Hybrid Multicast-Unicast Video Streaming over Heterogeneous Cellular Networks

by

Saleh Almowuena

M.Sc., University of Victoria, Canada, 2010
B.Eng., King Saud University, Saudi Arabia, 2006

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2016

Approval

Name: **Saleh Almowuena**

Degree: **Doctor of Philosophy**

Title: ***Hybrid Multicast-Unicast Video Streaming over Heterogeneous Cellular Networks***

Examining Committee: **Chair:** Nick Sumner
Assistant Professor

Mohamed Hefeeda
Senior Supervisor
Professor _____

Jiangchuan Liu
Supervisor
Professor _____

Joseph G. Peters
Internal Examiner
Professor
School of Computing Science _____

Lin Cai
External Examiner
Professor
Department of Electrical and Computer Engineering
University of Victoria _____

Date Defended: 19 December 2016

Abstract

The demand for multimedia streaming over mobile networks has been steadily increasing in 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 in their capacities with recent generations. It has also caused a significant increase in the energy consumption at mobile terminals. To overcome these challenges, we first present a novel hybrid unicast and multicast streaming algorithm to minimize the overall energy consumption of mobile terminals as well as the traffic load within cellular networks. Next, we introduce the idea of dynamically configuring cells in mobile networks to form single frequency networks based on the video popularity and user distribution in each cell. We formulate the transmission scheduling problem in such complex networks, and we then present optimal and heuristic solutions to maximize the number of served multimedia streams. Through detailed packet-level simulations, we assess the performance of the aforementioned 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. Finally, we extend our research to formulate the transmission scheduling problem for adaptive streaming in the emerging heterogeneous cellular networks. We propose an algorithm for the cells in a heterogeneous network to self organize their radio resource allocations in order to minimize the inter-cell interference and increase the average data rate received by mobile terminals. Then we evaluate its performance through extensive simulations of various heterogeneous configurations.

Keywords: Mobile multimedia; single frequency network; video streaming; hybrid unicast-multicast; bandwidth utilization; energy saving; heterogeneous network; adaptive streaming;

To my parents and my wife, with love

Acknowledgements

First, it has been a great honor for me to be supervised by Professor Mohamed Hefeeda. His broad knowledge, enthusiasm in research and pleasant personality has been a continued source of inspiration. I am incredibly grateful to Prof. Hefeeda for his insights and valuable feedback that contributed significantly to this thesis. He was always available for my questions and guiding me to the right direction and the proper perspective.

I would like to extend my sincerest gratitude to Prof. Jiangchuan Liu for his valuable comments and suggestions during my graduate studies. My deepest thanks also go to the members of my examining committee: Prof. Joseph Peters, my internal thesis examiner, and Prof. Lin Cai, my external thesis examiner, for reading the thesis and providing insightful feedback. I would like as well to thank Dr. Nick Sumner for taking the time to chair my thesis defense.

I am indebted to my colleagues at the Network Systems Lab who created a dynamic and an exciting research environment in which this thesis is evolved. I am really glad to work with such talented individuals. Special thanks go to Ahmed Hamza, Md. Mahfuzur Rahman, Khaled Diab, and Hamed Ahmadi. I am also grateful for my collaboration with Dr. Cheng-Hsin Hsu, who enlightens me about many of the concepts and assumptions associated with multimedia streaming over wireless cellular networks.

My sincere appreciation is expressed as well to the Ministry of Higher Education in Saudi Arabia and the Saudi Arabian Cultural Bureau in Canada for their financial support during my graduate studies at Simon Fraser University.

Finally, no words are sufficient to express my heartfelt gratitude and love to the special people in my life: my wonderful parents, whose encouragement and support are key factors in all my achievements; my beloved wife, Amani, whom I am forever indebted to for her continuous love and patience; and my children, Deem and Bassam, who are the sunshine in my universe. This thesis is dedicated to them.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, we describe the challenges of providing high-quality multimedia streaming services over wireless cellular networks. We then present several research problems to overcome these challenges and summarize our contributions to solve them. Finally, we conclude this chapter with the organization of the thesis.

1.1 Introduction

With the increasing popularity of mobile terminals such as smartphones and tablets, the past few years have witnessed a tremendous growth of multimedia applications in wireless systems. The proliferation of wireless video applications will most likely reshape the wireless traffic in the near future. For instance, Cisco Visual Index reports that mobile video traffic is predicted to grow over 15 times from 2014 to 2019, and it will take up nearly 75% of the world's mobile data traffic by 2019 [32]. Such explosive growth in mobile video traffic poses a significant challenge for 4G mobile broadband networks (e.g., LTE-Advanced) and even future 5G networks. It is hence critical to introduce efficient mobile video streaming mechanisms in order to better utilize the limited network radio resources.

To respond to the rising demand of data traffic, mobile providers can make use of the multicast capabilities in current and future cellular networks whenever possible. With these multicast-capable networks, a streaming server can significantly reduce the network load by serving mobile terminals interested in the same video stream using a single multicast session. Since the network is not aware of the channel quality conditions of each terminal, it should utilize a conservative modulation and coding scheme such that cell-edge terminals can successfully decode and stream the multicast video. Those cell-edge mobile terminals experience not only high signal loss but also suffer inter-cell interference from their neighboring cells. For this reason, the wireless cellular network typically considers the worst channel quality condition during the modulation and coding process of the multicast signal, thereby increasing the amount of radio resources required to deliver each video service.

To overcome such limitation, neighboring cells can cooperate and transmit the same video stream using identical and synchronized radio signals. That is, neighboring cells form the so-called single frequency network (SFN). In such cases, receivers at cell edges get multiple copies of the same data but from different base stations. These copies constructively interfere with each other, and they can be translated into useful signal energy. Hence, the strength of the received signal at the cell edge is enhanced, and the interference power at the same time is largely reduced. While using the SFN mode for multicast sessions helps in efficiently utilizing the radio resources in the network, it requires addressing several optimization questions. For instance, we need to determine the number of single frequency networks that should be created, which cell should belong to which SFN, and how to dynamically adjust cell configuration to make the best of single frequency networks.

Another approach to cope with the substantial demand for wireless bandwidth would be enabling mobile providers to transform their conventional homogenous networks into more heterogeneous systems, in which macrocells are deployed along with low-powered small cells. These small cells are placed to fill the coverage holes within macrocells and to increase the network capacity in densely populated regions. Since both macro and small cells share the same carrier frequency band, mobile terminals will likely suffer from inter-cell interference, which will negatively impact their achievable data rates. Thus, efficient interference management and power control algorithms need to be designed to mitigate the inter-cell interference problem without causing under-utilization of the available radio resources.

Multimedia streaming over wireless cellular networks also encounters several challenges due to the diverse nature of mobile receivers. For example, laptops connected to power outlets on commuter trains allow users to watch high-quality video streams without worrying about the consumption of their power resources. On the other hand, mechanisms for energy saving are usually required for battery-powered devices, such as smartphones and tablets.

In this thesis, we address multiple challenges of multimedia streaming over mobile networks including: limited bandwidth offered by wireless carriers, constrained power resources available at mobile terminals, managing heterogeneous networks, and supporting diverse mobile terminals. We propose a number of algorithms that outperform their corresponding state-of-the-art ones in the literature, with respect to various performance metrics, including the average service ratio, spectral efficiency, energy saving, frame loss rate, initial buffering time, and number of re-buffering events.

1.2 Thesis Contributions

In this section, we summarize the research problems addressed in this thesis, and we highlight our contributions for solving each problem.

1.2.1 Hybrid Unicast-Multicast Video Streaming

There exists a tradeoff between the different types of video transmission schemes: unicast connections allow the creation of short transmission bursts that help in lowering the energy consumption

of mobile terminals, whereas multicast sessions restrict the data rate of each group according to users with the worst channel conditions to increase the number of served users within cells. To balance such tradeoff, we propose group formation algorithms that concurrently utilize a mixture of unicast and multicast connections in order to construct sets of transmission bursts that result in increasing the overall energy saving at mobile terminals and the total number of admitted users. We consider a video streaming scenario with multiple base stations, mobile terminals, and resource allocators. Mobile terminals arrive asynchronously, and each user sends requests to a resource allocator asking for certain video streams. Every resource allocator periodically solves an optimization problem for leveraging both unicast and multicast to: (i) maximize the average energy saving across all mobile terminals, (ii) minimize the network resources consumed by video streaming, and (iii) ensure smooth playout on all mobile terminals. Once the optimization problem is solved, the allocator decides which chunks of videos should be sent, which transmission scheme must be utilized, when they should be delivered, and which modulation and coding schemes must be chosen. Our contributions in this topic can be summarized as follows [15]:

- For the single-cell configuration, we propose optimal and heuristic transmission scheduling algorithms to form unicast and multicast subgroups with the objective of maximizing the average energy saving at mobile terminals.

