CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	RTOP: optimal user grouping and SFN clustering for multiple eMBMS video sessions
Authors	Khalid, Ahmed;Zahran, Ahmed H.;Sreenan, Cormac J.
Publication date	2019-05
Original Citation	Khalid, A., Zahran, A. H. and Sreenan, C. J. (2019) 'RTOP: optimal user grouping and SFN clustering for multiple eMBMS video sessions', IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, Paris, France, 29 April-2 May, pp. 433-441. doi: 10.1109/INFOCOM.2019.8737643
Type of publication	Conference item
Link to publisher's version	https://ieeexplore.ieee.org/abstract/document/8737643 - 10.1109/INFOCOM.2019.8737643https://infocom2019.ieee-infocom.org/
Rights	© 2019, 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.
Download date	2024-07-06 01:27:19
Item downloaded from	https://hdl.handle.net/10468/8170

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

RTOP: Optimal User Grouping and SFN Clustering for Multiple eMBMS Video Sessions

Ahmed Khalid
Dept. of Computer Science
University College Cork, Ireland
Email: a.khalid@cs.ucc.ie

Ahmed H. Zahran
Dept. of Computer Science
University College Cork, Ireland
Email: a.zahran@cs.ucc.ie

Cormac J. Sreenan
Dept. of Computer Science
University College Cork, Ireland
Email: cjs@cs.ucc.ie

***Abstract*—Evolved Multimedia Broadcast Multicast Service (eMBMS) is a 3GPP standard that improves the utilization of scarce wireless resources and the quality of the received content. eMBMS uses a Single Frequency Network (SFN) to transmit real-time videos over synchronized resources across neighboring base stations (eNBs) and allows users to share wireless spectrum across multiple cell sites. However the user with the worst channel condition and the eNB with the least available resources limit the throughput of a session. To overcome such limitations, the SFN can be divided into non-overlapping clusters of eNBs and in each cluster users can be split into groups. We formulate an optimization problem that maximizes an operator-defined utility for multiple eMBMS sessions served at multiple bitrates by choosing the optimal set of SFN clusters and user groups for each session. We propose an algorithm, RTOP, that finds the optimal or a near-optimal solution in real-time regardless of the number of eMBMS users. Our extensive simulations indicate that, in comparison to state-of-the-art schemes, RTOP improves the system utility and average user bitrate by up to 14% and 90% respectively. Additionally, we show that the utility of RTOP always stays within a 1% gap from the optimal solution.**

I. Introduction

Mobile data traffic is increasing rapidly at a 47% annual growth rate and video accounts for more than 60% of this traffic[1]. Live streaming services like Periscope, Facebook Live and Twitch[2] are becoming more popular, which further elevates the demand for high definition (HD) video streaming over cellular networks. The delivery of highly popular content using traditional unicast method leads to inefficient resource utilization and poor user experience.

For users subscribed to the same content, Content delivery networks (CDN) or network layer multicast can reduce resource consumption in the backbone, core and wired access network [1], but the wireless last hop, where the resources are scarce, still suffers from redundant unicast transmissions. Recent experiments and trials [2] illustrate that these drawbacks can be alleviated using Evolved Multimedia Broadcast Multicast Service (eMBMS).

eMBMS [3] is a 3GPP standard that enables multicast over the wireless spectrum by grouping users watching the same content and transmitting the content to a group just once. This results in effective spectrum utilization, particularly when the number of active users is high. Furthermore, to improve the channel condition of users, eMBMS allows base stations (eNBs) in a spatially local area to transmit the same content at a common frequency and time, hence creating a Single Frequency Network (SFN). Interested users combine the signal received from each eNB in the SFN, improving their Signal to Interference-Noise Ratio (SINR). A Multicast Coordination Entity (MCE) manages the eMBMS users and resource allocation for all eNBs in an SFN (Figure 1).

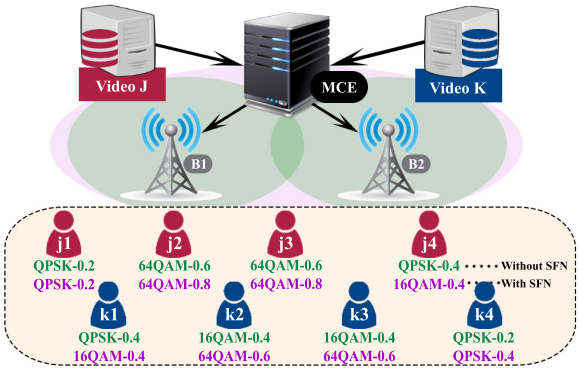


Fig. 1: eMBMS Architecture. *Using SFN can improve the MCS (shown as Modulation-Coding) of users.*

To ensure that all the users can decode the transmitted signal, the modulation and coding scheme (MCS) of a group is restricted to the user with the worst channel condition. Similarly, synchronizing eNBs in an SFN brings forth two limitations: the eNB with the least available resources limits the amount of resource blocks (RBs) available for an eMBMS session and users of one eNB with low MCS values can adversely affect users of other eNBs when creating user groups. To overcome these limitations, state-of-the-art solutions propose partitioning eNBs in the eMBMS service area into multiple SFN clusters [4] depending on the user distribution and available RBs at each eNB. Alternatively, users are split into groups based on their channel conditions [5] with each group receiving an appropriate video bitrate.

To maximize the benefits and potential of eMBMS, network operators need a solution that can run in real-time and solve the user grouping and SFN clustering problems. The existing models [6][7][8] ignore the inter-dependence of these two problems hence yielding sub-optimal results. Also most of these models are either too complex to solve in real-time for a large number of users [5]; do not consider multiple videos served by eMBMS at the same time [6][8]; aim to

[1]https://goo.gl/ySYurJ
[2]https://www.twitch.tv/year/2017

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number: 13/IA/1892.

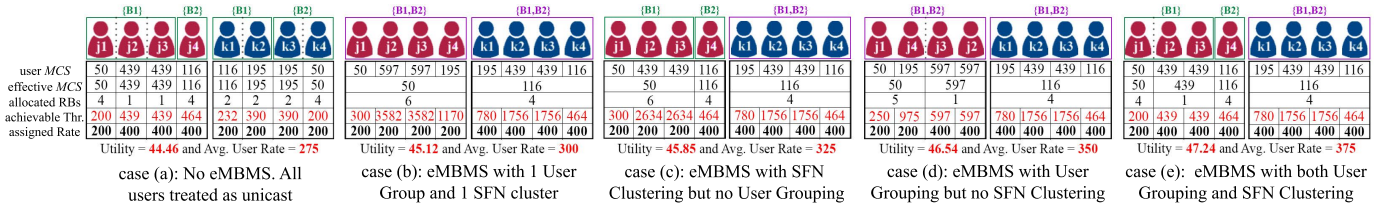


Fig. 2: Possible eMBMS Configurations. *Grouping users and creating SFN clusters can help improve the system utility and must be jointly optimized to maximize the utility as in case (e).*

maximize the network throughput instead of the application-level video bitrates [7][8] or; ignore the impact of eMBMS resource allocation on unicast users [9].

