Mobile Video Streaming over Dynamic Single-Frequency Networks

SALEH ALMOWUENA and MOHAMED HEFEEDA, Simon Fraser University

The demand for multimedia streaming over mobile networks has been steadily increasing over the past several years. For instance, it has become common for mobile users to stream full TV episodes, sports events, and movies while on the go. Unfortunately, this growth in demand has strained the wireless networks despite the significant increase of their capacities with recent generations. Hence, efficient utilization of the expensive and limited wireless spectrum remains an important problem, especially in the context of multimedia streaming services that consume a large portion of the bandwidth capacity. In this article, we introduce the idea of dynamically configuring cells in wireless cellular networks to form single-frequency networks based on the multimedia traffic demands from users in each cell. We formulate the resource allocation problem in such complex networks with the goal of maximizing the number of served multimedia streams, and we prove that this problem is NP-Complete. Then we present an optimal solution to maximize the number of served multimedia streams within a cellular network. This optimal solution, however, may suffer from an exponential time complexity in the worst case, which is not practical for real-time streaming over large-scale networks. Therefore, we propose a heuristic algorithm with polynomial running time to provide faster and more practical solution for real-time deployments. Through detailed packet-level simulations, we assess the performance of the proposed algorithms with respect to the average service ratio, energy saving, video quality, frame loss rate, initial buffering time, rate of re-buffering events, and bandwidth overhead. We show that the proposed algorithms achieve substantial improvements in all of these performance metrics compared to the state-of-the-art approaches. For example, for the service ratio metric, our algorithms can serve up to 11 times more users compared to the unicast approach, and they achieve up to 54% improvement over the closest multicast approaches in the literature.

CCS Concepts: • **Networks** → **Network architectures**; **Mobile networks**

Additional Key Words and Phrases: Mobile multimedia, single-frequency network, wireless streaming

ACM Reference Format:
Saleh Almowuena and Mohamed Hefeeda. 2016. Mobile video streaming over dynamic single-frequency networks. ACM Trans. Multimedia Comput. Commun. Appl. 12, 5s, Article 81 (November 2016), 26 pages.
DOI: http://dx.doi.org/10.1145/2983635

1. INTRODUCTION

The small size of digital integrated circuits accompanied with their reasonable cost has helped in introducing mobile terminals equipped with high processing capabilities, improved graphical user interfaces, and multiple radio antennas. Such hardware improvement has eventually enabled users to enjoy an enormous number of attractive features, including the ability to leisurely watch high-quality video streams on their

This work is supported in part by the National Science, Technology and Innovation Plan (NSTIP) of the Kingdom of Saudi Arabia, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and by the Qatar National Research Fund (grant # [NPRP8-519-1-108]).
Authors' addresses: S. Almowuena and M. Hefeeda, School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6; emails: salmowue@sfu.ca, mhefeeda@cs.sfu.ca.

DOI: http://dx.doi.org/10.1145/2983635

mobile phones. For instance, traffic analysis reports, such as Ciscos Visual Networking Index [Cisco 2015], show that the data traffic over mobile networks was equivalent to 2,500 petabytes per month in 2014. It is expected that this traffic will increase almost 10 times to reach 24,000 petabytes per month by the end of 2019. The same report predicted that around 72% of this traffic will most likely carry videos, because of the emergence of innovative media services and the accelerated adoption of other media services. For example, Netflix added 13 million new members in 2014, bringing its global total to 57.4 million paid subscribers [Netflix 2014], whereas YouTube nowadays has more than 1 billion users watching hundreds of millions of hours every day [YouTube 2014]. Two main cellular operators in Canada also launched their video-on-demand streaming services at the end of 2014 [Shomi 2014; Bell 2014] to provide more than 10,000 hours of videos to their subscribers. To cope with this growing demand, cellular service providers may need to rely on multicast capabilities of current and future cellular networks whenever possible. Currently, the Worldwide Interoperability for Microwave Access (WiMAX) standard defines the Multicast and Broadcast Service (MBS) in the data link layer in order to facilitate the process of initiating multicasting and broadcasting sessions [Cicconetti et al. 2006]. Similarly, the evolved Multimedia Broadcast Multicast Services (eMBMS) allows Long Term Evolution (LTE) cellular networks to deliver video streams over multicast groups [3GPP 2010].

With these multicast-capable networks, a streaming server can substantially reduce the wireless network load by serving mobile devices interested in the same video stream using a single multicast session. For example, a major telecommunication operator in the United States used multicast during the 2014 Super Bowl in New Jersey to serve multimedia content to more than 30,000 customers, which consumed about 1.9TB [Verizon 2014]. Other applications, including video-on-demand streaming, time-shifted events, and mobile video recorders, may benefit from the concept of multicast, since modern mobile devices have increasingly high storage space and can pre-stage some video data for later consumption. More specifically, for pre-staging, popular videos, such as episodes of latest TV shows and highlights of recent sports events, would be requested by users at different times, for example, in the evening of the release day. Because these videos are not immediately played back, their requests can be grouped into multicast sessions [Finamore et al. 2013]. These and similar applications offer optimization rooms to save the radio resources of mobile networks.

As defined in recent 4G standards, for example, Dahlman et al. [2013], multicast can be provided in two modes, which we refer to as *independent* and *single-frequency network (SFN)*. The independent mode provides multicast transmission within a single cell without any coordination or cooperation from neighboring cells. The SFN mode, however, represents a coordinated effort made by a set of base stations in order to transmit multimedia streams while minimizing the consumed wireless network resources. All base stations use the same frequency for the multicast sessions. Transmitting using SFN leads to significant improvements in the utilization of the wireless resources compared to transmitting using the independent mode. This is because, in the SFN mode, the coordinated cells are sending using identical radio signals, and thus receivers at the cell edges can get multiple copies of the same data but from different base stations. While these copies are considered *inter-cell interference* in independent cells, they are translated into useful signal energy in SFN. Hence, the strength of the received signal at the cell edge is enhanced, and the interference power at the same time is largely reduced. More information on how a single-frequency network manages its resources and operates in general can be found in 3GPP [2014]. An example for an existing deployment of a single-frequency network can be found in Nokia [2014], where a single frequency in the 700MHz LTE band has been utilized for TV broadcasting over a 200Km^2 square area in Munich, Germany.

Achieving the potential gains from multicast transmissions over SFN is, however, a challenging research problem. This is because the solution depends on finding the optimal configuration of cells within the SFN as well as adapting this configuration to handle the dynamic nature of the multimedia traffic and the users requesting this traffic. Although several works addressed various aspects of SFNs, such as coverage and modulation schemes [Rong et al. 2008; Talarico and Valenti 2014], and the size of an SFN and its impact on packet delivery [Alexiou et al. 2012], none of the previous works considered the much more challenging problem of managing the resources of multi-cell single-frequency networks in dynamic environments where the network traffic and users distribution change with time, which is the problem we address.

In this article, we consider a general model for wireless cellular networks that support multicast services, such as LTE and WiMAX. In this model, the network is composed of multiple cells. These cells can work independently from each other, so each cell can provide unicast and multicast services to users in its range. Cells can also collaborate by forming one or more SFNs. If a subgroup of cells forms an SFN, then a portion of the bandwidth is reserved for the multicast service in all participating cells, and the multicast service will be provided to all users within the range of this SFN. The cell membership in an SFN is dynamic, which means a cell can join or leave an SFN based on the demand from its current users. The specific problem we address is as follows: Given user demands for different video streams in various cells, determine the optimal configuration of the wireless network that maximizes the number of users served and the energy saving for mobile devices. More specifically, decide the number (zero or more) of SFNs that should be created and which cell should belong to which SFN. Furthermore, for each user request for a video session, decide whether it should be served using unicast, multicast in a single cell, or multicast across an SFN. This is a challenging problem; in fact, we prove that it is NP-Complete.

We simplify this problem and propose an optimal and a two-stage heuristic algorithms to solve it, which substantially improves the service ratio (i.e., the fraction of served requests to the number of received requests within the system) compared to current algorithms used in cellular networks. The contributions of this article can be summarized as follows:

—We introduce the novel idea of dynamically configuring cells in wireless cellular networks to form single-frequency networks based on the traffic demands from users in each cell.
—We formulate the resource allocation problem in multi-cell SFNs to serve multiple multimedia streams using various combinations of unicast and multicast sessions within each cell and across SFNs. We show that this problem is NP-Complete.
—We present an optimal algorithm to solve the multi-cell SFN resource allocation problem. To reduce its computational complexity and minimize its control overheads, we also introduce a heuristic algorithm consisting of two stages. The first stage is used by each base station to independently decide whether to form an SFN or join an existing one. The second stage computes the best option to serve each multimedia stream, whether unicast, multicast, or a combination thereof. We show that the heuristic algorithm achieves near-optimal results with respect to the achieved service ratio.
—We conduct an extensive simulation study using a detailed packet-level simulator Optimized Network Engineering Tools (OPNET) [OPNET 2010] to evaluate the proposed algorithms. Our results show that the proposed heuristic algorithm can serve up to 11× more users than the unicast streaming approachs. Compared to multicast approaches that do not use SFN, our algorithm can achieve up to 51% improvements in the number of users served. Even compared to approaches that do use SFN but

do not configure them dynamically, our algorithm achieves up to 14% improvement in the number of users served. In addition, our heuristic algorithm achieves better performance in terms of video quality, frame loss rate, and number of re-buffering events when it is compared against the state-of-the-art approaches in Araniti et al. [2013], Monserrat et al. [2012], and Lee et al. [2011]. The heuristic algorithm also runs in real time: It terminates in a *few milliseconds* on a commodity workstation. In real deployment, such algorithms run on servers once every *few seconds*; therefore, our algorithm is practical and efficient.

We note that a preliminary conference version of this work appeared in Almowuena and Hefeeda [2015]. The conference version focused on proposing a heuristic algorithm to reconfigure the areas of single-frequency networks, whereas this article introduces an optimal solution to maximize the service ratio within each cell and to act as an upper bound on the highest possible number of served users in the mobile system. This article also presents a theoretical analysis for the computational complexity of both the optimal and heuristic algorithms, and it shows the importance of taking into consideration the bandwidth overhead caused by both channel quality feedbacks and SFN control signals. Furthermore, this article presents more comprehensive simulations to evaluate the proposed algorithms in terms of several important performance metrics such as video quality, frame loss rate, initial buffering time, and rate of re-buffering events.

The rest of this article is organized as follows. Section 2 summarizes the related works in the literature. Section 3 describes the system model used in this article, and Section 4 states and formulates the considered problem. Sections 5 and 6 present the proposed optimal and heuristic algorithms, receptively. Section 7 presents our simulation results to assess the performance of our algorithms and compare them against others. Section 8 concludes the article.