- We extend the group formation problem to include the cases for multi-cell single frequency networks, and we introduce two algorithms and analyze their complexity.

- Through extensive simulations using OPNET, we demonstrate the effectiveness of the proposed algorithms and their impact on the battery life of mobile terminals as well as the traffic load within cellular networks. For example, the proposed algorithms in the single-cell configuration enable cellular networks to achieve high energy saving, close in the outcomes of those unicast-only approaches. The obtained simulation results show that our proposed algorithms consume only 6.5% more energy than the state-of-the-art unicast-only approaches [51], and they outperform the latest multicast-only approaches [19, 67, 76] by up to 20%. Our algorithms also achieve better performance with respect to video quality, frame loss rate, and number of re-buffering events, even in dense networks with 1,000 mobile users in each cell.

- Our algorithms proposed for the multi-cell SFNs leverage the enhanced coverage gained by the coordinated efforts among adjacent cells to increase both service ratio and energy saving. To show the potential impacts of multi-cell SFNs, we simulate the performance on base stations of a leading Canadian cellular operator deployed in downtown Vancouver. The evaluation results show that our heuristic algorithm for multi-cell SFNs achieves up to 13% higher service ratio and 6% higher energy saving, compared to the independent cell networks.

1.2.2 Dynamic Configuration of Single Frequency Networks

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 that each cell can provide unicast and multicast services to users in its range. Cells can also collaborate by forming one or more single frequency networks (SFNs). If a subgroup of cells forms an SFN, 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: 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 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. Our contributions can be summarized as follows [12, 13]:

- 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 transmission scheduling 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 transmission scheduling problem. Because of the high computational complexity of the optimal algorithm, 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 combination thereof.
- We conduct an extensive simulation study using OPNET to evaluate the proposed algorithms. Our results show that the proposed heuristic algorithm can serve up to 11X 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 [19,67,76]. The heuristic algorithm runs in real-time: it terminates in a *few milliseconds* on a commodity workstation, and it achieves near-optimal results with respect to the achieved service ratio. In real deployment, such algorithms run on servers once every *few seconds*; therefore, our algorithm is practical and efficient.

1.2.3 Adaptive Video Streaming over Heterogeneous Cellular Networks

User distribution and mobility behavior typically vary based on their environment nature. For this reason, wireless cellular systems are most likely to move toward the deployment of low-power nodes within macro-cells, thereby forming heterogeneous networks. In such configuration, inter-cell interference causes significant impacts on the signal strength at mobile terminals, thereby reducing the network throughput and minimizing the average transmitted data rate. We mitigate such problems by carefully grouping the incoming video requests and allocating the available sub-carriers in a way such that the quality of experience is enhanced at the end users. In video streaming over heterogeneous networks, we need to consider a large set of network decisions including the resource block allocation of each base station and the transmission power needed for every subcarrier. Finding the optimal decisions is not only computationally challenging (i.e., it is an NP-complete problem), but it also requires the solution of joint problems within different time scales. We present an efficient and heuristic algorithm to solve the transmission scheduling problem suitable for real-time mobile video streaming services. The proposed transmission scheduling algorithm receives incoming video requests, divides mobile terminals into groups, determines the data bit rate for each group, and performs interference-aware radio transmission scheduling. Each base station independently determines the resource block and power assignment to each video session. These decisions are taken by solving a scheduling problem and are updated based on the feedback about the channel conditions of users. The main technical contributions to this field can be summarized as follows [14]:

- We introduce a transmission scheduling formulation for adaptive video streaming over heterogeneous networks. The objective is to balance the trade-off between minimizing the inter-cell interference within a cell (in terms of the signal strength received at mobile terminals) and maximizing the quality of delivered video streams (in terms of average data rate and minimum quality switching).

- Using the proposed interference-aware group construction approach, we divide mobile terminals into subgroups, choose the suitable resource blocks for each subgroup, and adjust the transmission power of each subcarrier, with an objective of minimizing the inter-cell interference among cells.

- We introduce an adaptive bit rate allocation process, in which the base station is responsible for determining the assigned data rate of each mobile terminal. This process will then help mobile terminals to refine their selection of the available video quality representations, such

that the overall average data rate is maximized and the frequent switching among quality representations is minimized.

- We conduct extensive simulation studies using OPNET to evaluate the proposed algorithm with respect to the signal strength at mobile terminals, average data rate of video streams, and number of quality switches among video representations. From the obtained results, we show that the proposed algorithm outperforms the closest related works in [76] and [27] with significant margins. For instance, the proposed algorithm succeeds in increasing the signal strength at mobile terminals by average gains of 252% and 25% compared to the join unicast-multicast (JUM) algorithm in [76] and the fair optimal (FO) algorithm in [27]. Regarding the average data rate, employing our algorithm also helps in providing almost ten times the average data rate than JUM and up to 37% higher than those obtained in FO. Such improvement is provided without causing any significant disruption in the smooth video streaming since end-users experience no video quality variations for more than 98% of their streaming time.

1.3 Thesis Organization

The rest of this thesis is organized as follows. We present some background about multimedia streaming services as well as wireless cellular networks in Chapter 2. We propose and evaluate energy-aware video streaming techniques in Chapter 3 to minimize the power consumption at mobile terminals using hybrid unicast and multicast video sessions. Then we formulate and solve the problem of mobile video streaming over dynamic single frequency networks in Chapter 4. In Chapter 5, we present a self-organized transmission scheduling approach to provide adaptive video streaming over heterogeneous cells. Last, we conclude the thesis and outline future research directions in Chapter 6.

Chapter 2

Background

In this chapter, we provide a background about multimedia streaming over wireless cellular networks. We first present the current video streaming protocols and distinguish the differences among them. We also discuss the possible video streaming models and how to incorporate them into our designs. We then focus on the network architecture of the most dominant wireless standard. We present the various transmission methods that can be exploited to deliver date traffic to mobile terminals, and we indicate both advantages and challenges of creating single frequency networks. Last, we show the main components of a data transmission scheduling unit and how these components interact with one another.

2.1 Introduction

Due to the introduction of touch screens and smart phones, the traffic load on mobile networks has dramatically increased in the recent years. A significant portion of this traffic can be referred to the high consumption of mobile videos. Market research [32] reports that video streaming represents more than 50% of mobile Internet traffic since 2012, and this fraction will increase to 72% by 2019. This will create a challenge for wireless network operators, because of the limited wireless bandwidth and the substantial bandwidth requirements for each video session.

A video streaming service over mobile networks consists of three essential components: content servers, wireless cellular networks, and mobile terminals. Commonly, every content server is managed by a provider which might include television channels, online video rental stores, multimedia e-learning centers, entertainment websites, or even cellular operators. As a matter of fact, some U.S. cellular operators have recently delivered live events, such as Super Bowl [41, 98] and Indy 500 [99], using multicast over their commercial LTE networks to a vast number of mobile users. Two leading cellular operators in Canada have also launched their video-on-demand streaming services at the end of 2014 [24, 90] to provide more than 10,000 hours of videos to their subscribers. Providing multimedia services at this high volume places enormous demands on the computational capability, storage facility as well as distributed bandwidth of cellular networks.

Cellular networks consist of access gateways, backhaul links, and cellular base stations. Even though video streaming has been extensively addressed in wired networks, these conventional methods cannot easily be applied as solutions for the streaming problem over cellular networks for several reasons. For instance, a wireless channel is vulnerable to physical phenomena such as multi-path fading and interference. Furthermore, a mobile network always suffers from the instability of connections for an extended period due to the dynamic movement of its users. These factors may make terminals in a wireless cellular network experience high and variable round trip time, link outage, rate fluctuations, and occasional burst losses. In addition to these challenges, mobile terminals have constrained power supplies, low computational abilities, and limited buffer spaces.

Based on this ground, novel algorithms and systems have been designed to enable current and future cellular networks to increase their bandwidth capacities, optimize the quality of their delivered videos, and extend the lifetime of energy resources at their mobile terminals. We present a concise background of multimedia streaming protocols and its models in Section 2.2. We then review the architecture of wireless cellular networks and their transmission modes in Section 2.3.

