Hybrid Multicast-Unicast Streaming over Mobile Networks

Md. Mahfuzur Rahman[1], Cheng-Hsin Hsu[2], Abdul Hasib[1], and Mohamed Hefeeda[1]
[1]School of Computing Science, Simon Fraser University, Surrey, Canada
[2]Department of Computer Science, National Tsing Hua University, Hsin-Chu, Taiwan

***Abstract*—Mobile on-demand videos are getting tremendously popular and incurring staggering overhead on cellular networks. Fortunately, next generation cellular networks support video streaming over either unicast or multicast, but how to capitalize both unicast and multicast for optimal on-demand video streaming remains an open question. In this paper, we consider a resource allocation problem that concurrently utilizes unicast/multicast in order to support many more mobile streaming users and minimize the energy consumption of the battery-powered mobile devices. We formulate this problem as a Binary Integer Programming (BIP) problem. We present an optimal algorithm, SCOPT, for this problem. We also develop an efficient heuristic algorithm, SCG, for lower overhead. We conduct detailed packet-level simulations to evaluate the algorithms in LTE networks using OPNET. Our simulation study shows that the proposed algorithms: (i) result in lower energy consumption than multicast-only approach, (ii) scale to many more mobile users than unicast-only approach, and (iii) are more energy efficient with more network bandwidth or fewer videos. In addition, we discuss how our solution can be extended to support Single Frequency Networks in which multiple adjacent base stations operate on the same frequency.**

I. Introduction

The demand for multimedia streaming over mobile networks has been steadily increasing: worldwide mobile traffic amount reached 885 petabytes per month in 2012, and about 51% of that traffic carries videos [1]. Current 3G cellular networks only support unicast, which is not optimized for streaming a video to many mobile devices. This is because the video will be sent multiple times over a shared air medium, which consumes excessive mobile network bandwidth and may negatively affect voice call quality. There are several 4G technologies to enable multicast in mobile networks, e.g., the Multicast and Broadcast Service (MBS) in WiMAX [2] and Evolved Multimedia Broadcast Multicast Service (eMBMS) in LTE [3]. In fact, some U.S. cellular operators plan to launch eMBMS service in their LTE networks [4], which will allow operators to efficiently stream a video to numerous mobile devices. Although multicast in cellular networks seems to be only useful for live events broadcast, many other applications can potentially benefit from it. These applications include on-demand video streaming, timeshifted events, and mobile video recorders. These and similar applications can benefit from multicast because modern mobile devices have increasingly larger storage space, which can be used to prefetch some video segments that will be consumed later. More specifically, for live streaming, such as sports events, mobile users can arrive at different times, but they start receiving from the current moment. This creates a natural case for grouping users in multicast sessions. For prefetching, popular videos, such as latest TV episodes and highlights of recent sports events, will be requested by many users at different times, e.g., in the evening of the release day. Since those videos are not immediately played back, the requests can be grouped into multicast sessions as well. Video streaming applications impose tremendous loads on the mobile networks, which will likely force the cellular network operators to seriously consider the multicast supports in next generation mobile networks that will be deployed in the near future.

To cope with the staggering number of requests, the cellular base stations have to carefully determine whether to serve each request using unicast or multicast, in order to minimize the network load and prolong the mobile devices' battery life. Making such decisions is challenging because: (i) the user demands and network conditions are diverse and dynamic, and (ii) there exists a clear tradeoff between network load and energy saving. This tradeoff is due to two common features of modern mobile networks. First, network interfaces on mobile devices support multiple Modulation and Coding Scheme (MCS) modes to cope with different channel conditions. For example, mobile devices closer to the base station may use more aggressive, i.e., higher, MCS modes for higher transfer rates, whereas mobile devices at the cell edge can only use lower MCS modes at lower transfer rates. Second, network interfaces on mobile devices may be turned off[1] when not receiving in order to save energy. That is, when mobile devices use higher MCS modes, they receive at higher rates and finish earlier. This in turn results in higher energy savings. Therefore, streaming videos using unicast allows individual mobile devices to use the highest MCS modes allowed by their channel conditions for the highest possible energy saving at the expense of higher network load due to duplicated video streams. In contrast, streaming using multicast forces some mobile devices to use lower MCS modes to cope with the mobile device with the worst channel condition, and thus suffers from lower transfer rate and lower energy saving. Such a tradeoff between network load and energy saving motivates

[1]In this paper, *turning off* network interface loosely refers to putting the network interfaces into low-power states. Next generation cellular networks all support such power saving features, although they may employ slightly different terminologies.

ISBN 978-3-901882-58-6

us to study a *hybrid* on-demand video streaming system that concurrently leverages unicast and multicast to maximize the overall energy saving of mobile devices under various resource constraints.

In this paper, we study the resource allocation problem in a hybrid on-demand streaming system over both unicast and multicast in next generation mobile networks. The base station(s) concurrently serves multiple videos with diverse popularity to mobile devices, and different mobile devices may start watching at different times. The operator is assumed to reserve a fixed amount of network bandwidth for the on-demand video streaming service. Our problem is to schedule which chunks of videos should be sent over multicast (or unicast) and when to send them, in order to maximize the overall energy saving of mobile devices without consuming excessive network bandwidth. We prove that the resource allocation problem is NP-Complete, and then mathematically formulate it as a Binary Integer Programming (BIP) problem. The optimization problem can be optimally solved using optimization problem solvers, which however may be computationally expensive for on-demand streaming services. Therefore, we develop algorithms for close to optimal solutions.

While our solution is general for all cellular networks that support multicast, we use LTE networks in our evaluation for concrete discussion. In particular, our extensive simulation results, using OPNET [5], lead to the following observations.

- The proposed solution allows cellular networks to support a large number of mobile devices, as if in multicast-only networks, *and* achieve almost-optimal energy saving, as if in unicast-only networks. In some experiments, compared to 71% energy saving of the multicast-only solution, our proposed solution results in 89% energy saving, which is the same as the unicast-only network. On the other hand, unicast-only solution can only support very few mobile devices, and thus does not scale to large networks.
- The proposed algorithms can easily run in real time, as they have polynomial time complexities. In our experiments, they terminate in less than few milliseconds on a commodity workstation. In real deployment, they would typically run on powerful servers and periodically invoked once every several seconds.

In addition, next generation cellular networks support *Single Frequency Networks* [6], in which multiple base stations transmit the same wireless signals on the same frequency to increase the received signal strength of mobile devices. We also extend our formulation for Single Frequency Networks, which significantly reduce interference and allow mobile devices to receive combined signals for better reception.