2. RELATED WORK

Several works have been introduced to assess and improve the performance of multimedia multicast streaming over single-frequency networks. For instance, Rong et al. [2008] and Talarico and Valenti [2014] present analytical models to determine the coverage of a given configuration for single-frequency networks and how to utilize these models to choose the best-suitable modulation and coding scheme as well as the appropriate configuration for SFN areas. Having such knowledge prior to the network deployment helps in achieving a target bandwidth utilization. Urie et al. [2013] extend this assessment and provide a comprehensive evaluation of SFN performance under more realistic conditions. Alexiou et al. [2012] estimate the number of neighboring cells that should be enrolled into an SFN area such that a specific average signal-to-noise ratio is achieved and a minimum communication cost is incurred. To accomplish this goal, they calculate the cost of both packet delivery and signaling procedures under a set of different network topologies and user distributions. The works in Rong et al. [2008], Talarico and Valenti [2014], Urie et al. [2013], and Alexiou et al. [2012] assume a *static* SFN configuration in which cells are registered into a set of zones at early stages of deployment, and the enrollment of these cells do not change over time even if variations have been occurred for users distribution and network traffic. In contrast, we consider dynamic configuration of SFN areas, which is more useful in practice.

Given a particular configuration of SFN, the available radio resources should be allocated to a mixture of unicast and multicast services to optimize network utilization. As an example, Chen et al. [2013] optimize the unicast multimedia connections over Dynamic Adaptive Streaming over HTTP (DASH) with respect to fairness, stability, and efficiency. Elsherif et al. [2013] propose a resource allocation algorithm in

heterogeneous networks to minimize inter-cell interference and then maximize the system throughput. Their algorithm is relying on the concept of the shadow chasing technique, in which a feedback mechanism for link adaptation is exploited and interference is avoided through a probabilistic manner. Besides overcoming the inter-cell interference within the network, Lu et al. [2013] and Liang et al. [2012] aim at reaching additional objectives of achieving fairness among mobile terminals and providing adequate quality of service, respectively. To achieve these two goals, a graph is constructed to represent the possible interferences between every pair of base stations, and then the theory of vertex coloring is utilized to solve the problem of radio resources allocation within the network.

The concept of multicast over mobile network has been explored by a number of research works [Afolabi et al. 2013; Araniti et al. 2013, 2014; Won et al. 2009; Keller et al. 2012; Xu et al. 2010]. For instance, Araniti et al. [2013] and Won et al. [2009] address the issue of transmission scheduling over Orthogonal Frequency-Division Multiple Access (OFDMA), which is mainly related to the different data rate and quality requirements of users in the same multicast group. Keller et al. [2012] investigate the idea of data transmission via concurrent radio interfaces to enhance the connectivity of terminals in high-mobility scenarios and to improve the streaming bit rate in the downlink channels of multicast sessions. Xu et al. [2010] aim at maximizing the system capacity of a wireless network under a given total transmit power constraint by exploiting multiple input and output (MIMO) antennas at both transmitters and receivers. Although a number of algorithms have addressed the transmission scheduling problem for multicast service, a few research efforts have considered the hybrid unicast-multicast approach in the allocation problem. For example, Monserrat et al. [2012] and Lee et al. [2009] present two schemes in which both unicast and multicast connections are served to maintain fairness among mobile users and reduce the service blocking probability. On the other hand, Rahman et al. [2014] and Deng et al. [2012] utilize the hybrid approach to minimize the average power consumption of mobile terminals and guarantee a certain level for the quality of service, respectively.

In LTE networks, the discontinuous reception mode for energy saving is supported in both idle and connected states [Dahlman et al. 2013; 3GPP 2014]. Based on this concept, Hoque et al. [2014] introduce an energy-efficient video delivery system that relies on sending short data bursts using unicast conncetions, and these bursts are constructed in a way to reduce the power consumption and avoid any buffer violations. Hefeeda and Hsu [2010] also study the burst scheduling problem for optimal energy saving in mobile TV broadcast networks with arbitrary channel bit rates. They introduce an optimal algorithm for those special cases in which the bit rate of each TV channel is equivalent to the power of 2 times the lowest channel bit rate. Hsu and Hefeeda [2010] overcome this limitation and propose a near-optimal solution in which the energy saving in the system is maximized, transmission bursts are not overlapped, and buffer levels at each receiver are not violated.

Our proposed resource allocation algorithm in this article aims at increasing the bandwidth utilization of a cellular network, but it differs from the aforementioned works in two main aspects: (a) It utilizes an adaptive and flexible scheduling process, and (b) it takes an advantage of three transmission modes: unicast, multicast over SFN, and multicast within the local coverage of a cell. The fraction of radio resources reserved for multicast services is assumed to be constant in most existing scheduling algorithms, including those hybrid unicast-multicast methods in Monserrat et al. [2012], Lee et al. [2009], Rahman et al. [2014], and Deng et al. [2012]. In our case, the resource distribution between unicast and multicast connections is done *dynamically* based on which served request leads to better efficiency. Wireless cellular networks have reserved sub-frames for SFN transmission, and these standards cannot easily be

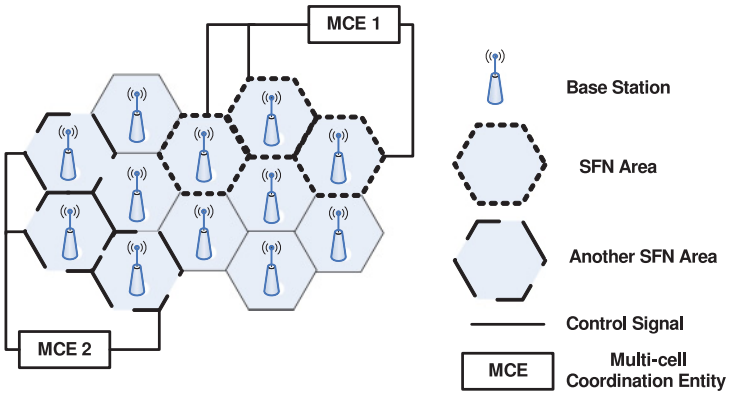


Fig. 1. The considered model for a mobile network.

changed. Therefore, we increase the transmission scheduling for multicast services by delivering some multicast connections through the available resources for unicast. Our proposed transmission scheduler is also an adaptive allocator because the modulation and coding scheme used for multicast services may not be suitable for those users with the worst channel conditions, especially in scenarios where their number is low.

The closest works to our proposed algorithms can be found in Araniti et al. [2013], Monserrat et al. [2012], and Lee et al. [2011], since they employ a mixture of multicast and unicast, allow splitting a multicast group into subgroups, and apply subgroup-based adaptive modulation and coding schemes. We compare our algorithms against these works, and we show that our algorithms outperform them with respect to the service ratio, energy saving, frame loss rate, and rate of re-buffering events.

3. SYSTEM MODELS

We first describe the cellular network model considered in this article and then present the assumed multimedia streaming model. We list all symbols and their definitions in Appendix A (Appendices are available online).

3.1. Wireless Network Model

We consider a wireless cellular network with base stations, mobile devices, and multi-cell coordination entities as illustrated in Figure 1. In addition to transmitting via unicast connections, our network utilizes two modes for multicast transmissions. The first mode is the *single-cell point-to-multipoint* transmission. This mode allows feedback from mobile terminals on their channel conditions to be sent to base stations, which are then used to dynamically adjust the modulation and coding schemes. The advantage of such mode is its adaptation to changes in the current distribution of users within a cell. Multicast services using this mode can be turned off within a particular cell in which there are no active users. The second mode of multicast transmission is the *multi-cell point-to-multipoint* approach, and it is called multicast over SFN. A single-frequency network represents a coordinated set of base stations in order to broadcast multimedia streams over a region of the network utilizing the same physical radio resources. To achieve such objective, a fixed modulation and coding scheme is applied to match the decoding requirements of the edge-user with the worst channel condition. Transmitting multicast services through SFN leads to significant improvements in the total spectral efficiency as compared to multicasting within a single-cell [Dahlman et al. 2013]. Since the coordinated cells are sending using an identical radio signals, receivers at cell edges get multiple copies of the same data but from different base stations. Hence, the

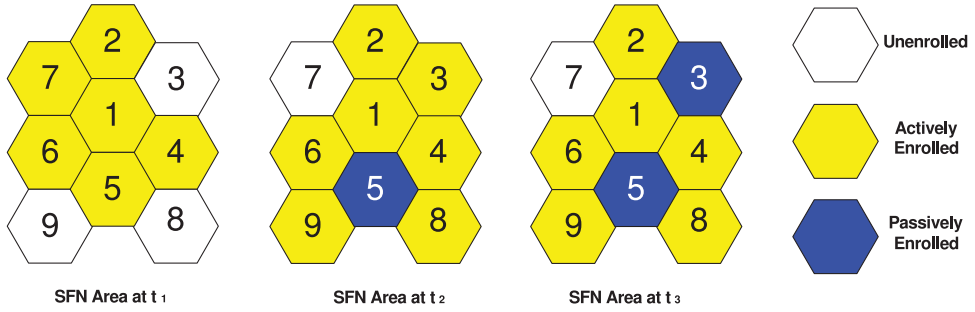


Fig. 2. Three possible types of cells within an SFN area.

strength of the received signal at the cell edge is enhanced, the interference power is largely reduced, and the overall performance remains consistent even if a user moves from one cell to another.

The mobile system shown in Figure 1 is divided into a number of SFN areas. A multi-cell coordination entity (MCE) ensures the full functionality of an SFN area by performing time synchronization as well as coordinating the usage of the same radio resources and transmission parameters across all cells belonging to a particular area. The cellular network can have many SFN areas at the same time, where a base station can join multiple SFNs up to a maximum of A_L areas. The radio resources of a base station are divided along both time, represented by sub-frame, and frequency, represented by sub-carrier, domains. Let the number of sub-carriers allocated to each cell be S. The resource allocation window has T sub-frames, indexed by t, and has a duration of Γ seconds. The smallest resource unit (resource block) in frequency-time space is identified by (s, t), where $s \in [1, S]$ and $t \in [1, T]$. Each allocation window consists of TS resource blocks, and we use d fraction of bandwidth, that is, dTS resource blocks, for video services. The user equipment is informed about which sub-frames are assigned to its video stream via a control channel broadcasted from its nearest base station, and the allocation can be changed dynamically at specified intervals. We assume that the maximum power for each base station is P and these base stations allocate equal power to their sub-carriers [Kim et al. 2012]. Hence, the sub-carrier power allocation is P/S.

In this article, we propose a novel approach to reconfigure SFN areas in a wireless cellular network based on video popularity and user distribution. To accomplish this objective, we assume that there are three possible types of cells within an SFN area, as shown in Figure 2. A cell can be either actively enrolled, un-enrolled, or passively enrolled in a certain area. Both actively and passively enrolled cells transmit a video stream through an identical radio resource channel. The main difference between both cells is that there are a number of users receiving this stream in the active cell, whereas passively enrolled cells have no interested users in this video stream. Passively enrolled cells are considered as helpers; they are formed to enhance the quality of streaming sessions in active SFN areas. Passively enrolled cells are only used when they themselves have a low amount of traffic. Un-enrolled cells do not belong to the multi-cell transmission, and they should have no impact on the signals broadcasted within their neighboring SFN cells. For this reason, an un-enrolled cell always tries to avoid causing any inter-cell interference by using the radio resources utilized for its neighboring SFN area for only unicast connections or single-cell point-to-multipoint services with a limited power. To determine the type of a certain cell in a mobile network, we assume that each base station works along with its nearest

multi-cell coordination entity to make such decision, and this decision is made regularly based on certain conditions as explained in Sections 5 and 6.1.