2.2 Multimedia Streaming

Multimedia streaming represents the case when video contents are decoded and then displayed at end-users while they are being received from the servers. Different from the download mode, streaming a video allows users to enjoy lower start-up time and consume fewer storage requirements. Traffic analysis reports, such as Cisco's Visual Networking Index [32], show that video content has become the prevalent traffic over mobile networks. These reports also predict that this consumption will continue its rapid increase in the upcoming years. Such prediction can be easily validated since various innovative media services and devices have been widely deployed. For example, Netflix added 12 million new members in 2016 bringing its global total to 86.7 million paid subscribers [77], whereas YouTube nowadays has more than 1 billion users watching hundreds of millions of hours every day, and more than half of its views come from mobile terminals [104].

To cope with this excessive need for bandwidth, video applications should compress its streams in an efficient way prior to the transmission phase. Higher compression ratios enhance the bandwidth utilization, but it requires additional computational processing, leading to increasing the power consumed at mobile terminals. Current video coding standards, H.264 and VP8, claim to provide efficient compression with low bit rates. Comparing the two techniques, Seeling et al. [86] demonstrate that H.264 yields better video quality and higher bandwidth utilization than VP8. High-Efficiency Video Coding (HEVC or sometimes H.265) has been standardized with the aim of providing an average of 50% bit rate reduction compared to prior standards at equal perceptual video quality [94]. In addition to efficient video compression, novel protocols for streaming over mobile networks are essential in order to cope with the fluctuations in the available radio resources and user channel conditions.

2.2.1 Multimedia Streaming Protocols

This section focuses on two classes of protocols that can be used for delivering video contents. These streaming protocols can be classified into two main categories: push-based and pull-based mechanisms [23, 68]. In **push-based streaming protocols**, a connection is established between a content server and an interested end-user at the beginning of a communication session, and then the video stream is continuously sent to the end-user with no need for any additional explicit requests until the session is stopped. The Real-time Streaming Protocol (RTSP) is the most deployed push-based streaming protocol, and it is applied on the top of the Real-time Transport Protocol (RTP) to enable delivering video frames over the unreliable User Datagram Protocol (UDP). A video session in RTSP begins once a video stream is requested from the content server. The content server adjusts its streaming parameters, including the bit rate, and then labels the session with an identifier to represent the shared state between both server and end-user. An RTSP session can be carried on any underlying transport layer protocol. RTP provides best-effort transmission for those applications with low-latency requirements. However, video packets in RTSP are usually sent using RTP over UDP, whereas the control commands are transmitted over TCP, as their arrival and order are essential to the quality of an ongoing video session.

On the other hand, content servers in **pull-based streaming protocols** schedule a set of incoming video requests explicitly generated from end-users in a way such that their quality of experience is satisfied. The Hypertext Transfer Protocol (HTTP) is the most common protocol for pull-based video delivery. Streaming over HTTP works on top of TCP, requires a simple web server, offers authentication and authorization infrastructure, and is capable of exploiting the existing HTTP infrastructure. To stream over HTTP, a **progressive download** can be utilized in which an end-user sends an HTTP request to the content server and then starts pulling the video stream from there. The video application at the client-side waits for its buffer to reach a predetermined minimum level. Once this buffer level is reached, the end-user starts playing the video stream. Meanwhile, the progressive download continues in the background. Several drawbacks can be noticed here. For example, the progressive download method may negatively impact the bandwidth utilization of a network in the cases when an end-user early terminates her session despite that video has been completely transferred. Furthermore, the progressive download does not provide support for real-time and live streaming and does not offer adaptation for video bit rates [23].

Adaptive streaming over HTTP is a hybrid scheme of both progressive download and traditional streaming. Its concept was introduced to support instant streaming and bit rate adaptation, and it can still be classified as a pull-based protocol. To achieve adaptive streaming over HTTP, each video content is divided into segments with a short length (i.e., between 2 and 20 seconds [64]), and every video segment is encoded at multiple bit rates. These bit rates are affected by the video resolution, quantization level, and frame rate. Bit rates are also impacted by the content itself such as its motion and structure. Prior to the beginning of a streaming process, the interested client downloads a manifest file for its requested video, in which the versions of qualities for each segment are

specified. Once the manifest file is retrieved, the end-user generates requests to download a number of initial segments according to the implementation of her video application. For instance, the end-user in a wireless cellular network might start with asking the content server for the first segment to be delivered using the version with the lowest bit rate. If the available radio resources permit a mobile terminal to choose segments with higher bit rates, the end-user can request her subsequent video segments to be sent in their better quality versions. During rush hours, it would be predicted that the bandwidth becomes limited so that end-users may readjust the bit rates of their video requests. Following such adaptive approach in streaming helps in improving the quality of experience for users. There are various implementations for adaptive streaming over HTTP, including Adobe HTTP Dynamic Streaming (HDS) [5], Microsoft Live Smooth Streaming [75], and Apple HTTP Live Streaming (HLS) [17]. Recently, an international standard was accepted for this streaming class, which is called Dynamic Adaptive Streaming over HTTP (DASH) [93].

2.2.2 Multimedia Streaming Model

We consider a general streaming model that can be used for live as well as on-demand streaming 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 same video and functions that may disrupt this synchrony, e.g., 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. The work in this thesis 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 various moments in the videos. This general asynchronous model for streaming is hard 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, which is 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 model, multicast alone will not be sufficient to provide true-on-demand service without imposing long waiting times on users. To solve this problem, existing delay mitigation mechanisms, such as patching videos using temporary unicast connections [22, 39, 50, 54] may be adopted. For example, it is possible to 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 temporary unicast session is shorter than a predetermined threshold; otherwise, a new multicast group for the requested video might be created.

In our streaming model, mobile terminals within a cell can request V different video streams. Mobile terminals receiving the same video stream 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 to the allocation window duration. Let $r_{vz} : v \in [1, V], z \in [1, Z_v]$ be the encoding rate of the video's segment (v, z). Because of the variable bit rate nature of video streams, r_{vz} may vary across the scheduling allocation windows.

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 ISPs' natural gateways such as Broadband Remote Access Servers (BRAS) or Packet Data Network Gateways (PGW). We also assume that these content servers may be part of a content delivery network, a proxy cache [42], or a transparent proxy [49]. Such architectures are gaining increasing interest as it is shown in [8] and [74]. As explained by Nicosia in [74], 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. However, 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 ULL (Unbundling of the Local Loop) agreement.

2.3 Wireless Cellular Networks

Wireless cellular networks have evolved from simple one-to-one push-to-talk to the Fourth-Generation (4G) systems within few decades. These mobile networks will continue to evolve with future generations of technologies bringing new capabilities. Agyapong et al. [7] and Dahlman et al. [35] discuss the challenges of the Fifth-Generation (5G) mobile networks. To support the increasing demand for data traffic and multimedia streaming over cellular networks, a radio access technology known as 3GPP Long Term Evolution (LTE) system was introduced. It improves on the prior UMTS mobile standard by enhancing the system capacity and transmission coverage and allowing both data and voice to be provided in an integrated fashion using the Internet Protocol. It also provides greater user experience by offering higher data rates, reduced propagation delay, lower deployment cost, and seamless integration with already deployed cellular system [36]. LTE networks are currently being deployed around the globe, and this technology has become the most dominant wireless standard. For this reason, this section focuses on the network architecture for LTE systems as an example of a mobile network.

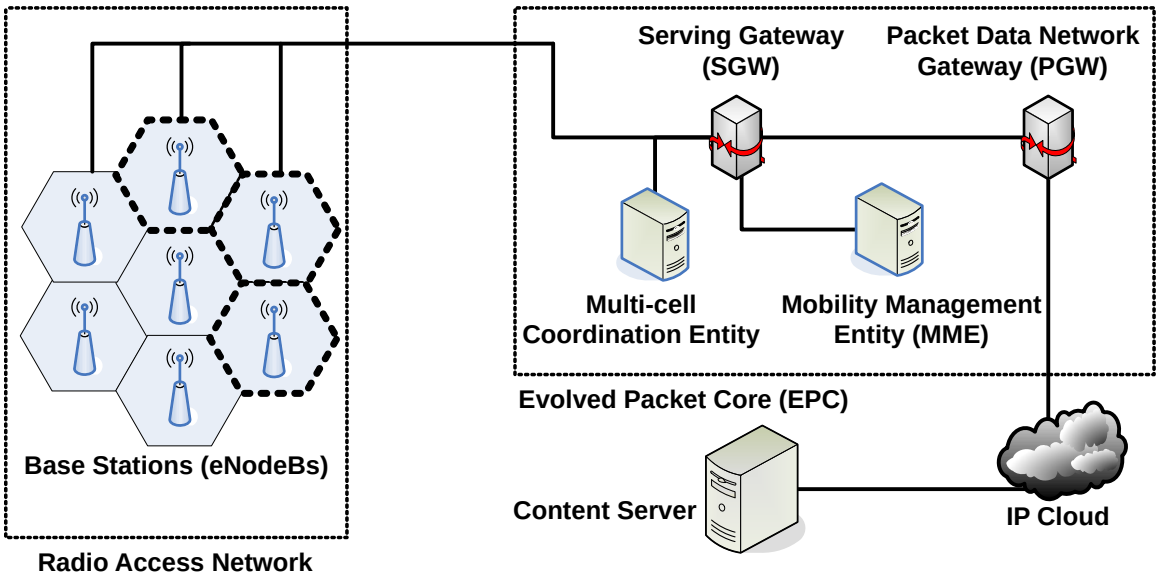


Figure 2.1: Architecture of LTE networks.