In this paper, we propose RTOP, a novel scalable resource management framework for eMBMS that jointly optimizes resource allocation, SFN clustering and user grouping to maximize an operator-defined utility. We evaluated RTOP and results show that the joint optimization can achieve up to a 14% improvement in the system utility and a 90% increase in the average bitrate received by users in comparison to state-of-the-art techniques [7][8]. Additionally, our framework guarantees a minimum video bitrate for all eMBMS users with most users receiving higher bitrates. Our contributions towards RTOP are multi-fold:

- We formulate the joint optimization problem for SFN clustering and user grouping for multiple video sessions. The solution of this problem determines the performance bound on a rate-based utility and presents a practical mechanism to handle the impact of eMBMS decisions on unicast users.
- We propose a scalable heuristics-based algorithm that produces optimal or near-optimal results in real-time independent of the number of users in typical eMBMS settings.
- We perform extensive evaluation with various network configurations, user distributions and number of videos served by eMBMS with different bitrates. Our results indicate the benefit of the joint optimization in comparison to state-of-the-art techniques [7][8]. We also show that the utility achievable by RTOP is always within a 1% gap from the globally optimal solution.

The rest of the paper is organized as follows. In Section 2, we present the system model and formulate the optimization problem. In Section 3, we present RTOP and the proposed heuristics. In Section 4, we present performance results of our extensive evaluation and comparison. In Section 5, we share related work and finally, in Section 6, we conclude our work.

II. Optimal eMBMS Clustering and User Grouping

A. System Model

We consider a cellular system consisting of a set of eNBs B in the eMBMS service area, serving some unicast users and a set of videos V to multicast users in set M using eMBMS. For each video v, eNBs B can be grouped to form one or more non-overlapping clusters of SFN. Each video v is encoded to a set of bitrates R_v and the MCE may transmit v at one or more distinct bitrates in each cluster at a chosen modulation and coding scheme (MCS). Note that the real-time variations in bitrates are handled by network buffers and we consider an average value over time. Based on its channel quality indicator (CQI) level, a user may select a bitrate, and consequently an MCS, that is best suitable for its channel condition.

TABLE I: Notations For Optimization Model

Symbol	Description
INPUTS	
B	Set of one or more eNBs in the eMBMS service area
C	Set of possible clusters of eNBs (non-empty subsets of B)
P	Set of all possible eNB configurations i.e. ways to configure eNBs B into non-overlapping SFN clusters
b_{pc}	Binary variable to inform if b is in cluster c for $p \in P$
E	Total number of available CQI levels (15 for LTE)
S_e	Achievable spectral efficiency from a CQI level e
T	Total number of resource blocks available at any eNB
α	Maximum fraction of resources allowed for eMBMS
$f(r)$	Operator-defined utility function that takes rate r as input
Y_b	Number of RBs requested by eNB b for its unicast users
V	Set of videos served by eMBMS in the service area
R_v	Set of bitrates available for video v
M, M_v	Set of multicast users (M) subscribed to video v (M_v)
M_{vpce}	Number of users of video v in cluster c with MCS e when eNB configuration p is used
VARIABLES	
P_{vp}	Binary variable to determine if eNB configuration p has been chosen for video v
X_{vpcr}	Number of RBs allocated by eNBs in cluster c of eNB configuration p to video v for bitrate r
M_{vpcer}	Binary variable to determine if users of video v with MCS e in cluster c of eNB configuration p are assigned bitrate r

As each MCS produces a certain spectral efficiency, the MCS chosen by MCE for a bitrate of a video determines the number of frequency-time resource blocks (RBs) needed to achieve that bitrate. Each eNB has T RBs which are used to serve both unicast and multicast users. The maximum fraction of RBs allowed for eMBMS is α [10] and each eNB b, needs Y_b RBs to serve its unicast users. Hence, the available RBs for eMBMS users at any eNB b equals $\min(\alpha T, T - Y_b)$. Table I summarizes the notations used in this paper. In such a system, different configurations of eMBMS are envisioned, leading to different achievable bitrates for users.

To illustrate this, we look at an example scenario (Figure 1) for two videos (J and K) served at two different bitrates (200kbps and 400kbps) by two eNBs (B1 and B2) and four users interested in each video. In Figure 1, the MCS values on top are what users can achieve from eNB configuration $\{\{B1\}, \{B2\}\}$, i.e. eNBs split into two clusters and the bottom MCS values are the achievable MCS from $\{\{B1, B2\}\}$, i.e. both eNBs in one SFN cluster. Both eNBs have 20 total RBs ($T = 20$) and 10 RBs reserved for unicast users ($Y_b = 10$)

leaving 10 RBs for users of Video J and K. We explore five possible configurations (Figure 2) and look at the average user bitrate and *sum-log* utility of user bitrates.

No eMBMS (Case a): All users are scheduled separately as unicast users and due to the limited resources only three users could be served with higher bitrate (400kbps). The sum-log utility is 44.46 and the average user bitrate is 275kbps.

Standard *eMBMS* (Case b): All eNBs are in one cluster and all users of a session in one group. As RBs are shared, each user gets more RBs than in unicast. However, the user with the lowest MCS value (*j*1 for Video J) restraint the rate for all users, therefore even though *j*2 and *j*3 have high spectral efficiency, they receive lower bitrate, limiting the total users served with 400kbps to four. The system utility is 45.12 and the average bitrate increases by 10% in comparison to unicast.

***eMBMS* with SFN Clustering** (Case c): For Video J, splitting eNBs into two clusters places *j*4 and the low-end user *j*1 in separate clusters. This allows *j*4 to receive a higher MCS and hence the higher bitrate. With five users receiving 400kbps, the system utility increases to 45.84 and the average bitrate by 8% in comparison to standard eMBMS.

***eMBMS* with User Grouping** (Case d): Instead of SFN clustering, user grouping is applied to standard eMBMS case. Users *j*2 and *j*3 form a separate group than *j*1 and *j*4 and receive the higher bitrate. Now six out of eight users are served with the higher bitrate and the average bitrate increases by 17% in comparison to standard eMBMS. The sum-log utility of the system increases to 46.54.