Similarly to any wireless communication, our mobile network is vulnerable to environmental phenomena such as shadowing, multipath fading, and interference. The channel qualities between mobile terminals and their nearby base station vary due to these phenomena as well as the mobility of viewers. To adapt to different channel qualities during data transmission and to allow the reception with an acceptable bit-error-rate, multiple modulation and coding schemes (MCS) are used as defined in OFDMA standards. The MCS mode is determined based on the wideband channel quality index report [Parruca and Gross 2013]. Higher MCS modes require good channel quality and lead to higher per-resource block capacity. On the other hand, smaller MCS modes are more robust and usable for diverse (both good and bad) channel qualities. For our system, the MCS mode is denoted by $m \in [1, M]$. Let n_m be the number of mobile terminals in a cell that can receive the data at MCS mode m or better. Here, $n_1 \geq n_2 \geq \cdots \geq n_M$. Similarly, $n_{vzm} : v \in [1, V], m \in [1, M]$ is the number of mobile terminals in a viewers' subgroup that can receive (buffer) the video segment (v, z) at MCS m mode or better, where $n_{vzm} \geq n_{vzm'}$ if $m < m'$. The per-resource block capacity (c_m) is non-decreasing quantity on $m \in [1, M]$ such that $c_1 \leq c_2 \leq \cdots \leq c_M$.

3.2. Streaming Model

We consider a general model that can be used for live as well as on-demand streaming but with some constraints. Live streaming is useful in several scenarios such as streaming sports events, live concerts, news, political debates, talks, seminars, and popular TV episodes. Live streaming is naturally suitable for multicast services as users are mostly synchronized: They are watching at the same moment in the video, and functions that may disrupt this synchrony, for example, fast forward, are not applicable. In addition, live streaming of popular events typically attracts a large number of *concurrent* users, which can put a huge load on the cellular network and may result in denying some users the streaming service because of the limited capacity. Our work in this article optimizes the wireless resources for live streaming by carefully creating a mixture of unicast and multicast sessions, which can be within single cells or across multiple coordinated cells.

In on-demand streaming, on the other hand, users arrive asynchronously to the system. That is, users may request the same video at different times and they can be watching at different moments in the video. This general asynchronous model for streaming is difficult to achieve using pure multicast, as very few users can form a multicast session, especially if they are requesting videos that are not popular. We consider a less general model, useful for requesting popular videos on relatively short periods of time such as requesting news clips during morning or afternoon commute times and streaming TV episodes during the evening peak watching times. This model is also useful for time-shifted viewing of various events and videos, where some users opt to watch such videos at different times than their original scheduled times. Even for such limited asynchronous models, multicast alone will not be sufficient to provide true-on-demand service without imposing long waiting times on users. To solve this problem, we adopt existing delay mitigation mechanisms, such as patching videos using temporary unicast connections [Hua et al. 1998; Hua and Sheu 1997; Paris and Long 2001; Eager et al. 2001; Hlavacs and Buchinger 2008]. For example, we implement an efficient patching algorithm in which a user can join any existing multicast group and receive the missing leading portions of this stream over a separate unicast connection. This unicast session is shorter than a predetermined threshold; otherwise, a new multicast group for the requested video is created.

In our streaming model, mobile terminals within a cell can request V different video streams. Mobile terminals receiving the same video streams may watch at different timestamps relative to that video. To facilitate these demands, each video stream $v \in [1, V]$ is divided into a number of segments $z \in [1, Z_v]$, and each video segment has a playout duration Γ, which equals the allocation window duration. Let $r_{vz} : v \in [1, V], z \in [1, Z_v]$ be the encoding rate of the video segment (v, z). Because of the variable-bit-rate nature of video streams, r_{vz} may vary across allocation windows. However, it is important to indicate that each video segment is non-adaptive (i.e., not DASH-like) in the sense that it has only a single quality representation and single data rate. Since the aggregate data rates for the requested segments, $\sum_v \sum_z r_{vz}$, in an allocation window may exceed the video service capacity, dTS, we need to decide which segments to transmit for each video with the objective of maximizing the service ratio among users.

We note that unicast and multicast sessions are in the last-hop of the cellular networks, between the mobile terminals and servers or close to the Internet Service Provider (ISP) natural gateways such as Broadband Remote Access Servers or Packet Data Network Gateways. We also assume that these content servers may be part of a content delivery network, a proxy cache [Finamore et al. 2013], or a transparent proxy [Hefeeda et al. 2011]. Such architectures are gaining increasing interest, as shown in Akamai [2013] and Nicosia [2010]. As explained by Nicosia [2010], many telecommunication providers see the rise of Internet streaming services as a nightmare for their businesses. They used to bill for every minute of voice transported or data carried. Instead, they are currently managing companies that transport more and more data traffic without any associated incremental revenues. For this reason, these providers have to either (a) impose traffic caps on their customers to avoid the negative impact of the added transport costs on their bottom line or (b) offer high-quality video delivery services through owning content severs or using their access infrastructure under a Unbundling of the Local Loop (ULL) agreement.

4. PROBLEM STATEMENT

In this section, we define and mathematically formulate our problem, which we divide into two sub-problems: the first determines the best configuration of single-frequency networks, and the second specifies the transmission scheduling for multimedia streams using a mixture of unicast and multicast sessions.

4.1. Problem Definition

Several wireless cellular networks consider static configurations for single-frequency networks [Dahlman et al. 2013]. However, static configurations are unaware of user distribution and video popularity across the mobile network. Therefore, these approaches may waste the radio resources of cells, especially in those scenarios where no mobile terminals are interested in a cell belonging to a predetermined multicast service. Dynamic configuration of SFNs provides more flexibility and thus can yield higher efficiency in using the radio resources. We consider an initial SFN configuration consisting of a number of hexagonal cells in which a set of mobile terminals are distributed within their transmission coverages. The information of transmitted videos in each cell along with the number of active viewers is periodically delivered to the nearest multi-cell coordination entity in the given wireless network. Our first problem in this article can be stated as follows:

Sub-Problem 1 (SFN Configuration). *For a given cell c, select the optimal subset of SFNs to join so the bandwidth utilization within this cell is maximized and the number of SFN zones in which c is enrolled does not exceed a predetermined limit.*

Under any given SFN configuration, base stations and their nearest multi-cell coordination entities should cooperate to allocate the available resource blocks for both multicast and unicast connections. Here, there is another issue that still exists for the transmission scheduling over OFDMA, which is mainly related to the different data rate and quality requirements of users in the same multicast group. Generally, mobile terminals close to the base station can obtain higher data rate, while cell-edge users are forced to reduce the data rate in order to minimize the bit error rate in data reception. Conventional multicast schemes adopt a conservative approach, which restricts the rate of the multicast session to the user with the worst channel condition. This approach introduces severe inefficiencies when some users (even if they are just a few) experience poor channel conditions. The objective of our proposed transmission scheduling algorithm is to address the aforementioned inefficiency problem of multicast communications in the presence of link quality differences among users within a multicast group. Instead of transmitting the video stream to a multicast group at a very low bit rate, it could be more efficient to eliminate some users from the multicast group and serve these users with unicast streams so they do not slow down all users in the multicast session.

Our second problem in this article is to consider a joint multicast-unicast transmission scheduling for bandwidth-efficient delivery of the requested videos to mobile terminals. In this approach, a base station dynamically handles an asynchronous incoming request for a certain video segment by either initiating a unicast stream or extending the number of participants in a multicast session. In particular, a base station along with its assigned multi-cell coordination entity need to compute a schedule that specifies: (i) which video streams to multicast, (ii) who can be enrolled into the multicast groups, and (iii) which video streams to unicast. This joint multicast-unicast problem can be stated as follows:

Sub-Problem 2 (Hybrid Transmission Scheduling). *Given an allocation window of T sub-frames and S sub-carriers, determine the optimal transmission schedule for video requests submitted by mobile terminals of diverse channel conditions and using a hybrid multicast-unicast streaming approach so the total service ratio across all mobile terminals is maximized.*

4.2. Problem Formulation

Sub-problems 1 and 2 have a common objective since they are aiming at maximizing the number of served users in the given mobile network. On this ground, we can formulate both of them into a single optimization problem. We use the Boolean decision variable x_{vzm} ($v \in [1, V],\ z \in [1, Z_v],\ m \in [1, M]$) to denote whether segment z of video v transmitted using MCS mode m. That is, $x_{vzm} = 1$ if the video segment (v, z) is transmitted using MCS mode m and $x_{vzm} = 0$ otherwise. We present the problem formulation in Equation (1), in which we consider the tradeoff between the number of served users at a certain modulation and coding scheme, that is, n_{vzm} and their resource requirements, that is, $r_{vz}\Gamma/c_m$. The objective function of Equation (1a) is to maximize the total number of served users within a cell. Equation (1b) implements the resource constraint. Equation (1c) ensures that at most one MCS mode is selected for each video segment. Equation (1d) ensures that the obtained bandwidth utilization at a certain instant of SFN reconfiguration (i) is greater than its value at a preceding configuration ($i - 1$). This condition helps in avoiding ineffective reconfigurations for SFN and limits the amount of signaling control overhead. Finally, Equation (1e) makes sure that the number of SFN zones in which c is enrolled does not exceed the maximum limit of zones A_L in which cell c is allowed to join, taking into account that A refers to the number of active SFN areas, $y_{a,c}$ is a binary number introduced to determine whether

c is enrolled in the area a. Solving this optimization problem, we can find the decision matrix $\mathbf{X}$ that contains the set of video segments selected for transmission and their corresponding MCS modes.

$$\max_{\mathbf{X}} \quad O = \sum_{v=1}^{V} \sum_{z=1}^{Z_v} \sum_{m=1}^{M} x_{vzm} n_{vzm} \tag{1a}$$

$$s.t. \qquad B = \sum_{v=1}^{V} \sum_{z=1}^{Z_v} \sum_{m=1}^{M} x_{vzm} \frac{r_{vz}\Gamma}{c_m} \leq dTS \tag{1b}$$

$$\sum_{m=1}^{M} x_{vzm} \leq 1 \tag{1c}$$

$$SFN_i \left(\frac{O}{B}\right) > SFN_{i-1} \left(\frac{O}{B}\right) \tag{1d}$$

$$\sum_{a=1}^{A} y_{a,c}^{t_i} \leq A_L \tag{1e}$$

$$x_{vzlm} \in \{0, 1\}, \forall\ v \in [1, V], z \in [1, Z_v], m \in [1, M] \tag{1f}$$

4.3. Problem Complexity

The formulation of Equation (1) is an integer programming problem, which is NP-Complete. To prove that, let us define the set of decision variables $a_{i,j,k}$ for all $i \in [1, S]$, $j \in [1, T]$, and $k \in [1, V]$ in the transmission scheduler such that $a_{i,j,k} = 1$ if the resource block in ith subchannel and jth symbol is allocated to video k, and 0 otherwise. Given an allocation assignment represented by the above variables, we can easily verify whether it is a valid assignment. We need to count the number of 1's and make sure that the count is $\leq dTS$. In order to ensure that the same resource block is not allocated to more than one video, we need just to check that for any $i \in [1, S]$, $j \in [1, T]$, and $\sum_{k=1}^{V} a_{i,j,k} \leq 1$. This checking operation can be done in polynomial time.

