed MSML^{++} technique achieves the highest throughput in highly loaded scenarios, i.e., when more than 5 groups are served in the cell. MSML is outperformed by the CMS, the MRA and the MSML^{+} policies when the number of multicast flows is low. Nevertheless, in case of a single group, the throughput is reduced by a factor almost equal to 50% compared to CMS and MRA, although MSML^{++} allows a reduction in terms of needed RBs equal to 75% compared to these policies. Also in this scenario, the high throughput performance for the CMS and MRA policies does not correspond to a high spectral efficiency (Fig. 10(a)).

Finally, the throughput standard deviation σ_T is analyzed (Fig. 11). From the achieved results, σ_T of the MRA policy turns out to be affected by the number of multicast flows served in the cell. Indeed, the σ_T of MRA increases up

TABLE VII
PERFORMANCE COMPARISON OF MSML^{++} WITH REFERRED POLICIES

	CMS	MUS	MQS	OLM	MRA	MSML^{++}
Spectral Efficiency	Low	Medium	Very Low	High	Medium	High
Throughput	Medium-High (large bandwidth scenarios); Low (high number of groups)	Low	Very Low	Low	Medium-High (large bandwidth scenarios); Low (high number of groups)	Medium-High
Resource Consumption	Very High	Low	Very High	Low	Very High	Low
Satisfaction standard deviation σ_T	Medium	Low	Low	Medium	High	Medium

to 0.32, and a similar trend (up to 0.21) can be outlined for our proposed MSML^{+}. Our MSML^{++} policy achieves a maximum value equal to 0.11 when four groups are served in the cell; then its performance does not vary with an increasing number of served groups.

The results of this simulation campaign demonstrate that the proposed MSML^{++} is well designed also to support several multicast flows in the LTE-A cell. It saves the system resource for other cellular services and efficiently exploits the multi-user diversity without affecting the satisfaction level of members in different multicast groups.

D. Discussion of results

The main results in the performance analysis of the surveyed algorithms are summarized in Table VII, to the purpose of better highlighting the measured relationship between throughput, spectral efficiency and resource consumption and the performance of our most performing scheme, i.e., MSML^{++}.

The figures in the Table underline that CMS suffers in terms of spectral efficiency because it can achieve a good throughput performance (in large bandwidth scenarios) only at the expense of a very high resource consumption. Compared to CMS, the proposed MSML^{++} is able to offer similar throughput results also in scenarios with limited bandwidth. At the same time, it significantly increases the spectral efficiency and reduces the resource consumption.

The MUS policy has poor performance in terms of both spectral efficiency and throughput, while it saves the allocated resources. Our MSML^{++} scheme achieves better throughput and spectral efficiency with respect to MUS and also it consumes a lower amount of resources.

MQS has the worst behavior according to all considered metrics. MSML^{++} outperforms MQS under all the addressed aspects.

The OLM scheme shows interesting results in terms of spectral efficiency and resource consumption while, on the other side, it suffers from poor throughput performance. The MSML^{++} scheme designed in this paper enhances the performance of OLM by guaranteeing higher spectral efficiency and higher throughput, while achieving similar resource consumption performance.

MRA guarantees the highest throughput, on average, but this is attained at the expense of high resource consumption and low fairness among the served groups. The proposed MSML^{++}, which shows throughput values close to MRA, drastically reduces the resource consumption and is able to guarantee adequate inter-group fairness.

Finally, our MSML^{+} shows an interesting behavior, but suffers of several inefficiencies when compared to MSML^{++} in scenarios with huge multicast loads (i.e., high number of multicast flows) and in terms of satisfaction fairness.

By summarizing, it clearly emerges the effectiveness of the proposed MSML^{++} policy with respect to other policies in achieving *(i)* high spectral efficiency, *(ii)* improved video session quality, *(iii)* fairness in terms of "satisfaction" among the multicast destinations, and *(iv)* low resource consumption.

VIII. CONCLUSIONS

In this work we presented the Multicast Subgrouping for multi-layer video applications (MSML) approach, designed to support real-time multicast video services in enhanced LTE and LTE-A networks. The proposed algorithm, designed to cope with different cost functions for enhancement layer group selection, exploits the multi-user diversity by organizing the multicast members into different subgroups according to the channel conditions of involved users. Moreover, MSML takes advantage of the frequency selectivity to achieve high spectral efficiency and a meaningful reduction in terms of resources needed to deliver multicast streams. This latter feature makes MSML able to serve, when coupled with a well targeted cost function tailored to take into account the satisfaction fairness, four video streams with a base layer ranging from 121 and 189 kbps and about 30% of users additionally to get the enhancement layers resulting in a total bitrate from 1.3 to 2 Mbps. Furthermore, by looking at the resource saving guaranteed by our MSML approach, we can conclude that it is also suitable for implementation in real systems, where multicast services coexist with unicast flows.

Future works will address *(i)* the efficient joint resource allocation of multicast and unicast service classes with different QoS constraints and *(ii)* the design of mechanisms to avoid unicast starvation due to the presence of heavy multicast load.