***eMBMS* with User Grouping and SFN Clustering** (Case e): The users of Video J are distributed in such a way that they achieve little (*j*2, *j*3 and *j*4) or no (*j*1) benefit by synchronizing B1 and B2. Splitting the eNBs in two clusters for Video J places *j*1, *j*2 and *j*3 in B1 and *j*4 in B2 which now has less users and hence more RBs available for *j*4. This enables *j*4 to receive the higher bitrate as well, giving us the optimal solution with seven out of eight users receiving the higher bitrate. The system utility increases to 47.24 and the average bitrate by 25% in comparison to standard eMBMS.

This example shows that the total utility of the system depends on the eNB configuration chosen for each video, the number of user groups created in each SFN cluster, the number of users placed in a group and the number of RBs and bitrate assigned to each user group. We define an optimization model with a goal to maximize an operator-defined utility by considering all of these factors.

B. Problem Formulation

The typical use cases of eMBMS service, such as sporting events, involve a large number of users. Hence, there is a need for scalable optimization framework to identify the performance bounds of resource allocation schemes. Existing work in literature, e.g. [8][9], employ optimization variables that increase with the number of users. Hence, the solution time grows exponentially as the number of users increases.

In our optimization framework, we eliminate the dependence on the number of users by relying on the fact that there are a limited number of distinct CQI values, e.g., 15 CQI levels in LTE networks. We leverage this fact and define CQI groups per video per cluster. Instead of handling users individually, we formulate our optimization problem to find the optimal bitrate for each CQI group. For each eNB configuration p and video v, we denote the number of users that belong to cluster c and report a CQI e as M_{vpce} and pass it as an input to the optimization model. This simple modeling trick reduces the time-complexity and enables us to find the optimal solution of our problem in a reasonable time when evaluating the effectiveness of our proposed heuristics.

In addition to eMBMS users, eNBs may also serve unicast users. In practice, an MCE has no control over how many RBs are allocated to a unicast user which is instead handled by the scheduler of the associated eNB. For each eNB b, we take Y_b as an input from the operator. An operator can choose the mechanism to calculate Y_b based on the priority of eMBMS over unicast [7] or the number of unicast and multicast users in a cell, e.g. with a multicast weight function [8].

We formulate our optimization problem with the objective (Equation 1a) to maximize an operator-defined utility of all multicast users in all the CQI groups. This is illustrated in *Problem 1* as follows:

$$max \sum_{v\in V}\sum_{p\in P} P_{vp}\cdot\sum_{c\in p}\sum_{r\in R_v}\left(f(r)\cdot\sum_{e=1}^{E} M_{vpcer}\cdot M_{vpce}\right) \quad (1a)$$

subject to

$$\sum_{p\in P} P_{vp}=1,\ \forall v\in V \quad (1b)$$

$$\sum_{r\in R_v} M_{vpcer}\leq 1, \forall e\in\{1,2,..,E\}, v\in V, c\in p\in P \quad (1c)$$

$$\sum_{p\in P}\sum_{c\in p}\sum_{r\in R_v}\sum_{e=1}^{E} M_{vpcer}\cdot M_{vpce}=M_v,\ \forall v\in V \quad (1d)$$

$$X_{vpcr}\geq\max_{e}\left(\frac{M_{vpcer}\cdot r}{S_e}\right), \forall r\in R_v, c\in C \quad (1e)$$

$$\sum_{v\in V}\sum_{p\in P}\sum_{c\in p}\sum_{r\in R_v} b_{pc}\cdot X_{vpcr}\leq\min\left(\alpha T, T-Y_b\right), \forall b\in B \quad (1f)$$

where P_{vp} is a binary variable to determine if eNB configuration p has been chosen for video v, M_{vpcer} is a binary variable to determine if users of CQI group M_{vpce} are assigned bitrate r and X_{vpcr} is the number of RBs assigned to r for v in cluster c of eNB configuration p.

Constraint 1b ensures that each video chooses only one eNB configuration. Constraint 1c and 1d guarantee that each CQI group (and hence user) is assigned one and only one bitrate. Constraint 1e ensures that the number of RBs used to transmit a video bitrate in a cluster are enough to decode it properly for users of any CQI group assigned that bitrate. Constraint 1f limits the total RBs used by eMBMS at any eNB to what's left after satisfying unicast resource request by each eNB. It also limits the percentage of RBs allowed for eMBMS to α, which is usually set to 60% [10] in LTE networks. This constraint can be tuned or relaxed in extremely congested networks to

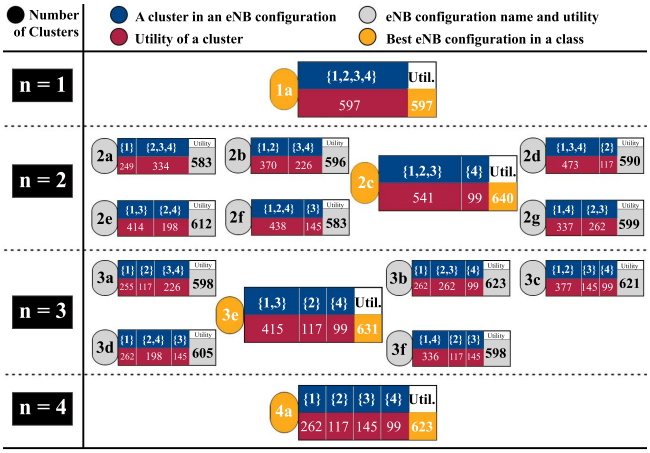


Fig. 3: Heuristic Example. *A candidate eNB configuration is chosen in each class. eNB4 is more congested than other eNBs, so configurations with 4 in a separate cluster perform better.*

adjust the amount of resources that an operator wants to allow for unicast users and leave for eMBMS users.

Time Scale of Optimization: Our problem is a Linearly-Constrained Quadratic Program (QP) with all integer (mostly binary) variables and a global maximum bound (at highest bitrate for all users), hence it is NP-complete [11]. However, even with the notion of CQI groups, when practically computing the optimal solution, the numerous possible eNB configurations and the dependence of each video's choice on other videos makes the problem complicated and time consuming to solve. Based on our experiments (Figure 6c), it can take up to 100s to compute the global maximum, which is not sufficiently fast to operate in real-time. Hence, we propose an efficient heuristic-based algorithm that achieves optimal or near-optimal solution in real-time.

III. RTOP: Solution Design

The optimization problem involves four inter-dependent decisions to make:

1) Identifying an eNB configuration for each video and assigning users to the best cluster of that configuration
2) Creating user groups based on channel conditions
3) Assigning an appropriate video bitrate to each group
4) Allocating RBs to various videos and underlying groups

In this section, we first present how these decisions are taken when one video is served by eMBMS. We then present the additional steps needed to handle the multiple video scenario.