Now, we will use the 0-1 knapsack problem in order to show that an NP-Complete instance is reducible to our problem. Our instance of 0-1 knapsack is defined as follows: There are n items x_l such that $l \in [1, n]$ and $x_l = 1$ if the item is chosen and 0 otherwise. The value and weight of item x_l are defined by p_l and b_l, respectively. The capacity of the knapsack is W. We assume non-negative values and weights, and we would like to maximize $\sum_{l=1}^{n} x_l p_l$ subject to $\sum_{l=1}^{n} x_l b_l \leq W$ and $x_l \in \{0, 1\}$. To reduce this instance of 0-1 knapsack to an instance of our problem, we set $n = TSV$. We define a new variable $x'_{i,j,k}$ for each x_l for any $i \in [1, S]$, $j \in [1, T]$, and $k \in [1, V]$ such that $x'_{i,j,k} = 1$ if the resource block in ith subchannel and jth symbol is allocated to video k and 0 otherwise. We introduce another variable $p'_{i,j,k}$ for each p_l such that $p'_{i,j,k}$ denotes the number of mobile devices served by allocating resource block $x'_{i,j,k}$. We replace each b_l with new $w'_{i,j,k}$ and set $w'_{i,j,k} = 1$ for all $i \in [1, S]$, $j \in [1, T]$, and $k \in [1, V]$. Finally, we set the knapsack capacity to be $W = dTS$. This reduction can be done in polynomial time. The reduced 0-1 knapsack problem will have a solution if and only if our considered problem has a solution.

5. PROPOSED OPTIMAL ALGORITHM

The goal of our proposed optimal algorithm is to maximize the service ratio during a transmission scheduling window. The proposed solution is shown in Figure 8 in Appendix C. The algorithm starts by dividing the incoming requests into subgroups based on their required videos, the time instances of the requested segments, and the best suitable modulation and coding scheme for interested users. In other words, mobile terminals asking for the same video segment and sharing similar channel conditions are clustered together into a subgroup. The algorithm then examines all possible combinations of these subgroups using a dynamic programming approach. Once the optimal set of subgroups is computed, each cell decides which users would be served through multicast sessions, which users would be not admitted during this window, and which users would be served via unicast connections. The cell also reconfigures its enrollment in SFN areas based on the obtained solution and establishes the procedures of shrinking and joining with the help of its multi-cell coordination entities.

The proposed optimal transmission scheduling algorithm produces a feasible allocation with a time complexity in the order of $O(NdTS)$, where N is the number of mobile terminals in a cell generating requests for video streams and dTS is the number of radio resource blocks reserved for video services. We note that dTS, unlike N, is pseudo-polynomial and potentially exponential in the length of the input (i.e., the number of bits required to represent dTS). For real networks, the maximum number of videos that can be concurrently streamed on the most recent LTE network is 170 [3GPP 2014], assuming an average video bit rate of 300Kbps [Adobe 2009] and maximum bandwidth of 20MHz [Dahlman et al. 2013]. Based on these values, we can notice that the drawback of using an optimal solution would be its exponential running time in the worst cases. Another drawback would be the absence of constraints on the usage of control signals needed for reconfiguring the areas in single-frequency networks. That means an excessive overhead on the bandwidth might occur if the video popularity and user distribution are changing in a dynamic way.

For these two reasons, the next section introduces a faster algorithm to solve the problem of maximizing the average service ratio in video streaming over mobile network. The proposed heuristic algorithm also takes the control overhead into consideration and tries to minimize its occurrence without impacting the number of served users within cells. As shown in Section 7, both running time and bandwidth overhead are reduced by around 360.7% and up to 48.8%, respectively, while our heuristic algorithm maintains near-optimal results ($<$1.38% on average) with respect to the achieved service ratio.

6. PROPOSED HEURISTIC ALGORITHM

We propose a heuristic algorithm to solve the problem defined in Equation (1). The algorithm performs two main steps: (1) dynamic configuration for the single-frequency network and (2) transmission scheduling of incoming video requests received within a pre-defined scheduling window. The first step reconfigures a network and reconstructs its SFN areas dynamically by taking into consideration the popularity of videos and the signal-to-noise ratios of served terminals. The second step schedules incoming requests with an objective of maximizing the service ratio in a given system. This scheduling is done for every resource allocation window. The details of each step are described in the following subsections.

6.1. Dynamic SFN Configuration

To reconfigure the areas within a single-frequency network, it is possible to allow every multi-cell coordination entity to collect current traffic and user distribution within

Algorithm 1: Dynamic Configuration of SFN Areas

Inputs: $M_c \leftarrow$ A set of served multicast streams in c
$U_c \leftarrow$ A set of unserved streams in a cell c
$W \leftarrow$ The bandwidth still available for multicasting over SFN
$A \leftarrow$ A set of active SFN areas in which cell c can join
$A_c \leftarrow$ A set of active SFN areas in which cell c is enrolled
$\{\alpha, \lambda\} \leftarrow$ The parameters used in the user behavior model
$Bandwidth(v, c) \leftarrow$ computes the required bandwidth to stream video v in cell c
$Weight(v, c) \leftarrow$ computes the ratio of users receiving v in cell c to its required bandwidth
Output: Decisions for re-configuring the SFN areas of cell c

Sort M_c ascendingly based on their weights;
$v_m \leftarrow M_c.getHead()$; // Get the served video with the minimum weight
Sort U_c descendingly based on their weights;
$v_u \leftarrow U_c.getHead()$; // Get the unserved video with the maximum weight
while ($Weight(v_u, c) > Weight(v_m, c)$) or ($W > 0$ and $U_c = \emptyset$) **do**
  if ($W > 0$ and $U_c \neq \emptyset$) **then**
    // Case 1: expand an SFN area to include cell c as it is given in Figure 9 in Appendix D
    $[W, M_c] =$ **Expand**(A, c, v_u, W, M_c);
  else if ($W > 0$ and $U_c = \emptyset$) **then**
    // Case 2: passively enroll cell c into an SFN area as it is given in Figure 10 in Appendix D
    $[W, M_c] =$ **Support**(A, c, W, M_c, λ);
  else if ($W = 0$ and $U_c \neq \emptyset$) **then**
    // Case 3: replace the video v_m with video v_u as it is given in Figure 11 in Appendix D
    $[W, M_c] =$ **Replace**(A_c, c, v_m, v_u, W, M_c, α);
  end if
  $v_m \leftarrow M_c.getHead()$; // Get the served video with the minimum weight
  $v_u \leftarrow U_c.getHead()$; // Get the unserved video with the maximum weight
end while

Fig. 3. Proposed algorithm for reconfiguring an SFN.

its coverage area and then perform a *centralized* process to search for the optimal configuration for SFNs. We avoid such centralized operations and propose a *coordinated* algorithm in which each base station can dynamically help in determining whether an SFN reconfiguration is required. This algorithm is illustrated in Figure 3. Four different decisions are performed: (a) expanding the number of areas in which cell c is enrolled to accommodate additional multicast services, (b) passively joining a multicast session to strengthen the transmission coverage of neighboring cells, (c) replacing an existing video stream with another one, and (d) shrinking the number of areas in which cell c is enrolled.

In the proposed algorithm, cell c periodically sorts its multicast sessions based on the estimated bandwidth utilization of each video stream. The sorting is conducted in an ascending order for its ongoing multicast sessions and in a descending order for those unserved incoming video requests. Once this phase is accomplished, cell c tries to improve the service ratio within its cell by rearranging its enrollment in the current SFN areas but without causing frequent usage of its control signals. Four types of control overhead are considered in the computation of signaling cost: C_{syn} represents the cost of conducting a synchronization process for coordinated cells, C_{poll} refers to the cost of counting interested clients for a certain multicast service, C_{init} defines the cost of

initiating a new multicast session, and C_{stop} is the cost of ending an existing multicast service and releasing its allocated resources. To achieve our objective, cell c begins its examination by checking if there is enough bandwidth for multicast services over SFN (i.e., $W > 0$) in order to expand its offered multicast sessions. In the cases where c does not exceed the upper limit of allowed SFN areas and $W > 0$, cell c starts its attempts with the unserved video request whose bandwidth utilization is the highest. Two possible options in this scenario can be predicted: (1) cell c joins an active area where this video is broadcasted and (2) c enrolls in a zone where there are enough resource blocks such that a new multicast session can be initiated. Enabling cell c to go with either options necessitates the use of both synchronization and initiation control signal, thereby costing the network C_{sys} and C_{init}, respectively. Polling signals are also required in the latter option to announce the new service in all enrolled cells and count how many users are interested in receiving it. This polling process is expected to cost C_{poll}. We assume these signaling values vary from an area to another based on the number of its active cells and their distances from the corresponding multi-cell coordination entity. When cell c explores all possible cases, it selects the SFN area that maximizes Equation (1a).

Sometimes, a cell c experiences low traffic volume. Our algorithm utilizes this unexploited bandwidth to improve the overall video quality within the mobile system. Cell c in such scenarios would be called a passively enrolled cell, as explained in Figure 2. To ensure taking a full advantage of this passive enrollment, the multi-cell coordination entity retrieves the most spectrally efficient multicast streams within each area a of the available SFN zones, A, and then analyzes the gain achieved by joining cell c into the SFN area a, where $a \in A$. In our calculation for this gain, we are following the same formula in Equation (1a). Yet, two additional parameters are introduced to avoid any extensive calls for reconfiguration. The first parameter is related to the remaining duration of the most spectrally efficient stream v within the SFN area a. We denote this time by $T(v_a)$ and use this parameter to give multicast streams with longer estimated playing time higher priorities than short video sessions. The second parameter depends on the arrival rate of users within the cell c. Here, we model the request arrival process in a video service using a Poisson distribution with an arrival rate λ, which is defined by $P(k) = \lambda^k e^{-\lambda}/k!$. To allow c to act as a passively enrolled cell, it should have no outstanding traffic within that period of time. In other words, k should be equal to 0, leading $P(k = 0) = e^{-\lambda}$. We note that the core idea of our algorithm is serving mobile users using combination of multicast and unicast sessions created over multiple cells where some of them form an SFN, which is independent of the specific distribution of user arrivals. We use the arrival distribution to optimize for the performance by reducing the chances of frequently re-configuring the network. In addition, our algorithm runs periodically; thus, the parameters of the user distribution (e.g., λ for the Poisson distribution) can be dynamically adjusted to capture the changing patterns of user arrivals. For example, we can have multiple values for λ over different periods of time, which can allow the system to support bursty arrivals of users during these periods.