ACKNOWLEDGEMENT

Authors would like to thank Jörg Huschke, Ericsson Eurolab, Germany, for the helpful comments and suggestions that allowed to definitively improve the contribution and the technical strength of the paper.

The research of Massimo Condoluci is supported by European Union, European Social Fund and Calabria Regional Government. This paper reflects the views only of the authors, and the EU, and the Calabria Regional Government cannot be held responsible for any use which may be made of the information contained therein.

References

[1] 3GPP, TS 36.300, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN),” Rel. 11, Sep. 2012.

[2] N. A. Ali, A.-E. M. Taha, and H. S. Hassanein, “Quality of Service in 3GPP R12 LTE-Advanced,” *IEEE Communications Magazine*, vol. 51, no. 8, pp. 103-109, Aug. 2013.

[3] M. Phan, and J. Huschke, “Adaptive point-to-multipoint transmission for Multimedia Broadcast Multicast Services in LTE,” *Computing Research Repository, CORR*, 2009.

[4] 3GPP, TS 23.246, “Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description,” Rel. 11, Mar. 2012.

[5] J. Huschke, J. Sachs, K. Balachandran, and J. Karlsson, “Spectrum requirements for TV broadcast services using cellular transmitters,” *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, pp. 22-31, May 2011.

[6] A. Richard, A. Dadlani, and K. Kim, “Multicast scheduling and resource allocation algorithms for OFDMA-based systems: A survey,” *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 240-254, Feb. 2013.

[7] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, Sep. 2007.

[8] H. Radha, M. Schaar, and Y. Chen, “The MPEG-4 fine-grained scalable video coding method for multimedia streaming over IP,” *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 53-68, Mar. 2001.

[9] 3GPP, TS 36.440, “General aspects and principles for interfaces supporting Multimedia Broadcast Multicast Service (MBMS) within E-UTRAN,” Rel. 11, Sep. 2012.

[10] 3GPP, TS 36.213, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures,” Rel. 11, Sep. 2012.

[11] K. I. Pedersen, T. E. Kolding, F. Frederiksen, I. Z. Kovács, D. Laselva, and P. E. Mogensen, “An Overview of Downlink Radio Resource Management for UTRAN Long-Term Evolution,” *IEEE Communications Magazine*, vol. 47, no. 7, pp. 86-93, Jul. 2009.

[12] L. Zhang, Y. Cai, Z. He, C. Wang, and P. Skov, “Performance evaluation of LTE MBMS baseline,” *5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*, pp. 1-4, Sep. 2009.

[13] S. Lu , Y. Cai , L. Zhang , J. Li , P. Skov , C. Wang , and Z. He, “Channel-Aware Frequency Domain Packet Scheduling for MBMS in LTE,” *IEEE 69th Vehicular Technology Conference (VTC- Spring)*, pp. 1-5, Apr. 2009.

[14] Y. Cai, S. Lu, L. Zhang, C. Wang, P. Skov, Z. He, and Kai Niu, “Reduced Feedback Schemes for LTE MBMS,” *IEEE 69th Vehicular Technology Conference (VTC-Spring)*, pp. 1-5, Apr. 2009.

[15] A. Alexious, C. Bouras, V. Kokkinos, and G. Tsichritzis, “Communication cost analysis of MBSFN in LTE,” *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pp. 1366-1371, Sep. 2010.

[16] K. Bakanoglu, Wu Mingquan, Liu Hang, and M. Saurabh, “Adaptive resource allocation in multicast OFDMA systems,” *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6, Apr. 2010.

[17] 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 Transactions on Wireless Communications*, vol. 9, no. 2, pp. 791-801, Sep. 2009.

[18] A. Alexious, C. Bouras, V. Kokkinos, A. Papazois, and G. Tsichritzis, “Efficient MCS selection for MBSFN transmissions over LTE networks,” *Wireless Days (WD), IFIP*, pp. 1-5, Oct. 2010.

[19] L. Zhang, Z. He, K. Niu, B. Zhang, and P. Skov, “Optimization of coverage and throughput in single-cell E-MBMS,” *IEEE 70th Vehicular Technology Conference Fall (VTC-Fall)*, pp. 1-5, Sep. 2009.

[20] Chih-Wei Huang, Shiang-Ming Huang, Po-Han Wu, Shiang-Jiun Lin, and Jenq-Neng Hwang, “OLM: Opportunistic Layered Multicasting for Scalable IPTV over Mobile WiMAX,” *IEEE Transactions on Mobile Computing*, vol. 11, no. 3, pp. 453-463, Mar. 2012.

[21] P. K. Gopala, and H. E. Gamal, “Opportunistic multicasting,” *Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, pp. 845-849, Nov. 2004.

[22] J. Huschke, “Max-Min Throughput-Optimal Multicast Link Adaptation for Non-Identically Distributed Link Qualities,” *IEEE 72nd Vehicular Technology Conference Fall (VTC-Fall)*, pp. 1-6, Sep. 2010.

[23] Yao Ma, K. Letaief, Zhengdao Wang, R. Murch, and Zhiqiang Wu, “Multiple description coding-based optimal resource allocation for OFDMA multicast service,” *IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 1-5, Dec. 2010.