2.3.1 An OFDMA System: LTE as an Example

The system architecture of an LTE network is comprised of two components: a) access network named Evolved UMTS Terrestrial Radio Access Network (E-UTRAN), and b) core network called Evolved Packet Core (EPC), as shown in Figure 2.1. Both E-UTRAN and EPC are responsible for offering quality-of-service (QoS) control within the mobile system. E-UTRAN is also responsible for managing the access of available radio resources and providing user- and control-plane support to mobile terminals. A user plane refers to a group of protocols utilized to support data transmission throughout the entire network, while a control plane represents a set of protocols needed to control the transmission and manage the connection between a network and mobile terminals within its coverage. Examples of these connection management functions include handover, service establishment, and resource controlling. An E-UTRAN consists only of a number of base stations (eNodeBs or eNBs in LTE terminologies). On the other hand, an EPC is a mobile core network whose primary responsibilities include management of mobility, policy, and security. EPC consists of a Mobility Management Entity (MME), a Serving Gateway (S-GW), and a Packet Data Network Gateway (P-GW). Compared with previous 3GPP architectures, LTE has fewer nodes; therefore, smaller user-plane latency is obtained. This architecture, however, requires eNodeBs to perform additional user-plane functions that are not traditionally done at base stations, such as ciphering.

LTE networks use Orthogonal Frequency-Division Multiple Access (OFDMA) as a transmission technique for its downlink channels and Single-Carrier Frequency-Division Multiple Access (SC-FDMA) for its uplink. In OFDMA, the frequency is divided into parallel sub-carriers, where each sub-carrier is capable of carrying one modulation symbol. Different sub-carriers are grouped together to form a sub-channel that serves as the basic unit of data transmission. The main reasons why OFDMA was selected as the transmission scheme for LTE are its high spectral efficiency,

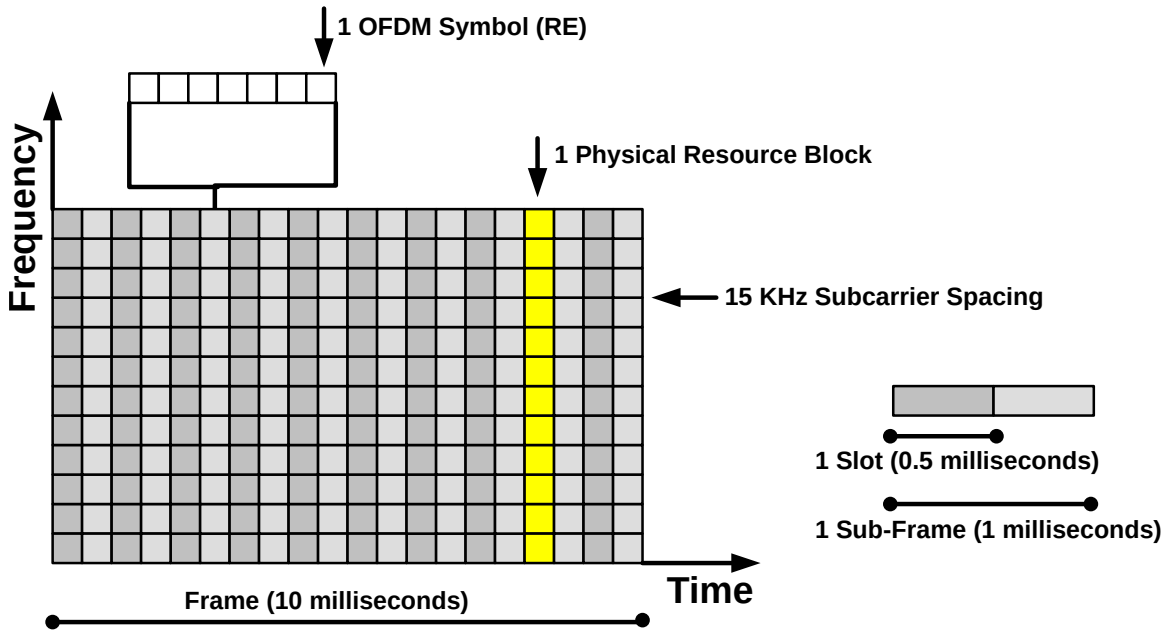


Figure 2.2: The conventional structure of LTE downlink frames.

low-complexity implementation, and ability to easily support advanced features such as frequency selective scheduling, multiple-input-multiple-output (MIMO) transmission, and interference coordination [36]. In its uplink channels, SC-FDMA was selected due to its ability to provide similar advantages to OFDM such as orthogonality among users, frequency-domain scheduling, and robustness with respect to multipath operation. However, SC-FDMA has a lower requirement for low-power amplifier back-off or de-rating. As a consequence, the average transmission power can be much higher using SC-FDMA than those obtained in the cases with OFDMA. Such increase consequently improves the coverage in uplink channels and provides higher uplink data rates to mobile terminals at the cell edge.

Both frequency-division duplex (FDD) and time-division duplex (TDD) transmission are supported in LTE systems. Furthermore, a base-band structure is common between FDD and TDD, thus making LTE to be flexible and efficient technology that can be deployed in either paired or unpaired spectrum. In LTE, the differences between TDD and FDD are mostly at the physical layer. As a result, identical network architecture can be used to support both modes, thus substantially reducing both deployment cost and complexity. In LTE specifications, the two modes have been designed to share as much functionality and as many features as possible, while the main design difference is the need to support different TDD downlink and uplink allocations and provide co-existence with other TDD systems.

Physical resources in the radio interface of an LTE network are organized into frames along the time domain, and sub-carriers along the frequency domain, as shown in Figure 2.2. A radio frame is ten milliseconds long in the time domain, and it is subdivided into ten sub-frames with 1-millisecond length in time. Each sub-frame is further split into two consecutive time slots, each of

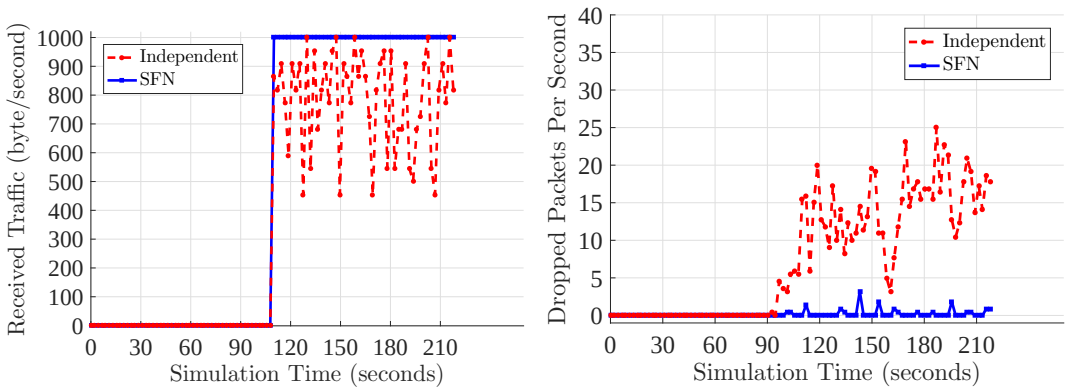
0.5-millisecond duration long as illustrated in Figure 2.2. The smallest physical resource in an LTE system is called a resource element, which represents one sub-carrier during one OFDM symbol. The sub-carrier spacing for OFDM in LTE equals 15 kHz. However, the minimum resource unit that can be allocated by a transmission scheduling algorithm to a mobile terminal consists of two consecutive physical resource blocks within one sub-frame, and it is referred to as a resource block pair. Each resource block pair corresponds to 12 consecutive sub-carriers resulting in a bandwidth of 180 kHz in the frequency domain and one half millisecond slot in the time domain.

