Fair and Optimal Resource Allocation for LTE Multicast (eMBMS): Group Partitioning and Dynamics*

Jiasi Chen†1 Mung Chiang† Jeffrey Erman‡ Guangzhi Li‡ K.K. Ramakrishnan§1 Rakesh K. Sinha‡
†Princeton University ‡AT&T Labs §University of California, Riverside
Princeton, NJ 08544 Bedminster, NJ 07921 Riverside, CA 92521

***Abstract*—With recent standardization and deployment of LTE eMBMS, cellular multicast is gaining traction as a method of efficiently using wireless spectrum to deliver large amounts of multimedia data to multiple cell sites. Cellular operators still seek methods of performing optimal resource allocation in eMBMS based on a complete understanding of the complex interactions among a number of mechanisms: the multicast coding scheme, the resources allocated to unicast users and their scheduling at the base stations, the resources allocated to a multicast group to satisfy the user experience of its members, and the number of groups and their membership, all of which we consider in this work. We determine the optimal allocation of wireless resources for users to maximize proportional fair utility. To handle the heterogeneity of user channel conditions, we efficiently and optimally partition multicast users into groups so that users with good signal strength do not suffer by being grouped together with users of poor signal strength. Numerical simulations are performed to compare our scheme to practical heuristics and state-of-the-art schemes. We demonstrate the tradeoff between improving unicast user rates and improving spectrum efficiency through multicast. Finally, we analyze the interaction between the globally fair solution and individual user's desire to maximize its rate. We show that even if the user deviates from the global solution in a number of scenarios, we can bound the number of selfish users that will choose to deviate.**

I. Introduction

The 3GPP standard specifies Evolved Multimedia Broadcast Multicast Service (eMBMS) for multicast over LTE networks [1]. eMBMS is particularly effective when requests for content have significant spatial and temporal locality, as is likely to be the case when users consume live video and music as well as popular on-demand content including news, advertisements and software distribution, in addition to video and music [2], [3].

Content delivery over cellular networks has been a challenge, with the last hop RAN being the bottleneck and a significant source of latency. CDNs relieve the backbone bandwidth utilization and reduce latency to deliver content to the end-user, but only provide a relatively small amount of relief. Similarly, traditional multicast solutions typically save resource consumption in the backbone (i.e., further up the multicast tree) [4]. With multicast on the cellular "last mile," the improvement in last-hop bottleneck utilization, end-to-end latency, and the cost of deploying resources to deliver content over cellular networks may be more effectively addressed.

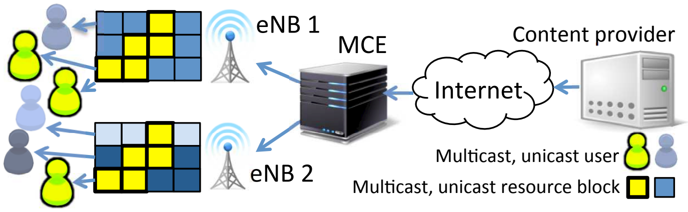


Fig. 1: eMBMS architecture. The Multicast Coordination Entity (MCE) reserves resources for multicast users in the eMBMS service area, which may encompass more than one eNB. The eNBs schedule unicast users in the remaining resource blocks.

Cellular network operators are continuing to evolve their strategies for deploying eMBMS. They also seek to understand how best to utilize scarce wireless resources, especially when there are users with heterogeneous channel conditions, and in the presence of a number of users consuming unicast traffic. These unicast users may be serviced by different eNBs, as shown in Fig. 1, due to the envisaged large service area of eMBMS. (1). The network operator must decide how many resources to allocate to each unicast user at each eNB and to the multicast group(s). Allocating more resources for multicast has the potential to reduce the quality of experience (QoE) of a unicast user, while allocating more resources for unicast reduces the efficiency with which wireless spectrum is utilized. (2). The network operator must decide how many multicast groups to offer, and which users interested in multicast content should band together to form a group. Grouping users with disparate channel conditions into one multicast group is unfair to the users with good channel conditions, as all group members will be constrained by the receivable rate of the worst user in the group (to ensure all users receive the transmitted data with high probability). Splintering potential multicast users into a large number of individual groups is similar to unicast and fails to take advantage of eMBMS.

Fig. 2 shows a toy example of our problem for three potential multicast users (A,B,C) and one unicast user (D) who are served by one base station with a total of 12 resource blocks (RBs). User (A:1) means user A has coding scheme that achieves 1 bit/RB. The user grouping is shown by the red boxes. The user rate is calculated as the coding scheme × number of RBs. *Case (a)*: All users are unicast and are allocated an equal share of resources, receiving rates proportional to their coding schemes. *Case (b)*: With eMBMS, we could group all multicast users together and combine their resources. This causes user A's rate to increase because she

*The authors are in alphabetical order except for the 1st author.

[1] This work was done when Jiasi Chen and K.K. Ramakrishnan were at AT&T Labs.

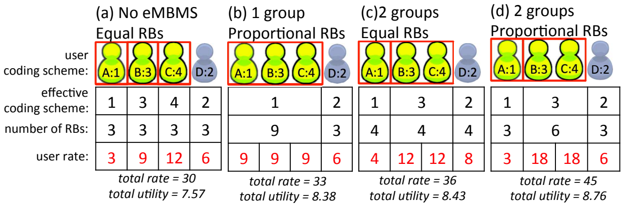


Fig. 2: Toy example. The multicast resource blocks (RBs), coding scheme, and user grouping must be jointly optimized to achieve maximum utility. In the optimal solution of case (d), all of the multicast users achieve equal or better rate compared to the unicast case (a), so they will not defect from the multicast group, but this may not be generally true.

shares resources with the group, but user C's rate suffers because she is forced to use the same coding scheme as user A. Were this the final solution, user C would prefer to leave her multicast group and switch to unicast, as in case (a), in order to receive 12 bits instead of 9. *Case (c)*: Separating users B and C, who have similar channel conditions, from user A and putting them in their own sub-group improves the utility. *Case (d)*: Allocating more RBs for multicast increases system efficiency in terms of total rate and utility. Both case (c) and (d) ensure that all of the multicast users receive higher rate than if they switched to unicast in case (a).

Thus, a user's rate is determined by the number of RBs (variable), her own coding scheme (fixed, based on her channel conditions), the user grouping (variable), and the group coding scheme (variable). To capture both fairness of the user rates and efficiency of the system, we choose the *proportional fair metric* [5] to measure global utility, which is defined as the sum of the log of the long-term rates of the users. We consider a set of multicast users who are interested in accessing one particular multicast content. **We jointly optimize the multicast resources and user grouping to maximize the proportional fair utility for multicast and unicast users across multiple cells in the eMBMS service area. For a single cell, we also analyze the disparity between the globally optimal solution and the local preference of the greedy user who is trying to maximize her own rate.**

Our system is designed to meet several goals:

- **Constraint:** Users subscribing to a multicast group should receive all the content multicast by eMBMS. This might not happen if, for example, eMBMS chooses too good a coding scheme. Then, a user with poor channel conditions will experience a higher bit-error rate and consequent data loss, which is especially intolerable for streaming content such as music and video.
- **Fair and efficient:** The received rate across all users in all cells covered by the eMBMS service area should be fairly distributed; i.e., a user with very good channel conditions should not consume all of the resources to the detriment of users with poor channel conditions. Conversely, users with good channel conditions should not suffer unduly by being placed in a multicast group that receives a very poor rate. We quantify this fairness using the proportional fair metric [5].
- **Dynamics**: The users should be satisfied with their rates under the globally fair resource allocation and user grouping decided by the network operator. Few users should be tempted to selfishly maximize their own rate by leaving their multicast group, thereby decreasing the rate of the remaining users in the group; or vice versa.

We further face several challenges in designing the optimization framework: (1) The optimization problem is an integer problem with non-linear constraints, for which brute-force approaches require exponential time to solve; (2) The network operator should encourage users to follow the optimal solution by offering them a better receive rate and overall improved resource usage, rather than forcing users to join or leave a multicast group; (3) The eMBMS architecture only allows for multicast resource allocation and must respect the unicast resource allocation decisions already made by the eNB.

