Dynamic Resource Allocation for Scalable Video Multirate Multicast over Wireless Networks

Shuangwu Chen, Bowen Yang, Jian Yang, *Senior Member, IEEE*, and Hanzo Lajos, *Fellow, IEEE*

***Abstract*—Aided by scalable video coding, multirate multicast has become a promising technique of providing differentiated quality of experience (QoE) for massive numbers of video subscribers operating in heterogeneous channel conditions. Nevertheless, due to the time-varying nature of wireless channels and the subscribers' diverse requirements, it is challenging to dynamically control the video rate in the light of the available radio resource to achieve the best QoE. To elaborate a little further, the time scale of resource scheduling is of short-term nature, which determines the short-term video quality variation, but from a service provider's perspective the design objective is to optimize the long-term QoE for all subscribers. Despite its importance, this problem has not been considered before. Explicitly, we formulated this problem as a time-averaged stochastic optimization problem which avoids the impact of both the short-term channel quality fluctuation and that of the video bitrates, whilst maintaining both inter- and intra- group fairness. The stratified structure of the problem inspires us to decompose it into a two-phase optimization: coarse grained assignment for each user group and fine grained assignment for each subgroup. We propose an adaptive multicast algorithm based on Lyapunov's optimization theory for solving this problem, by striking a compelling trade-off between the system's utility and its queue stability. We quantify the achievable performance of our proposed solution based on realistic video traces.**

***Index Terms*—multirate multicast, scalable video coding, resource assignment, wireless network.**

I. Introduction

THE overwhelming growth of demands for mobile multimedia services imposes enormous challenges on the provision of reliable quality of experience (QoE), when supporting a large number of users competing for the scarce radio resources in cellular networks [1]. This situation is aggravated by the demanding specifications of the emerging vehicle-to-everything (V2X) communications [2], imposed by resource-hungry applications [3] such as ultra-high-definition (UHD) video, augmented/virtual reality, 3D/multi-view television, which will put a heavy tele-traffic burden on today's already congested cellular networks. Video data is poised to dominate the mobile network traffic, which will account for 79 percent of the global mobile data by 2022 [4]. Furthermore, most of the tele-traffic are generated by group-oriented services and applications. In this context, multicast has been identified as a promising technique of addressing this challenge both in the 4G Long Term Evolution (LTE) and LTE-Advance (LTE-A) systems as well as in next generation systems [5]. By exploiting the properties of point-to-multipoint (PTMP) communication [6], multicast delivers the same multimedia content to an unlimited number of users over the same frequency band. The significantly improved channel capacity and energy efficiency will boost the potential of multicast in V2X scenarios, where the vehicles can receive multicast information from other vehicles, Internet and road-side infrastructure. Accordingly, the evolved multimedia broadcast multicast services (eMBMS) will still be adopted [7] in 5G networks for PTMP services.

However, high-quality multimedia multicast services suffer from a lot of problems. Due to the stochastic nature of user movement and channel fading, the channel quality as well as the transmission rate is time-varying. Different subscribers of a multicast group may experience different channel qualities caused by their geographically separated locations. Additionally, the source video rate may fluctuate strongly when the video scene changes or the motion activity occurs. It is challenging to dynamically control the video rate in the light of the available radio resource to achieve the best QoE for multicast services, while considering the fairness. Conventional multicast is usually single-rate, where the transmission rate is determined by the subscribers having the lowest channel quality, typically located at the edge of a cell. The single-rate multicast has a low complexity, but at the cost of eroding the video quality of the cell-centre subscribers having good channel conditions. As a result, the high potential of multimedia multicast is only partially exploited.

Aided by scalable video coding (SVC) [8], multirate multicast become a promising technique of supporting flexible video delivery for heterogeneous devices and time-varying channel conditions. An SVC stream comprising a base layer (BL) and several enhancement layers (ELs) provides a judiciously differentiated video quality by appropriately scaling the spatial resolution, the temporal frame rate and the reconstructed video definition. Multirate multicast allows different SVC layers to be delivered via different system resources. As a result, the multimedia stream is split into multiple substreams, each of which contains an SVC layer and serves a subgroup of users having similar channel conditions. The users in a multicast group can access the same content with differentiated QoE by subscribing to appropriate SVC layers based on their respective channel conditions. Specifically, the base layer has to be received by all subscribers in order to ensure a minimal

S. Chen, B. Yang and J. Yang are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, Anhui, China. E-mail: chensw@ustc.edu.cn, jianyang@ustc.edu.cn

L. Hanzo is with the Department of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK. E-mail:lh@ecs.soton.ac.uk.

L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/N004558/1, EP/P034284/1, EP/P034284/1, EP/P003990/1 (COALESCE), of the Royal Society's Global Challenges Research Fund Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom.

QoE. Afterwards, the subscribers having higher channel gains are capable of decoding more SVC layers and acquiring a higher QoE.

Nevertheless, combining SVC with multicast poses additional challenges. Explicitly, owing to the shared nature of the multicast channel, it is difficult to optimally control the video rate and radio resource for achieving the best QoE for group members having heterogeneous channel conditions, which is in stark contrast to the unicast transmissions [9] [10]. Furthermore, the co-existence of several multicast subgroup/group increases the difficulty of system design, since the rate balance among both the intra- and inter- group users must be considered. Giving preference to large groups with more subscribers would improve the system's overall utility, but at the expense of undermining the QoE perceived by small groups. The potential dynamics of group size is determined by the unknown time-variant requirements of cellular users. This situation makes it more difficult to find the optimal multicast scheduling and resource allocation (MSRA) solution in real time, which has been proved to be an NP-hard problem [11]. In fact, the complexity of MSRA increases exponentially with the number of users, multicast groups and SVC layers.

To overcome these issues, most MSRA schemes proposed rely on heuristic algorithms [12]–[14]. They tend to have low-complexity implementation, but provide sub-optimal solutions. As a result, they are either too conservative for attaining optimal video quality or too aggressive for guaranteeing fairness among different subscribers. Therefore, the model-based optimization methods are introduced to address this problem. All the legitimate solutions [15]–[18] are determined with the objective of finding the optimal resource assignment. The existing schemes aim for optimizing an instantaneous utility in each scheduling slot, such as the throughput, spectral efficiency or video quality, based on the near-instantaneous channel state information (CSI) and the individual requirements of the subscribers. Furthermore, most of them are based on the idealistic assumptions of a fixed channel model or an over-simplified subscriber distribution model. The time scale of near-instantaneous channel-quality-driven resource scheduling is of short-term, which also directly controls the video quality variation. However, from a service provider's perspective, the long-term video quality ought to be optimized. This problem has not yet been considered in the literature. Motivated by this open problem, we maximize the time-averaged QoE for all subscribers, while taking the dynamics of video bitrates and channel conditions into account. The new contributions of this paper are summarized as follows:

- We formulate the problem of adaptive resource assignment for scalable video multicast as a stochastic optimization problem, which maximizes the long-term QoE of all subscribers weighted by the group size and constrained by the time-averaged transmission rate. This formulation avoids the influence of the short-term fluctuation of both the channel conditions and the video bitrates, while maintaining proportional fairness among the groups by ensuring the long-term stability of the transmission queue for each SVC layer.
- Inspired by the layered structure of the initial time-averaged problem, we decompose it into a twin-phase optimization: coarse-grained assignment for each group and fine-grained assignment for each subgroup. Based on the classic Lyapunov's optimization theory, we develop an online algorithm for solving the problem, which strikes a trade-off between the system's utility and the queue stability. An efficient algorithm is proposed for finding the optimum strategy, which mitigates the computational burden imposed. Furthermore, we derive an analytical performance bound for the proposed solution.
- We quantify the achievable performance of our proposed scheme based on realistic video traces. The experimental results demonstrate that the proposed scheme achieves a higher long-term QoE than the benchmarks, while maintaining the most appropriate queue length, even in the face of dynamically fluctuating channel quality and non-uniformly distributed subscribers.

The remainder of this paper is organized as follows. Section II summarizes the related contributions on multicast. After presenting the models of wireless channel, video multicast and system utility, Section III formulates a time-averaged constrained optimization problem. In Section IV, we propose a layered solution derived from Lyapunov optimization theory for overcoming the aforementioned problem. Finally, we illustrate the performance of our proposed scheme in Section V, whilst a conclusive discussion is offered in Section VI.

II. Related Work

Multicast is an efficient technique of providing group-oriented services by feeding all the subscribers by a single transmission. Extensive studies have been conducted to address the adaptive resource allocation problem of video multicast over wireless networks [19]–[23]. The conventional multicast scheme [19] conservatively selected a common modulation and coding scheme (MCS) for each group based on the lowest channel quality indicator (CQI) of all subscribers. Although this scheme is easy to implement, it is obviously inefficient, because all cell-centre subscribers having good channel qualities have to endure a bad video quality. To overcome this drawback, opportunistic multicast schemes [20]–[22] were developed for providing more aggressive MCS selection. Specifically, a threshold-based solution was proposed in [20] that dynamically selected a subset of subscribers to be served within each scheduling slot. The authors of [21] have made a step forward by striking a trade-off between error resilience and transmission rate. In [22], rate selection was optimized by maximizing the minimum throughput in a multicast group, even when the channel qualities were non-identically distributed. In general, maximizing the spectral efficiency by relying on opportunistic strategies fails to provide any fairness guarantee among subscribers [23], since some of the subscribers experiencing low channel quality may be denied service. This is unacceptable for delay-constrained flawless lip-synchronized video services.