2.3.2 Multimedia Multicast Services

The demand for multimedia streaming over mobile networks has been steadily increasing in the past several years. According to [32], 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. In order 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 WiMAX standard defines Multicast and Broadcast Service (MBS) in the data link layer in order to facilitate the process of initiating multicasting and broadcasting sessions [31]. Similarly, the evolved Multimedia Broadcast Multicast Services (eMBMS) allows LTE cellular networks to deliver video streams over multicast [1]. 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.

As defined in 4G standards, e.g., [36], 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. It allows feedback from mobile terminals on their channel conditions and then dynamically adjusts their modulation and coding schemes based on these feedbacks. The advantage of such mode is its adaptation to any change in the current distribution of users within a cell. For instance, multicast services using this mode can be turned-off within a particular cell in which there are no active users. The **SFN mode** 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 cell edges can get multiple copies of the same data from different base stations. While these copies are considered as *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, the interference power at the same time is largely reduced, and the overall performance remains consistent even if a user moves from one cell to another.

To demonstrate the impact of multicast modes, we have utilized OPNET to simulate an LTE network consisting of two cells. We have employed the multipath fading channel model used in



(a) Received Traffic

(b) Dropped Packets

Figure 2.3: The impact of multicast modes on the received traffic at a mobile terminal.

pedestrian environments, as an example. The two base stations multicast data using independent and SFN modes. We let a mobile terminal located at the cell edge join the system at time 100 second and start receiving the video stream. Figure 2.3 shows the amount of received traffic and dropped packets by this mobile terminal over time. We can observe that all the video traffic is received by this terminal when the SFN mode is applied, whereas some packets are dropped in the independent multicast mode due to the negative effect of inter-cell interference and multipath components. Different from the independent mode, these packets are recovered in the SFN mode because of the constructive interference and useful signal energy gained from the two synchronized cells. More information on how a single frequency network manages its resources and operates in general can be found in [2, 36].

As shown in Figure 2.1, a multimedia multicast service consists of four network elements: Broadcast/Multicast Service Center (BMSC), Gateway (GW), Mobile Management Entity (MME), and Multi-cell/multicast Coordination Entity (MCE) [36]. First, the service center (BMSC), which is located within the core network, is responsible for authenticating and authorizing the content providers, managing the charging process, and controlling the overall configuration of data flow through the core network. Second, a gateway is considered as a logical node that helps in multicasting any IP packet generated from BMSC to all base stations located in a certain SFN area. Moreover, the gateway handles further session control signaling via the mobile management entity (MME). MME is a main entity for the LTE access network since it plays an important role in performing a number of controlling procedures such as: user tracking, paging, and bearer activating. Finally, MCE ensures the full functionality of an SFN area by performing the time synchronization as well as coordinating the usage of the same radio resources and transmission parameters across all cells belonging to a particular area.

Multicast services are offered on a time-shared basis with unicast connections. The frame structure in an LTE network is subdivided into 10 equal sub-frames whose lengths are equal to 1 millisecond. Some of these sub-frames (numbering 0, 4, 5, and 9) are reserved for unicast connections and

cell specific information. Any or all of the remaining six sub-frames may be allocated to multicast service. A mobile terminal is informed about which sub-frames are assigned to its multicast session via a broadcast channel and this allocation can be changed dynamically at specified intervals. A multicast sub-frame consists of two slots with four to five OFDM symbols in the first slot and six in the second slot. Each symbol is composed of a useful symbol duration of approximately 66.7μs, and it is preceded by an extended cyclic prefix of approximately 16.7μs. This is in contrast to unicast sub-frames which consist of 14 OFDM symbols with the normal cyclic prefix of approximately 4.7μs. As it can be seen in the case of multicast services, provision has been made for a longer cyclic prefix than unicast connections in order to accommodate a longer guard time, thus enabling more SFN signals from distant base stations to contribute to useful signal energy.

Several works have been introduced to assess the performance of multimedia multicast streaming over wireless cellular networks. For instance, Rong et al. [83] and Talarico and Valenti [95] 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. [97] extend this assessment and provide more comprehensive evaluation of SFN performance under more realistic conditions. Alexiou et al. [11] 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 objective, they calculate the cost of both packet delivery and signaling procedures under a set of different network topologies and user distributions. The works [11, 83, 95, 97] assume a static SFN configuration in which each cell is registered into certain set of zones at an early stage of deployment, and the enrollment of these cells do not change over time even if variations have been occurred for both users distribution and network traffic.

2.3.3 Date Transmission Scheduling

Data transmission scheduling in wireless cellular networks consists of three essential components: channel quality estimator, adaptive modulator, and radio resource allocator [25]. These three components are placed at both physical and data link layers, and they periodically interact with each other to provide better utilization of the available radio resources and to achieve certain optimization objectives (including minimizing the service blocking probability, maximizing the user quality experience, and saving the power consumption at both base stations and mobile terminals). In this section, we briefly describe these three components and their roles in the transmission scheduling process.

Channel Quality Estimator: It is a procedure that enables base stations to estimate the quality conditions of the downlink channels at a particular mobile terminal. Each channel quality condition is determined as a quantized and scaled measure of the Signal to Interference plus Noise Ratio (SINR) experienced at the mobile terminal. Several issues can be raised during the estimation process [80, 88]. For instance, it is important to take into consideration how such estimation is

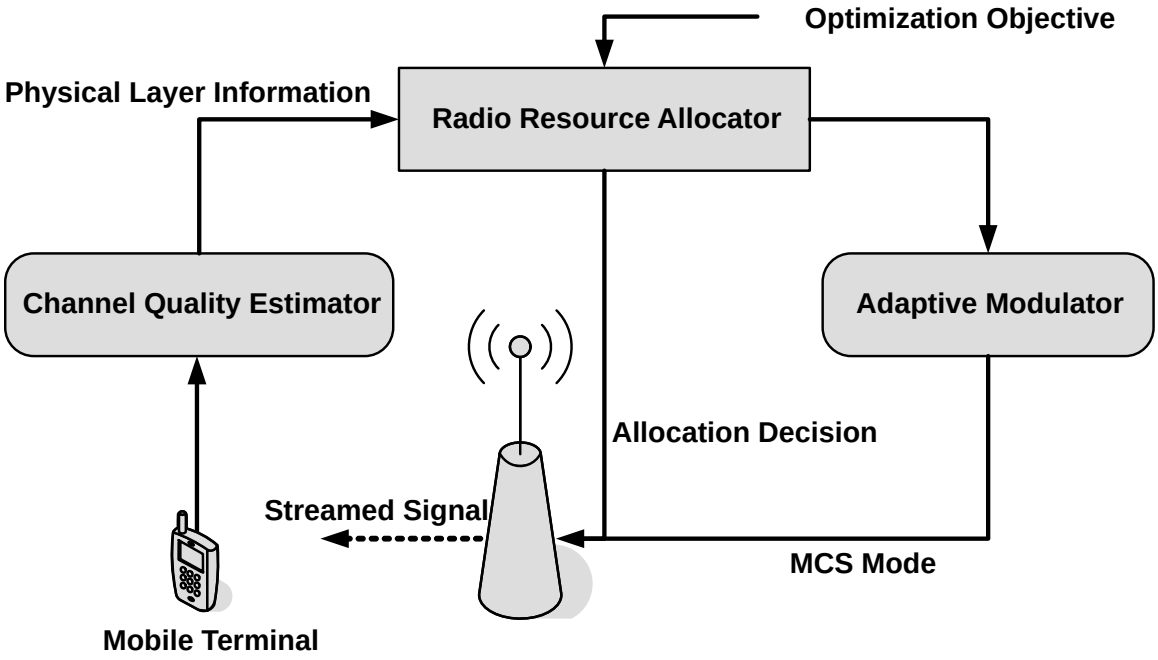


Figure 2.4: Components of a data transmission scheduling unit.

acquired with respect to both computational complexity and running time, how accurate the obtained measurement is, and how much control overhead is initiated during the estimation process.