A. Handling A Single Video

In a single video scenario, all the eMBMS resources in the service area are accessible to this video. Hence, the total utility mainly depends on the eNB configuration and underlying user grouping for that video.

eNB Configurations: Each eNB configuration is completely defined by identifying groups of eNBs acting as SFN clusters and assigning users to these clusters. The number of possible eNB configurations is equal to the Bell number[4] of the eNBs

[4]http://mathworld.wolfram.com/BellNumber.html

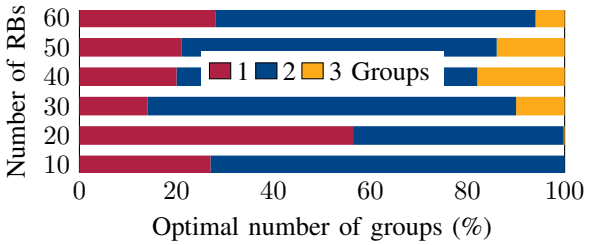


Fig. 4: Number of User-groups to achieve optimal utility from different system configurations.

in the service area. Figure 3 shows an example scenario for a video with four eNBs (1,2,3,4), where eNB4 has higher unicast load (85%) than other eNBs (55%). A 100 users are interested in the video and the eNBs produce different spectral efficiency values when clustered differently. There are 15 possible eNB configurations (fourth Bell number) ranging from one cluster (all eNBs synchronized as a single SFN) to four clusters (each eNB in a separate cluster). The configurations are divided into $|B|$ classes defined by the number of clusters in each configuration. These classes are more relevant to the multiple video scenario and will be discussed in the next section. In each configuration, to maximize the system utility, users are assigned to the cluster that provides them with higher MCS values. In the single video case, we explore user grouping for all the possible eNB configurations. Note that in general, eMBMS is used in a limited number of neighboring eNBs [10] serving a highly populated area.

User Grouping: For an eNB configuration, users in each SFN cluster can have disparate channel conditions. User grouping would enable improving the total utility by splitting users into groups based on their channel conditions and assigning an appropriate video bitrate to each group. On doing so, the achievable utility depends on the number of users in each group, the minimum user MCS value per group, and the RBs available to each group. Exhaustively searching through all possible groups of M_v users takes $\mathcal{O}(|R_v| \cdot |M_v|^{|M_v|})$ to solve. Hence, there is a scalability issue, especially for large number of users. We solve this issue by creating CQI groups as explained in Section II-B and deciding which CQI groups should aggregate to form a user group.

Theoretically, the maximum number of user groups equals the number of distinct video bitrates available. However splitting users in too many groups reduces the share of RBs per group, and may reduce the achievable bitrate by a group and hence the total utility. We conducted simulations by distributing 10, 100 or 1000 users in the service area and varied the number of available RBs. We ran each setup for 1000 runs. Results (Figure 4) show that in 90% of the cases, the optimal utility was achieved by one or two user groups. Therefore, we design our grouping algorithm that aggregates CQI groups and either places all users in one group or creates two user groups.

Algorithm 1 presents the user grouping algorithm and involves three main steps. First, we identify the number of RBs needed to assign a bitrate to the first (lowest) CQI group

Algorithm 1 User Grouping Algorithm

Input: RBs, R_v, CQI Groups in cluster (in ascending order) with CQI values Q and number of users N

Output: Best Utility U_{max} (initialized with 0), User Groups G with number of users and RBs for each bitrate

1: **for** $i \leftarrow 0$ *to* $|R_v|$ **do**
2: $r1 \leftarrow R_v[i]$
3: $r1RBs \leftarrow r1 \ / \ Q[0]$ ▷ *Lowest CQI*
4: **if** $r1RBs > RBs$ **then** *break* ▷ *Can't increase rate*
5: $U \leftarrow f(r1) \times sum(N)$ ▷ *f * No. of users*
6: **if** $U > U_{max}$ **then** $U_{max} \leftarrow U$
7: $G[r1, users] \leftarrow sum(N);\ G[r1, rbs] \leftarrow r1RBs$
8: $r2RBs \leftarrow RBs - r1RBs$ ▷ *Remaining RBs*
9: $thr \leftarrow [r2RBs \times cqi$ **for** cqi *in* $Q[1:]]$
10: **for** $j \leftarrow i+1$ *to* $|R_v|$ **do** ▷ *bitrates higher than* $r1$
11: $r2 \leftarrow R_v[j]$
12: $k \leftarrow index\ of\ first\ CQI\ group\ with\ thr \geq r2$
13: **if** *no* k **then** *break* ▷ *Can't increase 2nd G's rate*
14: $G1 \leftarrow sum(N[0:k]);\ G2 \leftarrow sum(N[k+1:])$
15: $U \leftarrow f(r1) \times G1 + f(r2) \times G2$
16: **if** $U > U_{max}$ **then** $U_{max} \leftarrow U$
17: $G[r1, users] \leftarrow G1;\ G[r1, rbs] \leftarrow r1RBs$
18: $G[r2, users] \leftarrow G2;\ G[r2, rbs] \leftarrow r2RBs$

and calculate the utility achievable by placing all users in one group (Line 1-7). Then we measure the throughput that can be achieved by higher CQI groups with the remaining RBs (Line 8-9). If some CQI groups have enough throughput to support the next bitrate then we calculate the utility for splitting users in two groups and assigning the higher bitrate to those CQI groups (Line 10-19). We repeat the process for all bitrates and choose the user grouping option with the maximum utility. This approach takes only $O(|R_v|^2)$ to solve.

We find the utility of each eNB configuration by running Algorithm 1 on each of its cluster. The configuration with the highest utility is the optimal choice and the eNBs can be configured to form clusters accordingly. Results obtained from Algorithm 1 tell us the number of groups to create in each cluster, number of users to place in each group and also the bitrate and number of RBs to assign to each group.

B. Handling Multiple Videos

In multiple video scenarios, the resources must be distributed optimally among all videos and underlying groups. Such distribution has an impact on the choice of eNB configuration and user grouping. Hence, the problem of choosing an eNB configuration for each video is combinatorial in nature and can result in exponentially increasing outcomes. To solve the problem in real-time, we first narrow down the choices of eNB configurations for each video to a subset of candidate configurations. Then, we identify the best combination of configurations for the videos. Finally, we determine the optimal resource allocation and user grouping for each video in their chosen eNB configurations.

Candidate eNB Configurations: For each video, we use the process of Section III-A to obtain the maximum utility of each eNB configuration based on the total available eMBMS resources. Additionally, we classify each configuration based on the number of clusters in it and then choose the best configuration for each class as a candidate configuration. Continuing with the example in Figure 3, we divide the eNB configurations into four classes and choose the configuration with the highest utility in each class. This gives us four candidate configurations (1a, 2c, 3e, 4a) for the considered video. This process is repeated for all the videos served by eMBMS in the service area to obtain a set of candidate eNB configurations for each video.