When the allocated bandwidth for multicast services over SFN is not sufficient to serve a set of outstanding video streams, our algorithm enables cell c to assess both active multicast sessions and incoming requests. It then observes its gains and losses from the perspective of achieved service ratio within its coverage. For instance, in the occasions where an incoming request v_u gives better bandwidth utilization than an existing multicast stream v_m, cell c should examine the benefit of broadcasting v_u rather than v_m by employing the ratio of control signaling cost to the projected spectral efficiency of each operation. Substituting a multicast service over SFN with another video stream requires initiating, synchronizing, and announcing the new video v_u, and it also needs stopping the ongoing transmission of v_m. Meanwhile, dividing the number of interested

users in v_u by its required resource blocks gives the estimated spectral efficiently of the new video. For a realistic comparison between the two streams, the remaining playing time as well as the popularity of both videos are also taken into account. To model the video popularity in a system, Zipf distribution is often used to characterize the access of viewers. If the video popularity is sorted in a decreasing order, then we can assume that, among the available titles, the stream v_u has an access probability given by $1/(v_u)^{\alpha}$, where α is the skew factor of the Zipf distribution. Once the cell c finds that it is not economical to replace v_m with v_u, it will test another alternative choice in which the radio resources allocated for v_m is released and switched to a single-cell mode. Changing the transmission mode from SFN to single-cell involves less amount of control signals, and it is most likely not going to increase the inter-symbol interference if these radio resources are used with low power. We call this operation of switching mode a shrinking process. If both replacing and shrinking approaches are still not cost-effective, then cell c is going to discard v_u and look for another outstanding stream.

6.2. Transmission Scheduling

The transmission scheduler works at radio base stations, and it is responsible of assigning portions of the shared bandwidth to users. At the beginning of every allocation window, the transmission scheduler determines its allocation decisions and then informs each mobile terminal which resource blocks it has been assigned for its multimedia streaming. The duration of this allocation window is recommended to be equal to the size of video chunks produced by video encoders (i.e., around 2s as in Microsoft Live Smooth Streaming [Microsoft 2010]).

The proposed transmission scheduling algorithm is presented in Figure 4. Once the transmission scheduler receives a set of incoming requests from mobile terminals within its transmission coverage, it creates a number of subgroups where each subgroup is identified by the segment number, its video identification, and its best suitable modulation and coding scheme. Each subgroup is also given a weight that is determined by two parameters: (1) the number of resource blocks needed to transmit this segment and (2) the number of possible users who are able to receive at this modulation and coding scheme and at the same time requesting this particular segment of video. Multiplying the former with the latter parameter gives a weight that is used in prioritizing the segments and then determining which set of them are chosen to be served during the current scheduling window. After constructing the subgroups of every required segment in the scheduling window, the proposed algorithm merges these small subgroups into more confined groups where the users of every group are requesting the same segment but might have different channel conditions. Since those larger groups may have diverse users regarding the channel conditions, the video streams of these groups are transmitted accordingly to the the member with the worst channel condition. These groups are also given weights, where the weight of each group is equivalent to the weight of the subgroup whose modulation and coding scheme is the lowest among its peers.

The number of available resource blocks is usually limited, so it is likely to have scenarios where some of the merged groups cannot be admitted during the current scheduling window. In these cases, our algorithm aims at choosing the best set of groups that maximizes the number of served users, and thereby minimizing the service blocking. To achieve such an objective, our proposed algorithm selects the group with the smallest weight and then eliminates its subgroup whose modulation and coding scheme is the lowest. Once this subgroup is removed, the weight of its own group is recalculated. The scheduler tries again to accommodate the groups in hand. If the bandwidth is still not enough, then the process of removing a subgroup is repeated until a solution is found.

Algorithm 2: Hybrid Transmission Scheduling

Inputs: $\{V, Z, R\} \leftarrow$ A set of requests for video streams and their data rates
$M \leftarrow$ The set of available modulation and coding schemes in the wireless network
$\Gamma \leftarrow$ The duration of the transmission scheduling window
$W_c \leftarrow$ The bandwidth still available for video services over cell c
$Bandwidth(v, c) \leftarrow$ computes the required bandwidth to stream video v in cell c
$Weight(v, c) \leftarrow$ computes the ratio of users receiving v in cell c to its required bandwidth
Output : $X \leftarrow$ The set of video segments to be served during the current allocation window

1: $W_r = 0$; // Initialize the required bandwidth to serve incoming video requests
2: $X = \{\emptyset\}$; // Initialize the set of video segments to be served during this allocation window
3: **for** each required segment (v, z) **do**
4: $n_{(v,z)} = 0$; // Initialize the number of users interested in this video segment
5: $g_{(v,z)} = \{\emptyset\}$; // Initialize the group of users interested in this video segment
6: **for** $m \in [M_{Max}, M_{Min}]$ **do**
7: $n_{(v,z,m)}$ = the number of viewers interested to receive this segment using MCS m;
8: $n_{(v,z)}$ += $n_{(v,z,m)}$;
9: $g_{(v,z,m)} = n_{(v,z,m)} \times Weight(v_{(z,m)}, c)$; // Calculate the weight of this subgroup
10: $g_{(v,z)}$ += $g_{(v,z,m)}$; // Merge this subgroup into its larger streaming group
11: **end for**
12: X += $g_{(v,z)}$; // Update the set of video segments to be served during this window
13: W_r += $Bandwidth(g_{(v,z)}, c)$; // Update the required bandwidth to serve video requests
14: **end for**
15: **while** $(W_c < W_r)$ **do**
16: Sort X ascendingly based on their weights;
17: $g_{(v,z)} \leftarrow X.getHead()$; // Get the group with the minimum weight
18: W_r −= $Bandwidth(g_{(v,z)}, c)$; // Update the required bandwidth to serve video requests
19: $g_{(v,z)}$ −= $g_{(v,z,m_{Min})}$; // Eliminate the subgroup with lowest MCS
20: X += $g_{(v,z)}$; // Update the set of video segments to be served during this window
21: W_r += $Bandwidth(g_{(v,z)}, c)$; // Update the required bandwidth to serve video requests
22: **end while**
23: **return** X

Fig. 4. Proposed transmission scheduling algorithm to maximize the service ratio for a video service over mobile networks.

The proposed transmission scheduling algorithm terminates in polynomial time: $O(N^2 Log(N))$, where N is the number of mobile terminals in a cell generating requests for video streams. The while loop in Figure 4 ensures the feasibility of the produced solution by satisfying the constraint in Equation (1b). Moreover, in each iteration it removes the least profitable streaming ensuring that the algorithm is not trapped in an infinite loop. The dominating computational complexity of the algorithm occurs in the third loop: (i) The while-loop there iterates at most N times and (ii) sorting the priority queue consumes $Nlog(N)$ times in its worst-case. Therefore, the time complexity of our proposed algorithm is $O(N^2 Log(N))$.

7. EVALUATION

In this section, we present an extensive trace-driven simulation performed in a detailed packet-level simulator. From the obtained results, we demonstrate the near-optimality of our heuristic algorithm, in which the number of served users is significantly

increased and the overall energy consumption at mobile terminals is reduced, while it imposes minimal overhead on the cellular network. We also show that our proposed algorithms outperform the closest three solutions in the literature [Araniti et al. 2013; Monserrat et al. 2012; Lee et al. 2011] as well as the energy-saving scheme introduced in Hoque et al. [2014]. In addition, we study the bandwidth overhead initiated by the channel quality reports and the SFN control signals, and we analyze the impact of user behavior on the performance of our proposed algorithms.

7.1. Simulation Setup

Simulator and Algorithms: We have implemented simulator for mobile video streaming systems using the OPNET modeler and its associated LTE Specialized module. OPNET is a detailed packet-level commercial simulator [OPNET 2010]. It consists of a suite of protocols and technologies that facilitate the test and demonstration of network designs in realistic scenarios prior to the production phase. The LTE module implements all details of realistic LTE wireless networks. To evaluate the proposed algorithms, we have also implemented the maximum-throughput algorithm [Araniti et al. 2013], proportional fair algorithm [Monserrat et al. 2012], combined unicast-multicast algorithm [Lee et al. 2011], and energy-saving algorithm [Hoque et al. 2014], and we refer to them as MT, PR, COMB, and ES, respectively, in our simulation results. We compare the results obtained by our solutions with these four methods from different perspectives. In addition to the resource allocation algorithms, we employ some heuristics to make our simulation setup more realistic. For example, when a mobile device cannot be admitted due to resource limitations, it does not leave the system immediately. Instead, it retries for the intended video up to 3 times with an exponential back-off waiting period starting at 2s. After being rejected 3 times, it finally stops requesting its desired video. Another heuristic is batching of requests such that all requests for videos within the duration of an allocation window are grouped together to be served at the beginning of the next window. These heuristics are likely to be incorporated in real video streaming systems.

Wireless Network Configuration: Although the proposed algorithms are applicable to any wireless networks with multicast support, we use the LTE Release-9 standard to evaluate the performance of the proposed algorithms [OPNET 2012]. In Appendix B, we list the LTE configuration parameters used for the simulations. Other parameters are set to the default values of the OPNET LTE module. More details about LTE networks and their configurations can be found in OPNET [2012] and Zaki et al. [2011]. We configure the LTE downlink with Evolved Packet System (EPS) bearers. We define an EPS bearer as a transmission path of defined quality, capacity, and delay [OPNET 2010]. The EPS bearer in LTE delivers bursty data at regular intervals, as scheduled, within Common Subframe Allocation (CSA) period and thus allows mobile devices to turn off the radio circuits between two bursts for saving energy. Moreover, the EPS bearer can be configured with specific quality of service attributes. For each bearer, we adjust the time intervals between any two adjacent bursts per the standard [3GPP 2014] in order to prevent overflow and underflow of ingress link-layer buffers. We adjust the quality-of-service attributes of EPS bearers to ensure specific MCS mode and bit rate of the video for transmission. Depending on the MCS mode of the bearer, the play time of the burst varies. We choose four MCS modes, that is, MCS 4, 8, 14, and 22, to support all possible channel qualities [3GPP 2014]. We define four types of bearers with respect to these MCS modes for each of the video streams. According to the proposed algorithms, each video can be transmitted using one bearer. For the assumed bearer configurations and MCS modes, depending on the channel conditions of the mobile devices: (i) MCS 4 to MCS 7 are served by the bearer of MCS 4,

(ii) MCS 8 to MCS 13 are served by the bearer of MCS 8, (iii) MCS 14 to MCS 21 are served by the bearer of MCS 14, and (iv) MCS 22 to MCS 28 are served by the bearer of MCS 22. The simulator runs the resource allocation algorithm once every allocation window of 2s. The obtained solutions are then mapped to the bearers, that is, we map a general resource allocation to an LTE-specific allocation for OPNET. We set the cell size to be around 10Km $\times$ 10Km by controlling the power of the base station. Each cell is served by one non-sectorized base station, called eNodeB in the LTE standard. The video server has the capability of both multicast and unicast services. The server can be directly connected to the Evolved Packet Core, or it may be located in the Internet.

User Distribution and Mobility: We assume a population between 200 and 1,000 users joining the system following a Poisson process with mean λ. λ is a simulation parameter that we set to 20 users per second by default for our simulations. We choose this value to allow users arrive over some time to cover different possible situations. We configure users to move following the random waypoint model in which mobility speed is randomly chosen between 0 and 72km/hr. This mobility model stresses our algorithms since it is difficult to predict the path of receivers and plan ahead of time. We configure mobile devices to send a Channel Quality Indicator report to the associated base stations every 100ms, which allows the base stations to determine the MCS mode depending on the channel condition. We choose this reporting interval to ensure that we do not miss any channel condition changes, and, at the same time, we do not receive unnecessary frequent reports. Mobile users are randomly distributed within each cell such that more users, about 90% of the total number of users, are densely populated within one third of cell radius and the rest of them are sparsely scattered around the rest of the cell area. This is done to mimic realistic scenarios as mobile operators, usually install base stations in crowded areas, to serve most users with strong signals.