II. Related Work

Several studies consider on-demand streaming services over mobile networks in a more restricted sense. For example, Hillestad et al. [7] proposed an adaptive algorithm for streaming scalable on-demand videos over fixed WiMAX networks. However, the authors did not exploit multicast to reduce the network load. Majumdar et al. [8] proposed a multimedia streaming approach leveraging both Forward Error Correction (FEC) and Automatic Repeat ReQuest (ARQ). The appropriate parameters of source and channel coding were determined so that the overall transfer rate is maximized. They also presented an algorithm for multicasting scenarios, which employs FEC only as ARQ is less applicable in multicast. Lee et al. [9] described a scheme that uses both unicast and multicast communications to reduce service blocking probability and bandwidth consumption. Different from our work, they did not take energy consumption into consideration. Hlavacs and Buchinger [10] proposed a patching system for mobile networks. Their system may suffer from low energy saving as eventually all videos are multicast. Yoon et al. [11] concentrated on the implementation details of a multicast video service in LTE networks. These studies [7], [8], [9], [10], [11] did not consider energy conservation, which is crucial to battery-powered mobile devices.

Tremendous research efforts have been devoted to saving energy on mobile devices. One of the earliest works in this area is called STPM [12] which proposes a self-tuning operating system module that adapts itself to the network access patterns and intent of applications to enable power management only when appropriate. E-Mili [13] adaptively downclocks the network radio to reduce the amount of energy consumed during idle listening. Zhu and Cao [14] presented a new scheduling algorithm, called rate-based bulk scheduling (RBS) for the base station to determine data flows to be served at different times. The concept of proxy server is employed to buffer data for the mobile devices so that the wireless network interface can sleep for a long time period to save power. Luna et al. [15] considered the selection of source coding parameters jointly with transmitter power and rate adaptation to reduce power consumption. These approaches [12], [13], [14], [15] try to reduce energy consumption of mobile devices from different perspectives, but do not jointly use unicast and multicast.

In summary, our work leverages both unicast and multicast in mobile networks to serve more users and maximize the overall energy saving of mobile devices under a given bandwidth constraint. To the best of our knowledge, this problem has not been rigorously investigated in the literature. Last, we note that there are more advanced mechanisms, such as MU-MIMO [16], which may result in better broadcast spectrum efficiency. These advanced mechanisms however are too complex and out of the scope of our work.

III. System Model and Problem Statement

We consider an on-demand streaming scenario with base stations, mobile devices, and resource allocators as illustrated in Figure 1. Mobile devices arrive asynchronously, and each mobile device sends requests to a resource allocator to receive video streams. These requests may be driven by mobile users' current demands or by some prediction logics running on mobile devices, e.g., a background mobile application may prefetch videos that are most likely to be watched in near future [17], [18]. Since the requests are driven by mobile devices/users, user inputs like delay, fast forward, and rewind

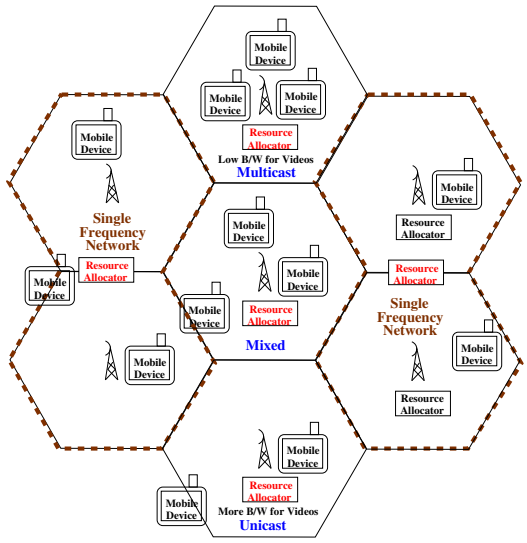

Fig. 1. The considered resource allocation problem in mobile networks.

can be supported, which enable diverse applications, including on-demand video streaming, live or time-shifted sports events, and mobile personal video recorders. Each resource allocator periodically solves an optimization problem for leveraging both unicast and multicast to: (i) maximize the average energy saving across all mobile devices, (ii) minimize the network resource consumed by video streaming, and (iii) ensure smooth playout on all mobile devices. Upon the optimization problem is solved, the resource allocator notifies the base stations to stream videos accordingly.

Figure 1 demonstrates the generality of our considered problem in two aspects. First, the resource allocator may manage one or multiple base stations. For example, the base stations of a Single Frequency Network must be managed by the same resource allocator for optimal allocations. For clarity, we first assume that each resource allocator manages a base station, and generalize the problem for Single Frequency Networks in Section V. Second, depending on: (i) channel conditions of individual mobile devices and (ii) reserved bandwidth for video streaming, resource allocators may decide to stream videos over multicast, unicast, or a mixture of both. For example, the top cell in Figure 1 consists of mobile devices with similar channel conditions and only has little bandwidth available for on-demand video streaming, which renders multicast only decisions. In contrast, the bottom cell suffers from heterogeneous channel conditions, but has more spare bandwidth, which in turns leads to unicast only decisions for higher energy saving. Our considered problem covers these two scenarios and any mixture of them such as the center cell in this figure.

Several types of cellular wireless networks adopt Orthogonal Frequency Division Multiple Access (OFDMA) modulation scheme, which divides the wireless medium along both time and frequency domains [19]. We consider an *allocation window* with T columns of *symbols* and S rows of *subchannels*. A pair of $t \in [1, T]$ and $s \in [1, S]$ uniquely determines a *resource block*, which is the unit of resource allocation in the network.[2] Let d denote the fraction of resource blocks that is reserved for video streaming, which can be adjusted based on the voice loads. Thus, the considered resource allocation problem is to distribute the dTS blocks of an allocation window among all mobile devices. Once an allocation is computed, it is used in several consecutive allocation windows until some mobile devices' channel conditions change. Note that the system parameter T affects the length of allocation windows: larger T leads to longer allocation windows for higher allocation flexibility, and smaller T results in shorter allocation windows for shorter video *service delay*. The service delay refers to the time difference between a mobile device switches to a video and the mobile device starts to render that video. Shorter service delay also results in faster adaptation to network dynamics. In the true on-demand case with real time constraints on the service delay, a *patching* solution [20], [21], [22] can be used. That is we define a threshold for a new request to join an on-going multicast session of video and at the same time create a separate, temporary unicast session for that user to receive the earlier parts of the video. This new request will be considered in the next allocation window, and potentially be merged into a multicast session.