Due to the inefficiency of single-rate multicast, substantial research efforts have been invested in designing multirate multicast. By delivering the same video content at different

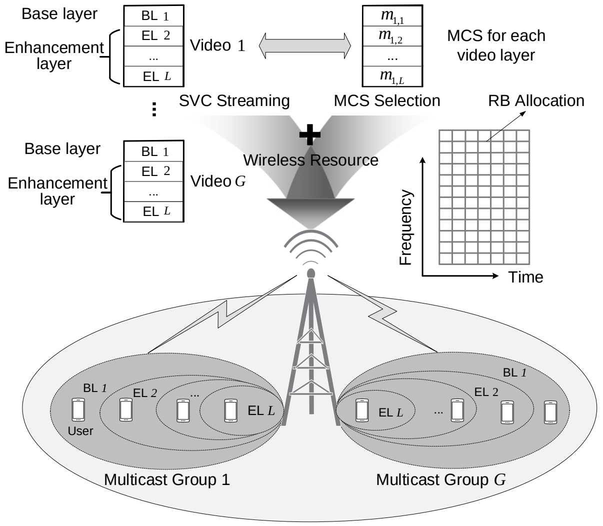


Fig. 1. SVC Multicast system

coding rates, multirate multicast becomes capable of serving subscribers having diverse channel conditions at different video qualities [24]. In [12], the opportunistic schemes were improved for supporting multirate multicast by exploiting multiple description coding (MDC) [25]. The original video was coded into multiple representations having various rates so that the subscribers could adaptively select the most appropriate representation according to their near-instantaneous channel conditions. This work was further discussed in [13] where the radio resources were assigned for maximizing the system throughput, while taking the fairness between groups into account. Since the different video representations are independent, redundant video delivery may substantially reduce the spectral efficiency. Fortunately, SVC circumvents this in video multicast by using the efficient layer-dependent coding structure. Based on SVC a conservative multicasting scheme (CMS) was proposed in [14]. Explicitly, the CMS first split the group members into multiple subgroups experiencing similar channel conditions and then adopted a greedy algorithm for allocating the resource among the subgroups for maintaining proportional fairness. The opportunistic layered multicast scheme (OLM) proposed in [13] minimized resource usage for basic video layer delivery, while maximizing the utility for the optional enhancement video layers delivery. In [15], the multicast subgrouping for multilayer video applications (MSML) was proposed, which adopted aggressive MCS assignments for improving the spectral efficiency. By maximizing a heuristic cost function formulated in terms of spectral efficiency, the MSML guaranteed the basic video quality for all subscribers while choosing the best subset of subscribers to receive the ELs. These heuristic solutions are simple to use in practice, but they tend to be suboptimal, which motivates the employment of formal optimization models. Using convex and dynamic programming, the authors of [16] struck a trade-off between fairness and spectral efficiency. In [17], the resource allocation of SVC multicast services was formulated as an optimization problem under resource constraints, which was also proved to be NP-hard. Then a two-step dynamic programming solution was proposed, but its excessive complexity hindered its actual deployment. To bridge the gap between the theoretical optimality and implementation concerns, the authors of [18] provided a closed-form solution for finding the optimal resource allocation based on a convex optimization model. This model relied on the idealistic simplifying assumption that the subscribers were uniformly distributed in the cell area and the user's perceived signal-to-noise (SNR) was only related to the distance from the base station (BS). However, the aforementioned contributions tend to focus on the short-term system utility without giving cognizance to the time-varying fluctuation of video frame size and the available radio resources.

TABLE I
NOTATION USED IN THE PAPER

l	index of subgroup and/or number of video layers
g	index of multicast group or index of video contents
$\mathcal{S}$	set of multicast subscribers
$\mathcal{S}_{g,l} \subset \mathcal{S}$	set of subscribers in the lth subgroup of group g
k	index of subscribers
u_k	utility of the subscriber k
r_k	reception rate of the subscriber k
$a_{g,l}$	video arrival rate of subgroup $\mathcal{S}_{g,l}$
$b_{g,l}$	channel transmission rate of subgroup $\mathcal{S}_{g,l}$
$m_{g,l}$	index of MCS for subgroup $\mathcal{S}_{g,l}$
$n_{g,l}$	number of RBs for subgroup $\mathcal{S}_{g,l}$
N	available RB for multicast service
$\bar{U}_{g,l}$	time-averaged utility for subgroup $\mathcal{S}_{g,l}$
$R_{g,l}$	the average video rate for the g-th video with l layers

III. System Model

Against this backdrop, we consider the OFDM-based single-cell PTMP scenario of Fig. 1, where the BS provides multimedia services to multiple user groups via multicast. The multimedia contents are encoded into multiple layers by SVC, providing an ever-improving video quality upon increasing the number of video layers. Thereby, an SVC layer is treated as a separate substream that may serve a subgroup of users experiencing similar channel quality. Given the time-varying and error-prone nature of wireless channels, the BS is expected to dynamically assign the physical resource blocks (RBs) and the MCSs for each SVC layer based on the CSI feedback, in order to achieve the best possible QoE. We note that no retransmission can be used for multicast services, since a single user's channel condition cannot represent that of all users sharing the multicast channel. In fact, the multicast channel uses the unacknowledged mode of Radio Link Control (RLC) to deliver the eMBMS payload [26]. Therefore, we can focus our attention on the problem of finding an efficient RB allocation and MCS assignment for the BS. Let us now discuss our mathematical models and problem formulation.

A. Wireless Channel Model

Let $\mathcal{S}$ represent the set of multicast service subscribers. The users subscribing to the same video content compose a multicast group, which is indexed by g $(1 \leqslant g \leqslant G)$. Accordingly, all the subscribers are split into G groups. Let $\mathcal{S}_g$

denote the subscribers in the group g, which share the same spectrum resources of the BS. For simplicity, assume that each SVC video content consists of L layers, including a BL and $L-1$ ELs. The subscribers that are able to decode the l-th $(1 \leqslant l \leqslant L)$ video layer in the group $\mathcal{S}_g$ compose a subgroup $\mathcal{S}_{g,l}$. Due to the stochastic nature of both the user requests and the wireless channel, the number of subscribers in each subgroup, i.e. $|\mathcal{S}_{g,l}|$, would dynamically change over time.

The BS operates in a time-slotted manner. The size of time slot is dependent on the specific communication system. For instance, in the context of LTE networks, a wireless time slot has a duration of 0.5 milliseconds (ms), which is exactly the time interval of 6 or 7 OFDM symbols, depending on whether an extended cyclic prefix is used or not. The minimal radio resource unit of an OFDM system is a single RB, which consists of 12 consecutive sub-carriers. Due to the fact that the two transmission modes of multicast and unicast share the physical (PHY) layer resources, LTE/LTE-A employs time division multiplexing (TDM) to divide the radio resources between multicast and unicast services and the eMBMS is restricted to occupy no more than 60% time [27]. We assume that the radio resource is scheduled at an interval of a radio frame which is indexed by t. Let $N(t)$ denote the total number of RBs available for multicast services during the slot t.

Each subgroup is assigned for an independent channel. The transmission rate over a certain channel is determined by the MCS. Let $m \in \{0, 1, \ldots, M\}$ denote the index of MCS, where $m = 0$ represents no data transmission. For instance, there are 16 MCSs in the state-of-the-art LTE system. Naturally, a greater MCS exhibits a higher transmission rate as well as a higher block error rate (BLER) [28]. If the BLER is too high, some subscribers having bad channel quality cannot receive the video layer successfully. Actually, each user evaluates the channel SNR and feeds it back to the BS through the CQI. A mapping between the channel quality and the MCS is defined by 3GPP [29] for attaining an acceptable BLER, which can be expressed as:

$$\tilde{m} = f(q), \quad f : CQI \to MCS, \tag{1}$$

where q is the user's perceived CQI and $\tilde{m}$ is the corresponding MCS index. A higher CQI q corresponds to a higher MCS $\tilde{m}$. Given a modulation scheme m, the coding rate of a modulated symbol is denoted by $c(m)$. Then, the data rate carried by a single RB under the MCS m is represented by:

$$d(m) = N_{\text{subcarriers}} \times N_{\text{symbols}} \times c(m). \tag{2}$$

where $N_{\text{subcarriers}}$ is the number of subcarriers per RB and N_{symbols} is the number of symbols over a subcarrier per slot. The essential problem is to assign the optimal MCS for each subgroup, while considering the heterogeneous channel quality.

Assume the MCS and RB assigned for the subgroup $\mathcal{S}_{g,l}$ in the time slot t are denoted by $m_{g,l}(t)$ and $n_{g,l}(t)$, respectively. Naturally, the total number of RBs assigned for the subscribers is restricted by the radio resource available for multicast services, which is formulated as:

$$\sum_{g=1}^{G} \sum_{l=1}^{L} n_{g,l}(t) \leqslant N. \tag{3}$$

Here, we assume that the channel is block fading, where the channel gain remains constant during each time slot, but potentially changes from one slot to another. Based on the users' channel CQIs, the BS has to assign an appropriate number of RBs and a suitable MCS for each subgroup at the aim of optimizing users' QoE and ensuring fairness.

B. Video Multicast Model

Due to the interdependent structure of the SVC, a higher layer is only decodable if all the lower layers are available. In other words, before delivering the l-th video layer to a specific subscriber, we have to ensure that the layers $\{0, \ldots, l-1\}$ are successfully received. To achieve this, the lower SVC layer ought to have a lower MCS so that the subscribers having better channel conditions are able to decode higher layers and thus obtain a better video quality. Therefore, the MCSs should meet the following constraint:

$$m_{g,1}(t) \leqslant m_{g,2}(t) \leqslant \cdots \leqslant m_{g,L}(t). \tag{4}$$

Notably, the MCS of the BL (i.e., $m_{g,1}$) depends on the specific group members having the worst channel quality, so that as a basic video service at least, the BL is adequately demodulated by all the subscribers in the group.