- We pose the resource allocation problem that maximizes the end-user's QoE based on the user rate, in accordance with eMBMS specifications and the existing LTE architecture (§II). We develop an optimal and efficient algorithm that decides on the resource allocation for the multicast group and the unicast users across multiple cell sites, and optimally groups together multicast users with similar channel conditions. We propose a weighting function that trades-off between efficiency and fairness of the rate received by multicast and unicast users. (§III).
- We present detailed simulations across a variety of different cellular conditions, and compare our solution with three alternatives: (a) a default solution that puts all users into one multicast group and chooses the worst coding scheme; (b) a more advanced heuristic that takes users' channel conditions into consideration and assigns users to one of four groups; and (c) a state-of-the-art approach from the literature. We show that our solution performs well across a variety of scenarios and automatically adapts to the channel conditions of the users (§V).
- Our analysis provides insights on the interaction between the user's selfish rate maximization and the globally optimal solution. We characterize the channel conditions that cause users to deviate from the globally optimal solution, and bound the number of users who would benefit by switching from multicast to unicast. We numerically quantify the tradeoff between giving enough resources to unicast vs. over-provisioning resources for multicast users, to prevent them from fleeing the group. (§IV).

II. Background and Problem Formulation

A. Background on LTE and eMBMS

In LTE, wireless radio resources are OFDMA frames. Frames are further divided into *resource blocks* (RBs) in the time and frequency domains, which are the basic minimum unit of resource allocation (Fig. 1). For unicast users, the job of the eNodeB (LTE base station, hereafter shortened to eNB) scheduler is to fill the RBs with data packets for its clients, often through some variant of a proportional fair scheduler [6].

	Symbol	Description
Inputs	M	number of users interested in multicast service
	B	number of eNBs in the MBSFN area
	U_b	set of unicast users at eNB b
	$N = \sum_{b=1}^{B} \lvert U_b \rvert$	total number of unicast users in the MBSFN area
	c_i	coding scheme of multicast user i (bits / RB)
	d_i	coding scheme of unicast user i (bits / RB)
	T	total number of resource blocks (RBs)
	α	maximum fraction of resources for multicast
Variables	K	number of multicast groups
	G_k	set of users in multicast group k
	$\hat{c}_k$	coding scheme of multicast group k (bits / RB)
	x_k	RBs of multicast group k (RBs)
	y_i	RBs of unicast user i (RBs)

TABLE I: Table of notation.

Since users have different channel conditions, the eNB chooses an appropriate *module and coding scheme (MCS)*, hereafter referred to as coding scheme, for each client, based on client feedback of the measured channel conditions. The coding scheme determines how many bits can be transmitted per resource block. Users with good signal strength can use more efficient coding schemes, and vice versa. If the eNB chooses a coding scheme that is higher than what the client's channel can support, then the client will not be able to decode the data and will suffer data loss due to high bit error rates.

In eMBMS, each multicast group uses resource blocks encoded with the same coding scheme. All users in a particular group receive the same data transmission. The multicast resource allocation and coding scheme are decided by the multi-cell/multicast coordinating entity (MCE) shown in Fig. 1 [1]. The typical envisaged eMBMS deployment covers multiple neighboring eNBs, called a multicast-broadcast single-frequency network (MBSFN), so as to exploit multicast across a sufficient number of recipients and improve QoE at the cell edge. The MCE provides the multicast resource allocation to the eNBs under its control. Each eNB reserves these resources for the multicast flows, and schedules its own unicast users with the remaining resources. Therefore, the resource block allocation for each eNB will be different depending on the unicast users, but the multicast resource block allocation will be the same across all eNBs.

B. Problem Formulation

Our problem is a mixed integer optimization problem with two parts. Prob. 1 is the global rate optimization for both multicast and unicast users. Prob. 2 is the eNB resource allocator at each eNB for the unicast users. These two problems are related because the global optimization should consider the impact of the multicast resource allocation on the unicast users. However, the eMBMS architecture does not provide the flexibility to modify the unicast scheduler, so we cannot directly control the unicast resource allocation. Instead, by understanding how the multicast resource allocation impacts unicast scheduling, we can optimize for both sets of users.

Optimizing multicast resources and user grouping: The control knobs are (a) to decide which users, out of those who are interested in a particular multicast content, should be put into the same multicast group, and (b) how many resources to give to each multicast group and to the remaining unicast users, in order to maximize the proportional fair rate for all users. Prob. 1 gives the complete optimization problem.

$$\begin{tikzcd}
 & & \{y_i\}, i \in U_1 \\
K, \{G_k\}, \{\hat{c}_k\} \arrow[r] & \{x_k\} \arrow[ur] \arrow[r] \arrow[dr] & \{y_i\}, i \in U_2 \\
 & & \{y_i\}, i \in U_B
\end{tikzcd}$$

Fig. 3: For each $K, \{G_k\}, \{\hat{c}_k\}$, we solve for the optimal $\{x_k\}, \{y_i\}$, calculate the total utility, and pick the best $K, \{G_k\}, \{\hat{c}_k\}$. We show our final solution is jointly optimal.

Problem 1. *Multicast and unicast resource allocator and multicast group partitioning for multiple eNBs*

$$\begin{aligned}
&\textit{maximize} && \sum_{b=1}^{B} \sum_{i \in U_b} \log(d_i y_i) + \sum_{k=1}^{K} |G_k| \log(\hat{c}_k x_k) && (1)\\
&\textit{s. t.} && \sum_{i \in U_b} y_i + \sum_{k=1}^{K} x_k \le T, \ \forall\, b && (2)\\
& && \hat{c}_k = \min_{i \in G_k} c_i, \ \forall\, k && (3)\\
& && x_k \le \frac{f(G_k)}{N + \sum_{\ell=1}^{K} f(G_\ell)} T, \ \forall\, k && (4)\\
& && 0 \le \sum_{k=1}^{K} x_k \le \alpha T && (5)\\
& && G_1 \cup \ldots \cup G_K = \{1, 2, \ldots, M\} && (6)\\
& && G_k \cap G_\ell = 0 \ \forall\, k, \ell && (7)\\
&\textit{variables} && \{y_i\}, \{x_k\}, \{G_k\}, \{\hat{c}_k\}, K &&
\end{aligned}$$

The objective (1) is to maximize the sum log-utility for multicast and unicast users. Constraint (2) says that the sum of the unicast and multicast RBs at each eNB must be less than the total RBs. Constraint (3) guarantees that all users in the group can receive all multicast data by setting the coding scheme of the multicast group to that of the worst user in the group. Constraint (5) limits the percentage of RBs available for multicast, which is 60% in today's eMBMS specification. Constraints (6-7) say that each multicast user is a member of a multicast group.

Multicast weighting function: A desirable feature of the multicast optimization is to ensure that users are allocated a proportionally fair rate, but also that both multicast and unicast users share the benefits of eMBMS. We introduce a weighting function $f(G_k)$ in constraint (4) that can be chosen to modulate the resource allocation of the multicast users to be more similar to either unicast or multicast. Specifically, we choose the weighting functions to be one of three classes:

$$\text{Constant: } f(G_k) = 1 \quad (8)$$

$$\text{Logarithmic:} f(G_k) = \log(|G_k| + 1) \quad (9)$$

$$\text{Linear: } f(G_k) = |G_k| \quad (10)$$

For the constant form $f(G_k) = 1$, the weighting function treats each multicast group similar to a single unicast user. This is because when all users are unicast, the regular proportional fair scheduler gives an equal share of the total available resources to each user. For the linear form $f(G_k) = |G_k|$, the resources given to the multicast group is similar to what resources each user in the group would have received as a unicast user and then pooled all their resources together. We also introduce the logarithmic form $f(G_k) = \log(|G_k| + 1)$ to be in between the linear and constant functions. This weighting function is a tunable parameter for the network operator.

eNB unicast resource allocator: In practice, the MCE cannot control the unicast RBs, which is a variable in Prob. 1. Instead, each eNB in the eMBMS service area schedules RBs for the unicast users in its cell. Any solution to Prob. 1 must take into account the behavior of the eNB's scheduler, which is based on the standard proportional fair scheduler [7]:

Problem 2. *eNB b's unicast resource allocator*

$$\begin{aligned} &\textit{maximize} && \sum_{i \in U_b} \log(d_i y_i) && (11) \\ &\textit{s.t.} && \sum_{i \in U_b} y_i \leq T - \sum_{k=1}^{K} x_k && (12) \\ &\textit{variables} && \{y_i\} \end{aligned}$$

The objective (11) is to maximize the proportional fair rate across all the unicast users. Constraint (12) says that the total unicast RBs cannot exceed the total RBs less the resources previously allocated by the MCE for multicast.

Time scale of optimization: When the channel conditions of the users change, or when users arrive and depart, the multicast resource allocation and user grouping previously chosen by the MCE may become sub-optimal. However, in the envisioned eMBMS use cases, such as sports events or stadiums with large audiences, users are fairly stationary, so the time scales of channel condition dynamics will be relatively long. We leave this as a design parameter for the network operator which can be chosen based on historical or expected user dynamics. It may be less feasible to make an optimal allocation in situations of high speed mobility, but we don't envisage the use of eMBMS such cases.