The on-demand streaming service offers V different videos. Let r_v denote the encoding rate of video v. We assume that each video v is watched by N_v mobile devices, and we let $N = \sum_{v=1}^{V} N_v$ be the total number of mobile devices. The network interface on each mobile device can be put into one of M Modulation and Coding Scheme (MCS) modes. We let per-block capacity c_m denote the amount of data that can be carried by a block with mode m, where c_m is non-decreasing in $m \in [1, M]$. Each mobile device is under a different channel condition, and can receive at a *maximum* MCS mode, which is determined by the firmware on the network interface to maintain reasonable bit error rates. Moreover, mobile devices may watch different parts of a video. We divide video v into Z_v consecutive parts in the length of allocation windows (a few seconds). We let $w_{v,m,z}$ ($v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$) be the number of mobile devices watching segment z of video v with maximum MCS mode m.

For a given video v, depending on the MCS mode, a mobile device needs to receive different number of blocks in each allocation window. This is because the amount of data to transmit is fixed at qTr_v, which can be carried by $\lceil qTr_v/c_m \rceil$ blocks, where q is the symbol time and m is the MCS mode. Allocating different number of blocks to satisfy such capacity demand could largely affect the *off time* of each mobile device, and thus its *energy saving*. We define the energy saving γ as the fraction of time each mobile device can turn off its network interface to save energy. We acknowledge that factors (e.g., MCS modes) other than the off time may slightly affect the energy consumption. We, however, only consider the dominating off time in this work for better tractability. Moreover, previous studies [19], [23]

[2]We interchangeably use resource blocks and blocks to refer to the resource allocation unit throughout this paper.

show that mobile device's energy consumption depends on the number of symbols it receives, and it is almost independent of the number of subchannels. Therefore, we assume that base stations first allocate blocks in the same column before considering different ones.

The considered problem can be formally written as:

Problem 1: We consider a cellular network with a single cell, in which a fraction d of the network resource blocks is reserved for an on-demand streaming service of V videos, where each video has N_v mobile devices in the allocation window. For video $v \in [1, V]$, there are $w_{v,m,z}$ mobile devices that can receive the video with the maximum MCS mode m and segment z, where $m \in [1, M]$ and $z \in [1, Z_v]$. An allocation specifies: (i) the mapping between each block and video, (ii) the multicast/unicast model of each block, and (iii) the MCS mode of each block. For each allocation window of T symbols and S subchannels, find the optimal allocation to transmit V videos to all $N = \sum_{v=1}^{V} N_v$ mobile devices, so that: (i) the average energy saving across all mobile devices is maximized, (ii) no more than dTS blocks are consumed by the on-demand streaming service, and (iii) all mobile devices watching video v receive at rate r_v for smooth playout.

Lemma 1 (Hardness): The considered resource allocation problem is NP-Complete.

Proof Sketch: We reduce the 0-1 knapsack problem to our problem, which yields the hardness of our problem. ■

The considered problem supports various applications, including live streaming, on-demand streaming, video prefetching, and mobile video recorders. For live streaming, mobile users naturally form multicast groups. However, some users may have poor channel conditions, which could degrade the performance for the whole multicast group. Solving our problem gives each user the optimal decision whether to join a multicast session or receive the live stream using unicast. Another case is prefetching videos for later playback, where mobile devices may signal the base stations to indicate less restricted time constraints. Solving our problem determines the optimal allocation of requests to multicast and unicast sessions, and we give the requests with closer deadline higher priority.

Furthermore, we note that the proposed hybrid on-demand video streaming approach may be readily augmented to satisfy different optimization criteria and resource constraints based on the requirements from cellular operators. For example, instead of minimizing the average energy consumption across all mobile devices, operators may prefer to minimize the maximal energy consumption among all mobile devices for fairness. Moreover, operators may specify energy budget for individual base stations, so as to control their operational costs. The possible optimization criteria and resource constraints are highly driven by *business policies*, and an exhaustive list of them is out of the scope of this paper.

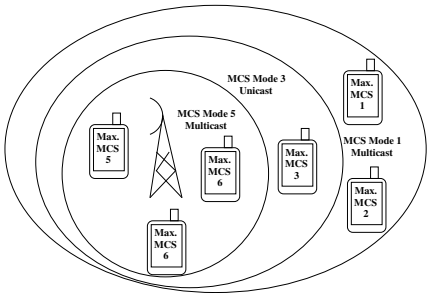

Fig. 2. An illustrative example showing that mobile devices receive videos using either unicast or multicast and different modulation modes.

IV. Proposed Solution

A. Mathematical Formulation

We next formulate the resource allocation problem (Problem 1), which assigns the resource blocks to individual videos, decides whether to use multicast or unicast, and determines the MCS modes of individual blocks, in order to maximize the overall energy saving while guaranteeing smooth playout. We use a boolean decision variable $x_{v,m,z}$ ($v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$) to denote whether the segment z of video v is unicast/multicast using MCS mode m. That is, $x_{v,m,z} = 1$ if segment z of video v is transmitted with MCS mode m, and $x_{v,m,z} = 0$ otherwise. When $w_{v,m,z} = 1$ ($w_{v,m,z} > 1$), the base stations stream video v using unicast (multicast). Figure 2 illustrates a sample solution of our resource allocation problem, in which we use multicast only when there are two or more mobile devices in the corresponding range. For example, the two mobile devices in the outer-most circle receive multicast signals with MCS mode 1. When $x_{v,m,z} = 0$, mobile devices with maximum MCS mode m receive z of v with the next *lower* MCS mode $n \in [1, M]$. For example, as Figure 2 shows, two mobile devices with maximum MCS mode 6 (which is m) only receive at MCS mode 5 (which is n). We define an *intermediate* boolean variable $y_{v,m,n,z}$ for each $v \in [1, V]$, $m, n \in [1, M]$, $n \leq m$, $z \in [1, Z_v]$ as follows. $y_{v,m,n,z} = 1$ when mobile device with maximum MCS mode m would receive segment z of video v with MCS mode n, and $y_{v,m,n,z} = 0$ otherwise. $y_{v,m,n,z}$ is determined by $x_{v,m',z}$, $m' \in [n, m]$ as follows:

$$y_{v,m,n,z} \leq 1 - x_{v,m',z}, \quad \forall\, m' \in [n+1, m], \tag{1}$$

$$y_{v,m,n,z} \leq x_{v,n,z}. \tag{2}$$

We present the formulation in Eq. (3). The objective function in Eq. (3a) is to maximize the average energy saving. The total size of video v in an allocation window is qTr_v, and the minimum number of symbols we need is $\lceil \frac{\lceil \frac{qTr_v}{c_m} \rceil}{S} \rceil$, where m is the MCS mode. The three summations iterate through all the videos, modes, and segments, respectively. The constraint in Eq. (3b) ensures that the on-demand streaming service only consumes up to d network resources. The constraint in Eq. (3c) guarantees that every mobile device receives its allocation window at a feasible MCS mode. This in turn ensures that all

$$\max_{\mathbf{x}} \quad \gamma = 1 - \frac{1}{N}\sum_{v'=1}^{V}\sum_{m'=1}^{M}\sum_{z'=1}^{Z_{v'}} w_{v',m',z'}\sum_{n'=1}^{m'} y_{v',m',n',z'}\lceil\lceil\frac{qTr_{v'}}{c_{n'}}\rceil/S\rceil \tag{3a}$$

$$\text{s.t.} \quad \sum_{v'=1}^{V}\sum_{m'=1}^{M}\sum_{z'=1}^{Z_{v'}} x_{v',m',z'}\lceil\frac{qTr_{v'}}{c_{m'}}\rceil \le dTS \tag{3b}$$

$$(1-\sum_{n'=1}^{m} y_{v,m,n',z})w_{v,m,z} = 0 \tag{3c}$$

$$y_{v,m,n,z} \le 1 - x_{v,m',z}, \ \ \forall\, m' \in [n+1, m] \tag{3d}$$

$$y_{v,m,n,z} \le x_{v,n,z} \tag{3e}$$

$$x_{v,m,z} \in \{0,1\}, y_{v,m,n,z} \in \{0,1\}, \ \forall\, v \in [1,V], m \in [1,M], n \in [1,m], z \in [1,Z_v].$$

1. **foreach** $v \in [1,V]$, $m \in [1,M]$, $z \in [1,Z_v]$
2. **initialize** $x_{v,m,z} = 1$ if $w_{v,m,z} > 0$; $x_{v,m,z} = 0$ o.w.
3. **let** $\Delta = \sum_{v=1}^{V}\sum_{m=1}^{M}\sum_{z=1}^{Z_v} x_{v,m,z}\lceil\frac{qTr_v}{c_m}\rceil - dTS$
4. **foreach** $v \in [1,V]$, $m \in [1,M]$, $n \in [1,m]$, $z \in [1,Z_v]$
5. **compute** $y_{v,m,n,z}$ using Eqs. (3d) and (3e)
6. **while** $\Delta > 0$
7. **foreach** $v \in [1,V]$, $m \in [1,M]$, $z \in [1,Z_v]$,
7. where $x_{v,m,z} = 1$
8. **update** $y_{v,m,n,z}$ and **compute** $\alpha_{v,m,z}$, $\beta_{v,m}$,
8. and $\tau_{v,m,z}$
9. **let** v^*, m^* z^* lead to the minimum τ_{v^*,m^*,z^*}
10. **let** $x_{v^*,m^*,z^*} = 0$
11. **let** $\Delta = \Delta - \beta_{v^*,m^*,z^*}$
12. **return** $\mathbf{x}$

Fig. 3. SCG:An efficient algorithm to solve the single-cell allocation problem.

mobile devices smoothly render the video. Last, the constraints in Eqs. (3d) and (3e) are from Eqs. (1) and (2).

B. Proposed Algorithms: SCOPT and SCG

The proposed resource allocation algorithms run on the resource allocators close to base stations to determine how to stream videos to maximize the overall energy saving of mobile devices. The formulation in Eq. (3) is a BIP problem, which may be solved by existing optimization problem solvers, such as CPLEX [24] and GLPK [25]. We use CPLEX to implement the optimal algorithm and refer to it as SCOPT (Single-Cell OPTimum). Although SCOPT gives us the optimum allocations, its worst-case running time is exponential. Therefore, we develop a greedy algorithm, called SCG (Single-Cell Greedy) in the following. We start from an ideal decision in which the number of blocks is more than enough to enable unicasts to all mobile devices. Setting up a unicast channel to each mobile device *maximizes* the overall energy saving. However, the constraint in Eq. (3b) may prevent us from setting up a unicast channel for each mobile device, which renders the ideal decision infeasible. To turn an infeasible allocation into a feasible one, we can reduce the number of unicast/multicast with different MCS modes of a video, and hope the constraint in Eq. (3b) can be satisfied. For example, by changing $x_{1,3}$ from 1 to 0, we reduce the network load attributed to the on-demand streaming service by $\lceil\frac{qTr_1}{c_3}\rceil$ blocks. Doing so, however, leads to a negative consequence: mobile devices watching v with MCS mode 3 have to *receive* at a lower MCS mode. This in turn leads to lower energy saving γ in Eq. (3a). This illustrative example demonstrates the trade-off between *profit* (Eq. (3a)) and *cost* (Eq. (3b)).

We let $\alpha_{v,m,z}$ and $\beta_{v,m}$ be the *offsets* of profit and cost after changing $x_{v,m}$ of an allocation from 1 to 0. Mathematically, we write $\alpha_{v,m,z} = \sum_{m'=m}^{M} w_{v,m',z} y_{v,m',m,z}\lceil\lceil\frac{qTr_v}{c_m}\rceil/S\rceil$ and $\beta_{v,m} = \lceil\frac{qTr_v}{c_m}\rceil$. Our greedy algorithm strives to *refine* an infeasible allocation by trading the minimum profit reduction (objective function) for the maximum cost reduction (constraint). In particular, our algorithms evaluate the ratio $\tau_{v,m,z} = \alpha_{v,m,z}/\beta_{v,m}$ of all $x_{v,m,z} = 1$ and drop the MCS mode m and video v with the smallest $\tau_{v,m,z}$ value in each iteration. Our algorithm stops once the constraint in Eq. (3b) is satisfied. The pseudocode of SCG is given in Figure 3.

Lemma 2 (Correctness and Complexity): The SCG algorithm gives a feasible allocation and terminates in polynomial time: $O(V^2M^3Z^2)$, where $Z = \max_{v=1}^{V} Z_v$.