For a certain subscriber k $(k \in \mathcal{S}_g)$ in the group $\mathcal{S}_g$, let $q_k(t)$ denote its CQI in the slot t. According to (1), the expected MCS is $\tilde{m}_k(t) = f[q_k(t)]$. The subscriber k in the group $\mathcal{S}_g$ can only demodulate the l-th layer if $\tilde{m}_k \geq m_{g,l}$. Here, we define an indicator function as:

$$I(\tilde{m}_k, m_{g,l}) = \begin{cases} 1, & if \quad m_{g,l} \leqslant \tilde{m}_k \\ 0, & otherwise. \end{cases} \tag{5}$$

Then, the number of video layers that user k could successfully demodulate is calculated as $l_k = \sum_{l=1}^{L} I(\tilde{m}_k, m_{g,l})$. If $l_k = l$, we have $k \in \mathcal{S}_{g,l}$ and vice versa. The actual reception rate r_k of user k is:

$$r_k(t) = \sum_{l=1}^{L} I[\tilde{m}_k(t), m_{g,l}(t)] d[m_{g,l}(t)] n_{g,l}(t). \tag{6}$$

Naturally, the reception rate is determined by the channel quality, the RB allocation and the MCS assignment.

C. System Utility Model

Traditional metrics [30], such as the throughput, delay, outage or BLER, are not adequate for characterizing the actual QoE, when human perception is involved. Most of the mean opinion score (MOS) [31] tests demonstrate that the perceived QoE of scalable video service saturates at higher video rates. Here, we adopt the widely used logarithmic QoE model of [32] to assess the users' utility. Let $u_k(t)$ denote the utility of user k in the time slot t, which is defined as a function of the received video rate [33],

$$u_k(t) = a \log b \frac{r_k(t)}{R_g}, \tag{7}$$

where r_k is the reception rate and R_g is the average video rate associated with the highest video layer; a and b are the coefficients to normalize $u_k(t)$ to stay in the range of 0 to 1, which could be empirically defined in practice.

In contrast to the unicast service, it is essential for the proposed scalable video multicast service to provide the best QoE for all the subscribers. The utility for an individual subgroup $\mathcal{S}_{g,l}$ is denoted by $U_{g,l}(t) = \sum_{k \in \mathcal{S}_{g,l}} u_k(t)$. Similarly, the utility for the group $\mathcal{S}_g$ is $U_g(t) = \sum_{l=1}^{L} u_{g,l}(t)$, where L is the number of subgroups. Furthermore, the overall utility $U(t)$ for all multicast subscribers can be defined as

$$U(t) = \frac{1}{|S|} \sum_{g=1}^{G} U_g(t) = \frac{1}{|S|} \sum_{g=1}^{G} \sum_{l=1}^{L} U_{g,l}(t), \tag{8}$$

where $|S|$ is the total numbers of multicast subscribers. Since $|S|$ would change over time, $\frac{1}{|S|}$ is used for normalizing the system utility. Notably, this definition ensures proportional fairness among different groups, because it prefers to guarantee the QoE of groups containing more subscribers.

Additionally, the instantaneous system utility $U(t)$ cannot reflect the impact of channel dynamics on the long-term video quality. Hence, we define a time-averaged system utility $\overline{U}$ as:

$$\overline{U} \triangleq \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[U(\tau)]. \tag{9}$$

Notably, according to (8), $\overline{U}$ can also be represented as

$$\overline{U} = \sum_{g=1}^{G} \frac{1}{|S|} \overline{U}_g = \sum_{g=1}^{G} \frac{1}{|S|} \sum_{l=1}^{L} \overline{U}_{g,l}, \tag{10}$$

where $\overline{U}_g$ and $\overline{U}_{g,l}$ denote the time-averaged $U_g(t)$ and $U_{g,l}(t)$, respectively.

D. Problem Formulation

In practice, the BS would cache the video received from remote servers before scheduling it for transmission over the wireless channel, whilst protecting it against the dynamics of video rate and channel conditions. If the video packet arrival rate exceeds the available transmission rate, the resultant cache overflow would cause packet loss. Hence, the channel transmission rate should be appropriately matched to the video arrival rate at a long time scale. In our system, each video layer can be treated as a separate substream serving a subgroup of users. Let $a_{g,l}(t)$ represent the arrival rate of the l-th layer in the group g. Indeed, affected by the fluctuations of video bitrate, $a_{g,l}(t)$ would change over time. The time-averaged video arrival rate is $\overline{a}_{g,l} = \lim\limits_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[a_{g,l}(\tau)]$. Given the MCS $m_{g,l}(t)$ and the RB $n_{g,l}(t)$, the transmission bitrate of the l-th layer in the group g is denoted by $b_{g,l}(t) = d[m_{g,l}(t)] n_{g,l}(t)$ and the time-averaged transmission rate is denoted by $\overline{b}_{g,l} = \lim\limits_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}[b_{g,l}(\tau)]$. Regarding each video layer substream, the following constraint should be satisfied:

$$\overline{a}_{g,l} \leqslant \overline{b}_{g,l}. \tag{11}$$

This constraint imposes restrictions on the average transmission rate for each subgroup. Although the constraint does not explicitly impose strict restrictions on the average queuing delay in the BS, they ensure the long-term stability of the transmission queues for each subgroup. That is, the average queuing delay is bounded. Hence, applying this constraint implicitly offers delay guarantee as a QoE metric.

By contrast, if the transmission rate stays above the video rate, the resultant buffer underflow would cause video jerking. Actually, the transmission rate of each subgroup is determined by the assigned MCS and RB. For a multicast channel relying on a limited number of RBs, an excessively high transmission rate means an excessively high MCS order. As a result, some subscribers having poor channel conditions cannot demodulate the OFDM signal, which would exclude a part of subscribers from successfully receiving the video stream. However, our objective is to maximize the system's utility, which is defined as a function of the video rate received by all subscribers. Naturally, this situation would lead to the reduction of system's utility, which is contrary to our objective. In order to obtain an improved utility, a lower-than-affordable MCS order is preferred for avoiding buffer underflow in case of high near-instantaneous channel qualities, because an excessively high transmission rate would deplete the buffer. Therefore, maximizing the system utility can mitigate the problem of buffer underflow.

Our objective is to maximize the time-averaged system utility (9) via appropriate MCS selection and RB assignment, while satisfying the RB constraint (3), the MCS constraint (4) and the transmission rate constraint (11). The problem can be formulated as:

$$\begin{aligned} \max_{m_{g,l}, n_{g,l}} \quad & \overline{U} \\ \text{Subject to} \quad & (3), (4), (11). \end{aligned} \tag{12}$$

Generally, addressing such a time-averaged constrained optimization problem hinges on the dynamics of users' requests and channel quality, but there is no priori knowledge about them. To overcome this difficulty, we will develop an online algorithm based on Lyapunov optimization theory.

IV. A Layered Solution for MCS and RB Assignment in Multirate Multicast

The problem (12) seeks to maximize the time-averaged utility subject to an additional time-averaged constraint. This motivates us to employ the Lyapunov optimization framework derived from [34] for addressing this problem. Meanwhile, finding the optimal RB and MCS assignment for each subgroup directly would have an excessive decision space, including $\sum_{g=1}^{G} L$ subgroups and $N \times M$ possible assignment choices for each group. Hence, it is unfeasible to make online decisions when both the user distributions and the channel conditions change frequently over time. To address this issue, we decompose the initial problem into a pair of sub-problems.

Then, we propose a layered solution for group and subgroup resource assignment.

A. Problem decomposition

According to (10), the system utility is the sum of group utility, while the group utility is the sum of the subgroup utility. The layered structure enables us to decompose the initial problem (12) into a two-phase optimization: coarse-grained assignment for each group and fine-grained assignment for each subgroup.

In the first phase, the subscribers belonging to the same group are considered as a whole, and the channel condition of the whole group can be characterized by mean or median CQI of the group members, which is denoted by $\overline{q}_g$. Actually, the BS can readily acquire the channel CQI q_k for any subscriber k ($k \in \mathcal{S}_g$). The corresponding MCS is $\overline{m}_g = f(\overline{q}_g)$. The only problem is how to assign the appropriate RBs for each group. Then, the initial problem is reformulated as:

$$\begin{aligned} &\max_{n_g} && \overline{U} = \frac{1}{|S|} \sum\nolimits_{g=1}^{G} \overline{U}_g \\ &\text{Subject to} && \overline{a}_g \leqslant \overline{b}_g \quad and \quad \sum\nolimits_{g=1}^{G} n_g \leqslant N, \end{aligned} \tag{13}$$

where n_g is the allocated RBs, $\overline{a}_g$ is the time-averaged video rate and $\overline{b}_g$ is the time-averaged channel transmission rate for the g-th group. Similar to (3) and (11), each group has to meet both the RBs constraint and the transmission rate constraint.

In the second phase, given the available RBs for each group, the problem is how to find the optimal MCS and RB assignment for each subgroup, which is represented by

$$\begin{aligned} &\max_{n_{g,l}, m_{g,l}} && \overline{U}_g = \sum\nolimits_{l=1}^{L} \overline{U}_{g,l} \\ &\text{Subject to} && (4), (11), and \sum\nolimits_{l=1}^{L} n_{g,l} \leqslant n_g^*, \end{aligned} \tag{14}$$

where n_g^* represents the optimally assigned RBs for the g-th group in the first phase. Additionally, each subgroup has to follow the MCS constraint in (4).

B. Coarse-grained assignment for each group

To tackle the problem (13), we model the video transmission process by a queue. Then, the time-averaged transmission rate constraint could be converted into the queue stability constraint through the Lyapunov method. In the time slot t, the dynamics of the transmission queue for the g-th group are described as:

$$H_g(t+1) = \left[H_g(t) + \tilde{a}_g(t) - \tilde{b}_g(t) \right]^+. \tag{15}$$

Given the different rates of the different video contents, we use the average video rate R_g of the group g to normalize the video arrival rate $a_g(t)$ and the transmission rate $b_g(t)$, namely $\tilde{a}_g(t) = \frac{a_g(t)}{R_g}$, $\tilde{b}_g(t) = \frac{b_g(t)}{R_g}$. Here, we define $(x)^+ \triangleq \max(x, 0)$. Indeed, $a_g(t)$ may be measured online, whereas $b_g(t)$ depends on how many resources are assigned for the g-th group. Since each group is treated as a whole using a unified MCS $\overline{m}_g$, the corresponding transmission rate is estimated by $b_g(t) = d(\overline{m}_g) n_g$. All the subscribers in a group are supposed to have the same reception rate, which is equal to $b_g(t)$. Hence, according to (7), the group utility is estimated by

$$\hat{U}_g(t) = |\mathcal{S}_g| a \log b \frac{b_g(t)}{R_g} \triangleq |\mathcal{S}_g| a \log b \frac{d(\overline{m}_g) n_g}{R_g}, \tag{16}$$

where $|\mathcal{S}_g|$ is the group size. Observe that $\hat{U}_g(t)$ is only related to the RB assignment $n_g(t)$.