[24] D. T. Ngo, C. Tellambura, and H. H. Nguyen, “Efficient resource allocation for OFDMA multicast systems with fairness consideration,” *IEEE Radio and Wireless Symposium (RWS)*, pp. 392-395, Jan. 2009.

[25] C. Suh, and J. Mo, “Resource Allocation for Multicast Services in Multicarrier Wireless Communications,”*IEEE 25th International Conference on Computer Communications (INFOCOM)*, pp. 1-12, Apr. 2006.

[26] G. Araniti, V. Scordamaglia, M. Condoluci, A. Molinaro, and A. Iera, “Efficient Frequency Domain Packet Scheduler for Point-to-Multipoint Transmissions in LTE Networks,” *IEEE International Conference on Communications (ICC)*, pp. 4405-4409, Jun. 2012.

[27] S. Deb, S. Jaiswal, and K. Nagaraj, “Real-Time Video Multicast in WiMAX Networks,” *IEEE 27th Conference on Computer Communications (INFOCOM)*, pp. 1579-1587, Apr. 2008.

[28] Wen-Hsing Kuo, Tehuang Liu, and Wanjiun Liao. “Utility-based resource allocation for layer-encoded IPTV multicast in IEEE 802.16 (WiMAX) wireless networks,” *IEEE International Conference on Communications (ICC)*, pp. 1754-1759, Jun. 2007.

[29] Peilong Li, H. Zhang, B. Zhao, and S. Rangarajan, “Scalable video multicast in multi-carrier wireless data system,” *IEEE 17th International Conference on Network Protocols (ICNP)*, pp. 141-150, Oct. 2009.

[30] M. Condoluci, G. Araniti, A. Molinaro, and A. Iera, “Exploiting Frequency-Selectivity in Real-Time Multicast Services over LTE Networks,” *IEEE 24rd International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1799-1803, Sep. 2013.

[31] 3GPP, TR 25.814, “Physical layer aspect for evolved Universal Terrestrial Radio Access (UTRA),” Rel. 7, Oct. 2006.

[32] C. Mehlführer, M. Wrulich, J. Ikuno, B. Colom, D. Bosanska, and M. Rupp, “Simulating the long term evolution physical layer,” *17th European Signal Processing Conference (EUSIPCO)*, pp. 1471-1478, Aug. 2009.

[33] R. Giuliano, and F. Mazzenga, “Exponential Effective SINR Approximations for OFDM/OFDMA-Based Cellular System Planning,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4434-4439, Sep. 2009.

[34] A. Urie, A. Rudrapatna, C. Raman, and J-M. Hanriot, “Evolved Multimedia Broadcast Multicast Service in LTE: An Assessment of System Performance Under Realistic Radio Network Engineering Conditions,” *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 57-76, Sep. 2013.

[35] S. Sharangi, R. Krishnamurti, and M. Hefeeda, “Energy-Efficient Multicasting of Scalable Video Streams Over WiMAX Networks,” *IEEE Transactions on Multimedia*, vol. 13, no. 1, pp. 102-115, Feb. 2011.

Massimo Condoluci received the B.Sc. and the M.Sc. degrees in telecommunications engineering in 2008 and 2011, respectively, from the University Mediterranea of Reggio Calabria, Italy, where he is currently pursuing the Ph.D. degree in information technology. In 2014, he has been a visiting Ph.D. student with the Centre for Telecommunications Research, King’s College London, UK. His current research interests include machine-type communications, radio resource management, multicasting and device-to-device over 4G/5G cellular networks.

Giuseppe Araniti is an Assistant Professor of Telecommunications at the University Mediterranea of Reggio Calabria, Italy. From the same University he received the Laurea (2000) and the Ph.D. degree (2004) in Electronic Engineering. His major area of research includes personal communications systems, enhanced wireless and satellite systems, traffic and radio resource management, multicast and broadcast services, device-to-device and machine-type communications over 4G/5G cellular networks.

Antonella Molinaro is an Associate Professor of Telecommunications with the University Mediterranea of Reggio Calabria, Italy. She was an Assistant Professor with the University of Messina, Italy, from 1998 to 2001, with the University of Calabria, Cosenza, Italy, from 2001 to 2004, and with the Polytechnic of Milan, Italy, from 1997 to 1998, with a research contract. She was with Telesoft, Rome, from 1992 to 1993, and at Siemens, Munich (DE), from 1994 to 1995, with an European fellowship contract. Her current research interests include vehicular networking and future Internet architectures.

Antonio Iera graduated in Computer Engineering at the University of Calabria, Italy, in 1991 and received a Master Diploma in Information Technology from CEFRIEL/Politecnico di Milano, Italy, in 1992 and a Ph.D. degree from the University of Calabria in 1996. Since 1997 he has been with the University of Reggio Calabria and currently holds the position of full professor of Telecommunications and Director of the Laboratory for Advanced Research into Telecommunication Systems (www.arts.unirc.it). IEEE Senior Member since 2007. His research interests include, next generation mobile and wireless systems, RFID systems, and Internet of Things.