Adaptive Modulator: Once the channel quality condition of a mobile terminal is retrieved, the base station chooses its modulation and coding scheme, with the objective of maximizing its possible received data rate without exceeding a target Block Error Rate (BLER) [36]. On this ground, mobile terminals experiencing high channel conditions are capable of decoding more aggressive coding schemes and receiving higher data rates, whereas terminals at cell edges (i.e., who are suffering from weak channel quality conditions) would be served with lower data rate using more conservative modulation and coding schemes. The number of possible modulation and coding schemes is limited; therefore, the capacity of a resource block at certain transmission power has an upper bound. In other words, there is a particular threshold at which an increase in the SINR value would not lead to any gain in the throughput of the wireless system.

Radio Resource Allocator: The data channel of a wireless cellular network is shared among the mobile terminals, so base stations are responsible for allocating these users and distribute their radio resources in every scheduling window. The resource allocation is periodically performed with a time granularity of a few milliseconds. Hence, finding the optimal scheduling decision through complex and non-linear optimization problems would cause excessive computational and memory consumption on base stations. Therefore, several scheduling techniques have been proposed to balance the trade-off between both optimally and complexity [25].

Figure 2.4 summarizes the main transmission scheduling components and their interaction with each other. As it is illustrated in the figure, each mobile terminal decodes the reference signal, deter-

mines its channel quality condition, and then sends its estimation back to the base station. To take advantage of this feedback, the base station solves an optimization problem under a predetermined objective, obtains the resource allocation decisions, and then chooses the suitable modulation and coding scheme for each mobile terminal. The information about mobile terminals, their allocated radio resources, and their selected modulation and coding schemes are delivered to the terminals using some control channels in the downlink channel. When a mobile terminal retrieves such information, it will tune to the corresponding radio channel and start receiving its requested video stream.

2.3.4 Position of Proposed Algorithms

In this subsection, we discuss the positions of the proposed algorithms in the wireless cellular network. In Chapter 3, we introduce hybrid unicast-multicast streaming algorithms in order to minimize the power consumption at mobile terminals and increase the bandwidth utilization within cells. The proposed algorithms can operate in either single-cell configuration or multi-cell single frequency networks. In the single-cell configuration, the proposed hybrid streaming approaches are implemented on base stations, in which multicast and unicast sessions are scheduled and transmitted with no coordination among cells. On the other hand, each base station and its associated multi-cell coordination entity cooperate together to apply the proposed hybrid streaming algorithms in the multi-cell SFN configurations. Here, the multi-cell coordination entity is responsible for the video selection, session initiation, and session termination procedures belonging to the multicast streams broadcasted across its SFN areas. It is also responsible for ensuring the synchronization of cells within its SFN areas to improve the coverage at cell edges.

Each base station reserves the resource blocks needed for SFN operations, as instructed by its associated multi-cell coordination entity. Then it schedules the unicast streams and those multicast sessions operating in the independent mode. The base station also informs mobile terminals about which radio resources are allocated for their requested video streams. Since the hybrid streaming algorithms in Chapter 3 enable the discontinuous reception mode [2], the base station determines the necessary parameters for this mode to allow mobile terminals to control their radio transceivers and decrease their energy consumption.

The dynamic SFN configuration algorithms in Chapter 4 are applied in the wireless cellular network in a similar way to the hybrid unicast-multicast streaming algorithms. The main difference in Chapter 4 is that SFN configuration is not static. Instead, it is a dynamic configuration, in which a cell joins and leaves SFN areas over time. This dynamic SFN configuration is performed by cooperation between base stations and multi-cell coordination entities. To maximize the number of admitted terminals in cells, the proposed dynamic configuration algorithms rely on the user distribution and video popularity within the network. The bandwidth overhead caused by the control signals between base stations and multi-cell coordination entities is also considered in our proposed heuristic algorithm.

In Chapter 5, we present a self-organized transmission scheduling algorithm to be implemented on the base stations in heterogeneous cellular networks. Each base station is responsible for receiving incoming video requests from its associated mobile terminals. It then independently divides these terminals into subgroups, allocates the required radio resources for each subgroup, and adjusts the transmission power of each subcarrier, with the objective of strengthening the received signals at mobile terminals. Since mobile terminals in the proposed algorithm utilize an adaptive video streaming approach, the base station also provides bandwidth estimations to help these terminals in facilitating their video adaptation processes.

Chapter 3

Hybrid Unicast-Multicast Video Streaming

In this chapter, we introduce a novel hybrid unicast-multicast video streaming over wireless cellular networks, with the objectives of reducing the power consumption at mobile terminals and increasing the bandwidth efficiency within cells. We describe and formulate the transmission scheduling problem studied in this chapter, and we show its hardness. We then propose optimal and heuristic hybrid streaming algorithms for both independent cells and multi-cell single frequency networks. These algorithms are thoroughly evaluated using simulations with respect to various performance metrics, and they are also compared against their closest state-of-the-art approaches.

3.1 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 mobile phones. Unfortunately, the technology of batteries has not been improving at the same pace as the processing speeds and powers of mobile terminals. With nearly 1500 mAh battery power in current smartphones, if a typical video application consumes roughly 300 mW of energy in a hour, neglecting all other power consumption on that phone, the battery is expected to last between 4 and 5 hours [57].

In this chapter, we study the transmission scheduling problem for large-scale video streaming over multi-cell networks, which is one of the most challenging research problems in 4G/5G multicast networks. Base station(s) concurrently serve multiple videos with diverse popularity to mobile terminals, and these mobile terminals may start watching at different time instants. Our problem is to determine which chunks of videos should be sent, when to send them, and with what modulation and coding scheme (MCS) modes, in order to minimize the overall energy consumption of mobile

devices and maximize the number of served users without consuming excessive network bandwidth. The considered transmission scheduling problem directly affects both cellular network's load and mobile devices' battery life. For example, transmitting at higher MCS modes allows the mobile devices to receive at higher rates and then finish earlier. This in turn results in higher energy saving because the mobile devices may turn off their wireless interfaces for longer time durations. We first focus on non-scalable videos, which can be decoded by most mobile devices and require lower coding complexity. Then we discuss extending our work to scalable videos.

To maximize the energy saving, the base station may set up a unicast connection to each mobile device using the best MCS mode determined by each device's channel conditions. Using a unicast-only approach, e.g., [45, 51, 59], consumes excessive network resources. To cope with this issue, a base station may put mobile devices into multiple multicast groups based on their requested videos. Using a multicast-only approach, however, may result in higher energy consumption, because each video is transmitted with the MCS mode suitable to the mobile device with the worst channel condition. This unnecessarily increases the energy consumption of some mobile devices, even if they are under better channel conditions. To get the merits of both multicast and unicast, we consider a *hybrid* video streaming system that concurrently leverages unicast and multicast to maximize the energy saving of mobile devices under various resource constraints. We first address the transmission scheduling problem in a hybrid video streaming system within independent cell networks, which is simpler yet useful in its own right. We prove that the problem is NP-Complete and mathematically formulate it as a Binary Integer Programming problem. The optimization problem can be solved by general optimization solvers (such as CPLEX [33] and GLPK [43]), which however are computationally expensive for real-time video streaming services. Hence, we develop a heuristic algorithm, which gives close-to-optimal solutions. Next, we extend the solution to multi-cell SFNs for better channel conditions and overall performance. We also propose optimal and heuristic algorithms for the extended problem in multi-cell SFNs.

The rest of this chapter is organized as follows. We survey the literature in Section 3.2. The problem statement is presented in Section 3.3. We formulate and solve the optimization problems for independent cell networks and multi-cell SFNs in Sections 3.4 and 3.5, respectively. We then evaluate our proposed solutions in Section 3.6 and conclude the chapter in Section 3.7.

3.2 Related Work

To increase the energy saving at mobile terminals, a number of middle-ware and application-level approaches were proposed in the literature. For instance, an early work was presented by Anand et al. [16] to introduce a self-tuning operating system module adapting itself to the access pattern of networks and the expected intent of applications. If an application is capable of tolerating a certain delay for its incoming packets, the network interface will be carefully disabled for a certain period of time. Applications in [16] are also allowed to specifically disclose the start and end of their data transmission such that the power management is accordingly enabled. Zhang and

Shin [107] observe that a majority of energy consumption is spent on idle listening. Hence, their system adaptively downclocks its radio interfaces in order to reduce the amount of consumed energy during idle listening periods. Pyles et al. [79] prioritize applications into a set of class and then manage the network interface based on the classes of these applications. When a high-priority application is running on the mobile terminal, radio resources are put into continuous awake mode. Otherwise, these interfaces are switched into a power-saving mode in the presence of application with low priorities.