From the theorem in [34], the time-averaged constraint in (13) is only satisfied if the queue $H_g(t)$ is mean-rate-stable, namely $\lim_{t \to \infty} \frac{\mathbb{E}[H(t)]}{t} = 0$. Here, we define a quadratic Lyapunov function as

$$L(t) = \frac{1}{2} \sum\nolimits_{g=1}^{G} H_g^2(t). \tag{17}$$

Let $H(t) \triangleq \{H_g(t), 1 \leqslant g \leqslant G\}$ denote a vector concatenating $H_g(t)$. The conditional Lyapunov drift is then defined as the expected variation of the Lyapunov function:

$$\Delta(t) \triangleq \mathbb{E}[L(t+1) - L(t) | H(t)]. \tag{18}$$

According to the classic theory of Lyapunov drift, minimizing the Lyapunov drift $\Delta(t)$ achieves the stability of the queue $H_g(t)$, and thus asymptotically meeting the time-averaged transmission rate constraint. Here, we define a "drift-plus-penalty" function by combining the Lyapunov drift and the instantaneous system utility $U(t)$, which is as follows:

$$\Delta(t) - V\mathbb{E}[U(t) | H(t)], \tag{19}$$

where the parameter V is a non-negative value invoked for striking a trade-off between the system utility and the queue stability in our control strategy. A large V is beneficial for improving the system utility, while a small V means focusing more on the stability of the queue. In Section V we have conducted an experiment to analyze the performance sensitivity to the trade-off factor V. Minimizing the "draft-plus-penalty" of (19) implies maximizing the system utility and minimizing the Lyapunov drift. This is consistent with our objective of maximizing the system utility while complying with the time-averaged constraint. Hence, the optimization problem (13) is further reformulated as making a decision on the RB assignment for each group in order to minimize the "draft-plus-penalty" of (19).

Squaring the two sides of the queue dynamics (15) yields

$$H_g^2(t+1) \leqslant H_g^2(t) + a_g^2(t) + b_g^2(t) + 2H_g(t)[a_g(t) - b_g(t)]. \tag{20}$$

Since $d_g \leqslant d(M)$ and $\sum_{g=1}^{G} n_g(t) \leqslant N$, we have

$$\sum\nolimits_{g=1}^{G} b_g^2(t) \leqslant d^2(M) \sum\nolimits_{g=1}^{G} n_g(t)^2 \leqslant N^2 d^2(M). \tag{21}$$

Then, after substituting (18), (20) and (21) into (19), we can derive a "loose" upper bound of the drift-plus-penalty as follows

$$\begin{aligned} &\Delta(t) - V\mathbb{E}[U(t) | H(t)] \\ &\leqslant C_0 + C_1 - \sum\nolimits_{g=1}^{G} [H_g(t) b_g(t) + \frac{V}{|S|} U_g(t)], \end{aligned} \tag{22}$$

where $C_0 = \frac{1}{2} \sum_{g=1}^{G} a_g^2(t) + \mathbb{E}[\sum_{g=1}^{G} H_g(t) a_g(t) | H(t)]$ and $C_1 = \frac{1}{2} N^2 d^2(M)$. Notably, C_1 is a constant. At the beginning of slot t, it is viable for the BS to measure the current queue state $H_g(t)$ as well as the video arrival rate $a_g(t)$, and thus C_0 is also a constant. As mentioned, $U_g(t)$ can be estimated by $\hat{U}_g(t)$. Hence, the second item at the right side of (22) only depends on the decision variables n_g.

According to the Lyapunov optimization framework [34], instead of directly minimizing the "drift-plus-penalty" (19), we can solve the time-averaged optimization problem by minimizing the "loose" bound (22), which is as follows:

$$\begin{aligned} &\max_{n_g} && \sum\nolimits_{g=1}^{G}[H_g(t)b_g(t) + \frac{V}{|\mathcal{S}|}U_g(t)] \\ &\text{Subject to} && \sum\nolimits_{g=1}^{G} n_g \leqslant N. \end{aligned} \tag{23}$$

In (23), $H_g(t)b_g(t)$ is interpreted as the queue-length-weighted transmission rate. A higher $H_g(t)$ implies a long backlog of video data in the queue of group g. In this situation, it prefers to improve the transmission rate by assigning more RBs for this group, in order to maintain queue stability. Similarly, a higher mean CQI $\bar{m}_g(t)$ implies a better channel condition for the group g, hence more RBs will be allocated to the group g for maximizing the objective function in (23). Consequently, the subscribers in the group g perceive a better QoE. At every slot, by substituting all possible n_g values into (16) and (23), we can find the optimal RB assignment n_g^* for each group.

Due to the fact that we seek to minimize the "loose" bound (22), the solution of (23) is a sub-optimal solution of the original problem (13). Thus, we present **Lemma 1** for theoretically quantifying the performance bound of our sub-optimal solution.

Lemma 1: Assume that the queue length is initially zero. For any position value V, the average system utility $\overline{U^*}$ obtained by solving the problem (23) satisfies

$$\overline{U^*} \geqslant \overline{U^{opt}} - \frac{C_0 + C_1}{V}, \tag{24}$$

where $\overline{U^{opt}}$ is the optimal utility of the problem (13).

Proof: Please see Appendix A. ■

Lemma 1 implies that the suboptimal utility $\overline{U^*}$ asymptotically approaches the optimal value $\overline{U^{opt}}$ as V increases.

C. Fine-grained assignment for each subgroup

After assigning appropriate RBs for each multicast group in the first phase, the remaining problem is to find the optimal fine grained assignment of RBs and MCS for each subgroup, as seen in (14). The transmission process for each subgroup is also modeled by a queue. The dynamics of the queue for the subgroup $\mathcal{S}_{g,l}$ are characterized by

$$Q_{g,l}(t+1) = \left[Q_{g,l}(t) + \tilde{a}_{g,l}(t) - \tilde{b}_{g,l}(t)\right]^+, \tag{25}$$

where $Q_{g,l}$ is the queue length, $\tilde{a}_{g,l}(t) = \frac{a_{g,l}(t)}{R_{g,l}}$ is the normalized arrival rate and $\tilde{b}_{g,l}(t) = \frac{b_{g,l}(t)}{R_{g,l}}$ is the normalized transmission rate. Here, we define a quadratic Lyapunov's function as

$$F_g(t) = \frac{1}{2}\sum\nolimits_{g=1}^{G} Q_{g,l}^2(t). \tag{26}$$

Let $Q_g(t) \triangleq \{Q_{g,l}(t), 1 \leqslant l \leqslant L\}$ denote the vector of concatenated $Q_{g,l}(t)$ values. The conditional Lyapunov's drift is defined as:

$$\Delta_g(t) \triangleq \mathbb{E}[F_g(t+1) - F_g(t)|\boldsymbol{Q}_g(t)]. \tag{27}$$

Similar to (19), a "drift-plus-penalty" function is given by

$$\Delta_g(t) - W_g\mathbb{E}[U_g(t)|\boldsymbol{Q}_g(t)], \tag{28}$$

where the non-negative parameter W_g facilitates a trade-off between the group utility and the queue stability. In practice, W_g may be set to the same value as V for different groups. If the queue $Q_{g,l}(t)$ is mean-rate-stable, the time-averaged transmission rate constraint in (14) may be satisfied. Then, minimizing the "draft-plus-penalty" of (19) implies maximizing the group utility and minimizing the Lyapunov's drift, which is consistent with the problem (14).

Since the RBs available for each group are decided by n_g^* in the first phase, the RBs for each subgroup ought to satisfy $\sum_{l=1}^{L} n_{g,l}(t) \leqslant n_g^*$, and then we have

$$\sum\nolimits_{l=1}^{L} b_{g,l}^2(t) \leqslant d^2(M)\sum\nolimits_{l=1}^{L} n_{g,l}^2(t) \leqslant (n_g^*)^2 d^2(M). \tag{29}$$

Using a similar derivation as (22), we can derive a loose upper bound of (28), which is as follows:

$$\begin{aligned} &\Delta_g(t) - W_g\mathbb{E}[U_g(t)|\boldsymbol{Q}_g(t)] \\ &\leqslant C_2 + C_3 - \sum\nolimits_{l=1}^{L}[Q_{g,l}(t)b_{g,l}(t) + W_g U_{g,l}(t)], \end{aligned} \tag{30}$$

where $C_2 = \frac{1}{2}\sum_{l=1}^{L} a_{g,l}^2(t) + \mathbb{E}[\sum_{l=1}^{L} Q_{g,l}(t)a_{g,l}(t)]$ and $C_3 = \frac{1}{2}(n_g^*)^2 d^2(M)$. Since the queue state $Q_{g,l}$ and the video arrival rate $a_{g,l}$ for each subgroup is explicitly known, both C_2 and C_3 are constants. By the definition in Section III-C, we have $U_{g,l}(t) = \sum_{k\in\mathcal{S}_{g,l}} u_k(t)$, which is a function of our decision variables $m_{g,l}$ and $n_{g,l}$. By minimizing the "loose" bound of (30), the optimization problem (14) is interpreted as

$$\begin{aligned} &\max_{n_{g,l}, m_{g,l}} && \sum\nolimits_{l=1}^{L}[Q_{g,l}(t)b_g(t) + W_g U_{g,l}(t)] \\ &\text{Subject to} && (4) \quad \textit{and} \quad \sum\nolimits_{l=1}^{L} n_{g,l}(t) \leqslant n_g^*. \end{aligned} \tag{31}$$

This formulation prefers to assign more resources to the subgroup having a larger $Q_{g,l}$. Since a large $Q_{g,l}$ implies a long backlog of data for the l-th video layer, improving the transmission rate for the subgroup $\mathcal{S}_{g,l}$ is capable of maintaining the queue's stability. Similar to Lemma 1, it may be proved that the time-averaged utility acquired by solving the problem (31) asymptotically approaches the optimal value.