Multimedia content delivery: eMBMS is intended for multimedia streaming content such as video. Traditional videos require a fixed rate for content delivery, but with the prevalence of adaptive bitrate streaming protocols [8], content providers encode and store multiple versions of each video at different rates. The appropriate video rate should be selected to fully utilize the reserved multicast resources.

III. Solutions for Resource Allocation & Group Partitioning

The solution to Prob. 1 has two steps: (Step 1), an outer loop where we efficiently search across possible user groups, with associated optimal utility, and find the best user grouping, described in §III-B; (Step 2), an inner loop where for fixed user grouping, we solve for the optimal resource allocation and utility, described in §III-A. These two steps give us the jointly optimal solution to Prob. 1. Fig. 3 illustrates our approach, and the full algorithm is shown in Alg. 1. All proofs can be found in the technical report [9].

A. Resource Allocation

In this section, we discuss the inner loop of the solution to Prob. 1, where the multicast user grouping $G = \{G_1, \ldots, G_K\}$ of users is fixed (which also fixes the number of groups, K, and multicast group coding schemes $\{\hat{c}_k\}$ of all groups). Then the variables are the multicast and unicast RBs $\{x_k\}, \{y_i\}$. Under the eMBMS architecture, the unicast RBs $\{y_i\}$ cannot be controlled by the MCE; instead, we must determine the correct optimization problem to be solved by the MCE which, in combination with the eNB proportional fair scheduler, induces the solution to Prob. 1 for fixed user grouping. Therefore, we further decompose the resource allocation solution into two steps: (Step 2a) on the multicast MCE, optimize the multicast RBs (Prob. 3); and (Step 2b) on each eNB, optimize the unicast RBs (Prob. 2).

Problem 3. *MCE multicast resource allocator*

$$\begin{aligned} &\textit{maximize} && \sum_{b=1}^{B} \sum_{i \in U_b} \log(d_i y_i^*(\{x_k\}) + \sum_{k=1}^{K} |G_k| \log(\hat{c}_k x_k) && (13) \\ &\textit{s. t.} && \hat{c}_k = \min_{i \in G_k} c_i, \ \forall\, k && (14) \\ & && x_k \leq \frac{f(G_k)}{N + \sum_{\ell=1}^{K} f(G_\ell)} T, \ \forall\, k && (15) \\ & && 0 \leq \sum_{k=1}^{K} x_k \leq \alpha T && (16) \\ &\textit{variables} && \{x_k\} \end{aligned}$$

The MCE adjusts the multicast RBs, taking into account the impact on the unicast RBs. The objective function (13) is similar to the objective function (1) from Prob. 1, but contains $y_i^*(\{x_k\})$, which is the solution of Prob. 2. The constraints (14) 15) (16) are the same as (3) (4) (5) from Prob. 1.

At the eNB, the proportional fair scheduler allocates unicast RBs equally to each user, as given by Lemma 1. This means that users with good channel conditions receive higher rate and users with poor channel conditions receive lower rate.

Lemma 1. *The solution of Prob. 2 is:*

$$y_i^*(\mathbf{x}) = \frac{T - \sum_{k=1}^{K} x_k}{|U_b|}, \ \forall\, i \in U_b, \ \forall\, b$$

Finally, we show in Lemma 2 and Prop. 1 that the MCE optimization plus the eNB optimization together result in the jointly optimal solution to Prob. 1, for fixed user grouping.

Lemma 2. *Given $K, \{G_k\}, \{\hat{c}_k\}$, the feasible solution set of Prob. 1 is equal to the feasible solution set of Prob. 2 and Prob. 3.*

Proposition 1. *Given $K, \{G_k\}, \{\hat{c}_k\}$, the solution of Prob. 1:*

$$x_k^* = \begin{cases} \frac{f(G_k)}{N+\sum_l f(G_l)}T & \textit{if } \alpha \geq \frac{\sum_l f(l)}{N+\sum_l f(G_l)} \\ \min\left(\frac{|G_k|}{\frac{U}{(1-\alpha)T}+\lambda}, \frac{f(G_k)}{N+\sum_k f(G_k)}T\right) & \textit{if } \alpha < \frac{\sum_l f(l)}{N+\sum_l f(G_l)} \end{cases}$$

where λ satisfies $\sum_l \min\left(\frac{|G_l|}{\frac{U}{(1-\alpha)T}+\lambda}, \frac{f(G_l)}{N+\sum_k f(G_k)}T\right) = \alpha T$

The solution to the multicast resource allocation problem has two cases depending on the total fraction of resources α allowed for multicast. When α is large, then the solution gives resources proportional to the multicast constraint function $f(G_k)$. When α is small, i.e. the system wants to give more resources than allowed to multicast, the solution can be found by a variant of a water-filling algorithm. For the linear multicast weighting function $f(G_k) = |G_k|$ the solution can be further simplified as in Corollary 1. If α is large, the system allocates resources to each multicast group proportional to the size of the group. If α is small, the solution is to split αT RBs between the multicast groups, and leave the remaining RBs for unicast users. The overall resource allocation solution is computed by the ResourceMulti function in Alg. 1.

Corollary 1. *For fixed $K, \{G_k\}, \{\hat{c}_k\}$, and $f(G_k) = |G_k|$, the solution of Prob. 1 is:*

$$x_k^* = \begin{cases} \frac{|G_k|}{N+\sum_l |G_l|}T & \textit{if } \alpha \geq \frac{\sum_l |G_l|}{N+\sum_l |G_l|} \\ \frac{|G_k|}{\sum_l |G_l|}\alpha T & \textit{if } \alpha < \frac{\sum_l |G_l|}{N+\sum_l |G_l|} \end{cases} \tag{17}$$

B. Group Membership Assignment

Even with the partial solution to Prob. 1 given in Prop. 1, finding the user grouping is not straightforward. The system must decide how to optimally partition users into multicast groups based on their channel conditions. Intuitively, users with identical channel conditions should be grouped together, but what about users with similar channel conditions? The brute force solution is to try all possible partitions of the users and pick the partition that gives the maximum utility. The complexity of this approach is $\Theta(M^M)$. Instead, we present an optimal dynamic programming solution that runs in $O(M^4)$.

The dynamic program uses the intuition that users with similar channel conditions should be in the same group, and first sorts the multicast users by their coding schemes.

Lemma 3. *Let the users be sorted in ascending order according to coding scheme. Define an unordered grouping as a group that contains user i and user $i+j$ but not some user $i+k, 1 < k < j$. Then there is an optimal solution without any unordered grouping.*

With the list of users sorted by increasing channel conditions, we can think of the problem as placing partitions between users on the list, with users between two consecutive partitions forming a multicast group. The decision is where to place the partitions, and how many partitions to place. The problem is complicated because the coding scheme of the multicast group, and thus the rate and overall utility we are trying to maximize, is determined by the placement of the partitions. The user immediately to the right of a partition determines the coding scheme and utility of her multicast group. Exhaustively trying all possible partitions takes $\Theta(2^M)$.

Algorithm 1 Resource Allocation & User Partitioning

Global input: Number of unicast users N, Number of multicast users M, Multicast user channel conditions sorted in non-decreasing order $\hat{c}_1 \leq \ldots \leq \hat{c}_M$, Total resource blocks T, Fraction of resources blocks allowed for multicast α
Output: Number of multicast groups K^*, Multicast groups $\{G_k^*\}$, Multicast resources $\{x_k^*\}$

```
Partition
for $K \leftarrow 1$ to $M$ do                                ▷ $K$ total groups
    for $i \leftarrow 1$ to $1 - K + M$ do                        ▷ initialize
        $U(K, i, 1) = \textsc{Utility}(K, \{1, \ldots, i\})$
    for $k \leftarrow 2$ to $K$ do                             ▷ first $k$ groups
        for $i \leftarrow k$ to $k - K + M$ do                 ▷ first $i$ users
            for $j \leftarrow k - 1$ to $i - 1$ do ▷ partition at user $j$
                $G_k \leftarrow \{j+1, \ldots, i\}$
                $u \leftarrow U(K, j, k-1) + \textsc{Utility}(K, G_k)$
                if $u > U(K, i, k)$ then
                    $U(K, i, k) \leftarrow u$; $p(K, i, k) \leftarrow j$
    if $U(K, M, K) > U_{max}$ then ▷ check for max utility
        $U_{max} \leftarrow U(K, M, K)$; $K^* \leftarrow K$
$j \leftarrow M$                          ▷ backtrack to find optimal solution
for $k \leftarrow K_{opt}$ to 1 do
    $i \leftarrow j$; $j \leftarrow p(K, i, k)$; $G_k^* \leftarrow \{j+1, \ldots, i\}$
    $x_k^* \leftarrow \textsc{ResourceMulti}(K^*, G_k^*)$
```