Proof: The while-loop starts from line 6 ensures the correctness. Let $Z = \max_{v=1}^{V} Z_v$. The dominating complexity occurs in lines 6–8: (i) the while-loop starts from line 6 iterates VMZ times in the worst-case, (ii) the for-loop starts from line 7 repeats up to VMZ times, and (iii) line 8 updates up to M $y_{v,m,n,z}$ values. Collectively, the time complexity of the SCG algorithm is $O(V^2M^3Z^2)$. ■

We note that for real networks, V, M, Z are not large numbers and the complexity does not depend on the number of users, which can be large. For example, the maximum number of videos that can be concurrently streamed on the most recent LTE network is 23 [26], assuming average video bit rate of 1736 Kbps [27] and maximum wireless bandwidth of 20 MHz [28]. Similarly, the largest value for M is 28 [26], and for Z is 5 [29] assuming an allocation window of 10 seconds. Moreover, all computations are simple scalar operations. Thus, the algorithm can easily run in real time. In the evaluation section, we show that SCG produces solutions close to those of SCOPT and terminates in a few milliseconds.

V. Extended Formulation for Single Frequency Networks

The formulation in Eq. (3) considers a single cell. In real deployments, Single Frequency Networks help mobile devices

$$\max_{\mathbf{x}} \quad \gamma = 1 - \frac{1}{\sum_{h'=1}^{H} N^{h'}} \Big[\sum_{h'=1}^{H} \sum_{v'=1}^{V} \sum_{m'=1}^{M} \sum_{z'=1}^{Z_{v'}} \hat{w}_{v',m',z'}^{h'} \sum_{n'=1}^{m'} y_{v',m',n',z'}^{h} \Big\lceil \Big\lceil \frac{qTr_{v'}}{c_{n'}} \Big\rceil / S \Big\rceil \Big] \tag{4a}$$

$$\text{s.t.} \quad \sum_{v'=1}^{V} \sum_{m'=1}^{M} \sum_{z'=1}^{Z_{v'}} x_{v',m',z'}^{h} \Big\lceil \frac{qTr_{v'}}{c_{m'}} \Big\rceil \leq dTS, \ \ \forall\, h \in [1,H] \tag{4b}$$

$$(1 - \sum_{n'=1}^{m} y_{v,m,n',z}^{h}) \hat{w}_{v,m,z}^{h} = 0, \ \ \forall\, h \in [1,H] \tag{4c}$$

$$y_{v,m,n,z}^{h} \leq 1 - x_{v,m',z}^{h}, \ \ \forall\, m' \in [n+1, m], \ h \in [1,H] \tag{4d}$$

$$y_{v,m,n,z}^{h} \leq x_{v,n,z}^{h}, \ \ \forall\, h \in [1,H] \tag{4e}$$

$$\hat{w}_{v,m,z}^{h} = w_{v,m,z}^{h} + \sum_{h' \in [1,H] \setminus \{h\}} x_{v,m,z}^{h'} \delta_{v,m,z}^{h,h'}, \ \ \forall\, h \in [1,H] \tag{4f}$$

$$x_{v,m,z}^{h} \in \{0,1\}, y_{v,m,n,z}^{h} \in \{0,1\}, \ \forall\, v \in [1,V], m \in [1,M], n \in [1,m], h \in [1,H], z \in [1,Z_v].$$

to improve the Signal-Interference-plus-Noise-Ratio (SINR). This allows some mobile devices to receive at higher MCS modes in order to increase the off time for higher energy saving. In this section, we consider H hexagonal cells that form a *dynamic* Single Frequency Network, where each block can be assigned to a Single Frequency Network independently. Such an extension requires two major enhancements: (i) expanding the solution space to multiple cells and (ii) modeling Single Frequency Network gains from neighboring cells. We explain each of the enhancements below.

Expanding Solution Space. We concurrently consider H cells, and add a *superscript* h ($h \in [1,H]$) to variables whenever applicable. For example, N_v^h denotes the number of mobile devices in cell h ($h \in [1,H]$) who watch video v ($v \in [1,V]$). As another example, we let $x_{v,m,z}^h$ ($v \in [1,V]$, $m \in [1,M]$, $z \in [1,Z_v]$, and $h \in [1,H]$) be the decision variable in the extended formulation. Adding the superscript allows us to expand the solution space for all H cells.

Modeling Single Frequency Network Gains. In the single-cell formulation, we assume that $w_{v,m,z}$ ($v \in [1,V]$, $m \in [1,M]$, $z \in [1,Z_v]$) is an input to our problem. In real systems, $w_{v,m,z}$ is a function of the SINR levels of individual mobile devices. The precise function depends on the MCS adaptation algorithm, which can be as simple as a stair-wise function to guarantee a certain bit error rate, say $< 5\%$. The actual MCS adaptation algorithm belongs to the link layer, and is out of the scope of this paper. Without loss of generality, we model the Single Frequency Network gain of mobile devices watching allocation window z ($z \in [1,Z_v]$) of video v ($v \in [1,V]$) with maximum MCS mode m ($m \in [1,M]$), from cell h' ($h' \in [1,H]$) to cell h ($h \in [1,H]$, $h \neq h'$) by $\delta_{v,m,z}^{h,h'}$, which represents the number of more/fewer mobile devices in h that have maximum MCS mode m if cell h' would transmit allocation window z of video v with MCS mode m as well. Upon considering the Single Frequency Network gains from all cells, the number of mobile devices with maximum MCS mode m in cell h is written as: $\hat{w}_{v,m,z}^h = w_{v,m,z}^h + \sum_{h' \in [1,H] \setminus \{h\}} x_{v,m,z}^{h'} \delta_{v,m,z}^{h,h'}$.

Combining these two enhancements, we get the formulation for a Single Frequency Network in Eq. (4). The objective function in Eq. (4a) maximizes the average energy saving across all H cells. The constraint in Eq. (4b) makes sure that each cell is not overloaded. The constraint in Eq. (4c) ensures that every mobile device receives at an MCS mode, which is equal to or smaller than its maximum MCS mode. The constraints in Eqs. (4d) and (4e) relate variables $y_{v,m,n}^h$ and $x_{v,m}^h$. The constraint in Eq. (4f) takes the Single Frequency Network gains into consideration.

Eq. (4) is a BIP problem and can be solved by existing optimization solvers [24], [25] for an optimal algorithm like SCOPT. Moreover, our heuristic algorithm (SCG) may be extended to solve Eq. (4). One possible extension is to start with transmitting each video with as many MCS modes as possible, and iteratively reduce the network load of the cell that suffers from the largest excessive network load. We notice that there may be other ways to extend SCG for Eq. (4). However, the detailed design is out of scope of this paper due to the space limitations.