The power consumption in wireless cellular networks depends mainly on the signal strength of radio resources. Weak signals eventually lead to higher transmitting power and lower transfer rate. Moreover, the communication energy per bit is found to be as much as six times higher when the signal is weak compared to those cases when it is strong [85]. On this ground, in order to save energy, applications in [85] communicate only if the radio signal is strong, either by deferring non-urgent traffic or advancing anticipated communications to coincide with periods of strong signals. Zhu and Cao [109] present a transmission scheduling algorithm for base stations to determine their needed data flows at different time instants. To facilitate such decisions, the concept of proxy servers are employed to buffer data traffic sent to mobile devices such that cellular network interfaces can sleep for longer time periods. Moreover, an adaptive technique is presented in [109] to adjust the sleep time according to the channel condition of each mobile device. Luna et al. [72] propose a technique to transmit a video frame using its minimal required transmission energy without violating any delay and quality constraints. They carefully select the parameters for both source coding and transmitter power and then adapt the rate in data delivery to increase the overall energy saving.

In LTE networks, the discontinuous reception mode for energy saving is supported in both idle and connected states [2, 36]. In an idle state, the radio connection between a mobile terminal and its nearby base station is released, while this terminal context is still maintained by the mobility management element as well as the serving and packet data network gateways. On the other hand, both radio connection and terminal context are active in the connected state. The discontinuous reception mode is activated by a mobile terminal after a time period of its successful data transmission or packet reception. In other words, this mobile terminal is not going to receive and does not have outstanding packets to transmit during the discontinuous reception mode. However, this mobile terminal entering the DRX mode would still be in the connected state, and it begins periodically turning off its radio interfaces for a duration known as the short cycle. At the end of this cycle, the terminal wakes up again to check its downlink control channel to observe whether it would receive any incoming data. Once there is no transmission or reception scheduled, the terminal goes into another sleep mode but for longer cycle. During this long cycle, radio connections are released, and the mobile terminal listens only to the downlink broadcast channel for paging procedures. Based on this concept, Hoque et al. [51] introduce an energy-efficient video delivery system which relies on sending short data bursts using unicast connections, and these bursts are constructed in a way to reduce the power consumption and avoid any buffer violations. Hefeeda et al. [48] also study the burst scheduling problem for optimal energy saving in mobile TV broadcast networks with arbitrary

channel bit rates. Then 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 et al. [52] 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 is not violated. The energy saving in multicast over cellular networks has also been addressed in [30, 53, 55, 65, 89, 92, 105], where they utilize the concept of scalable video coding.

The closest works to our proposed algorithms are [19, 67, 76], because 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 average service ratio, spectral efficiency, energy saving, frame loss rate, and number of re-buffering events. Since our main objective is minimizing the overall energy consumption of mobile devices, we also compare our algorithms against the energy-efficient video delivery system introduced in [51]. We show that our algorithms admit more users than this unicast scheme and achieve close results regarding the amount of energy saving.

3.3 Problem Statement

The symbols used in this chapter are given in Table 3.1. Several cellular networks adopt the Orthogonal Frequency Division Multiple Access (OFDMA) modulation scheme, which divides a wireless medium along both time and frequency domains [105]. We consider an *allocation window* with T columns of *symbols* and S rows of *subchannels*. A pair of $t \in [1, T]$ and $s \in [1, S]$ uniquely determines a *resource block*, which is the minimum unit of data transmission in the network. Let d denote the fraction of resource blocks that is reserved for video streaming, which can be adjusted based on the loads of voice and data applications. Thus, the considered transmission scheduling problem is to distribute the dTS blocks of an allocation window among all mobile devices. Note that the system parameter T affects the length of allocation windows: larger T leads to longer allocation windows for higher allocation flexibility, and smaller T results in shorter allocation windows for shorter video *service delay*. The service delay refers to the time difference between a mobile device switches to a video and the mobile device starts rendering that video. Shorter service delay also results in faster adaptation to network dynamics. To support the true on-demand streaming cases with real time constraints on the service delay, a *patching* solution [39, 50, 54] may be used. That is, we define a threshold for a new request to join an on-going multicast session of a video and at the same time create a separate, temporary unicast session for that user to receive the earlier parts of the video. This new user will be considered when solving the transmission scheduling problem in the next allocation window, and potentially be assigned to a multicast session.

A video streaming service offers V different videos. Let r_v denote the encoding rate of video v. We assume each video v is watched by N_v mobile devices, and we let $N = \sum_{v=1}^{V} N_v$ be the total number of mobile devices. The network interface on each mobile device can be put into one

Table 3.1: Symbols Used in This Chapter.

Symbol	Description
T	Number of symbol columns in an allocation window
S	Number of subcarrier rows in an allocation window
d	Fraction of resource blocks reserved for videos
Γ	Duration of an allocation window
V	Number of videos available for streaming within the system
r_v	The encoding rate of a video v
N_v	Number of mobile terminals watching video v
N	Total number of mobile terminals in the system
M	Number of modulation and coding scheme (MCS) modes
c_m	Per-block capacity with the modulation and coding scheme m
Z_v	Number of segments into which a video v is divided
$w_{v,m,z}$	Number of mobile terminals watching segment z of video v with m
q	The length of symbol time used to transmit the data burst
γ	The average amount of energy saving at mobile terminals
$x_{v,m,z}$	Whether segment z of video v is sent using MCS mode m
$y_{v,m,n,z}$	Whether mobile terminals with maximum MCS mode m receive segment z of video v with MCS mode n
H	Number of hexagonal cells in an SFN
N_v^h	Number of mobile terminals in cell h watching v

of M MCS modes. We let per-block capacity c_m denote the amount of data that can be carried by a block with mode m, where c_m is non-decreasing in $m \in [1, M]$. Each mobile device is under a different channel condition and can receive at a *maximum* MCS mode, which is determined by the firmware on the network interface to maintain reasonable bit error rates. Moreover, mobile devices may watch different parts of a video. We divide video v into Z_v consecutive parts in the length of allocation windows (a few seconds). We let $w_{v,m,z}$ ($v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$) be the number of mobile devices watching segment z of video v with maximum MCS mode m.

For a given video v, depending on the MCS mode, a mobile device needs to receive different number of blocks in each allocation window. This is because the amount of data to transmit is fixed at qTr_v, which can be carried by $\lceil qTr_v/c_m \rceil$ blocks, where q is the symbol time and m is the MCS mode. Allocating different number of blocks to satisfy such capacity demand could largely affect the *off time* of each mobile device, and thus its *energy saving*. We define the energy saving γ as the fraction of time each mobile device can turn off its network interface to save energy. Other factors are less crucial as explained in [56], and then they can be ignored for better tractability. Moreover, previous studies [63, 105] show that mobile device's energy consumption depends on the number of symbols it receives, and it is almost independent of the number of subchannels. Therefore, we assume that base stations first allocate blocks in the same column before considering different ones. The considered problem can be formally written as:

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

The following lemma states the hardness of our problem.

Lemma 1 (Hardness)**.** *The considered transmission scheduling problem is NP-Complete.*

Proof. We reduce the 0-1 knapsack problem to Problem 1. In the 0-1 knapsack problem, we consider O objects, where object o ($1 \leq o \leq O$) has a weight θ_o and a value ϕ_o. The problem is to select a subset of objects for maximizing the total value without exceeding the weight limit $\hat{\theta}$. Given a 0-1 knapsack problem, we generate a corresponding problem instance as follows. For each object o, we create a new MCS mode, and we: (i) add ϕ_o mobile devices in that MCS mode, and (ii) set the per-block capacity to be proportional to the weight θ_o. Last, we set the dTS value based on the weight limit $\hat{\theta}$. This results in a proper instance of Problem 1 in polynomial time. In addition, a solution to Problem 1 can easily be verified in polynomial time. Therefore, Problem 1 is NP-Complete. □