Input: Number of groups K, Multicast group G_k
Output: Utility of group u_k

```
Utility$(K, G_k)$
$x_k \leftarrow \textsc{ResourceMulti}(K, G_k)$
$u_k \leftarrow |G_k| \log\left(\min_{i \in G_k} \{\hat{c}_i\} x_k\right)$
return $u_k$
```

Input: Number of groups K, Multicast group G_k
Output: Resources of group x_k

```
ResourceMulti$(K, G_k)$
if $f(G_k) = 1$ then $x_k \leftarrow \frac{1}{K+N}T$
if $f(G_k) = |G_k|$ then
    if $\alpha \geq \frac{M}{M+N}$ then $x_k \leftarrow \frac{|G_k|}{M+N}T$
    else $x_k \leftarrow \frac{|G_k|}{M}\alpha T$
return $x_k$
```

We can efficiently search through the possible partitions using the recurrence relation given in Prop. 2. The idea is that the maximum utility from k groups can be found by considering the optimal utility of $k-1$ groups, plus the utility of the new group formed by users at the end of the list.

Proposition 2. *When $f(G_k) = |G_k|$, $\forall \alpha$ or $f(G_k) = 1$, $\alpha \geq \frac{M}{N+M}$, the recurrence relation for the dynamic program is:*

$$U(g, i, k) = \max_{1 \leq j < i} \{U(K, j, k-1) + \textsc{Utility}(K, \{j+1, \ldots, i\})\}$$

where $U(K, i, k)$ is the sum utility from the first i users in the first k partitions, when there are K multicast groups.

Scenario	User switch	Better coding scheme	Analysis results
IV-A	Unicast → multicast	Unicast	User never wants to switch from the operator's solution
IV-B	Unicast → multicast	Multicast	User always wants to switch, but operator prevents
IV-C	Multicast → unicast	Unicast	Number of switching users is bounded
IV-D	Multicast → unicast	Multicast	User never wants to switch

TABLE II: Users may selfishly try to deviate from the network operator's recommended multicast grouping.

The final solution to Prob. 1 is given by Alg. 1 and is a combination of dynamic programming to find the user grouping and convex optimization to find the optimal resource allocation. PARTITION builds up a table of utility values $U(K, i, k)$ and selects the entry that gives the maximum utility value. UTILITY(K, G_k) gives the optimal utility obtained from users in group G_k in a multicast group, when they are assigned resources using RESOURCEMULTI based on Prop. 1. The running time is $O(M^4)$ when $f(G_k) = 1$ and $O(M^3)$ when $f(G_k) = |G_k|$.

IV. USERS' SELFISH SWITCHING BEHAVIORS

In §III, the network operator allocates spectrum resources to users to achieve a globally fair and efficient distribution of rates. However, this misses an important practical consideration: an individual user might selfishly maximize her rate by switching between multicast and unicast, deviating from the operator's desired solution. For example, if a user in a multicast group is forced to use a low coding scheme by the worst member of her multicast group, she might leave her multicast group to obtain a better rate on unicast. Similarly, a user interested in multicast content who has been placed in unicast might see that other multicast users who are grouped together are receiving more RBs, and try to join the group. The switching behavior is further complicated by the fact that if the user switches, the MCE re-allocates resources according to the new user groups created by the switch. The cellular operator should ensure that the user's rate under the globally optimal solution is high enough so that the user has no desire to switch from multicast to unicast, or vice versa.

We analyze all possible scenarios, as summarized in Table II. We use $f(G_k) = |G_k|$ and consider a single eNB. We assume that the multimedia content is available both through multicast and regular unicast data channels, and that the user can freely choose choose between them (possibly through a dedicated multicast mobile application). Our analysis shows that of the four scenarios, only one will result in users defecting from the globally optimal solution, and we bound the number of users who exhibit such undesirable behavior. We assume no collusion between users, so we focus on one particular user i and group k, and drop the subscripts from the user coding scheme c_i, group coding scheme $\hat{c}_k$, and group G_k for notational simplicity.

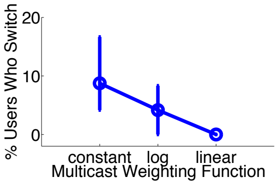


Fig. 4: Fewer users are incentivized to switch for the linear weighting function. Error bars indicate range of values.

A. Unicast user switches to multicast, worse coding scheme

In this scenario, the user has better coding scheme if she remains on unicast than if she joins the multicast group. What is the incentive for the user to switch? The user may wish to join the multicast group to take advantage of the greater amount of resources allocated to the multicast group. We will show that the global solution is such that this will never happen. The idea is to show that if the user would like to switch from unicast to a multicast group, the global solution would have already put the user in the multicast group. Lemma 4 gives the relationship between parameters that results in the user or network operator desiring the switch. Prop. 3 shows that this does not occur.

Lemma 4. *For $c > \hat{c}, f(G) = |G|$, a user will want to switch from unicast to multicast when $\frac{c}{\hat{c}} < m$. The global utility will increase from this switch when $\frac{c}{\hat{c}} < \left(1 + \frac{1}{|G|-1}\right)^{|G|-1} |G|$.*

Proposition 3. *For $c > \hat{c}, f(G) = |G|$, if the solution is globally optimal, there will be no unicast user who wants to switch to multicast.*

B. Unicast user switches to multicast, better coding scheme

A unicast user may wish to join a multicast group with a better coding scheme. Since the multicast group's coding scheme is set to the minimum coding scheme of its members, if the user joins the multicast group, the multicast group will be forced to lower its coding scheme. This will lead to decreased rate for the other users in the multicast group. The analysis is similar to §IV-A, which we omit due to space constraints, and says that user may indeed wish to deviate from the globally optimal solution. However, the network operator can avoid this undesirable scenario and maintain global fairness by either (a) refusing admission to the multicast group for the switching user, or (b) not lowering the coding scheme of the multicast group, rendering the user unable to decode the data and encouraging her to stay on unicast.

C. Multicast user switches to unicast, better coding scheme

This is arguably the most interesting scenario. The global solution may prefer a user to be in a multicast group, but the user would receive a better coding scheme and higher rate if she were to leave the multicast group and switch to unicast transmission. The situation is further complicated by the fact that if a few users leave the multicast group, other users in the group may receive reduced rate and also desire to switch, creating a "ripple" effect. However, we show that the number of potentially switching users per multicast group is bounded. We first give the conditions under which the user or network operator desires the switch:

Lemma 5. *For $\hat{c} < c, f(G) = |G|$, a user will want to switch from multicast to unicast when $\frac{c}{\hat{c}} > m$. The global utility will increase from this switch when $\frac{c}{\hat{c}} > \left(1 + \frac{1}{|G|-1}\right)^{|G|-1} |G|$.*

Prop. 4 bounds the number of switching users. The intuition is that users with better coding scheme are placed in the multicast group because even though they sacrifice their good coding scheme, by increasing the size of the multicast group, all the users in the multicast group benefit. However, if there are many users with good coding scheme, placing them in a separate multicast group with higher coding scheme outweighs the decrease in utility of the remaining users in the multicast group. So the optimal solution contains only contains a few users who might want to switch to unicast.

Lemma 6. *For $\hat{c} < c, 0 < \beta \leq 1$, if the solution is globally optimal, there are at most $\lceil \frac{e}{\beta} \rceil - 1$ users who have coding scheme $> \hat{c}\beta|G|$.*

Proposition 4. *For $\hat{c} < c$, if the solution is globally optimal, the number of users who switch is $\min n : n+1 = \left\lceil \frac{e}{1-\frac{n}{|G|}} \right\rceil$.*

For weighting functions other than $f(G) = |G|$, we numerically simulate the number of defecting users in Fig. 4. We find that no users tend to flee the multicast group when $f(G) = |G|$, which is better than our upper bound. The weighting functions $f(G) = 1$ and $f(G) = \log(|G|)$ result in less rate for multicast users, so the number of users wishing to leave the group is higher.