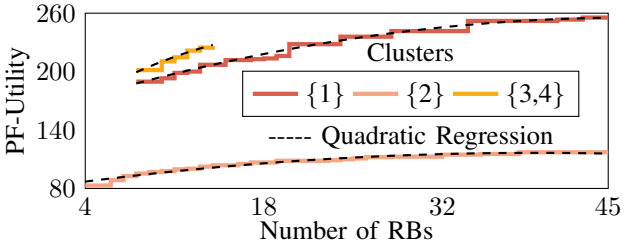


Fig. 5: A sample Utility vs RB graph with quadratic regression. *For any cluster, the lower bound of RBs is the minimum RBs needed to satisfy all users of a video and the upper bound is the RBs available for eMBMS in that cluster.*

Combination of eNB Configurations: In this step, we select one eNB configuration for each video from the previously identified candidates by maximizing an approximated utility function. First, we run the user grouping algorithm (Algorithm 1) on each cluster of an eNB configuration, to identify the achievable utility with up to two user groups per video. This step is repeated for all possible values of RBs, which can vary from one to the maximum RBs available for eMBMS in that cluster, i.e., $min(\alpha T, T - Y_b)$. Figure 5 shows the achievable utility as a function of available RBs for one eNB configuration (3a from Figure 3), which consists of three clusters ($\{1\}$, $\{2\}$ and $\{3,4\}$). We repeat the process for all candidate eNB configurations of each video in V.

We then proceed to find the best combination of eNB configurations for different videos using the measured utility data. To speed this search, we define an optimization problem with an approximated objective utility based on the quadratic regression of the utility values, as shown in Figure 5. We chose quadratic regression as it provided sufficiently accurate and fast solution in comparison to the less accurate linear regression and the slower higher-degree polynomials.

Problem 2 illustrates our optimization problem:

$$max \sum_{v \in V} \sum_{c \in P_v} i_{vc} \cdot X_{vc}^2 + j_{vc} \cdot X_{vc} + k_{vc} \tag{2a}$$

subject to

$$\sum_{v \in V} \sum_{c \in P_v} b_{pc} \cdot X_{vc} \leq \min\left(\alpha T, T - Y_b\right), \forall\, b \in B \tag{2b}$$

$$X_{vc} \geq L_{vc}, \forall\, v \in V, c \in C \tag{2c}$$

where X_{vc} represents the RBs allocated to video v in cluster c, L_{vc} is the lower bound for X_{vc} and i_{vc}, j_{vc}, k_{vc} are the quadratic

coefficients of the utility for video v in cluster c.

Constraint 2b is similar to Constraint 1f and limits the available RBs in a cluster for eMBMS users. Constraint 2c ensures that RBs allocated to a v in c are enough to attain the lowest bitrate of v. Solving *Problem 2* gives us the maximum achievable utility from a particular combination of eNB configurations of all the videos. We repeat this process for all combinations and choose the one with the highest utility as our final combination for eNB configurations.

Resource allocation and user grouping: With an eNB configuration chosen for each video, we backtrack to find the optimal resource allocation and the user grouping. We solve the resource allocation problem (Equation 2) one more time for the chosen combination of eNB configurations, but instead of using the quadratic regression, we use the actual discrete utility values. This gives us the optimal RB share for users in all clusters subscribed to each video. We pass this as an input to user grouping algorithm (Algorithm 1) to define our user groups, their allocated RBs and assigned bitrates. MCE can now configure the eNBs to create video-specific SFN clusters and transmit different bitrates with the chosen MCS values on a certain set of RBs. The total complexity of RTOP is $\mathcal{O}(|V|\cdot|B|\cdot|T'|\cdot|R_v|^2 + |B|^{|V|})$.

Algorithm 2 Complete RTOP algorithm

Input: S_{vpc}: CQI Groups in cluster c of eNB configuration p for video v; See Table I for all other inputs

Output: $Groups$ with no. of users and RBs for each bitrate

for $v \in V$ **do**
  for $p \in P$ **do**
    $U[v,n,p] \leftarrow 0$ ▷ *n is # of clusters in p (Figure 3)*
    for $c \in p$ **do** ▷ *For each cluster in p*
      $rb \leftarrow$ *min(αT, T-Y_b* **for** $b \in c$*)* ▷ *Available RBs*
      $utility, \_ \leftarrow$ **Algorithm 1**(rb, R_v, S_{vpc})
      $U[v,n,p]$ += $utility$
    if $|V| == 1$ *and* $U[v,n,p] > U_{max}$ **then**
      $U_{max} = U[v,n,p]$; $Optimal\_Solution = (p)$
    else if $U[v,n,p] > maxU[v,n]$ **then**
      $maxU[v,n] \leftarrow U[v,n,p]$
      $Candidates[v,n] \leftarrow p$ ▷ *For class n*
  if $|V| == 1$ **then** *go to* ***Line*** **24** ▷ *Only one video*
  for $p \in Candidates[v]$ **do** ▷ *eNB configurations for v*
    for $c \in p$ **do** ▷ *Get Graphs for each cluster in p*
      $rb \leftarrow$ *min(αT, T-Y_b* **for** $b \in c$*)* ▷ *Available RBs*
      $Graph[v,p,c] \leftarrow$ **RBGRAPH**(R_v, S_{vpc}, rb)
$Cartesian \leftarrow (p_1, .., p_v) \mid p_v \in Candidates[v]$ **for** $v \in V$
$U_{max} \leftarrow 0$
for $solution \in Cartesian$ **do** ▷ *A combination of p's*
  Get Utility U with regression from Equation 2
  if $U > U_{max}$ **then**
    $Optimal\_Solution \leftarrow solution; U_{max} \leftarrow U$
Get Optimal_RBs for each cluster in Optimal_Solution without regression (Section III-B)
for $v \in V$ **do** ▷ *Backtrack to find the optimal user groups*
  for $c, rb \in Optimal\_RBs[v]$ **do**
    $\_, Groups[v,c] \leftarrow$ **Algorithm 1**(R_v, S_{vpc}, rb)

RBGRAPH$(R_v, S, maxRBs)$ ▷ $S \rightarrow$ *CQI groups*
for $rb \leftarrow 1$ *to* $maxRBs$ **do**
  $Utilities[rbs], \_ \leftarrow$ **Algorithm 1**(rbs, R_v, S)
return $Utilities$

IV. Performance Evaluation

In this section, we first present our simulation setup followed by our performance evaluation results.

A. Simulation Setup

We consider an eMBMS service area consisting of a numbers of eNBs arranged in a hexagonal grid. Users are distributed normally or uniformly in the service area. We apply the commonly used parameters [10][12] to our simulation setup as listed in Table II. A normal distribution represents cases such as sporting events or concerts where most of the users are located in the center of the service area. A uniform distribution represents cases such as shopping malls where users are evenly located across the service area.