VI. Evaluation

A. Setup

Simulator and Algorithms. We have implemented an on-demand video streaming system in OPNET [5], which is a detailed packet-level simulator. We have also implemented the proposed SCG and SCOPT using a mixture of C/C++, Matlab, and CPLEX [24] in the simulator. The heuristic SCG algorithm is evaluated against the optimal solutions generated by SCOPT. In addition, we have implemented unicast- and multicast-only policies employed by the current systems, and we refer to them as CUR_u and CUR_m in simulation results. CUR_u sets up a unicast connection to each mobile device, while for each video, CUR_m selects the minimal MCS mode of all mobile devices receiving that video.

Wireless Network Configurations. Although our proposed algorithms are general, we use LTE networks in our simulations. Several enhancements on the OPNET LTE module have been made, as detailed in the following. To enable multicast, we employ eMBMS *bearers* in LTE downlinks. Each bearer periodically delivers data bursts within every Common Subframe Allocation (CSA) period for energy saving. More details about LTE networks and their configurations can be found in [30], [31]. We consider MCS modes 4, 8, 14, and 22 [26] to support diverse channel conditions, so that each bearer can carry a video with a minimal bit rate of 256 kbps, which is

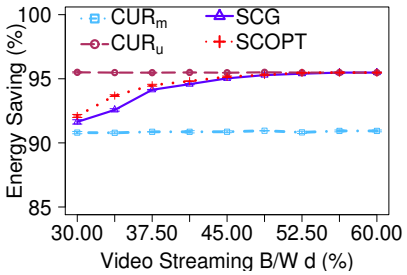

Fig. 4. Energy saving under different bandwidth constraints in an LTE 10 MHz eMBMS system.

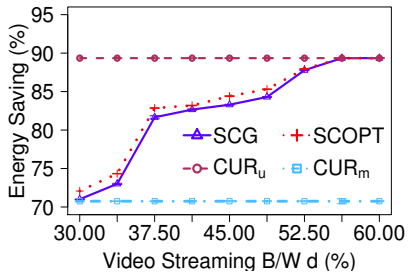

Fig. 5. Energy saving under different bandwidth constraints in an LTE 3 MHz eMBMS system.

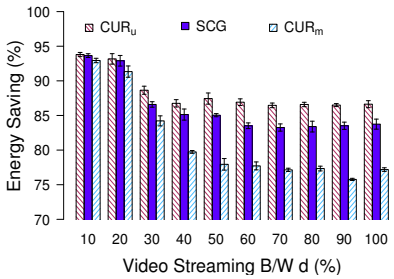

Fig. 6. Energy saving in large-scale simulations of 1,000 mobile devices.

TABLE I
LTE NETWORK CONFIGURATIONS

Parameter	Value
Physical Profile	LTE 10 MHz FDD
Maximal Transmission Power	0.01 Watt
eNodeB Antenna gain (dBi)	15 dBi
User Equipment Antenna Gain (dBi)	-1 dBi
Common Subframe Allocation (CSA) Period	8 frames
eMBMS Subframe Allocation per Frame	6 subframes (Max.)
Maximum Downlink Bit Rate	300 Kbps
Modulation and Coding Scheme (MCS)	4, 8, 14, 22
Evolved Packet System Bearer for Uplink	Best effort
Propagation Model	Free space, Walfisch-Ikegami line of sight
Video Stream Bit Rate	256 Kbps

a common bit rate for mobile devices. For each bearer, we adjust the time intervals between any two adjacent bursts per the standard [32], [26] in order to prevent overflow/underflow of the ingress link-layer buffer. The simulator runs the resource allocation algorithm once every allocation window of 10 s. The solutions are then mapped to the bearers, i.e., we map a general resource allocation to an LTE-specific allocation for OPNET. Table I gives the LTE configurations used in our simulations, all other configurations follow the defaults set by OPNET.

We consider one and multiple cells for single-cell and single frequency networks respectively, where each cell covers a 10×10 km^2 area. We consider up to 1,000 mobile users in the whole area, and the mobile users join the system following a Poisson process with mean λ, which is set to 20 s by default. The mobile users are randomly deployed in the covered area, so that more mobile users are close to the base stations as cellular operators build more base stations in more crowded areas. In particular, we assume 90% of mobile users are located within 1/3 of cell radius. These mobile users either: (i) are static or (ii) follow Random Waypoint mobility model. Upon joining the system each mobile user randomly requests for video.

Videos. For realistic video characteristics, we crawl YouTube to collect 1,000 videos and we sort them on popularity. We then employ the Zipf distribution with a skewness factor α to assign synthetic popularity to each video, so as to exercise a wider range of popularity distributions. We set $\alpha = 1.5$ if not otherwise specified. The YouTube videos are in 240p, and we scale the bit rates up by a factor of 9 to emulate 720p videos, which are popular on modern smartphones. The resulting popularity, video size, and video quality are used to drive our simulator.

Performance Metrics. We consider the following performance metrics and report average results with 95% confidence intervals whenever applicable.

- *Energy saving*: the fraction of time each mobile device can turn off its network interface to save energy.
- *Service ratio*: the fraction of mobile devices that can be served by the video streaming service under the given bandwidth constraint.
- *Service delay*: the time difference between a mobile device switch to a new video and the first packet of that video arrives at the mobile device. It also affects how fast our allocations adapt to network dynamics.

B. Results

Near Optimality. We first compare the results achieved by our SCG algorithm versus those computed by the optimal algorithm (SCOPT) in terms of energy saving. We simulate a small number of mobile users (25) in order to be able to compute the optimal results. Figures 4 and 5 present the sample results from 25 static mobile devices in 10 and 3 MHz LTE eMBMS networks, respectively, under diverse d between 30% and 60%. In both figures, the energy saving achieved by our algorithm is very close to the optimal. In the same figures, we plot the results achieved by the current multicast-only CUR$_\text{m}$ and unicast-only CUR$_\text{u}$ algorithms for comparison. We make three observations on these two figures. First, SCOPT/SCG outperform CUR$_\text{m}$ by 5% and 20% in energy saving. This is because SCOPT/SCG leverage unicast communications to apply more aggressive MCS modes on mobile devices closer to the base station. Second, when more bandwidth is allocated for video streaming service (larger d values), SCOPT/SCG achieve higher energy saving, which approaches that of CUR$_\text{u}$. However, different from SCOPT/SCG that support all 25 mobile devices, CUR$_\text{u}$ can only support 12 mobile devices when $d = 30\%$ (not shown in the figure). Third, while SCG achieves energy saving very close to SCOPT, SCG terminates in < 1 ms, while SCOPT may take as long as 200 ms.