In general, this problem can be solved by exhaustively searching through all the feasible values to find the optimum. Although the decision space is narrowed down to a single group, for a given RB assignment $\mathcal{N}_g \triangleq (n_{g,1}, \cdots, n_{g,l}, \cdots, n_{g,L})$ for each subgroup, the number of MCS options would still be $O(L^M)$. For instance in LTE we have $M = 16$, which yields an excessive search space. Here, we design a MCS assignment algorithm based on the block coordinate descent (BCD) method [35] to accelerate the search for the optimum MCS.

The basic principle of the BCD algorithm is to decompose the decision vector into multiple coordinate blocks.

Algorithm 1 The Optimum MCS Search Based on BCD

```
Input: Given a feasible subgroup RB assignment $\mathcal{N}_g$, the queue length $Q_{g,l}$, the subscribers' channel quality $q_k(k \in \mathcal{S}_g)$, the max iterations $Z$.
Output: $\mathcal{M}_g$
Initialization $\mathcal{M}_g^0 \triangleq (m_{g,1}, M, \cdots, M)$
$m_1 \leftarrow \min f(q_k)$ according to (4)
for $k = 1, 2, \ldots, Z$ do
    for $l = 2, \ldots, L$ do
        update $m_{g,l}^i$ according to (32)
    end for
    if stopping criterion is satisfied then
        return $\mathcal{M}_g^i$
    end if
end for
```

BCD minimizes the objective function cyclically over each coordinate block, while fixing the remaining blocks at their most recent updated values. Here, the decision vector is defined as $\mathcal{M}_g^i \triangleq (m_{g,1}, \cdots, m_{g,l}, \cdots, m_{g,L})$. The objective function is "drift-plus-penalty", denoted by $\varphi(\mathcal{N}_g, \mathcal{M}_g) \triangleq \sum_{1 \leqslant l \leqslant L}[-Q_{g,l}(t)b_g(t) - W_g U_{g,l}(t)]$, and the MCS of the l-th layer is regarded as the l-th coordinate block. Let $m_{g,l}^i$ denote the value of $m_{g,l}$ after i iterations. The MCS of the BL, *i.e.* $m_{g,1}$, is determined by the subscriber having the worst channel quality. $m_{g,l}(1 < l \leqslant L)$ is initialized by M so that the constraint of (4) can be satisfied, and it is updated by

$$m_{g,l}^i = \arg\min_{m_{g,l}}[\frac{1}{2}(m_{g,l} - m_{g,l}^{i-1})^2 - Q_{g,l}(t)b_{g,l}(t) - W_g U_{g,l}(t)]. \tag{32}$$

Since the variation of channel quality is unknown, we cannot ensure the convexity of the objective function. According to [35], we adopt the proximal minimization in (32), which allows the objective function to be non-convex over each block of variables. The iterations will end either when the algorithm converges or a certain number of iterations are reached. The procedure of searching for the optimum MCS based on BCD is summarized in Algorithm 1.

The procedure of the subgroup RB and MCS assignment is depicted in Algorithm 2. Under the constraint of $\sum_{l=1}^{L} n_{g,l} \leqslant n_g^*$, we may tentatively try all legitimate RB assignments for each subgroup, namely $\mathcal{N}_g \triangleq (n_{g,1}, \cdots, n_{g,l}, \cdots, n_{g,L})$. The next step is to search for the optimum $\mathcal{M}_g$ under a given $\mathcal{N}_g$ using Algorithm 1. We then repeat this process until the end of iterations. Finally, we can get the optimal $\mathcal{N}_g^*$ and $\mathcal{M}_g^*$ for minimizing the objective function $\varphi(\mathcal{N}_g, \mathcal{M}_g)$.

D. Complexity Analysis

The MCS and RB assignment will be conducted in every scheduling slot. Here, we will analyze the complexity of the proposed layered optimization algorithm for dynamic MCS and RB assignment in multirate multicast services. The algorithm is divided into two phases. In the first phase, we try all the legitimate RB assignments for each group. Since we have to assign N RBs to G groups, there is a total of $\binom{N-1}{G-1}$ options. Then, the complexity of the second phase would be $\mathcal{O}[\binom{N-1}{G-1}]$. In the second phase, we try all legitimate feasible RB assignments for each subgroup. Since we have to assign n_g^* RBs to L subgroups, there are totally $\binom{n_g^*-1}{L-1}$ options. Here, we consider a general case that the N RBs are assigned equally to each group, that is, $n_g^* = \lfloor \frac{N}{G} \rfloor$. For each subgroup RB assignment, we have to find the optimum MCS based on BCD. According to Algorithm 1, the search complexity is $\mathcal{O}(Z \cdot L)$. Notably, the subgroup RB and MCS assignment procedure of Algorithm 2 will be carried out for each group. Accordingly, the complexity of the second phase would be $\mathcal{O}[G \cdot Z \cdot L \cdot \binom{\lfloor \frac{N}{G} \rfloor - 1}{L-1}]$. The overall complexity of our proposed algorithm is $\mathcal{O}[\binom{N-1}{G-1} + G \cdot Z \cdot L \cdot \binom{\lfloor \frac{N}{G} \rfloor - 1}{L-1}]$.

Algorithm 2 Subgroup RB and MCS Assignment

```
Input: The allocated RB for each group $n_g^*$, the queue length $Q_{g,l}$, the subscribers' channel quality $q_k(k \in \mathcal{S}_g)$, the max iterations $Z$.
Output: $\mathcal{M}_g, \mathcal{N}_g$
Initialization $\varphi^*(\mathcal{N}_g, \mathcal{M}_g) = 0$
repeat
    Search for the optimum MCS $\mathcal{M}_g$ using Algorithm 1
    Calculate $\varphi(\mathcal{N}_g, \mathcal{M}_g)$
    if $\varphi(\mathcal{N}_g, \mathcal{M}_g) > \varphi^*(\mathcal{N}_g, \mathcal{M}_g)$ then
        Update the objective: $\varphi^*(\mathcal{N}_g, \mathcal{M}_g) \leftarrow \varphi(\mathcal{N}_g, \mathcal{M}_g)$
        Update MCS and RB: $\mathcal{N}_g^* \leftarrow \mathcal{N}_g, \mathcal{M}_g^* \leftarrow \mathcal{M}_g$
    end if
until All $\mathcal{N}_g$ are iterated
```

TABLE II
PARAMETERS SETUP

Parameter	Value
Cell Radius	500 m
Carrier Frequency	2 GHz
Transmission Power	43 dBm
Distance Attenuation	128.1 + 37.6*log(d), d [km]
Shadow Fading	Log-normal, 0 mean, $\sigma = 8$ dB
RB Size	12 Subcarriers, 0.5 ms
Total RBs	100
Scheduling Time Slot	A Radio Frame, 10 ms
Subcarrier Bandwidth	15 kHz
MIMO Configuration	1 Tx, 1 Rx
Noise	-174 dBm/Hz

V. Performance Evaluation

In this section, the performance of our dynamic RB and MCS assignment algorithm (DRMA) proposed for multirate multicast services is verified. To show the performance improvement of our solution, we also implement the OLM [13], the MSML [15] and the dynamic adaptive multicast (DAM) [18] scheme as benchmarks, which are the most recent and advanced multicast algorithm for OFDM systems.

A. Simulation setup

In the simulations, four video clips, namely *Crew*, *Football*, *City* and *Harbour* clips [36], at a 4CIF resolution (704×576) were encoded into a BL and 3 ELs by the H.264/SVC reference software JSVM [37]. We adopted the SNR scalability mode to generate the SVC videos. Therefore, all the layers had an identical spatial and temporal resolution, but the quantization parameters (QPs) were gradually reduced upon increasing the video quality, which were set to 36, 32, 30 and 28, respectively. Notably, the proposed method is also capable of supporting temporal and spatial scalability coding. The Group of Picture (GOP) size was fixed to 16 and the frame rate was set to 30 fps. Actually, even with fixed QPs, the bitrate of the video stream would fluctuate with time, especially when the video scene or the motion activity changed. The average bitrate of each video layer was summarized in Table III. In order to assess the subjective video quality, we computed the perceived quality scores of each SVC layer using the video multi-method assessment fusion (VMAF) [38] in Table III.