Videos: For realistic video characteristics, we crawled YouTube and collected 1,000 videos. For each video, we have retrieved its duration, view count, and bit rate. The first two values are obtained using the YouTube Application program interface (API), while the bit-rate information is embedded in the video meta-date. If the bit rate is not embedded, then we use the video length and size to calculate its average bit rate, in a way similar to the dataset in Cheng et al. [2008]. The video format is MPEG-4, and these videos are categorized in four resolution classes: 240p, 360p, 480p, and 720p (250 videos for each class), where each video belongs to a single resolution class (i.e., a non-adaptive video). We rank these videos based on the view count, and then we employ the Zipf distribution with a skewness factor α to assign synthetic popularity to each video, so it is possible to exercise a wider range of popularity distributions. We set $\alpha = 1.5$ if not otherwise specified.

Simulation Scenarios: We evaluate the proposed heuristic transmission scheduling algorithm in three scenarios: (1) seven independent cells serving unicast and multicast connections, (2) seven cells forming a dynamic single-frequency network, and (3) seven cells operating in a static single-frequency mode. We refer to them as *SC, DSFN, SSFN*, respectively, in our simulation results. In the scenario applying independent-cell traffic, we customize the resource allocator in eNodeB to schedule incoming requests and set up radio resources for multicast and unicast connections. Differing from the independent-cell case, we designed an SFN area where eNodeBs are only responsible about scheduling incoming unicast requests, whereas the multi-cell coordination entity performs the required admission control for multicast sessions and assigns uniform radio resources among its cells to ensure enhanced coverage and synchronized data transmission. Our optimal and heuristic algorithms follow a dynamic technique in which multicast groups and their MCSs are configured dynamically to adapt and accommodate any changes in video popularity and channel conditions.

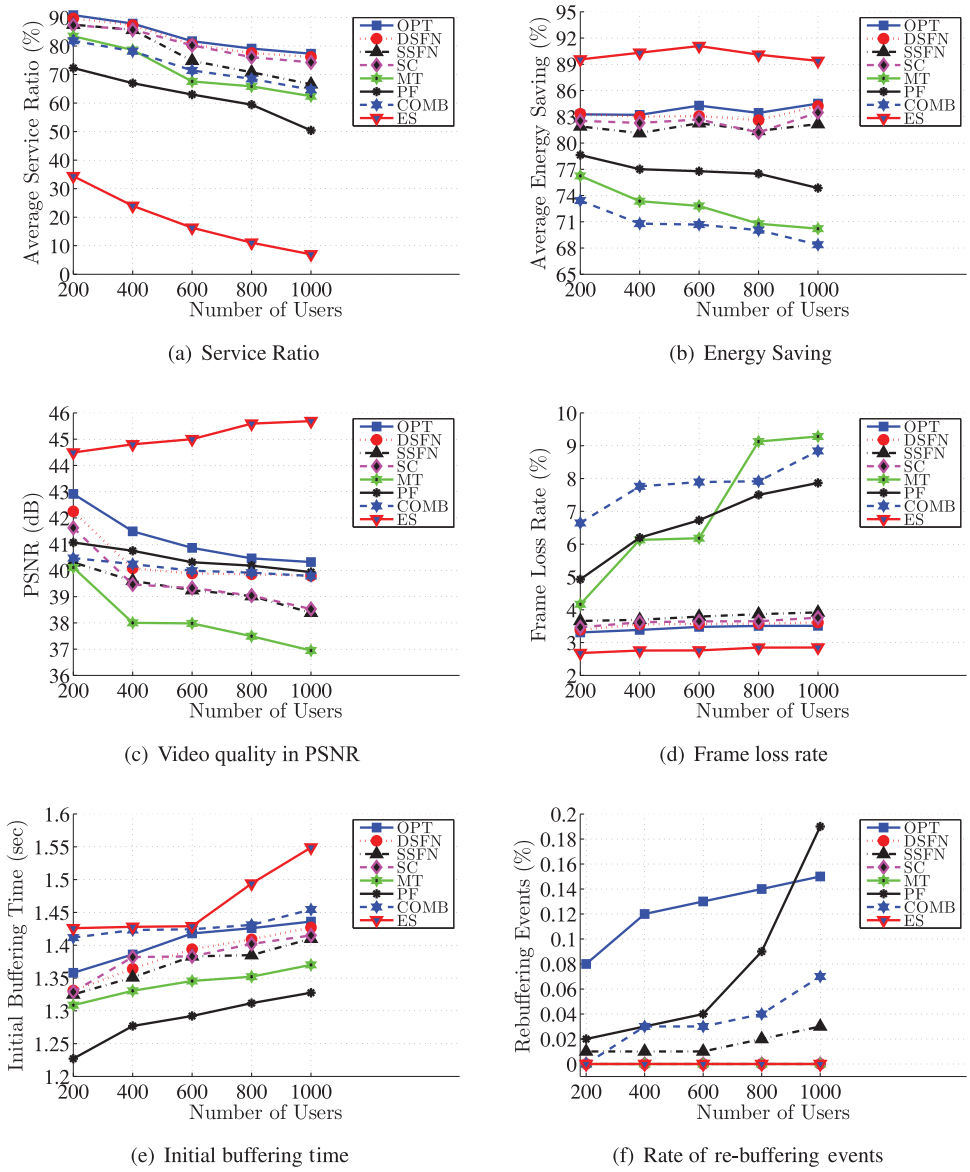


Fig. 5. Comparisons of the achieved performance of the proposed algorithms against the state-of-the-art approaches.

7.2. Comparison against Current Algorithms

We compare our proposed algorithms versus three multicast policies (MT, PR, and COMB) in addition to the unicast method (ES). The performance metrics used in this experiments are the average service ratio, energy saving, Peak Single-to-Noise Ratio (PSNR), frame loss rate, initial buffering time, and rate of re-buffering events. We simulate an LTE network where mobile terminals in each cell generate requests for a pool of 1,000 possible video streams. We vary the number of users in a cell from 200 to 1,000 and report the mean results from five simulation runs in Figure 5. Collectively,

these results indicate that our proposed algorithms not only outperform others with significant margins on the achieved average service ratio, but they also achieve much better energy saving without causing any violation in the buffer levels nor degrading the quality of video streams. The simulation results are comprehensively discussed below.

Service Ratio: Due to the limited radio resources in cellular networks, it may not be possible to serve all incoming video requests. For this reason, we estimate the service ratio by computing the fraction of served requests to the number of received requests within the system. Figure 5(a) indicates that our optimal and heuristic algorithms outperform other approaches on the achieved average service ratio. For instance, when there are 1,000 mobile users in each cell, our heuristic algorithm in the seven independent cells operating under the independent-cell configuration (denoted by SC) admits an average of about 75.5% of users at any given time, while systems employing MT, PF, COMB, and ES algorithms accept only an average of 62.4%, 50.4%, 64.7%, and 7% of users, respectively. This means that our heuristic algorithm in the independent-cell scenario provides a service ratio which is approximately 47%, 14%, 15% and 966% higher than the MT, PF, COMB, and ES, respectively.

It can be also shown in Figure 5(a) that applying the concept of single-frequency network improves the achieved service ratio by a significant gain. Figure 5(a) presents the achieved service ratio by our proposed heuristic algorithm in two types of SFN configurations: dynamic (DSFN) and static (SSFN). Applying our dynamic configuration results in an average of 76.18% service ratio. This improvement is 14% and 3% higher than the achieved ratios in both static SFN and independent-cell scenarios, while the gains over state-of-the-art techniques such as groups with MT, PF, COMB, and ES methods are 22.1%, 51.1%, 17.8% and 994%, respectively. Compared against the optimal solution denoted by OPT, our DSFN algorithm gives only 1.3% and 1.5% lower average service ratio when the numbers of users in each cell are 200 and 1,000, respectively.

Energy Saving: We define the energy saving as the percentage of time in which a served mobile device is able to turn off its network interface, thereby reducing its power consumption. The time required to switch the network interface from an active to idle is assumed to be negligible as shown in Yu et al. [2012]. Thus, it is sufficient to utilize the time duration where a network interface is turned off as a direct representation of the energy saving for such receiver. The unicast algorithm represents the maximum energy saving possible in an independent-cell configuration, since individual unicast connections are served according to their best-suitable modulation and coding schemes. Figure 5(b) illustrates that our proposed DSFN algorithm leads to 7.5% and 6.1% lower saving than the unicast ES algorithm when there are 200 and 1,000 users in a cell, respectively. However, compared to the multicast approaches (i.e., MT, PF, and COMB), our DSFN algorithm outperforms them by at least 12.5% and up to 23.2 in energy saving, when the number of users in a cell is 1,000. Our heuristic algorithm in the dynamic SFN configuration also succeeds in increasing the energy saving at mobile terminals by up to 2.4% and 1.7% when it is compared with the SSFN and the independent-cell (SC) scenarios, respectively. Comparing the results achieved by our DSFN algorithm versus those computed by the optimal algorithm, we notice that the energy saving obtained in our DSFN algorithm is close to the optimal with a small gap of 0.6% on average.

Video Quality: Figures 5(c) and 5(d) present the achieved video quality of the proposed algorithms against the latest algorithms in terms of PSNR and frame loss rate, respectively. We first observe that the unicast-only approach (ES) achieves the highest PSNR and the lowest frame loss rate. This is because it only admits very few mobile terminals at a time. In contrast, with 200 mobile users in each cell, our proposed

DSFN algorithm yields an average of 42.24dB in PSNR and 3.39% in frame loss rate. Even when the number of mobile users is increased from 200 to 1,000, the DSFN algorithm still achieves 39.79dB in PSNR and 3.61% in frame loss rate. Comparing with the related multicast policies, MT, PF, and COMP in the case of 1,000 users within a cell give a rate of 9.28%, 7.87%, and 8.84% in its frame loss, respectively. These values are higher than the results obtained when our proposed DSFN algorithm is applied by 157.2%, 117.9%, and 144.9%, respectively.

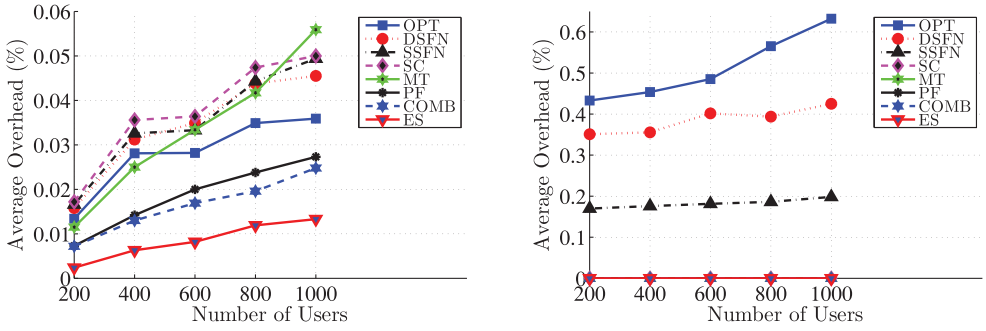
Initial Buffering Time: In video streaming systems, a playback starts after an initial buffering time and continues while the video is being downloaded. The initial buffering time in our algorithm depends mainly on the resource allocation window size. Intuitively, longer allocation windows provide more chances for expanding the multicast groups and thereby result in higher service ratios. Yet, larger allocation windows increase the initial buffering time. During our simulations, the window size is set to 2s, which is equal to the size of video chunks produced by video streaming solutions, such as Microsoft Live Smooth Streaming [Microsoft 2010]. At this window size, the initial buffering time is shown in Figure 5(e), which shows that our algorithms outperform the unicast-only approach in its initial buffering time and scale well with serving many mobile terminals.