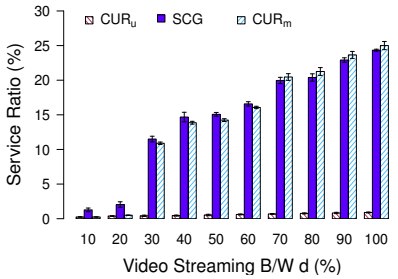

Fig. 7. Service ratio in large-scale simulations of 1,000 mobile devices.

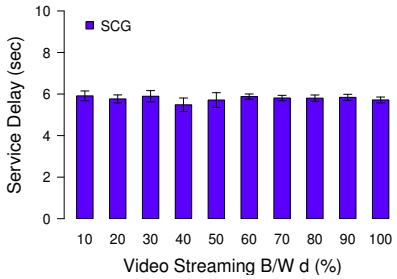

Fig. 8. Service delay in large-scale simulations of 1,000 mobile devices.

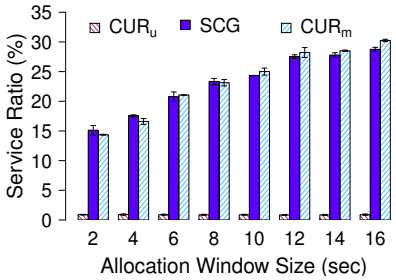

Fig. 9. Implications of window size on service ratio.

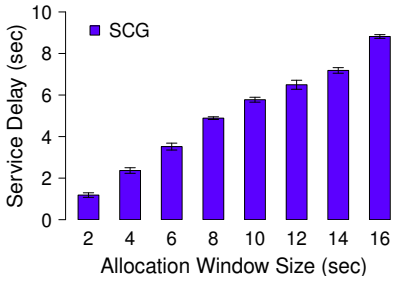

Fig. 10. Implications of window size on delay.

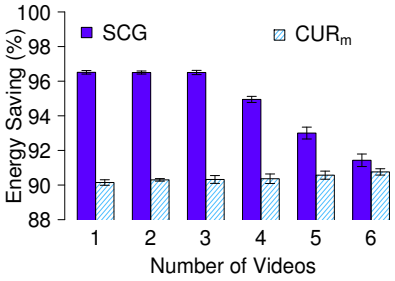

Fig. 11. Energy saving vs. number of videos.

In summary, SCOPT/SCG outperforms CUR_m and CUR_u in energy saving and/or service ratio, and SCG runs faster without compromising the optimality too much. Hence, in the rest of this paper, we no longer report results from SCOPT.

Scalability to Support More Users. We next consider 1,000 mobile devices in a 20 MHz LTE eMBMS network. Figure 6 plots the average energy saving achieved by different algorithms. This figure shows that the proposed SCG algorithm constantly outperforms CUR_m in terms of energy saving, and the gap is larger when more bandwidth is reserved for the video streaming service. Figure 7 presents the average service ratio, which shows that CUR_u only supports < 20 mobile devices, while the proposed SCG supports up to about 240 mobile devices concurrently: a 6X improvement in scalability. This figure also reveals that SCG outperforms CUR_m in service ratio when $d \leq 60\%$, which is because the proposed SCG algorithm intelligently allocates the resources among mobile devices, while CUR_m is first-come, first-serve. We note that reserving more than 60% of resource blocks for on-demand videos is not a typical setup. Even if that's the case, the SCG algorithm still outperform CUR_m in energy saving as indicated in Figure 6. Last, Figure 8 reports the service delay of SCG under diverse d values, which shows the resulting average delay is about 6.5 s, which is slightly over half of the allocation window length. If a shorter service delay is required, patching techniques [20], [21], [22] can be adopted.

We note that, different from the results from static users (Figures 4 and 5), the energy saving of the SCG algorithm in Figure 6 is not increasing along with larger d values. This is because the number of static users is much smaller, and the SCG algorithm always achieves 100% service ratio. Therefore, more reserved bandwidth enables more energy-efficient allocation. In contrast, the number of mobile users is large, and more reserved bandwidth leads to higher service ratio as indicated in Figure 7.

Implications of Window Size. We vary the allocation window size between 2 and 16 s. Intuitively, longer allocation windows lead to more rooms for optimization, and thus higher service ratios. We plot the service ratio of the proposed SCG algorithm in Figure 9, which shows a small increasing trend. On the other hand, Figure 10 gives the average service delay under different allocation window sizes. This figure shows that the service delay increases significantly with longer allocation windows. Combining these two figures, we conclude that the benefits of larger allocation window sizes are out-weighed by the longer service delay. Given that the SCG algorithm terminates in < 1 ms in our simulations, we recommend short allocation window size.

Mobility. We configure mobile devices to move following Random Waypoint mobility model. The mobility speed is randomly chosen between 0–72 km/hr. We consider up to 50 mobile devices in this experiment. With the proposed SCG algorithm, the resulting energy saving does not change much with the number of mobile devices, even when they are mobile.

The observed energy saving is about 96%, which is higher than 91% achieved by CUR_m. CUR_u can not support all 50 mobile devices, and thus we do not report the results from it.

Number of Videos. We next vary the number of videos in the video streaming service. We employ a 10 MHz LTE eMBMS network with $d = 60\%$. There are 36 mobile users. Figure 11 reports the energy saving achieved by different algorithms under diverse number of videos. This figure shows that SCG always outperforms CUR_m, and the performance gap is larger with fewer videos, which can be attributed to the larger optimization room leveraged by the SCG algorithm.

VII. Conclusions

We studied the resource allocation problem of a hybrid multicast-unicast video streaming service in cellular networks. The goal of employing the hybrid model is to support the ever increasing number of mobile users requesting video services. We formulated an optimization problem for the hybrid streaming service, which maximizes the number of supported mobile devices, and their energy saving in single-cell networks. We showed that our problem is NP-Complete, and described an optimal solution (SCOPT). To avoid exponential running time, we proposed an efficient algorithm (SCG), which terminates in polynomial time and produces near optimal results. Our simulation results indicate that: (i) SCG results in higher energy saving than the algorithms that only support either unicast or multicast, (ii) SCG achieves energy saving very close to that of SCOPT, while it can run in real time, and (iii) more bandwidth reserved for video streaming service or fewer supported videos result in higher energy saving.