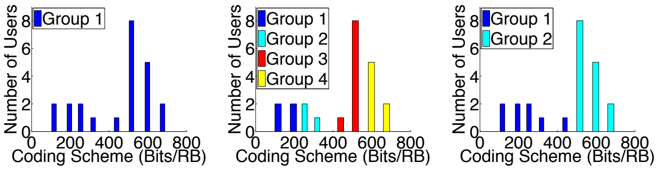
D. Multicast user switches to unicast, same coding scheme

The multicast user with the worst coding scheme in the group may consider switching out of the multicast group and going solo on unicast. We observe that the user always prefers to stay in the multicast group because she gets (a) more resources, and (b) same coding scheme, so she receives greater rate on multicast than unicast. The analysis is similar to §IV-C and is omitted due to space constraints.

V. Performance Evaluation

In this section, we compare our multicast resource allocation algorithm against several other approaches:

- **no-eMBMS (all unicast):** Without eMBMS service, all potential multicast users access the content via unicast.
- **1 Group (all multicast) (1G)**: Put all the multicast users into a single group, and use the worst user's coding scheme as the group coding scheme.
- **1 Group + Time-variable coding scheme (1G-V)**: Put all the multicast users into single group, and vary the multicast group's coding scheme over time [10]. The key difference from VG is to vary the coding scheme over *time* rather than *user partitions*.
- **4 Groups + Fixed coding scheme (4G)**: Put the multicast users into 4 groups by dividing the range of channel conditions into 4 equal bins, and place users in associated bins. We choose 4 groups as a heuristic [11], [12], but any fixed number of groups would have similar results.



(a) 1G, 1G-V: All users are placed in the same group.

(b) 4G: Users are divided into 4 groups based on their coding scheme.

(c) VG: Choose the intuitive grouping based on coding scheme similarity.

Fig. 6: For a fixed set of user channel conditions, each algorithm chooses different multicast user partitions.

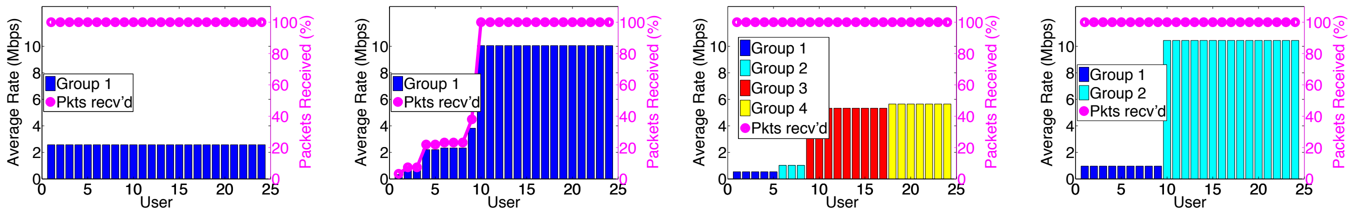
- **Variable Groups + Fixed coding scheme (VG)**: Our proposed scheme. We set $f(G_k) = |G_k|$ in §V-A and §V-B, but vary it in §V-C.

Simulation setup: We sweep across different parameters in §V-B and §V-C. The default parameters, unless otherwise stated, are 50 unicast users and 24 users at one eNB who are interested in one multicast content. 2/3 of the users have good channel conditions and can use a high-rate coding scheme, and 1/3 of the users have poor channel conditions and use a low-rate coding scheme. The discrete set of possible coding schemes is specified by LTE and maps to approximately $[20, 31, 50, 79, 116, 155, 195, 253, 318, 360, 439, 515, 597, 675, 733]$ bits/RB [1]. The distribution of bits/RB for the users is bi-modal normal with means 555 and 198 (1/4 and 3/4 of the range of coding schemes) and standard deviation 59 (to span the range of coding schemes).

A. Multicast grouping, user rate, and packet loss

Our algorithm chooses the optimal resource allocation, coding scheme, and user grouping to maximize the proportional fair metric. An example of the user grouping chosen by each algorithm is shown in Fig. 6. 1G and 1G-V put all the multicast users into one group, which does not take into account the heterogeneity of users' channel conditions, so 4G uses a heuristic to split the users into 4 groups. This sometimes forces *many* users with *better* channel conditions (e.g. coding scheme 438) to form a multicast group with a *few* users with *worse* channel conditions (e.g. coding scheme 515), which is sub-optimal. VG splits the users into two groups corresponding to the bi-modal distribution of channel conditions.

For the same experiment, we plot the rate of the multicast users in Fig. 5. 1G uses a fixed coding scheme for all users, so users with good channel conditions are forced to use the same coding scheme as users with poor channel conditions, leading to low rate for all users. Using heuristics to split the users into groups allows rate differentiation by 4G, but the proportional fair scheme of 1G-V can improve the throughput for all the users. However, users with poor channel conditions may be unable to decode transmitted data, which is unacceptable for streaming data such as video. VG, however, ensures that all users can receive all data and reasonable rate commensurate with their channel conditions. This is because VG is adaptive to the underlying channel conditions of the users, both in terms of number of users and their coding schemes.



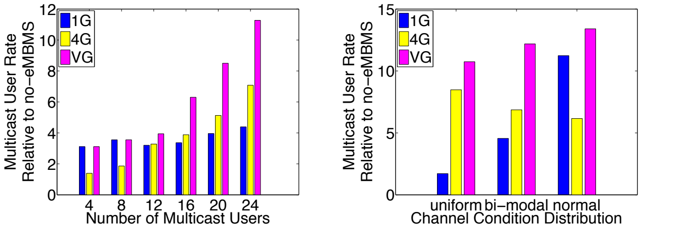
(a) 1G: All users receive the same low rate due to the poor channel conditions of the worst user in the group.

(b) 1G-V: Users with poor channel conditions suffer greatly by not receiving the majority of packets.

(c) 4G: User rate is commensurate with channel conditions, but suboptimal user grouping still lowers the rate of each group.

(d) VG: Users with good channel conditions receive a higher rate, and users with poor channel conditions receive reasonable rate.

Fig. 5: For the same setup and coloring as Fig. 6, we show the multicast user rate (bar plot) and the packet loss (line plot). VG gives users rate commensurate to their channel conditions, and ensures that all users receive all packets.



(a) Number of multicast users: with more multicast users, users pool their resources, resulting in higher rate.

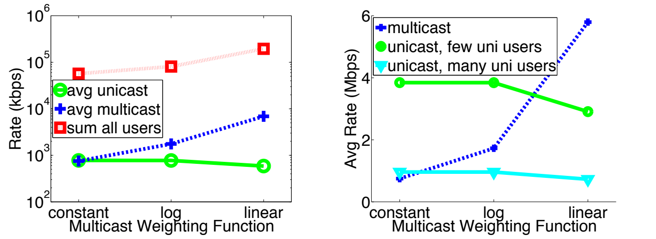
(b) Channel conditions: when multicast users have similar channel conditions, they can be better grouped to achieve higher rate.

Fig. 7: Comparison with no-eMBMS. Multicast rate gains are greatest when there are (a) more users with (b) similar channel conditions.

B. Rate gain over no-eMBMS

Impact of number of multicast users: To examine the impact on user rates, we fix the number of unicast users at 50 and vary the number of multicast users from 4 to 24. Intuitively, there might be two possible outcomes: more multicast users means more RBs can be pooled together to obtain higher rate, or more multicast users will share a fixed amount of RBs and leave the remaining RBs for the unicast users. Fig. 7a plots the average multicast user's rate and shows that the first case occurs, since our algorithm adapts the multicast RBs based on the number of multicast users. 1G provides minimal rate gain despite the increased RBs, because it is constrained by the coding scheme of the worst user in the group. The rate of unicast users under our VG scheme does not change, and is omitted due to space constraints.

Impact of channel distribution: We simulate three different distributions of channel conditions: (1) a uniform distribution of the entire range of LTE coding schemes; (2) a bimodal distribution; and (3) normal distribution with parameters $\mathcal{N}(377, 119)$. For each distribution, we compare the rate provided by each algorithms. There are 50 unicast users and 24 multicast users. Fig. 7b shows the average user rate for increasingly (along the horizontal axis) homogeneous users, in terms of channel conditions. We see that 1G, which forms one group, does well when the users are very similar, and 4G, which forms 4 groups, does best when the channel conditions are heterogeneous. VG outperforms the fixed group formation algorithms in all cases since it recognizes the underlying user diversity and chooses appropriate groups and resource allocation for the users.



(a) Single eNB: When multicast users are favored (linear), total rate is higher since more efficient multicast RBs are reserved.

(b) Two eNBs: When multicast users are favored (linear), unicast rate at the sparse eNB drops slightly.