We used the Matlab LTE Toolbox for performing video transmission and reception over the LTE system. The toolbox provides fully-implemented uplink and downlink PHY and MAC functions, such as the adaptive modulation and coding (AMC), the path loss measurements and the channel state information feedback, which are appropriate for us to simulate the wireless communications links. The evaluation was conducted in accordance with the standard LTE parameters, which are summarized in Table II. The channel quality experienced by a specific subscriber was characterized by SNR, which was defined as $\gamma = \frac{h^2 P}{d^\alpha N_0}$, where P was the transmission power, h was the channel gain, N_0 was the noise power, α was the path loss exponent and d was the distance from BS. In the simulation, the subscribers of each multicast group were randomly distributed (2-dimensional uniform distribution) in the coverage area of the BS. Hence, the probability that the subscriber was located within d from the BS is $Pr(d) = \frac{d^2}{R^2}$, where R was the cell radius. The path loss was set to $128.1 + 37.6 \cdot log(d)$ and the SNR experienced was exponentially distributed according to [18]. Meanwhile, the small scale fading of each subscriber's channel obeyed an i.i.d. Rayleigh distribution with a coefficient of 2. The noise was Gaussian distributed with a power of $-174\ dBm/Hz$. Since only 1 Tx antenna and 1 Rx antenna were used in our evaluations, the mapping of [39] was applied to map the SNR to the CQI, while keeping the BLER under 10%. Other MIMO configurations could be readily applied by modifying the mapping tables [40]. According to the LTE specifications, the Transmission Time Interval (TTI) was specified at 1 ms and the CQI reporting cycle was set to 2 TTIs. The scheduling slot was fixed to a radio frame, namely 10 ms. Initially, we generated 20 uniformly distributed subscribers for each multicast group. To emulate an actual multicast scenario with an ever-changing number of subscribers, we assumed that the subscribers' arrival/departure followed a Poisson process with a rate of $\lambda = 10^{-4}$ per TTI.

The modulated video stream was transmitted over the simulated wireless channels. According to the above channel models, the SNR values of each subscriber were updated every TTI. After demodulating the OFDM signals, different subscribers might experience different BLERs of the received data, due to their heterogeneous channel conditions. We decoded the received video stream into a video sequence in YUV format using JSVM. By comparing it to the raw video clips using VMAF, we assessed the quality of received video [41]. We simulated a video delivery period of 1000 scheduling slots, and all the results were averaged over 50 independent simulation runs. During each simulation run, the generated traces of channel condition were the same for different algorithms.

B. Performance Metrics

- *System Utility* quantifies the average QoE perceived by the multicast subscribers. Here, the coefficients, *i.e.* a and b, of the logarithmic-form user QoE model in (7) are set to 1, and then the utility ranges from $-\infty$ to 0. The higher the system utility, the better the service quality. The time-averaged system utility is defined as $\frac{1}{T}\sum_{t=1}^{T} U(t)$, where T is the simulation time and $U(t)$ is defined in (8).
- *Data Backlog* indicates the amount of video data cached by the BS. A large backlog of data would cause not only high latency, but also packet loss in case of cache overflow. The data backlog is quantified by $\frac{1}{TG}\sum_{g=1}^{G}\sum_{t=1}^{T} H_g(t)$, where H_g is defined in (15).
- *VMAF Score* quantifies the subjective quality of the received video. Due to the time-varying channel conditions, the video quality would fluctuate with time. We compute the VMAF score of each subscriber by comparing the decoded video sequence with the raw video sequence. Explicitly, the mean VMAF score is defined as
$$\bar{P} = \frac{\sum_{g=1}^{G}\sum_{k\in\mathcal{S}_g}\sum_{i=1}^{F} P_{k,i}}{F \cdot |\mathcal{S}|}, \tag{33}$$
where F is the total number of frames and $P_{k,i}$ is the VMAF score of the ith frame for the kth subscriber.
- *Spectral Efficiency* is defined as the ratio between the average number of bits received by all subscribers and the channel bandwidth consumed by multicast services, which is quantified by $\frac{1}{T|\mathcal{S}|}\sum_{t=1}^{T}\sum_{k\in\mathcal{S}} \frac{r_k(t)}{B(t)}$, where $B(t)$ denotes the channel bandwidth.

C. Simulation Results

1) performance for different channel bandwidth: We varied the number of RBs available for multicast services from 20 to 60 and quantified the performances of these multicast schemes, which were plotted in Fig. **??**. Here, the available bandwidth was fixed during each run of the experiment. The average channel SNR for all subscribers was set to 16 dB and the BS cache size was set to 0.5 Mbits. Moreover, the trade-off factors V and W were fixed to 100 for DRMA.

As shown in Fig. **??**(a), the system's utility increases with the growth of available bandwidth, which indicates that the proposed scheme is capable of accommodating different channel conditions by adaptively assigning the RB and MCS for each video layer. For a specific video layer, having more radio resources allows us to use lower channel coding rates, so that even the subscribers having poor channel conditions are able to perceive a better QoE. In detail, the proposed DRMA outperforms the others, since only DRMA optimizes the long-term system utility. By contrast, the OLM and the MSML

TABLE III
BITRATE AND VMAF SCORE OF THE SVC VIDEO CLIPS

Video Clips		Crew	Football	City	Harbour
BL	Bitrate (Kbps)	602	640	361	958
	VMAF Score	72.3	82.3	84.6	81.0
EL1	Bitrate (Kbps)	1068	1192	839	1879
	VMAF Score	81.4	90.5	32.3	90.3
EL2	Bitrate (Kbps)	1689	1798	1404	2945
	VMAF Score	87.5	95.2	93.5	91.5
EL3	Bitrate (Kbps)	2361	2475	2102	4247
	VMAF Score	90.7	97.2	94.9	93.7

solutions always adopt the specific RB and MCS assignment that maximizes the short-term system utility based on the instantaneous CQI. Due to the stochastic nature of the wireless channel, the current CQI is not always an accurate predictor of near-future variations of the channel, especially when the scheduling slot is much longer than the CQI reporting cycle. We can observe that the DAM achieves a slightly worse utility than the DRAM. In DAM, the subscribers in a multicast group are considered to be uniformly distributed in a single cell area, and the distribution of the subscriber's channel SNR is assumed to be entirely determined by the distance from the BS. These assumptions are naturally idealistic, because in practice the subscriber's distribution appears to be random and time-variant, and the subscriber's channel quality also suffers from stochastic fast fading and shadow fading. By contrast, the DRAM searches for the optimum MCS and RB assignment using the user reported CQIs. It can be also seen in Fig. **??**(a) that the MSML suffers from a poor utility, even when increasing the available bandwidth. The reason behind this trend is that the MSML is designed for maximizing the intra-group spectral efficiency rather than the QoE. It maintains a high MCS order for the delivery of ELs, which results in excluding a part of subscribers having moderate channel quality from obtaining a better video quality.

The average data backlog during the video streaming is characterized in Fig. **??**(b). The DAM backlogs the minimum amount of data in the transmission queue. It assigns the RB and MCS based on the average video bitrate of each layer. If the arrival rate remains constant at the average bitrate, there would be no data backlog. However, affected by the time-varying network conditions and video bitrates, the actual arrival rate is fluctuated with time. The proposed DRMA has a slightly larger data backlog, since the DRMA considers the long-term stability of the transmission queues. If massive amounts of data are waiting to be transmitted in the queue, the DRMA prefers to improve the transmission rate by adopting a higher-order MCS, and vice versa. By contrast, the decision-makings of both the OLM and MSML are merely based on the instantaneous channel condition, which cannot guarantee the transmission queue stable. If the available radio resources are insufficient to transmit all the arriving bits of any video layer in a single time slot, they would accumulate the video data in the transmission queue. Consequently, as shown in Fig. **??**(b), both the OLM and MSML lead to a large backlog of data. By comparing Fig. **??**(a) to Fig. **??**(c), we also observe that the scheme achieving a higher system utility can guarantee a much better video quality for the subscribers, which is consistent with our optimization goal.

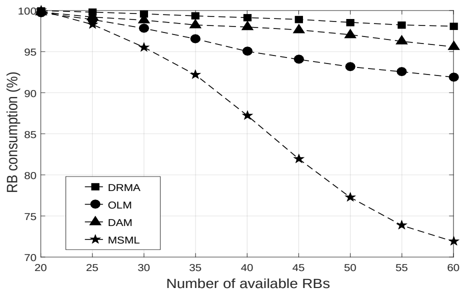
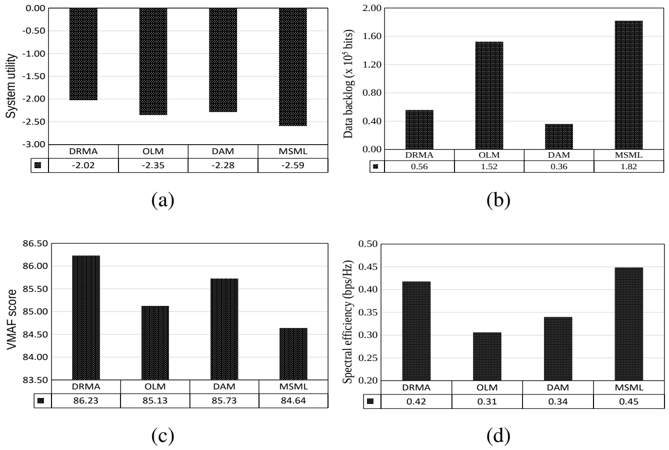


Fig. 3. RB consumption for different channel bandwidth.



(a)

(b)

(c)

(d)

Fig. 4. Performance results in the scenario with dynamic bandwidths. (a) System utility. (b) Data backlog. (c) VMAF score. (d) Spectral efficiency.

As shown in Fig. **??**(d), the spectral efficiency decays as the growth of available bandwidth, because more radio resources enable the BS to convey the same data using much lower MCS orders, which reduce the average number of bits carried by each OFDM signal. In detail, the DRMA achieves a higher spectral efficiency than both the OLM and DAM under the same channel conditions. Since by definition, the spectral efficiency characterizes how efficiently the radio resources are exploited, this result implies that our scheme can improve the reception rate as well as the video quality for all subscribers, which is consistent with the results in Fig. **??**(a). We can also observe that the MSML outperforms the other schemes in terms of spectral efficiency, when the number of RBs available is over 40. The MSML policy is designed for maximizing the spectral efficiency. In high-bandwidth scenarios, the further increase of bandwidth would no longer help to improve the spectral efficiency. As a result, in order to guarantee a high spectral efficiency, the MSML would only exploit a part of the available RBs for multicast services, even though the channel bandwidth is adequate. We plot the RB consumption under different channel bandwidths in Fig. 3. We find that the proportion of RB consumption of the MSML is far below that of the other schemes, especially in case of a high channel bandwidth. Since the spectral efficiency is defined as the ratio between the average bit rate received by all subscribers and the channel bandwidth consumed by multicast services, a lower

TABLE IV
STANDARD DEVIATION OF THE RESULTS.