Number of Re-buffering Events: We instrument our simulator to keep track of the buffer status of each mobile terminal. When the buffer of a mobile device receiving a video stream is empty or full, we declare a re-buffering event or an overflow event. We first verified through checking the logs of our simulation experiments that our proposed heuristic algorithm never leads to buffer overflow events. Then, we calculate the average rate of re-buffering events for the different algorithms by counting the number of re-buffering events per playback among viewers. These numbers are reported in Figure 5(f). This figure shows that our heuristic algorithms in both independent-cell and DSFN scenarios yield no re-buffering event. Since the optimal solution for the dynamic SFN configuration aims at increasing the number of served users within the entire system, it tries to achieve such an objective even if this goal may cause interruptions for certain video streams and result in downgrading the quality of experience for some users. According to the obtained outcomes in Figure 5(f), the optimal solution causes an average rate of around 0.10% and 0.15% in re-buffering events when the number of mobile terminals in a cell are 200 and 1,000, respectively.

7.3. Impact of Control Signals and Quality Reports

We assess the bandwidth overhead imposed on the system. Two types of overhead are considered: (1) frequent reports sent by each mobile terminal to update the nearest base station about the status of its channel quality condition and (2) control signals and messages sent to form a single-frequency network, coordinate its necessary operations, and perform time synchronization if needed.

Overhead of Feedback Channels: Mobile terminals in our proposed algorithms as well as the other state-of-the-art algorithms [Araniti et al. 2013; Monserrat et al. 2012; Lee et al. 2011; Hoque et al. 2014] are required to report their SNR values to the base station over a feedback channel. Having knowledge of the channel conditions of each mobile user helps in determining the highest possible MCS mode. In LTE Release 12 [3GPP 2014], two different reports can be obtained from mobile terminals: sub-band and wide-band feedback. Sub-band reports give the channel state information for each sub-band, whereas wide-band reports give the average channel quality information for the entire spectrum. We adopt wide-band reports during our simulations since they are sufficient, especially in large-scale scenarios. Moreover, since we activate the Discontinuous Reception feature [3GPP 2014] for energy saving, not all users utilize the dedicated upload control channels all the time. Instead, the wide-band reports



(a) Channel quality reports (b) SFN control signals

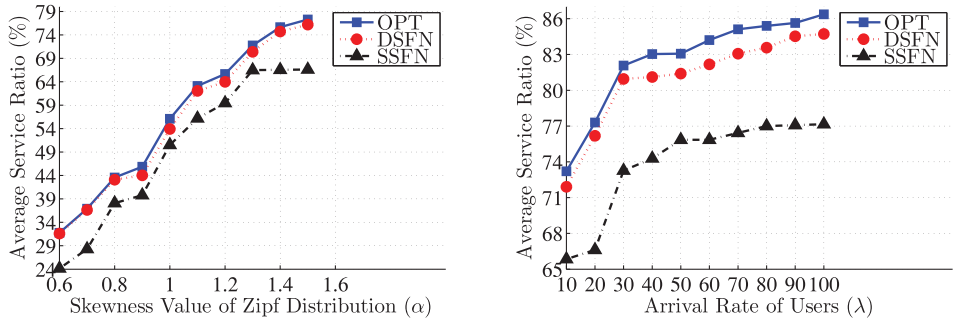
Fig. 6. Overhead caused by the feedbacks sent to base stations.

are sent by mobile terminals only when they receive video streams. We measure the overhead value as the fraction of bandwidth used to send feedback reports to the total bandwidth available for both data and control transmission. Figure 6(a) shows the overhead occurred in the six algorithms when the number of users within each cell is varied. Although the DSFN algorithm admits more users than other algorithms, its feedback overhead is less than 0.016% and 0.046% when the number of users are 200 and 1,000, respectively.

Overhead of SFN Control Signals: In a single-frequency network, the wireless bandwidth is mainly impacted by four types of control overheads: control signals to conduct a synchronization process for the coordinated cells, control signals to count the number of interested clients for a certain multicast service, control signals to initiate a new multicast session within a cell, and control signals to end an existing multicast service and release its allocated radio resources. Figure 6(b) presents the overhead caused by the SFN control signals in our algorithms during two types of configurations: DSFN and SSFN. The overhead value is measured as the fraction of bandwidth used to send these SFN control signals to the total bandwidth available for both data and control transmission. When the number of mobile terminals within a cell is 1,000, the signals required to manage the functionality of DSFN and SSFN consume approximately 0.43% and 0.20% of the bandwidth, respectively. These control overheads can be reduced to around 0.35% and 0.17%, respectively, once the number of users in each cell is decreased to 200. Compared against the optimal solution given by OPT, our heuristic algorithm (DSFN) outperforms by giving 23.3% and 48.8% less control overheads during the SFN reconfiguration process in the cases when the numbers of users in each cell are 200 and 1,000, respectively. We note that OPT produces optimal results in terms of service ratios, but it does not consider the overheads during its calculation.

7.4. Impact of User Behavior Model

We analyze the impact of user behavior on the performance of proposed transmission scheduling algorithms with respect to the achieved service ratio. Two important aspects of user behavior models are considered: videos popularity and request arrival. To study the effect of video selection policy on the proposed resource scheduler, we emulate a Zipf distribution to let 1,000 mobile terminals select streams from a pool consisting of 1,000 different videos. The skewness parameter α guides the selection strategy of these videos in a way that higher values of α is going to assign greater probability for most popular videos to be chosen and vice versa. We vary the value of α from 0.6 to 1.5 to study various policies for the video selection process. Figure 7(a) reports the impact



(a) Video popularity distribution (b) User arrival rate

Fig. 7. Impact of the user behavior model on the service ratio.

of varying the skewness parameter on the achieved service ratio. The figure shows that the average service ratio gradually increases with the increase in the skewness parameter α. Higher values of skewness lead more users to select from the top-ranked videos, thereby resulting in more possible multicast groups. This means larger chances to serve additional mobile terminals through multicast sessions and to decrease the service blocking probability in the system.

The effect of request arrival distribution on the proposed algorithm is also examined. We utilize the Poisson process and vary its mean in order to emulate variations in the user arrivals within cells. The arrival rate λ indicates the number of incoming requests per second, where higher values for this parameter are offering more opportunities for the creation of multicast sessions. Figure 7(b) shows the impact of varying the arrival rate of Poisson distribution on the average achieved service ratio. We vary the value of arrival rate from 10 to 100 requests per second. In Figure 7(b), it is shown that the service ratio increases for the proposed algorithm as the arrival rate increases. This is due to the fact that higher values of the arrival rate ensure larger numbers of request arrivals per second, with the same selections of video streams since the skewness parameter is kept unchanged during these experiments. This gives a chance to merge larger numbers of mobile devices into multicast groups, which eventually results in higher service ratio. Figure 7(b) also indicates the effectiveness of the proposed scheme under conditions of high loads. However, we can see from the figure that the service ratio increases significantly with the increase in arrival rates until it reaches 70, after which the service ratio becomes quite steady.

Figures 5(a), 7(a), and 7(b) demonstrate that our heuristic algorithm is close in its performance to the optimal solution under any given traffic load and any chosen user behavior models. On the other hand, Figures 7(a) and 7(b) point to a fundamental problem in the static deployment of SFN networks. Typically, the concept of a single-frequency network is employed to enhance the coverage and maximize the average signal-to-noise ratio within cells. Because the static configuration is pre-designed at an early stage of deployment, it is most probably unaware of any variation in the user distributions and video requests during the operation time. As a consequence, it may waste a substantial amount of radio resources reserved for SFN, especially in those scenarios where a few number of mobile terminals are interested in the multicast services offered by their cells. DSFN overcomes this limitation and adjusts its multicast zones according to both user distribution and video popularity. In extreme cases, cells in DSFN can remove themselves from all SFN areas and switch their settings to the independent-cell topology. In other words, our proposed transmission scheduler under

the dynamic SFN configuration adapts itself in a way so the best possible bandwidth utilization is reached.

8. CONCLUSIONS AND FUTURE WORK

Due to the introduction of mobile devices, traffic loads on mobile networks have dramatically increased during the recent decade, where a large portion of this traffic is due to the escalated consumption of videos. This trend of watching more multimedia content on mobile devices is expected to continue in the coming years. This creates a challenge for wireless network operators because of the constraint on their available radio resources and the substantial bandwidth requirements for each video session. This article proposed adaptive mobile multimedia streaming algorithms over single-frequency networks in which current user distributions and video popularities are taken into consideration during its network configurations and scheduling decisions. Differing from existing works, we do not assume a static configuration of single-frequency networks. Instead, we presented optimal and heuristic algorithms that dynamically rearrange SFN zones in a way that maximizes the total bandwidth utilization. We demonstrated through simulations that applying the concept of dynamic reconfiguration adds significant gain in the service ratio, as compared to those techniques with static SFN settings, and these obtained gains are independent of the amount of available bandwidth and the model of user behaviors.

Once a proper configuration for SFNs is reached, the available radio resources are allocated for both unicast and multicast services with an objective of increasing the average service ratio. Our proposed transmission schedulers achieve this goal by utilizing a flexible allocation process in which the resource distribution between unicast and multicast connections is done dynamically. To offer the flexibility of resource distribution, this article exploits three different types of transmission: unicast, multicast over an SFN, and multicast restricted within the coverage of a cell. According to our detailed simulation results obtained using a packet-level simulator (OPNET), the proposed transmission scheduling algorithms under a dynamic SFN configuration outperform the state-of-the-art multicast algorithms in the literature with respect to the service ratio, energy saving, video quality, frame loss rate, and number of re-buffering events. For instance, our algorithms serve up to 51.1% more users and consume up to 23.2% less power consumption compared to the state-of-the-art multicast-capable transmission algorithms.

This work can be extended in several directions. For example, we considered single-layer non-scalable videos. Scalable video streams can be supported to further improve the quality and performance of the cellular networks. Even with single-layer streams, DASH-like adaptability can be achieved by encoding each video segment into multiple quality representations. These different quality representations offer more optimization opportunities to dynamically adapt to the variations in the wireless channel conditions. Another extension can be enabling mobile terminals to adopt trajectory prediction algorithms to achieve proactive resource allocation across multiple cells.

REFERENCES

3GPP. 2010. Improved video support for Packet Switched Streaming (PSS) and Multimedia Broadcast/Multicast Service Services (3GPP TR 26.903 V9.0.0). Retrieved July 18, 2015, from http://tiny.cc/3GPP26.

3GPP. 2014. Evolved Universal Terrestrial Radio Access and Evolved Universal Terrestrial Radio Access Network: Overall Description (3GPP TS 36.300 V12.2.0). Retrieved July 18, 2015, from http://tiny.cc/3GPP36.