Fig. 8: Impact of multicast weighting function. The rate gains from multicast are distributed between multicast and unicast users by modulating the weighting function.

C. Impact of the multicast weighting function

For the preceding experiments, VG used the linear multicast weighting function (10). In this set of experiments, we examine the performance of VG in more detail to find that this weighting function sometimes over-provisions for multicast users at the expense of unicast users. Fig. 8a shows the per-user unicast rate, per-user multicast rate, and total rate across all users for different weighting functions. Although the linear weighting function makes the most efficient use of spectrum in terms of total rate, it achieves this primarily by giving the multicast users higher rate. The unicast rate is slightly lower when the linear multicast weighting function is used because the multicast resources are less constrained (4). The logarithmic and constant multicast weighting functions can achieve a better balance between unicast and multicast rate. Thus the multicast weighting function can be used to trade off the benefits of eMBMS between multicast and unicast users.

We also examine the impact of heterogeneity of user demand for multicast content across multiple eNBs. One might expect cells with few multicast and many unicast users to suffer, because resources are being unnecessarily allocated for multicast. In this experiment, we set up two eNBs, one with more unicast users (40) than multicast users (4), and one with fewer unicast user (10) than multicast users (20). In Fig. 8b, we plot the average rate of the unicast users in each cell.

Similar to the single eNB case, the unicast rate is lowered when the linear multicast weighting function is used. The rate drop is significant in the cell with few unicast users, but the absolute value of the rate is quite high. However, if this drop is considered too large to maintain acceptable user-experience, one may wish to adopt a modified weighting function. In the cell with many unicast users, the rate drop is minimal, but this small contribution from each user can provide a significant benefit to the multicast rate. In neither case do the unicast users suffer unduly when the linear weighting function is used, but the logarithmic or constant weighting functions can give them increased rate at the expense of multicast user rate.

VI. Related Works

A large number of works have studied multicast content delivery. We cover the most relevant categories.

Multicast optimization in the backbone: Multicast resource allocation that focuses on constructing efficient multicast trees in wireline networks [4], can reduce backbone network utilization and latency, but has limited overall performance improvement due to the last hop generally being the bottleneck. Wireline optimizations also do not take into account the limited spectrum resources in wireless networks. For example, [11] studies an efficient user partitioning scheme for wireline multicast users with heterogeneous bottleneck links, which is similar to heterogeneous channel conditions in wireless networks, but does not consider resource allocation.

Multicast in OFDMA networks: A survey of multicast resource allocation in OFDMA networks can be found in [13]. [10] optimizes the multicast group's coding scheme to achieve proportional fairness, but sometimes causes users at the edge of the cell to not receive packets when the coding scheme is better than the user can support. [14] studies proportional fair scheduling with heuristics for partitioning users. [15] leverages heterogeneous channel conditions by allowing users with good channel conditions to act as relays for the other users. Using layered multimedia data, [16] [17] [18] study resource allocation and coding scheme selection to maximize the utility for multicast users. However, eMBMS multiplexes unicast and multicast data on the same carrier, so the impact on the unicast users should also be considered.

eMBMS: Application-layer approaches such as [19] [20] incorporate forward-error correction in eMBMS to improve reliability in the absence of detailed user feedback. This is complementary to our link-layer approach of choosing the coding scheme that all users can decode. [21] performs a simulation-based study of how eMBMS parameters such as number of eNBs affect file transfer times.

VII. conclusions

Industry trials of multimedia content delivery using LTE eMBMS have sparked interest in re-visiting resource allocation for cellular multicast. Any such approach must incorporate the unique features of eMBMS, including synchronized multicast transmission across multiple cells, and limited control of unicast schedulers at the eNBs. Based on the eMBMS architecture, we designed a system to fairly allocate rate to multicast and unicast users across multiple cells. Through convex optimization and dynamic programming, we developed an efficient and optimal solution that is adaptive to the channel conditions of the users and allows the operator to trade off unicast user rates and multicast spectral efficiency. In the case of users at a single eNB trying to maximize their individual rate by switching between multicast and unicast, we found that such scenarios are limited and the impact can be bounded.

Acknowledgements

This paper benefited significantly from InfoCom 2015 reviewer comments, for which we are very thankful. We thank Dahai Xu for the initial discussion on this line of work. Jiasi Chen and Mung Chiang were supported in part by NSF NeTS grants CNS-1011962 and CNS-1318607.

References

[1] 3GPP, "E-UTRA and E-UTRAN, Ch. 15: MBMS," http://www.3gpp.org/DynaReport/36300.htm, 2013.

[2] J. Erman and K. Ramakrishnan, "Understanding the Super-sized traffic of the Super Bowl," *ACM IMC*, 2013.

[3] "CES 2013: Verizon Eyes Broadcast Over LTE for Super Bowl 2014," http://www.pcmag.com/article2/0,2817,2414062,00.asp.

[4] J. C.-I. Chuang and M. A. Sirbu, "Pricing multicast communication: A cost-based approach," *Telecom. Systems*, vol. 17, no. 3, 2001.

[5] S.-B. Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, "Proportional fair frequency-domain packet scheduling for 3gpp lte uplink," *INFOCOM*, 2009.

[6] S. Sesia, I. Toufik, and M. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice*. Wiley, 2011.

[7] F. Kelly, "Charging and rate control for elastic traffic," *Euro. Trans. Telecom.*, vol. 8, no. 1, pp. 33–37, 1997.

[8] "MPEG-DASH," http://dashif.org/mpeg-dash/.

[9] "Fair and Optimal Resource Allocation for LTE Multicast [Technical Report]," https://www.dropbox.com/sh/1prqdg5901j3qc7/AABIpj10_LiX7xTB0V7tidELa.

[10] H. Won, H. Cai, D. Y. Eun, K. Guo, A. Netravali, I. Rhee, and K. Sabnani, "Multicast scheduling in cellular data networks," *IEEE Trans. Wireless. Comm.*, vol. 8, no. 9, pp. 4540–4549, Sep. 2009.

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

[12] R.-H. Gau, "On group partition for wireless multicast flow control," *Communications Letters, IEEE*, vol. 16, no. 6, pp. 870–873, June 2012.

[13] R. Afolabi, A. Dadlani, and K. Kim, "Multicast scheduling and resource allocation algorithms for ofdma-based systems: A survey," *IEEE Comm. Surveys Tutorials*, vol. 15, no. 1, pp. 240–254, First 2013.

[14] C. H. Koh and Y. Y. Kim, "A proportional fair scheduling for multicast services in wireless cellular networks," *VTC*, 2006.

[15] F. Hou, L. Cai, P.-H. Ho, X. Shen, and J. Zhang, "A cooperative multicast scheduling scheme for multimedia services in ieee 802.16 networks," *IEEE Trans. Wireless Comm.*, vol. 8, no. 3, pp. 1508–1519, March 2009.

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

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

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

[19] T. Mladenov, S. Nooshabadi, and K. Kim, "Efficient incremental raptor decoding over bec for 3gpp mbms and dvb ip-datacast services," *IEEE Trans. Broadcasting*, vol. 57, no. 2, pp. 313–318, June 2011.

[20] E. Baik, A. Pande, and P. Mohapatra, "Cross-layer coordination for efficient contents delivery in lte embms traffic," *IEEE MASS*, 2012.

[21] J. Monserrat, J. Calabuig, A. Fernandez-Aguilella, and D. Gomez-Barquero, "Joint delivery of unicast and e-mbms services in lte networks," *IEEE Trans. Broadcasting,*, vol. 58, no. 2, pp. 157–167, 2012.

Appendix

Lemma 1

Proof: Prob. 2 is clearly concave. The Lagrangian is: $L(\{y_i\}, \lambda) = -\sum_{i \in U_b} + \lambda(\sum_{i \in U_b} y_i + \sum_{k=1}^{K} x_k - T)$. Writing the KKT conditions:

Stationarity: $y_i = \frac{1}{\lambda},\ forall i \in U_b$

Primal feasibility: $\sum_{i \in U_b} y_i \leq T - \sum_{k=1}^{K} x_k$

Dual feasibility: $\lambda \geq 0$

Complementary slackness: $\lambda \left(\sum_{i \in U_b} y_i + \sum_{k=1}^{K} x_k - T \right) = 0$

Clearly $\lambda > 0$, so from complementary slackness $\sum_{i \in U_b} y_i = T - \sum_{k=1}^{K} x_k$. Combining this with stationarity allows us to solve for $\lambda = \frac{|U_b|}{T - \sum_{k=1}^{K} x_k}$ and gives the desired result. ■

Lemma 2

Proof: Denote an optimal solution to Prob. 1 by $(\mathbf{x}_1^*, \mathbf{y}_1^*)$ and an optimal solution to Probs. 2,3 by $(\mathbf{x}_2^*, \mathbf{y}_2^*)$. The feasible sets are the same since the constraints are identical: (2) = (12), (4) = (15), (5) = (16).