Deviation Method Metric	DRMA	OLM	DAM	MSML
System utility	0.0849	0.1114	0.0965	0.0998
Data backlog ($\times 10^5$ bits)	0.0656	0.0771	0.0752	0.0821
VMAF score	0.8445	0.8075	0.7843	0.9265
Spectral efficiency (bps/Hz)	0.0292	0.0436	0.0279	0.0362

bandwidth consumption naturally leads to a higher spectral efficiency. Additionally, the results in Fig. 3 also imply that the DRMA can make the most of the available bandwidth.

2) performance for dynamic channel bandwidth: Different from the above experiment where the available channel bandwidth was fixed during video transmission, we conducted a further experiment to investigate the performance for a dynamic channel. Here, we assumed that the number of RBs available for multicast services was Poisson distributed with a mean of 40. The remaining parameters were the same as the above experiments and the results were plotted in Fig. 4.

As shown in Fig. 4(a), the DRMA outperforms the other schemes in terms of system utility. The DAM, OLM and MSML assign the RB and MCS for each subgroup merely based on the instantaneous available bandwidth, which varies randomly in our experiment. As a result, they may adopt a high-order MCS to increase the throughput once the available bandwidth decreases, even though there is a very small backlog of data in the transmission queue. By contrast, since the DRMA considers the long-term dynamics of the available bandwidth, it may adopt a low-order MCS so that the subscribers having bad channel quality can receive the ELs. This result indicates that by considering the stability of the transmission queue, the DRMA readily accommodates the bandwidth fluctuations and thus provides the best QoE for the subscribers.

From Fig. 4(b), we can observe that there is almost no data backlog for the DAM, since it always matches the transmission rate with the mean video rate. By maintaining the long-term stability of the transmission queue, the DRMA achieves a much smaller data backlog than the other schemes. The MSML has a large backlog of data, because the ELs may be discarded when the bandwidth is inadequate. Fig. 4(c) shows the video quality in face of a fluctuating bandwidth. As expected, a higher system utility can ensure a much better video quality for the subscribers. By comparing Fig. 4(d) to Fig. 4(c), we observe that the MSML achieves the highest spectral efficiency of 0.45 bps/Hz at the expense of eroded video quality. Since both the channel conditions and the number of available RBs are randomly generated for each simulation run, we also give the standard deviations of the results over 50 independent runs in Table IV. The deviation is small, which implies that the results of independent simulation runs are close to the mean values.

3) performance for non-uniformly distributed subscribers with dynamic bandwidths: We conducted a further experiment for investigating the performance for non-uniformly distributed subscribers, while the above experiments were based on a uniform distribution. The numbers of subscribers in the groups requesting the *Crew*, *Football*, *City*, *Harbour* video clips were set to 50, 40, 25, 15, respectively. The dynamics of the number of RBs available for multicast services were assumed to be Poisson distributed with a mean of 30. Since the utility was negative, we plotted its absolute value for a better representation in Fig. **??**, where a lower absolute value implied a higher utility.

TABLE V
UTILITY FOR THE SAME CHANNEL MODEL PARAMETERS.

Utility Group Criterion	Crew	Football	City	Harbour
Mean	-2.04	-1.98	-2.03	-2.12
Median	-2.05	-2.01	-2.11	-2.09
Minimum	-2.01	-1.96	-2.05	-1.98
Maximum	-1.95	-2.02	-2.10	-2.06

It is found in Fig. **??**(a) that the system utilities of both the DAM and DRMA increase in the order of *Crew*, *Football*, *City* and *Harbour*, which implies that the group containing more subscribers achieves a higher utility. Since the system's utility is defined as a weighted sum of the group utilities, where the weight is the normalized number of subscribers in the group, the DRMA can guarantee proportional fairness between different multicast groups. Consequently, the DRMA prefers to assign more resources to the group having more subscribers. More radio resources would result in a higher spectral efficiency and a lower backlog of video data. As shown in Fig. **??**(b) and Fig. **??**(d), the *Crew* group (containing most subscribers) has the highest spectral efficiency and lowest backlog of video data compared to the other groups. Due to the differences of the video scenarios, we compare the increase of the VMAF scores relative to the BLs, rather than directly comparing the VMAF scores. In Fig. **??**(c), the *Crew* group achieves a higher video quality improvement than the other groups relative to their basic video qualities (VMAF score of the BLs). As shown in Fig. **??**(d), for OLM, the ratio of the utility between the *Crew* group and the *Harbour* group is much higher than that of the other groups. The reason behind this result is that using the OLM, the group having more subscribers is given priority for allocating the limited radio resources. That is, when the data transmissions of the large groups are completed, the remaining radio resources are then allocated to the smaller groups. Naturally, the OLM cannot achieve proportional fairness between groups. By contrast, the system's utility of the MSML varies slightly between different groups, because the MSML invokes a heuristic cost function, which is defined as the ratio between the reception bitrate and the video bitrate, to schedule the radio resources without any consideration to the group size.

4) performance vs. assessment criteria of group channel conditions: In the above experiments, we adopted the mean CQI of all the group members to assess the channel condition of a group. Here, we evaluated the performance under different criteria, namely the mean CQI, the median CQI, the minimum CQI and the maximum CQI. Each group had the same number of subscribers and the available channel bandwidth changed dynamically following a Poisson distribution with a mean of

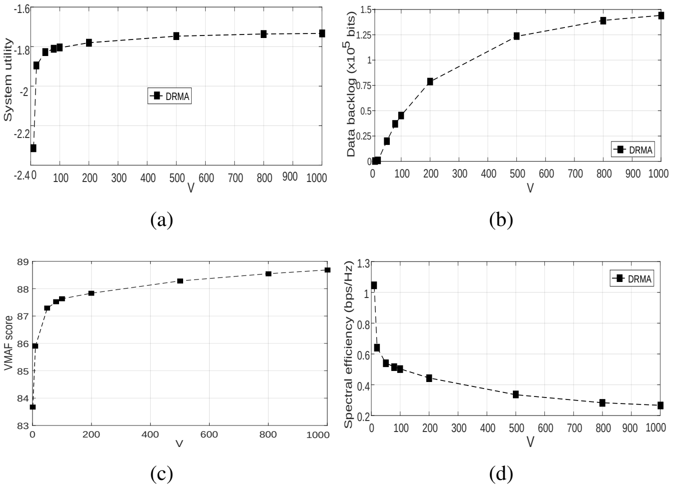
TABLE VI
UTILITY FOR DIFFERENT CHANNEL MODEL PARAMETERS.

Utility Group Criterion	Crew	Football	City	Harbour
Mean	-2.39	-1.82	-1.58	-1.35
Median	-2.45	-1.85	-1.56	-1.34
Minimum	-1.84	-1.83	-1.80	-1.81
Maximum	-1.82	-1.84	-1.93	-1.87

40 RBs. In the first case, we used the same channel model parameters to generate the channel SNR for the subscribers in different groups and the corresponding results were shown in Table V. Observe that the utilities of the different groups are quite similar. Using the same channel model parameters, different groups have the same channel SNR distribution across the subscribers. As a result, the mean, median, minimum or maximum CQI metrics of the different groups are almost the same. Hence, the radio resources would be equally shared among the groups, regardless of which criterion we use.

In the second case, we changed the channel model parameters of different groups, ensuring that the average SNRs of the group *Crew*, *Football*, *City* and *Harbour* increase in turn and the results were shown in Table VI. We find that the utility increases with the average SNR under the mean and median criteria, but it changes very little under the minimum and maximum criteria. This is because the mean or median CQIs of different groups would change with the average channel SNRs, but the minimum or maximum CQIs of different groups are almost unchanged. These results imply that both the mean and median criteria are capable of characterizing the channel condition changes of the whole group.

5) performance vs. the trade-off factors: We conducted further experiments for investigating the performance sensitivity of the proposed DRMA vs. the trade-off factor V in (23) and W in (31), where V was set to be equal to W and the value of V varied in the range of 1 to 1000. The number of RBs available for multicast services was fixed to 50 and the remaining parameters were set to the same values as in Fig. **??**. The simulation results were exhibited in Fig. 6. From Fig. 6(a) and (c), we observe that both the time averaged system utility and the video quality saturate after a period of rapid growth vs. V. Since the DRMA strikes a trade-off between the queue length and the utility according to (23) and (31), a large V improves the system utility by allocating more resources to the groups containing more subscribers or having better channel conditions. As shown in Fig. 6(d), upon increasing the value of V, the DRMA tends to a lower spectral efficiency by reducing the MCS order. In this way, the number of users that could receive a better video quality is increased, and the system utility is improved as well. However, this will impose a large backlog of data on the other groups. As shown in Fig. 6(b), the data backlog increases with V. By contrast, a small V increases the importance of queue stability in the decision making of the DRMA. As a result, the DRMA tends to assign more resource to the groups having lager backlog, which avoids the playback freezes. Hence, controlling V provides an efficient and flexible way for us to strike a trade-off between the system's utility and the playback smoothness.



(a)

(b)

(c)

(d)

Fig. 6. Performance vs. the trade-off factors V. (a) System utility. (b) Data backlog. (c) VMAF score. (d) Spectral efficiency.

VI. CONCLUSIONS

We conceived SVC based multicast to provide differentiated QoE for numerous video subscribers having time-varying and heterogeneous channel conditions. By dynamically scheduling the video rate and the radio resource, we optimized the long-term QoE for all subscribers, while taking both the inter- and intra-group fairness into account. The problem was formulated as a time-averaged optimization problem constrained by the limited radio resources. This formulation was not influenced by the instantaneous fluctuation of channel conditions and video bitrates. Based on Lyapunov's drift and optimization theory, we proposed a layered twin-phase solution for solving this problem. An efficient algorithm was proposed for accelerating the search for the optimum MCS, which mitigated the computational burden. We then derived an analytical performance bound to show its optimality. Our experiments conducted using realistic video traces demonstrated that the proposed solution achieved the best QoE, while additionally maintaining proportional fairness among groups. Our future research will focus on harnessing non-orthogonal multiplexing techniques, which is considered as a promising candidate for the next generation systems.