Users calculate Reference Signals Received Power (RSRP) from each eNB, measure the achievable SINR from various possible clusters and report the best CQI for each cluster based on AWGN BLER vs SINR curves [13] with 1% error margin. The target BLER is 1% in eMBMS [10], contrary to 10% in unicast, as there are no physical layer re-transmissions.

For the system utility, we use Proportional Fairness (PF), which is defined as the sum-log of rates assigned to all the users. PF is a widely-used utility [7][8][9] for measuring system fairness and efficacy. Based on PF-utility, we compare the performance of the following approaches:

Optimal: Optimal results of our optimization model that considers both SFN clustering and user grouping. We use Gurobi Optimizer[5] to solve the model.

BoLTE [7]: Creates SFN clusters to maximize PF-utility but does not consider user grouping. The proposed algorithm assumes single bitrate per video. For fair comparison, we use the same heuristics but calculate the utility of an eNB configuration by assigning the best achievable bitrate in each of its clusters.

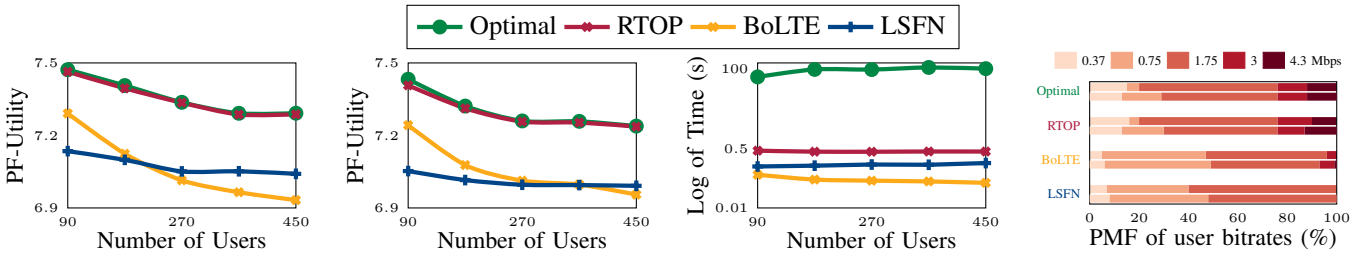
Variable Groups (**VG**) [8]: Creates user groups to maximize the PF-utility but does not consider SFN clustering. Also, the proposed algorithm solves the resource allocation problem for only a single video.

One Large SFN (**LSFN**): A scheme that considers only user grouping (no SFN clustering) based on our optimization model and constraints. We replace VG with LSFN when analyzing scenarios with multiple videos.

RTOP: Our proposed heuristics.

We evaluate the performance of these strategies in two key scenarios: A generic scenario with multiple videos and a mega event scenario with one video. In both cases, videos are encoded at five different bitrates, as shown in Table II. For

[5]http://www.gurobi.com/



(a) Utility for Normal Distribution (b) Utility for Uniform Distribution (c) Solution Time for Uniform User Distribution (d) PMF for Normal (top) and Uniform (below) distributions

Fig. 6: Results for generic scenario with multiple videos.

mega event scenario, we also look at the impact of varying resources available for eMBMS.

Our key performance metrics include the **PF-Utility** of the system, Probability Mass Function (**PMF**) of the bitrates assigned to users, **Degraded Users** (i.e. users with throughput less than the lowest available bitrate) and the **solution time** taken by an algorithm to perform resource allocation. We analyze the impact of various network parameters and algorithms on these metrics. For each configuration, we repeated the experiment 25 times by varying the user topology. The presented metrics are based on the average of these runs.

B. Generic Scenario: Multiple Videos

We present results for 3 videos and 4 eNBs in the service area, where 1 eNB has higher unicast load (75 RBs) than other 3 eNBs (50 RBs). We have tested networks with 2, 3 or 5 eNBs as well, and our conclusions are consistent among the tested scenarios. Since VG cannot handle multiple videos, we consider LSFN as a reference to analyze the benefits of SFN clustering in our scheme.

1) System utility: Figures 6a and 6b plot the system utility for normal and uniform user distribution, respectively. The figures illustrate that our proposed heuristics achieved optimal or near-optimal utility with a gap less than 1%. Additionally, the figures show that in comparison to LSFN and BoLTE, RTOP improves the utility by up to 8% which is achieved by an increase of average bitrate by up to 50%. Note that as LSFN does not consider SFN clustering, the available RBs for eMBMS were restricted by the eNB with least resources and all the users had to be satisfied with these RBs. On the other hand, BoLTE lacks user grouping and could not assign rates to users commensurate to their channel conditions.

For small number of users, a slight gap between RTOP and optimal results can be noticed, especially for uniform user distribution (Figure 6b). This is because, in such cases, the difference between the achievable utility from different eNB configuration can be very low and might not be detected by RTOP, as it uses heuristics and quadratic regression for faster approximation. However the difference in utility was marginal and solutions chosen by RTOP were always within a 1% gap from the optimal solution.

2) PMF of Assigned bitrates: Figure 6d shows the PMF for normal and uniform user distributions. RTOP assigned bitrates in almost the same manner as optimal results, unlike BoLTE or LSFN that allocated lower bitrates to users. The figure shows that RTOP assigned bitrates to users proportionate to their channel conditions and hence almost 75% of users received a bitrate of 1.75 Mbps or higher. On the contrary, this ratio dropped to around 50% for both BoLTE and LSFN.

TABLE II: Simulation Parameters

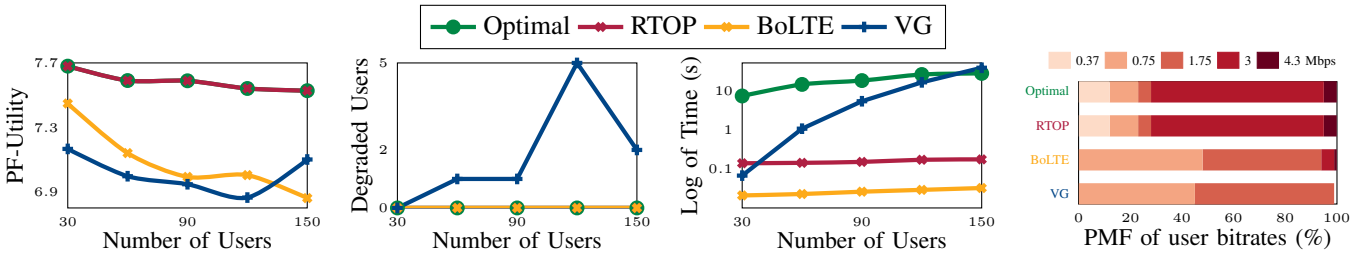
Parameter	**Value**
Cellular Layout	Hexagonal grid with up to 5 eNBs
Cell radius	500 m
eNB Tx Power	20 Watts
Carrier Frequency	2.1 GHz
System Bandwidth	20 MHz
Number of RBs in 20MHz	100
Path Loss Model	Log-Normal Shadowing n=4 (Urban)
White Noise Power Density	-174 dBm/Hz
User UE Noise Figure	7 dB
Number of Users per Video	30 to 150
DASH Video bitrates [14]	375, 750, 1750, 3000 and 4300 kbps
Channel Model	Multi-path Fading AWGN [13]
Spectral Efficiencies (bits/RB) from CQIs 1 to 15	[20, 31, 50, 79, 116, 155, 195, 253, 318, 360, 439, 515, 597, 675, 733]
Simulation Laptop Specs.	Dual Core Intel i7-5500U, 16GB RAM