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the British Columbia Innovation Council, and the National Science, Technology and Innovation Plan (NSTIP) of the Kingdom of Saudi Arabia.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 20122017," http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.

[2] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," *IEEE Network Magazine*, vol. 20, no. 2, pp. 50–55, March 2006.

[3] "Improved video support for Packet Switched Streaming (PSS) and Multimedia Broadcast/Multicast Service (MBMS) services (release 9). Third Generation Partnership Project (3GPP) standard TR 26.903 ver. 9.0.0."

[4] "Mobile TV is back: Ericsson launches broadcast video for 4G," http://www.theregister.co.uk/2013/02/25/broadcast_tv_yet_again/.

[5] OPNET Technologies, Inc., March. 2012, http://www.opnet.com.

[6] E. Dahlman, S. Parkvall, and J. Sköld, *4G LTE/LTE-Advanced for Mobile Broadband*. Elsevier Ltd., 2011.

[7] O. Hillestad, A. Perkis, V. Genc, S. Murphy, and J. Murphy, "Delivery of on-demand video services in rural areas via IEEE 802.16 broadband wireless access networks," in *ACM international workshop on Wireless multimedia networking and performance modeling*, Terromolinos, Spain, October 2006, pp. 43–52.

[8] A. Majumdar, D. Sachs, I. Kozintsev, K. Ramchandran, and M. Yeung, "Multicast and unicast real-time video streaming over wireless LANs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 524–534, June 2002.

[9] J. Lee, H. Park, S. Choi, and J. Choi, "Adaptive hybrid transmission mechanism for on-demand mobile IPTV over WiMAX," *IEEE Transactions on Broadcasting*, vol. 55, no. 2, pp. 468–477, June 2009.

[10] H. Hlavacs and S. Buchinger, "Optimal server bandwidth for mobile video on demand," *Annals of Telecommunications*, vol. 65, no. 1, pp. 31–46, February 2010.

[11] J. Yoon, H. Zhang, S. Banerjee, and S. Rangarajan, "MuVi: A multicast video delivery scheme for 4G cellular networks," in *Proc. of ACM International Conference on Mobile Computing and Networking (Mobicom'12)*, Istanbul, Turkey, August 2012, pp. 209–220.

[12] M. Anand, E. Nightingale, and J. Flinn, "Self-tuning wireless network power management," *Wireless Networks*, vol. 11, no. 4, pp. 451–469, July 2005.

[13] X. Zhang and K. Shin, "E-MiLi: energy-minimizing idle listening in wireless networks," *IEEE Transactions on Mobile Computing*, vol. 11, no. 9, pp. 1441–1454, September 2012.

[14] H. Zhu and G. Cao, "On supporting power-efficient streaming applications in wireless environments," *IEEE Transactions on Mobile Computing*, vol. 4, no. 4, pp. 391–403, July 2005.

[15] C. Luna, Y. Eisenberg, R. Berry, T. Pappas, and A. Katsaggelos, "Joint source coding and data rate adaptation for energy efficient wireless video streaming," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1710–1720, December 2003.

[16] R. Ghaffar, "LTE-Advanced Multi-User MIMO: Improved feedback and precoding design," in *IEEE International Vehicular Technology Conference (VTC Fall)*, Quebec City, Canada, September 2012, pp. 1–5.

[17] Y. Zhao, N. Do, S. Wang, C. Hsu, and N. Venkatasubramanian, "O^2SM: Enabling efficient offline access to online social media and social networks," in *Proc. of ACM/IFIP/USENIX International Conference on Middleware (Middleware'13)*, Beijing, China, December 2013.

[18] S. Wang, T. Lin, Y. Wang, C. Hsu, and X. Liu, "Poster: Fusing prefetch and delay-tolerant transfer for mobile videos," in *Proc. of ACM International Conference on Mobile Systems, Applications and Services (MobiSys'13)*, Taipei, Taiwan, June 2013, p. 525.

[19] Y. Yu, P. Hsiu, and A. Pang, "Energy-efficient video multicast in 4G wireless systems," *IEEE Transactions on Mobile Computing*, vol. 11, no. 10, pp. 1508–1522, October 2012.

[20] H. Hlavacs and S. Buchinger, "Hierarchical video patching with optimal server bandwidth," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 4, no. 1, pp. 8:1–8:23, January 2008.

[21] C. Griwodz, M. Liepert, M. Zink, and R. Steinmetz, "Tune to Lambda patching," *SIGMETRICS Performance Evalulation Review*, vol. 27, no. 4, pp. 20–26, March 2000.

[22] A. Bar-Noy, J. Goshi, R. Ladner, and K. Tam, "Comparison of stream merging algorithms for media-on-demand," *Multimedia Systems*, vol. 9, no. 5, pp. 411–423, 2004.

[23] J. Kim, T. Kwon, and D. Cho, "Resource allocation scheme for minimizing power consumption in OFDM multicast systems," *IEEE Communications Letters*, vol. 11, no. 6, pp. 486–488, June 2007.

[24] "IBM ILOG CPLEX," http://www.ibm.com/software/integration/optimization/cplex-optimizer.

[25] "GLPK (GNU Linear Programming Kit)," http://www.gnu.org/software/glpk/.

[26] "Evolved Universal Terrestrial Radio Access (E-UTRA); physical layer procedures (release 9). Third Generation Partnership Project (3GPP) standard TS 36.213 ver. 9.2.0," 2010.

[27] "Recommended bit rates for live streaming," http://www.adobe.com/devnet/adobe-media-server/articles/dynstream_live/popup.html.

[28] E. Dahlman, S. Parkvall, and J. Sk$\acute{}$old, *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, Inc., 2011.

[29] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," in *Proc. of the 3rd Multimedia Systems Conference (MMSys '12)*, New York, NY, USA, February 2012, pp. 89–94.

[30] "LTE Model User Guide, OPNET Modeler 9.1 Documentation."

[31] Y. Zaki, T. Weerawardane, C. Görg, and A. Timm-Giel, "Long term evolution (LTE) model development within OPNET simulation environment," in *OPNET workshop*, Washington, DC, August 2011.

[32] F. Hartung, U. Horn, J. Huschke, M. Kampmann, T. Lohmar, and M. Lundevall, "Delivery of broadcast services in 3G networks," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 188–199, March 2007.