APPENDIX A
PROOF OF LEMMA 1

By minimizing the "loose" bound of (22), we obtain a suboptimal RB assignment n_g^* for the problem (13). Let b_g^* and U^* denote the corresponding transmission rate and system utility, respectively. Then, we have

$$\begin{aligned}\Delta(t) - V\mathbb{E}[U^*(t)] &= \frac{1}{2}\mathbb{E}[\sum_{g=1}^{G} H_g^2(t+1) - H_g^2(t)] - \\ V\mathbb{E}[U^*(t)] &\leqslant C_0 + C_1 - \mathbb{E}[\sum_{g=1}^{G} H_g(t)b_g^*(t) + VU^*(t)].\end{aligned} \tag{34}$$

According to Theorem 4.5 in [34], there exists a stationary optimal ω-only policy that achieves the optimal system utility $\overline{U}^{opt}$, while satisfying the constraints in problem (13). The ω-only policy implies that the decision only depends on the observation of the random event ω, which characterizes the channel dynamics here. Since the right side of inequality (34) is the minimum of the "loose" bound of (22), any other assignment policy would increase its value, yielding

$$\begin{aligned}&\frac{1}{2}\mathbb{E}[\sum_{g=1}^{G} H_g^2(t+1) - H_g^2(t)] - V\mathbb{E}[U^*(t)] \leqslant \\ &C_0 + C_1 - \mathbb{E}[\sum_{g=1}^{G} H_g(t) b_g^{opt}(t) + VU^{opt}(t) | \boldsymbol{H(t)}],\end{aligned} \tag{35}$$

where b_g^{opt} is the transmission rate under the optimal ω-only policy. Since the above inequality holds for all slots, summing both sides over $t = 0, 1, \ldots, T-1$ yields

$$\begin{aligned}&\frac{1}{2}\mathbb{E}[\sum_{g=1}^{G} H_g^2(T) - H_g^2(0)] - V\sum_{t=0}^{T-1}\mathbb{E}[U^*(t)] \leqslant \\ &TC_0 + TC_1 - \sum_{t=0}^{T-1}\mathbb{E}[\sum_{g=1}^{G} H_g(t) b_g^{opt}(t) + VU^{opt}(t)].\end{aligned} \tag{36}$$

Upon dividing both sides by VT and taking $T \to \infty$, we have

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{2VT}\mathbb{E}[\sum_{g=1}^{G} H_g^2(T)] - \lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[U^*(\tau)] \leqslant \frac{C_0+C_1}{V} \\ &+ \lim_{T\to\infty}\frac{Nd(M)}{VT}\sum_{\tau=0}^{T-1}\mathbb{E}[\sum_{g=1}^{G} H_g^2(\tau)] - \lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[U^{opt}(\tau)].\end{aligned} \tag{37}$$

Here, the inequality $b_g \leqslant Nd(M)$ is used. Because the queue $H_g(t)$ is mean-rate-stable, namely we have $\lim\limits_{t\to\infty}\frac{1}{t}\mathbb{E}[H_g(t)] = 0$, both the first term on the left side and the second on the right side equal to 0. By definition, we have $\overline{U^*} = \lim\limits_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[U^*(\tau)]$ and $\overline{U^{opt}} = \lim\limits_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}[U^{opt}(\tau)]$. Hence, the above inequality (37) may be represented by

$$\overline{U^*} \geq \overline{U^{opt}} - \frac{C_0+C_1}{V}. \tag{38}$$

Acknowledgment

This work is supported by the Fundamental Research Funds for the Central Universities (No. WK2100000009), the Anhui Provincial Natural Science Foundation (No. 1908085QF266), the Youth Innovation Promotion Association CAS (No. CX2100107001) and the Anhui Special Support Program.

References

[1] L. Hanzo, P. J. Cherriman, and J. Streit, *Wireless video communications: second to third generation and beyond*. John Wiley & Sons, 2001.

[2] F. Mikael, A. Taimoor, and A. Sylvain, "Multicast and broadcast enablers for high-performing cellular V2X systems," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 454–463, Jun. 2019.

[3] J. Yang, B. Yang, S. Chen, Y. Zhang, Y. Zhang, and L. Hanzo, "Dynamic resource allocation for streaming scalable videos in SDN-aided dense small-cell networks," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2114–2129, Mar. 2019.

[4] C. White Paper, "Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022 white paper," Feb. 2019.

[5] J. Montalban, P. Scopelliti, M. Fadda, E. Iradier, and G. Araniti, "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.

[6] A. Biason and M. Zorzi, "Multicast via point to multipoint transmissions in directional 5G mmWave communications," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 88–94, Feb. 2019.

[7] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Netw.*, vol. 31, no. 2, pp. 80–89, Feb. 2017.

[8] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[9] J.-P. Sheu, C.-C. Kao, S.-R. Yang, and L.-F. Chang, "A resource allocation scheme for scalable video multicast in WiMAX relay networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 1, pp. 90–104, Jan. 2013.

[10] G. Jia, X. Gong, L. Jie, and et. al, "An optimized hybrid unicast/multicast adaptive video streaming scheme over MBMS-enabled wireless networks," *IEEE Trans. Broadcast.*, vol. 64, no. 4, pp. 791–802, Dec. 2018.

[11] J. Guo, X. Gong, J. Liang, W. Wang, and X. Que, "An optimized hybrid unicast/multicast adaptive video streaming scheme over MBMS-enabled wireless networks," *IEEE Trans. Broadcast.*, no. 99, pp. 1–12, Dec. 2018.

[12] D. Striccoli, G. Piro, and G. Boggia, "Multicast and broadcast services over mobile networks: A survey on standardized approaches and scientific outcomes," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1020–1063, Nov. 2018.

[13] C.-W. Huang, S.-M. Huang, P.-H. Wu, and et al, "OLM: Opportunistic layered multicasting for scalable IPTV over mobile WiMAX," *IEEE Trans. Mobile Comput.*, vol. 11, no. 3, pp. 453–463, Mar. 2012.

[14] 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.

[15] 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.

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

[17] P. Li, H. Zhang, B. Zhao, and S. Rangarajan, "Scalable video multicast with adaptive modulation and coding in broadband wireless data systems," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 57–68, Feb. 2012.

[18] J. Park, J.-N. Hwang, and et al, "Optimal DASH-multicasting over LTE," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4487–4500, May 2018.

[19] 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 Quarter 2013.

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

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

[22] J. Huschke, "Max-min throughput-optimal multicast link adaptation for non-identically distributed link qualities," in *2010 IEEE 72nd Vehicular Technology Conference-Fall*. IEEE, 2010, pp. 1–6.

[23] A. De La Fuente, G. Femenias, and et al, "Subband CQI feedback-based multicast resource allocation in MIMO-OFDMA networks," *IEEE Trans. Broadcast.*, vol. 64, no. 4, pp. 846–864, Dec. 2018.

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

[25] J. Kua, G. Armitage, and P. Branch, "A survey of rate adaptation techniques for dynamic adaptive streaming over HTTP," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1842–1866, 3rd Quarter 2017.

[26] 3GPP, "Multimedia broad cast/multicast service (MBMS)," 3GPP Standard TS 26.346, Release 13, 2015.

[27] 3GPP, “Transparent end-to-end packetswitched streaming service (PSS); progressive download and dynamic adaptive streaming over HTTP (3GP-DASH),” *Tech. Rep. TS 23.246 Rel. 11*, 2012.

[28] S. N. Donthi and N. B. Mehta, “An accurate model for EESM and its application to analysis of CQI feedback schemes and scheduling in LTE,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3436–3448, Oct. 2011.

[29] 3GPP TS 36.201, “Physical layer-general description,” Aug. 2019. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/36 series/36.201/.

[30] G. Ku and M. L. Walsh, “Resource allocation and link adaptation in LTE and LTE advanced: A tutorial,” *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1605–1633, 3rd Quarter 2017.

[31] A. Moldovan, I. Ghergulescu, and C. H. Mutean, “VQAMap: A novel mechanism for mapping objective video quality metrics to subjective MOS scale,” *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 610–627, Sep. 2016.

[32] K. Zheng, X. Zhang, Q. Zheng, W. Xiang, and L. Hanzo, “Quality-of-experience assessment and its application to video services in LTE networks,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 70–78, Feb. 2015.

[33] W. Zhang, Y. Wen, Z. Chen, and A. Khisti, “QoE-driven cache management for HTTP adaptive bit rate streaming over wireless networks,” *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1431–1445, Oct. 2013.

[34] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.

[35] Y. Xu and W. Yin, “A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion,” *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[36] Video Test Media, Accessed: Apr. 2020. [Online]. Available: https://media.xiph.org/video/derf/

[37] J. Reichel, S. Heiko, and W. Mathias, “Joint scalable video model 11 (JSVM 11),” Joint Video Team Doc. JVT-X202 (2007): 23.

[38] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward a practical perceptual video quality metric,” Accessed: Apr. 2020. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[39] R. Giuliano and F. Mazzenga, “Exponential effective SINR approximations for OFDM/OFDMA-based cellular system planning,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4434–4439, Sep. 2009.

[40] J. Niu, T. Su, G. Y. Li, D. Lee, and Y. Fu, “Joint transmission mode selection and scheduling in LTE downlink MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 2, pp. 173–176, Apr. 2014.

[41] VMAF Open-Source Project, Accessed: Apr. 2020. [Online]. Available: https://github.com/Netflix/vmaf