⇒: We first show that $(\mathbf{x}_1^*, \mathbf{y}_1^*)$ is a solution to Probs. 2, 3. First consider $\mathbf{y}_1^*$. Given $\mathbf{x}_1^*$, $\mathbf{y}_1^*$ maximizes the first term of the objective function $\sum_{b=1}^{B} \sum_{i \in U_b} \log(d_i y_i)$ of Prob. 1, so each term $\sum_{i \in U_b} \log(d_i y_i)$ must also be maximized, since there are no coupling constraints between b for users $\{i\}_{i \in U_b}$. This is exactly the objective function (11) of Prob. 2, so $\mathbf{y}_1^*$ solves Prob. 2. Now consider $\mathbf{x}_1^*$. Given $\mathbf{y}_1^*$, $\mathbf{x}_1^*$ maximizes the objective function of Prob. 1 in terms of $\mathbf{x}$, which is the same objective function and variable as Prob. 3, so $\mathbf{x}_1^*$ solves Prob. 3.

⇐: We now show that $(\mathbf{x}_2^*, \mathbf{y}_2^*)$ is a solution to Prob. 1. Given $\mathbf{x}_2^*$, $\mathbf{y}_2^*$ solves Prob. 2 by maximizing the objective function $\sum_{i \in U_b} \log(d_i y_i)$ for each b, so the sum $\sum_{b=1}^{B} \sum_{i \in U_b} \log(d_i y_i)$ must also be maximized, which is the first term in the objective function of Prob. 3. $\mathbf{x}_2^*$ maximizes the second term in the objective function, so the sum of the terms is also maximized. This is exactly the objective function (1) of Prob. 1. ■

Proposition 1

Proof: From Lemma 2, the solution of Probs. 2 and 3 gives us the solution for Prob. 1 for fixed $\{G_k\}, \{\hat{c}_k\}, K$. Lemma 1 gives us the solution to Prob. 2, $\{y_i^*(\mathbf{x})\}$, which we plug into the objective function (13) of Prob. 3 to solve for $\mathbf{x}$. Prob. 3 is concave since the objective function (13) is a sum of logarithms and constraints (15), (16) are linear. Writing the KKT conditions:

Stationarity: $\frac{N}{T - \sum_l x_l} - \frac{|G_k|}{x_k} - \lambda_1 + \lambda_2 + \mu_k = 0\ \forall k$

Primal feasibility: $0 \leq \sum_l x_l \leq \alpha T, x_k \leq \frac{f(G_k)}{N + \sum_l f(G_l)} \forall k$

Dual feasibility: $\lambda_1 \geq 0, \lambda_2 \geq 0, \mu_k \geq 0\ \forall k$

Complementary slackness: $\lambda_1 \sum_l x_l = 0, \lambda_2 \left(\sum_l x_l - \alpha T \right) = 0, \mu_k \left(x_k - \frac{f(G_k)}{N + \sum_l f(G_l)} \right) = 0\ \forall k$

$\lambda_1 > 0$ leads to degenerate solutions so we only consider $\lambda_1 = 0$.

Case 1: $\lambda_2 = 0$. We can verify that our solution $x_k^* = \frac{f(G_k)}{N + \sum_l f(G_l)} T$ satisfies the KKT conditions above, which is sufficient for optimality. Plugging x_k^* into the stationarity condition and rearranging, we have $\mu_k = \frac{(|G_k| - f(G_k))(N + \sum_l f(G_l))}{f(G_k) T} \geq 0$ since $|G_k| \geq f(G_k)$, so μ_k is dual feasible. The condition $\frac{\sum_l f(G_l)}{N + \sum_l f(G_l)} \leq \alpha$ comes from primal feasibility. The other KKT conditions also follow.

Case 2: $\lambda_2 > 0$. For some i, $\mu_i = 0$, so stationarity gives that $x_i = \frac{|G_i|}{\frac{U}{(1-\alpha)T} + \lambda_2}$. For some other j, $\mu_j > 0$, so complementary slackness of μ_j gives that $x_j = \frac{f(G_j)}{N + \sum_l f(G_l)} T$. In order to satisfy primal feasibility of x_k, then $x_k = \min \left(\frac{|G_l|}{\frac{U}{(1-\alpha)T} + \lambda}, \frac{f(G_k)}{N + \sum_l f(G_l)} T \right)$. Complementary slackness gives that $\sum_l x_l = \alpha T$, so substituting our equation in, we have the condition $\sum_l \min \left(\frac{|G_l|}{\frac{U}{(1-\alpha)T} + \lambda_2}, \frac{f(G_l)}{N + \sum_k f(G_k)} T \right) = \alpha T$. If $\frac{\sum_l f(G_l)}{N + \sum_l f(G_l)} > \alpha$, then $\lambda_2 > 0$ to satisfy this condition. This is a variant of a water-filling algorithm: since the first term is a decreasing function of λ_2, we search for the smallest λ_2 which satisfies the condition.

The other KKT conditions follow. ■

Corollary 1

Proof: We have the same KKT conditions as in Proposition 1, but replace $f(G_k)$ with $|G_k|$. $\lambda_1 > 0$ leads to degenerate solutions so we only consider $\lambda_1 = 0$.

Case 1: $\lambda_2 = 0$. Same as in Proposition 1, replacing $f(G_k)$ with $|G_k|$.

Case 2: $\lambda_2 > 0$. First, we will show that $\mu_k = 0\ \forall k$. We know that $x_k = |G_k| T \min \left(\frac{1}{\frac{U}{(1-\alpha)} + T\lambda_2}, \frac{1}{N + \sum_k f(G_k)} \right)$ and the condition $\sum_k |G_k| \min \left(\frac{1}{\frac{U}{(1-\alpha)} + T\lambda_2}, \frac{1}{N + \sum_k f(G_k)} \right) = \alpha$ must be satisfied. If the first term is selected, it means $\mu_k = 0$, and the second term means $\mu_k > 0$. The first term is a decreasing function of λ_2 and the second term is a constant, so the optimal λ_2^* that satisfies the condition must be the first term for all k, so $\mu_k = 0\ \forall k$.

Now, we can verify that our solution $x_k^* = \frac{|G_k|}{\sum_l |G_l|} \alpha T$ satisfies the KKT conditions with dual variables $\lambda_2 = \frac{\sum_k |G_k|}{\alpha T} - \frac{N}{(1-\alpha)T}$ and $\mu_k = 0\ \forall k$. Dual feasibility of λ_2 is because $\lambda_2 > 0$ iff $\alpha < \frac{\sum_l |G_l|}{N}$, which is true because $\alpha < \frac{\sum_l f(G_l)}{N + \sum_l f(G_l)} < \frac{\sum_l |G_l|}{N}$. Primal feasibility of x_k is because $x_k = \frac{|G_k|}{\sum_l |G_l|} \alpha T < \frac{|G_k|}{\sum_l |G_l|} \frac{\sum_l f(G_l)}{N + \sum_l f(G_l)} T \leq \frac{\sum_l f(G_l)}{N + \sum_l f(G_l)} T$. The other KKT conditions also follow. ■

Lemma 3

Proof: Without loss of generality, consider users $1, 2, 3$ with coding scheme $c_1 \leq c_2 \leq c_3$ respectively. Assume we have a globally optimal unordered grouping where 1 and 3 are in the same multicast group A (possibly with other users) and 2 is in a different multicast group B (possibly with other users or possibly just a unicast user). By switching users 2 and 3, we will create an ordered grouping and prove that the global utility stays the same or improves.

With the unordered groups, the utility is $\log(c_A x_A) + \log(c_B x_B)$, where x_A is the resources allocated to group A, x_B is the resource allocated to group B, c_A is the effective coding scheme of group A, and c_B is the effective coding scheme of group B. After the swap, the new coding scheme of group A is c'_A and the new coding scheme of group B is c'_B. Notice from Proposition 1 that the resource allocation only depends on the number of users in the group. Since we are only swapping users between groups and not changing the number of users in each group, it suffices to prove our claim on the group coding scheme.