Adobe. 2009. Bit Rates for Live Streaming. Retrieved July 18, 2015, from http://tiny.cc/Adobe.

Richard Afolabi, Aresh Dadlani, and Kiseon Kim. 2013. Multicast scheduling and resource allocation algorithms for OFDMA-based systems: A survey. *IEEE Commun. Surv. Tutor.* 15, 1 (Jan, 2013), 240–254.

Akamai. 2013. Press Releases: Swisscom and Akamai Enter into a Strategic Partnership. Retrieved January 25, 2016, from http://tiny.cc/Akamai.

Antonios Alexiou, Christos Bouras, Vasileios Kokkinos, and George Tsichritzis. 2012. Performance evaluation of LTE for MBSFN transmissions. *Wireless Netw.* 18, 3 (Apr. 2012), 227–240.

Saleh Almowuena and Mohamed Hefeeda. 2015. Dynamic configuration of single frequency networks in mobile streaming systems. In *Proceedings of the ACM Multimedia Systems Conf. (MMSys'15)*. 153–164.

Giuseppe Araniti, Massimo Condoluci, Antonio Iera, Antonella Molinaro, John Cosmas, and Mohammadreza Behjati. 2014. A low-complexity resource allocation algorithm for multicast service delivery in OFDMA networks. *IEEE Trans. Broadcast.* 60, 2 (Jun. 2014), 358–369.

Giuseppe Araniti, Massimo Condoluci, Leonardo Militano, and Antonio Iera. 2013. Adaptive resource allocation to multicast services in LTE systems. *IEEE Trans. Broadcast.* 59, 4 (Dec. 2013), 658–664.

Bell. 2014. Crave TV: Video-on-demand Service. Retrieved July 18, 2015, from www.cravetv.ca.

Jiasi Chen, Rajesh Mahindra, Mohammad Amir Khojastepour, Sampath Rangarajan, and Mung Chiang. 2013. A scheduling framework for adaptive video delivery over cellular networks. In *Proc. of ACM Conf. on Mobile Computing and Networking (MobiCom'13)*, 389–400.

Xu Cheng, Cameron Dale, and Jiangchuan Liu. 2008. Statistics and social network of YouTube videos. In *Proceedings of the IEEE Workshop on Quality of Service (IWQoS'08)*. 229–238.

Claudio Cicconetti, Luciano Lenzini, Enzo Mingozzi, and Carl Eklund. 2006. Quality of service support in IEEE 802.16 networks. *IEEE Netw. Mag.* 20, 2 (Mar. 2006), 50–55.

Cisco. 2015. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014-2019. Retrieved July 18, 2015, from http://tiny.cc/Cisco14.

Erik Dahlman, Stefan Parkvall, and Johan Skold. 2013. *4G: LTE/LTE-advanced for Mobile Broadband*. Academic Press, Waltham, MA.

Hui Deng, Xiaoming Tao, and Jianhua Lu. 2012. QoS-aware resource allocation for mixed multicast and unicast traffic in OFDMA networks. *EURASIP J. Wireless Commun. Netw.* 2012, 1 (Jun. 2012), 1–10.

Derek Eager, Mary Vernon, and John Zahorjan. 2001. Minimizing bandwidth requirements for on-demand data delivery. *IEEE Trans. Knowl. Data Eng.* 13, 5 (Sep. 2001), 742–757.

Ahmed Elsherif, Zhi Ding, Xin Liu, and Jyri Hamalainen. 2013. Resource allocation in two-tier heterogeneous networks through enhanced shadow chasing. *IEEE Trans. Wireless Commun.* 12, 12 (Dec. 2013), 6439–6453.

Alessandro Finamore, Marco Mellia, Zafar Gilani, Konstantina Papagiannaki, Vijay Erramilli, and Yan Grunenberger. 2013. Is there a case for mobile phone content pre-staging? In *Proceedings of ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT'13)*. 321–326.

Mohamed Hefeeda and Cheng-Hsin Hsu. 2010. On burst transmission scheduling in mobile TV broadcast networks. *IEEE/ACM Trans. Netw.* 18, 2 (Apr. 2010), 610–623.

Mohamed Hefeeda, Cheng-Hsin Hsu, and Kianoosh Mokhtarian. 2011. Design and evaluation of a proxy cache for peer-to-peer traffic. *IEEE Trans. Comput.* 60, 7 (Jul. 2011), 964–977.

Helmut Hlavacs and Shelley Buchinger. 2008. Hierarchical video patching with optimal server bandwidth. *ACM Trans. Multimedia Comput. Commun. Appl.* 4, 1 (Jan. 2008), 8:1–8:23.

Mohammad Hoque, Matti Siekkinen, Jukka Nurminen, Sasu Tarkoma, and Mika Aalto. 2014. Saving energy in mobile devices for on-demand multimedia streaming – a cross-layer approach. *ACM Trans. Multimedia Comput. Commun. Appl.* 10, 3, Article 25 (Apr. 2014), 23 pages.

Cheng-Hsin Hsu and Mohamed Hefeeda. 2010. Broadcasting video streams encoded with arbitrary bit rates in energy-constrained mobile TV networks. *IEEE/ACM Trans. Netw.* 18, 3 (Jun. 2010), 681–694.

Kien Hua, Ying Cai, and Simon Sheu. 1998. Patching: A multicast technique for true video-on-demand services. In *Proceedings of the ACM Multimedia Conference.* 191–200.

Kien Hua and Simon Sheu. 1997. Skyscraper broadcasting: A new broadcasting scheme for metropolitan video-on-demand systems. *ACM SIGCOMM Comput. Commun. Rev.* 27, 4 (October 1997), 89–100.

Lorenzo Keller, Anh Le, Blerim Cici, Hulya Seferoglu, Christina Fragouli, and Athina Markopoulou. 2012. MicroCast: Cooperative video streaming on smartphones. In *Proceedings of the ACM Conference on Mobile Systems, Applications, and Services (MobiSys'12)*. 57–70.

Hongseok Kim, Gustavo de Veciana, Xiangying Yang, and Muthaiah Venkatachalam. 2012. Distributed α-optimal user association and cell load balancing in wireless networks. *IEEE/ACM Trans. Netw.* 20, 1 (February 2012), 177 –190.

Jong Lee, Hyo Park, Seong Choi, and Jun Choi. 2009. Adaptive hybrid transmission mechanism for on-demand mobile IPTV over WiMAX. *IEEE Trans. Broadcast.* 55, 2 (June 2009), 468–477.

Seung Joon Lee, Yongjoo Tcha, Sang-Yong Seo, and Seong-Choon Lee. 2011. Efficient use of multicast and unicast channels for multicast service transmission. *IEEE Trans. Commun.* 59, 5 (May 2011), 1264–1267.

Yu Liang, Wei Chung, Guo Ni, Ing Chen, Hongke Zhang, and Sy Kuo. 2012. Resource allocation with interference avoidance in OFDMA femtocell networks. *IEEE Trans. Vehic. Technol.* 61, 5 (Jun. 2012), 2243–2255.

Zhixue Lu, Tarun Bansal, and Prasun Sinha. 2013. Achieving user-level fairness in open-access femtocell-based architecture. *IEEE Trans. Mobile Comput.* 12, 10 (Oct. 2013), 1943–1954.

Microsoft. 2010. Microsoft: Live Smooth Streaming. Retrieved July 18, 2015, from http://tiny.cc/MSSmooth.

Jose Monserrat, Jorge Calabuig, Ana Fernandez-Aguilella, and David Gomez-Barquero. 2012. Joint delivery of unicast and e-MBMS services in LTE networks. *IEEE Trans. Broadcast.* 58, 2 (June 2012), 157–167.

Netflix. 2014. Netflix: Letter to Shareholders. Retrieved January 13, 2015, from http://tiny.cc/Netflix2015.

Marco Nicosia. 2010. Internet Video: New Revenue Opportunity for Telecommunications and Cable Providers. Retrieved January 25, 2016, from http://tiny.cc/Cisco2010.

Nokia. 2014. Nokia: LTE for National TV Broadcasting. Retrieved July 18, 2015, from http://tiny.cc/NokiaLTE.

OPNET. 2010. Riverbed: OPNET Modeler Suite. Retrieved July 18, 2015, from http://tiny.cc/OPNET.

OPNET. 2012. Riverbed: LTE Model User Guide. Retrieved July 18, 2015, from http://tiny.cc/OPNETLTE.

Jehan-Francois Paris and Darrell Long. 2001. The case for aggressive partial preloading in broadcasting protocols for video-on-demand. In *Proceedings of the IEEE Conference on Multimedia and Expo (ICME'01)*, 113–116.

Donald Parruca and James Gross. 2013. Rate selection analysis under semi-persistent scheduling in LTE networks. In *Proceedings of the IEEE Conference on Computing, Networking and Communications (ICNC'13)*. 1184–1190.

Md. Mahfuzur Rahman, Cheng-Hsin Hsu, Abdul Hasib, and Mohamed Hefeeda. 2014. Hybrid multicast-unicast streaming over mobile networks. In *Proceedings of the IFIP Networking Conference (Networking'14)*. 1–9.

Letian Rong, Olfa Haddada, and Salah-Eddine Elayoubi. 2008. Analytical analysis of the coverage of a MBSFN OFDMA network. In *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM'08)*. 1–5.

Shomi. 2014. Shomi: Video-on-demand Service. (July 2014). Retrieved July 18, 2015 from www.shomi.com.

Salvatore Talarico and Matthew Valenti. 2014. An accurate and efficient analysis of a MBSFN network. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 6994–6998.

Alistair Urie, Ashok Rudrapatna, Chandrasekharan Raman, and Jean Hanriot. 2013. Evolved multimedia broadcast multicast service in LTE: An assessment of system performance under realistic radio network engineering conditions. *Bell Labs Tech. J.* 18, 2 (Sep. 2013), 57–76.

Verizon. 2014. Verizon Wireless: Customers Use 1.9 Terabytes of Data in Stadium at Super Bowl. Retrieved July 18, 2015 from http://tiny.cc/Verizon2014.

Hyungsuk Won, Han Cai, Do Eun, Katherine Guo, Arun Netravali, Injong Rhee, and Krishan Sabnani. 2009. Multicast scheduling in cellular data networks. *IEEE Trans. Wireless Commun.* 8, 9 (Sep. 2009), 4540–4549.

Jian Xu, Sang Lee, Woo Kang, and Jong Seo. 2010. Adaptive resource allocation for MIMO-OFDM based wireless multicast systems. *IEEE Trans. Broadcast.* 56, 1 (Mar. 2010), 98–102.

YouTube. 2014. YouTube: Product Statistics. Retrieved March 20, 2015, from http://tiny.cc/YouTube2015.

Ya-Ju Yu, Pi-Cheng Hsiu, and Ai-Chun Pang. 2012. Energy-efficient video multicast in 4G wireless systems. *IEEE Trans. Mobile Comput.* 11, 10 (Oct. 2012), 1508–1522.

Yasir Zaki, Thushara Weerawardane, Carmelita Görg, and Andreas Timm-Giel. 2011. Long term evolution (LTE) model development within OPNET simulation environment. In *OPNET Workshop*. 1–8.

Received August 2015; revised February 2016; accepted February 2016