3) Solution Time: Figure 6c plots the time taken to compute the final solution for uniform user distribution. Solving the problem optimally took around 100 seconds which is practically infeasible to implement in a real-time network. RTOP was able to consistently solve the same problem in 500ms, which is well within the limits of expected time constraints [15]. BoLTE and LSFN solved the problem faster but at a much lower utility as indicated above.

C. Mega Event Scenarios

A mega event refers to highly popular live events, such as world cup finals. In this section, we analyze the performance of RTOP for a mega event that is transmitted to users distributed normally in a 5-eNB service area where 1 eNB has a higher unicast load (90 RBs) than the other 4 eNBs (75 RBs).

1) System utility: Figure 7a shows the system utility of various algorithms. Similar to the generic scenario (Section IV-B), RTOP outperformed VG and BoLTE by increasing the utility up to 14% and average user bitrate up to 90%. Moreover, the solution of RTOP was identical to the optimal case. This is because the video had access to all the RBs available for eMBMS and RTOP did not need to perform regression for faster approximation.



(a) System Utility Per User (b) Number of Degraded Users (c) Solution Time (d) PMF of user bitrates

Fig. 7: Results for Mega Event Scenario.

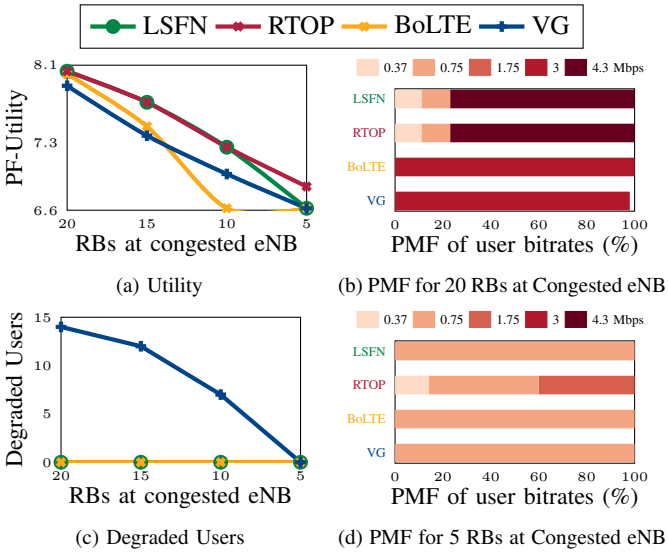
2) Degraded Users: While VG places each user in a group, it does not ensure that each group gets enough RBs to achieve at least the minimum bitrate. Hence, users with very low throughput are likely to experience frame skipping and video stalls. Figure 7b plots the number of such degraded users. With VG, in some cases up to 4% users were degraded. Our optimization model and RTOP define a constraint on the minimum allocated throughput to ensure a smooth streaming experience. BoLTE does not explicitly define such a constraint, however our implementation of BoLTE assigns the best possible bitrate based on user-throughput in a cluster, and hence users can get lower bitrates if throughput is low, therefore no users were degraded.

3) PMF of Assigned bitrates: Figure 7d shows the PMF of user bitrates. As the users were distributed normally, most of the users were in the center of the service area where the channel conditions were good. However, the number of users that received high bitrates of 3 or 4.3 Mbps was only 6% with BoLTE and 0% with VG, which was unfair to users with good channel conditions. With RTOP this ratio increased to 72% leading to the maximum utility.

4) Solution Time: Figure 7c plots the time taken to compute the final solution. As the number of users increased, the computation time for VG increased exponentially, which makes it unsuitable for mega events with large number of users. Our optimization model is independent of number of users and instead depends on number of CQI groups, which is usually limited to 15 values in LTE. Solving the model optimally still took more than 10 seconds which is not ideal for dynamic operating conditions. RTOP solved the same problem within 200ms for any number of users in the network. Hence, RTOP can be used in a dynamic network to accommodate changes in the network and solve the resource allocation problem in real-time.

D. Impact of Available Resources on Various Metrics

In this section, we explore the impact of available resources at eNBs on the performance in case of a mega event. As the RBs available to eMBMS decrease, subject to their channel conditions, the number of users receiving high bitrates decreases. We consider a service area comprising of 5 eNBs with 20 RBs for eMBMS. We analyze the impact of decreasing available RBs at one eNB when there are 1000 users normally distributed in the service area. Since RTOP always achieves the same result as the optimization model, we omit the optimization model results in the interest of space.



(a) Utility (b) PMF for 20 RBs at Congested eNB

(c) Degraded Users (d) PMF for 5 RBs at Congested eNB

Fig. 8: Impact of RBs available for eMBMS.

When there were 20 RBs available at each eNB, the best decision was to place all eNBs in one cluster and split users in two groups. LSFN and RTOP achieved maximum utility (Figure 8a) by making the right decisions. As VG is agnostic to video bitrates, it maximizes the system utility without assigning enough RBs to the user group with bad channel conditions. Hence, VG resulted in some degraded users (Figure 8c). BoLTE also achieved lower utility than RTOP as it is incapable of creating user groups and placed all users in one group.

As the available RBs on the highly congested eNB decreased, the utility started dropping. Until this eNB had 10 RBs available for eMBMS, the best decision was to keep all eNBs in one cluster and hence LSFN and RTOP achieved the same utility by creating one cluster and grouping users based on their channel conditions. Although VG can also create user groups, most of the times it achieved a lower utility because the constraint of serving all the users with a bitrate was not respected. BoLTE is unable to create user groups and hence achieved lower utility as well.

As the available RBs decreased further, separating the highly congested eNB in a second SFN cluster was the better decision. RTOP made this decision and achieved optimal

utility but LSFN being unable to create SFN clusters, kept all eNBs synchronized and achieved lower utility than RTOP. RTOP outperformed VG and BoLTE in these cases as well (Figure 8d).