The coding scheme of group A is $c_A \leq c_1$. Putting user 2 in group A will keep this effective scheme unchanged (because $c_2 \geq c_1$). Similarly, the coding scheme of group B was $c_B \leq c_2$, so putting user 3 there will either improve or keep unchanged its effective coding scheme (because $c_3 \geq c_2$). So the new utility after the swap is $\log(c_A x_1) + \log(c'_B x_2), c'_B \geq c_B$. Therefore we managed to create another grouping different from the unordered grouping with identical or improved global utility. ■

Proposition 2

Proof: With the users sorted by coding scheme in ascending order, the problem becomes where to place the (unknown number of) partitions so as to maximize total utility. We can construct a table $U(g, i, k)$, where each entry is the utility when there are $g = 1 \ldots M$ partitions, with the first $i = 1 \ldots M$ multicast users, and the first $k = 1 \ldots g$ groups. We then search across the entries $U(g, M, g)\ \forall g$ and pick the partition with the best utility.

$U(g, j, k)$ can be efficiently calculated by considering the optimal utility from $k-1$ partitions and growing the last partition containing the last user i. The last partition can have users $\{i\}$ or users $\{i-1, i\}$ or users $\{i-2, i-1, i\}$, and so on up to users $\{2, \ldots, i\}$. Let the function $\textsc{OptResource}(g, \{a, \ldots, b\})$ return the utility of a single partition of users $a, \ldots, b$ when there are g groups. Writing it out:

$$U(g,i,k) = \max\{ \tag{18}$$

$$U(g,1,k-1) + \textsc{OptResource}(g, \{2,\ldots,i\}), \tag{19}$$

$$U(g,2,k-1) + \textsc{OptResource}(g, \{3,\ldots,i\}), \tag{20}$$

$$\ldots, \tag{21}$$

$$U(g,i-1,k-1) + \textsc{OptResource}(g, \{i\})\} \tag{22}$$

Summing the utility of $k-1$ partitions plus the new partition is only possible if adding the new partition does not change the utility of the previous $k-1$ partitions. In other words, the utility of each partition cannot depend on the partitions that will be created after it. For $f(G_k) = |G_k|$ and $f(G_k) = 1, \alpha \geq \frac{M}{N+M}$, according to Prop. 1, the resource allocation and therefore utility of the partition depends only on the total number of users, the size of the group, and the total number of groups g, which is not affected by later partitions. ■

Lemma 4

Proof: (a) *User switch:* If a user j stays on unicast, her utility is:

$$c_j T \frac{1}{N + f(G \setminus j)} = c_j T \frac{1}{N + |G| - 1} \tag{23}$$

The unicast user has better coding scheme than the multicast group ($c_j > \hat{c}$), so the multicast group does not change its coding scheme when the unicast user joins. Therefore, if the user switches to multicast, the rate is:

$$\hat{c} T \frac{f(G)}{N - 1 + f(G)} = \hat{c} T \frac{|G|}{N - 1 + |G|} \tag{24}$$

Setting (24) > (23), then we can rearrange to get the result.
(b) *Global switch:* Before the switch, the global utility is:

$$\sum_{i \in U} \log\left(c_i T \frac{1}{N + f(G \setminus j)}\right) + |G \setminus j| \log\left(\hat{c} T \frac{f(G \setminus j)}{N + f(G \setminus j)}\right) \tag{25}$$

After the switch, the global utility is:

$$\sum_{i \in U \setminus j} \log\left(c_i T \frac{1}{N - 1 + f(G)}\right) + |G| \log\left(\hat{c} T \frac{f(G)}{N - 1 + f(G)}\right) \tag{26}$$

Setting (26) > (25), $f(G) = |G|$, $f(G \setminus j) = |G| - 1$, we can rearrange to get the result. ■

Proposition 3

Proof: By contradiction. Assume that the solution is optimal and there is a unicast user j who wants to switch to multicast. This means:

$$\frac{c}{\hat{c}} < |G| < \left(1 + \frac{1}{|G| - 1}\right)^{|G|-1} |G| \tag{27}$$

where the first inequality is from the user switching condition, and the second inequality is because $\left(1 + \frac{1}{|G|-1}\right)^{|G|-1} > 1$. This implies that the local switch also increases the global utility, and contradicts the assumption that the solution was optimal. ■

Lemma 5

Proof: (a) *User switch:* If the user stays in the multicast group, her rate is:

$$\hat{c} T \frac{f(G)}{N - 1 + f(G)} = \hat{c} T \frac{|G|}{N - 1 + |G|} \tag{28}$$

If a user j leaves the multicast group, then her rate improves since she is no longer contained by the low coding scheme of the multicast. Her rate on unicast is:

$$c_j T \frac{1}{N + f(G \setminus j)} = c_j T \frac{1}{N + |G| - 1} \tag{29}$$

Setting (28) > (29), then we can rearrange to get the result.

(b) *Global switch*: Before the switch, when the user is in multicast, the global utility is:

$$\sum_{i \in U} \log\left(c_i T \frac{1}{N + f(G)}\right) + |G| \log\left(\hat{c} T \frac{f(G)}{N + f(G)}\right) \tag{30}$$

After the switch, the global utility is:

$$\sum_{i\in U} \log\left(c_i T \frac{1}{N+1+f(G\setminus j)}\right) + \log\left(c_j T \frac{1}{N+1+f(G\setminus j)}\right) \\ +(m-1)\log\left(\hat{c}T\frac{f(G\setminus j)}{N+1+f(G\setminus j)}\right) \tag{31}$$

Setting (31) $>$ (30), $f(G) = |G|$, and $f(G\setminus j) = |G| - 1$, we can rearrange to get the result. ■

Lemma 6

Proof: By contradiction. Suppose we have an optimal solution, and split the users into two groups: L users who have coding scheme greater than $\hat{c}\beta|G|$, and $|G|-L$ users who will stay in the multicast group. We will compare the global utility if the L users leave the multicast group to form their own group, and show that global utility increases if $L > \lceil\frac{e}{\beta}\rceil$, which contradicts the fact that this is an optimal solution.

The initial utility with all the users in the multicast group is:

$$|G|\log\left(\hat{c}T\frac{f(G)}{N+f(G)}\right) = |G|\log\left(\hat{c}T\frac{|G|}{N+|G|}\right) \tag{32}$$

The new utility if the L users leave is:

$$(|G|-L)\log\left(\hat{c}T\frac{|G|-L}{N+|G|}\right) + L\log\left(\hat{c}\beta|G|T\frac{L}{N+|G|}\right) \tag{33}$$

Setting (33) $>$ (32), the new utility is greater than the initial utility iff:

$$1 > \left(1-\frac{L}{|G|}\right)^{L-|G|}\frac{1}{(\beta L)^L} \tag{34}$$

The RHS of (34) is decreasing in L and increasing in $|G|$. So if we find the smallest L, l, and largest $|G|$, g, for which the condition holds, it will also hold for $L \geq l$ and $|G| \leq g$. As $|G| \to \infty$, (34) becomes $1 > \frac{e}{\beta L}$, so $l = \lceil\frac{e}{\beta}\rceil$ works. This means there cannot be $L \geq \lceil\frac{e}{\beta}\rceil$ users in the multicast group with coding scheme $> \hat{c}\beta|G|$. ■

Proposition 4

Proof: In the first iteration, users with coding scheme $>$ $\hat{c}|G|$ may switch (Lemma 5), $n_0 \leq \lceil\frac{e}{1}\rceil - 1$ of them (Lemma 6), leaving $|G| - n_0$ users in the group. In the second iteration, since the number of users in the group has changed, users with coding scheme $> \hat{c}(|G| - n_0)$ may desire to switch, $n_1 \leq \lceil\frac{e}{1-\frac{n_0}{|G|}}\rceil - 1 - n_0$ of them, leaving $|G| - n_0 - n_1$ users in the group. In the third iteration, users with coding scheme $> \hat{c}(|G| - n_0 - n_1)$ may switch, $n_2 \leq \lceil\frac{e}{1-\frac{n_0+n_1}{|G|}}\rceil - 1 - n_0 - n_1$ of them; and so on. Clearly the stopping criterion is when the number of switching users in the tth iteration is 0, i.e. $n_t \leq \lceil\frac{e}{1-\frac{\sum_{\tau=0}^{t-1} n_\tau}{|G|}}\rceil - 1 - \sum_{\tau=0}^{t-1} n_\tau = 0$. Let $n = \sum_{\tau=1}^{t} n_\tau$. Then the stopping condition is $n + 1 = \lceil\frac{e}{1-\frac{n}{|G|}}\rceil$. Since both the RHS and LHS are increasing functions of integer n, and $n = 0$ does not work, we can search for the minimum n that satisfies the condition. ■