V. Related Work

A large amount of work focuses on using multicast at the network [1] or application [16] layers. Approaches such as mCast [17] reduce resource consumption in the wired network including backbone, core and wired access network. However, the last hop for cellular networks is wireless, where the spectrum is scarce and higher layer multicast gets converted back to unicast, resulting in redundant transmissions and wastage of physical resources. User grouping schemes [9] [18] focused on wired medium do not consider resource allocation in the wireless network.

eMBMS enables multicast at the physical and link-layer of cellular networks by configuring users to receive video content over shared wireless resources. Research has been conducted to incorporate forward-error correction in eMBMS [19] at the application layer to improve reliability in the absence of detailed user feedback. However, issues such as efficiently allocating resources at the physical layer are not addressed.

Some researchers have proposed multicast resource allocation techniques for cellular networks. Authors of [20] propose using users with good channel conditions as relays for users with bad channel conditions. Such methods are not realistic due to the greedy nature of users and low-latency requirements of live streaming. Muvi [6] uses scalable video coding in attempt to maximize the utility for multicast users but does not consider the impact on unicast users. [21] solves user grouping problem for a single-cell network and assigns an MCS value to each group with the goal to maximize proportional fairness, but sometimes places users in groups where the MCS value is higher than what they can decode.

Authors of [8] propose an optimization model that considers grouping users based on their channel conditions while considering the impact on unicast users. Although this approach ensures that all users are assigned some resources, it does not guarantee that those resources are enough to achieve at least the minimum bitrate of the transmitted video. Also, the model does not consider the presence of multiple videos or the possibility of SFN clustering.

In [12], the authors evaluate how the number of eNBs in an SFN cluster affect the eMBMS service, but they do not propose any solution for determining the best eNB configuration. BoLTE [7] addresses the SFN clustering problem for multiple broadcast sessions, however does not explore the possibility of grouping users based on their channel conditions, which is unfair to users with good conditions and limits the achievable utility.

VI. Conclusion

In this paper, we considered the joint optimization of user grouping, SFN clustering and resource allocation in an eMBMS network. We developed an efficient and scalable heuristic-based algorithm, RTOP, that finds optimal or near-optimal results in real-time with no more than a 1% gap from the optimal solution. Additionally, we compare RTOP with state-of-the art techniques using extensive simulations. In situations where other approaches could assign high bitrates to less than 10% users, RTOP was able to assign high bitrates to 75% of the users. We show that overall our approach improves the system utility by up to 14% and the average user bitrates by up to 90% while avoiding service degradation for the users.

References

[1] L. Lao, J. Cui, M. Gerla, and D. Maggiorini, "A comparative study of multicast protocols: top, bottom, or in the middle," *Joint Conference of the IEEE Computer and Communications Societies*, March 2005.

[2] GSA, "LTE Broadcast (eMBMS) Market Update," Global Mobile Suppliers, Tech. Rep. March, 2018. [Online]. Available: goo.gl/UQ5g61

[3] 3GPP, "Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description," Tech. Rep. 23.246, 2015. [Online]. Available: www.3gpp.org/DynaReport/23246.htm

[4] C. Borgiattino, C. Casetti, C. Chiasserini, and F. Malandrino, "Efficient area formation for LTE broadcasting," *IEEE International Conference on Sensing, Communication, and Networking (SECON)*, June 2015.

[5] R. Afolabi, A. Dadlani, and K. Kim, "Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, First 2013.

[6] J. Yoon, H. Zhang, S. Banerjee, and S. Rangarajan, "Muvi: A multicast video delivery scheme for 4G cellular networks," *ACM MobiCom*, 2012.

[7] R. Sivaraj, M. Arslan, K. Sundaresan, S. Rangarajan, and P. Mohapatra, "BoLTE: Efficient network-wide LTE broadcasting," *IEEE 25th International Conference on Network Protocols (ICNP)*, April 2017.

[8] J. Chen, M. Chiang, J. Erman, G. Li, K. Ramakrishnan, and R. Sinha, "Fair and optimal resource allocation for LTE multicast (eMBMS): Group partitioning and dynamics," *IEEE Conference on Computer Communications (INFOCOM)*, October 2015.

[9] Y. Yang, M. Kim, and S. Lam, "Optimal partitioning of multicast receivers," *ICNP*, November 2000.

[10] EBU, "Delivery of broadcast content over LTE networks," European Broadcasting Union, Tech. Rep. TR 027, 2014. [Online]. Available: https://tech.ebu.ch/docs/techreports/tr027.pdf

[11] S. A. Vavasis, "Quadratic programming is in NP," *Information Processing Letters*, vol. 36, no. 2, October 1990.

[12] J. Monserrat, J. Calabuig, A. Aguilella, and D. Barquero, "Joint Delivery of Unicast and E-MBMS Services in LTE Networks," *IEEE Transactions on Broadcasting*, vol. 58, no. 2, April 2012.

[13] J. Ikuno, M. Wrulich, and M. Rupp, "System level simulation of LTE networks," *IEEE 71st Vehicular Technology Conference*, May 2010.

[14] J. Quinlan, A. Zahran, and C. Sreenan, "Datasets for AVC (H.264) and HEVC (H.265) Evaluation of Dynamic Adaptive Streaming over HTTP (DASH)," *ACM Multimedia Systems Conference*, May 2016.

[15] Y. Bejerano, C. Raman, C. Yu, V. Gupta, C. Gutterman, T. Young, H. Infante, Y. Abdelmalek, and G. Zussman, "DyMo: Dynamic monitoring of large scale LTE-Multicast systems," *IEEE INFOCOM*, May 2017.

[16] M. Hosseini, D. T. Ahmed, S. Shirmohammadi, and N. D. Georganas, "A Survey of Application-Layer Multicast Protocols," *IEEE Communications Surveys & Tutorials*, vol. 9, no. 3, September 2007.

[17] A. Khalid, A. Zahran, and C. Sreenan, "mCast: An SDN-Based Resource-Efficient Live Video Streaming Architecture with ISP-CDN Collaboration," *Local Computer Networks (LCN)*, Oct. 2017.

[18] J. Chuang and M. Sirbu, "Pricing multicast communication: A cost-based approach," *Telecommunications Systems*, vol. 17, no. 3, July 2001.

[19] T. Mladenov, S. Nooshabadi, and K. Kim, "Efficient Incremental Raptor Decoding Over BEC for 3GPP MBMS and DVB IP-Datacast Services," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, June 2011.

[20] F. Hou, L. Cai, P. Ho, X. Shen, and J. Zhang, "A cooperative multicast scheduling scheme for multimedia services in IEEE 802.16 networks," *IEEE Trans. on Wireless Communications*, vol. 8, no. 3, March 2009.

[21] H. Won, H. Cai, D. Eun, K. Guo, A. Netravali, I. Rhee, and K. Sabnani, "Multicast scheduling in cellular data networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, September 2009.