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

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

3.4 Hybrid Streaming over Independent Cell Networks

3.4.1 Mathematical Formulation

We formulate the transmission scheduling problem stated in Problem 1, which assigns the available blocks to individual videos, decides whether to use multicast or unicast, and determines the MCS modes of individual blocks, in order to maximize the overall energy saving while guaranteeing smooth playout. We use the boolean decision variable $x_{v,m,z}$ ($v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$) to denote whether the segment z of video v is unicast/multicast using MCS mode m. That is, $x_{v,m,z} = 1$ if segment z of video v is transmitted with MCS mode m, and $x_{v,m,z} = 0$ otherwise. Recall that $w_{v,m,z}$ denotes the number of mobile devices watching segment z of video v with maximum MCS mode m. Therefore, when $w_{v,m,z} = 1$ the base stations stream video v using unicast; and when $w_{v,m,z} > 1$ the base stations stream video v using multicast. When $x_{v,m,z} = 0$, mobile devices with maximum MCS mode m receive z of v with the next *lower* MCS mode $n \in [1, M]$ that are available in the solution. We define an *intermediate* boolean variable $y_{v,m,n,z}$ for each $v \in [1, V]$, $m, n \in [1, M]$, $n \leq m$, $z \in [1, Z_v]$ as follows. $y_{v,m,n,z} = 1$ when mobile device with maximum MCS mode m would receive segment z of video v with MCS mode n, and $y_{v,m,n,z} = 0$ otherwise. $y_{v,m,n,z}$ is determined by $x_{v,m',z}$, $m' \in [n, m]$ as follows:

$$y_{v,m,n,z} \leq 1 - x_{v,m',z}, \quad \forall\, m' \in [n+1, m], \tag{3.1}$$

$$y_{v,m,n,z} \leq x_{v,n,z}. \tag{3.2}$$

We present the formulation in Eq. (3.3). The objective function in Eq. (3.3a) is to maximize the average energy saving. The total size of video v in an allocation window is qTr_v, and the minimum number of symbols we need is $\lceil \lceil \frac{qTr_v}{c_m} \rceil / S \rceil$, where m is the MCS mode. The three summations iterate through all the videos, modes, and segments, respectively. The constraint in Eq. (3.3b) ensures that the on-demand streaming service only consumes up to d network resources. The constraint in Eq. (3.3c) guarantees that every mobile device receives its allocation window at a feasible MCS mode. This in turn ensures that all mobile devices smoothly render the video. Last, the constraints in Eqs. (3.3d) and (3.3e) are from Eqs. (3.1) and (3.2).

3.4.2 Proposed Algorithms

The proposed algorithms run on the resource allocators close to base stations to determine how to stream videos in order to maximize the overall energy saving of mobile devices. The formulation in Eq. (3.3) is a Binary Integer Programming problem, which can be solved by existing optimization problem solvers, such as CPLEX [33] and GLPK [43]. We use CPLEX to implement the optimal algorithm and refer to it as SCOPT (Single-Cell OPTimum). Although SCOPT gives optimum allocations, its worst-case running time is exponential. Therefore, we develop a heuristic algorithm, called SCG (Single-Cell Greedy), whose pseudocode is given in Figure 3.1. We start from an ideal decision in which the number of blocks is more than enough to enable unicast to all mobile devices.

$$\max_{\mathbf{x}} \quad \gamma = 1 - \frac{1}{N} \sum_{v'=1}^{V} \sum_{m'=1}^{M} \sum_{z'=1}^{Z_{v'}} w_{v',m',z'} \sum_{n'=1}^{m'} y_{v',m',n',z'} \lceil \lceil \frac{qTr_{v'}}{c_{n'}} \rceil \Big/ S \rceil \tag{3.3a}$$

$$\text{s.t.} \quad \sum_{v'=1}^{V} \sum_{m'=1}^{M} \sum_{z'=1}^{Z_{v'}} x_{v',m',z'} \lceil \frac{qTr_{v'}}{c_{m'}} \rceil \leq dTS \tag{3.3b}$$

$$(1 - \sum_{n'=1}^{m} y_{v,m,n',z}) w_{v,m,z} = 0 \tag{3.3c}$$

$$y_{v,m,n,z} \leq 1 - x_{v,m',z}, \ \ \forall\, m' \in [n+1, m] \tag{3.3d}$$

$$y_{v,m,n,z} \leq x_{v,n,z} \tag{3.3e}$$

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

Setting up a unicast channel to each mobile device *maximizes* the overall energy saving. However, the constraint in Eq. (3.3b) may prevent us from setting up a unicast channel for each mobile device, which renders the ideal decision infeasible. To turn an infeasible allocation into a feasible one, we can reduce the number of unicast/multicast with different MCS modes of a video, so that the constraint in Eq. (3.3b) can be satisfied. For example, by changing $x_{1,3}$ from 1 to 0, we reduce the network load attributed to the video streaming service by $\lceil \frac{qTr_1}{c_3} \rceil$ blocks. Doing so, however, leads to negative consequences: devices watching v with MCS mode 3 have to *receive* at a lower MCS mode. This in turn leads to lower energy saving γ in Eq. (3.3a). This example demonstrates the trade-off between *profit* (Eq. (3.3a)) and *cost* (Eq. (3.3b)). Profit refers to any increase in the energy saving, whereas cost refers to any consumption of the radio resources of a base station.

We let $\alpha_{v,m,z}$ and $\beta_{v,m}$ be the *offset* of profit and cost after changing $x_{v,m}$ from 1 to 0. The offset parameters are used to balance between profit and cost. Mathematically, we write $\alpha_{v,m,z} = \sum_{m'=m}^{M} w_{v,m',z} y_{v,m',m,z} \lceil \lceil \frac{qTr_v}{c_m} \rceil / S \rceil$ and $\beta_{v,m} = \lceil \frac{qTr_v}{c_m} \rceil$. Our algorithm strives to *refine* an infeasible allocation by trading the minimum profit reduction (objective function) for the maximum cost reduction (constraint). In particular, our algorithm evaluates the ratio $\tau_{v,m,z} = \alpha_{v,m,z} / \beta_{v,m}$ of all $x_{v,m,z} = 1$ and drops the MCS mode m and video v with the smallest $\tau_{v,m,z}$ value in each iteration. The algorithm stops once the constraint in Eq. (3.3b) is satisfied. Figure 3.2 illustrates a sample solution of our transmission scheduling problem. For clarity, we assume a free space propagation model in which the distance between base stations and terminals is the major impact on the channel quality conditions of mobile terminals. Users located near the base station have higher reception qualities, whereas those users at the cell-edge suffer from lower reception qualities. Based on this, we give the numbers of the maximum modulation and coding scheme for each mobile terminal. We assume terminals in Figure 3.2 are requesting the same video, and the solution for the transmission scheduling problem is to subgroup the multicast session into three subgroups: multicast subgroup transmitted using MCS mode 5, unicast subgroup transmitted using MCS mode 3, and multicast subgroup transmitted using MCS mode 1.

Lemma 2 (Complexity)**.** *The SCG algorithm terminates in polynomial time:* $O(V^2 M^3 Z^2)$.

Algorithm 1: Hybrid Streaming Algorithm for Single-Cell (SCG)

Inputs: $\{V, Z, R\} \leftarrow$ A set of requests for video streams and their data rates
$W_c \leftarrow$ The bandwidth still available for video services over cell c
$Bandwidth(v_z, m) \leftarrow$ computes the required bandwidth to stream segment v_z using MCS m
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: **for** $m \in [M_{Min}, M_{Max}]$ **do**
5: $n_{(v,z,m)}$ = the number of viewers interested to receive this segment using MCS m;
6: **if** $(n_{(v,z,m)} > 0)$ **then**
7: $x_{(v,z,m)} = 1$; // Create a subgroup for this video segment
8: $W_r \ += Bandwidth(v_z, m)$; // Update the required bandwidth to serve videos
9: $X \ += x_{(v,z,m)}$; // Update the set of video segments to be served during this window
10: **for** $n \in [1, m]$ **do**
11: **Update** the intermediate variable $y_{(v,z,m,n)}$ using Eqs. (3.3d) and (3.3e);
12: **end for**
13: **end if**
14: **end for**
15: **end for**
16: **while** $(W_c < W_r)$ **do**
17: **for** each subgroup $x_{(v,z,m)}$ **do**
18: **Compute** $\alpha_{v,m,z}$, $\beta_{v,m}$, $\tau_{v,m,z}$; // Estimate both profit (Eq. 3.3a) and cost (Eq. 3.3b)
19: **end for**
20: Sort X ascendingly based on their ratios (i.e., $\tau_{v,m,z}$);
21: $g_{(v,z,m)} \leftarrow X.getHead()$; // Get the subgroup with the minimum ratio
22: $W_r \ -= Bandwidth(v_z, m)$; // Update the required bandwidth to serve videos
23: $x_{(v,z,m)} = 0$; //Eliminate this subgroup
24: **for** each subgroup $x_{(v,z,m)}$ **do**
25: **for** $n \in [1, m]$ **do**
26: **Update** the intermediate variable $y_{(v,z,m,n)}$ using Eqs. (3.3d) and (3.3e);
27: **end for**
28: **end for**
29: **end while**
30: **return** X

Figure 3.1: SCG: An efficient algorithm to solve the single-cell allocation problem.

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

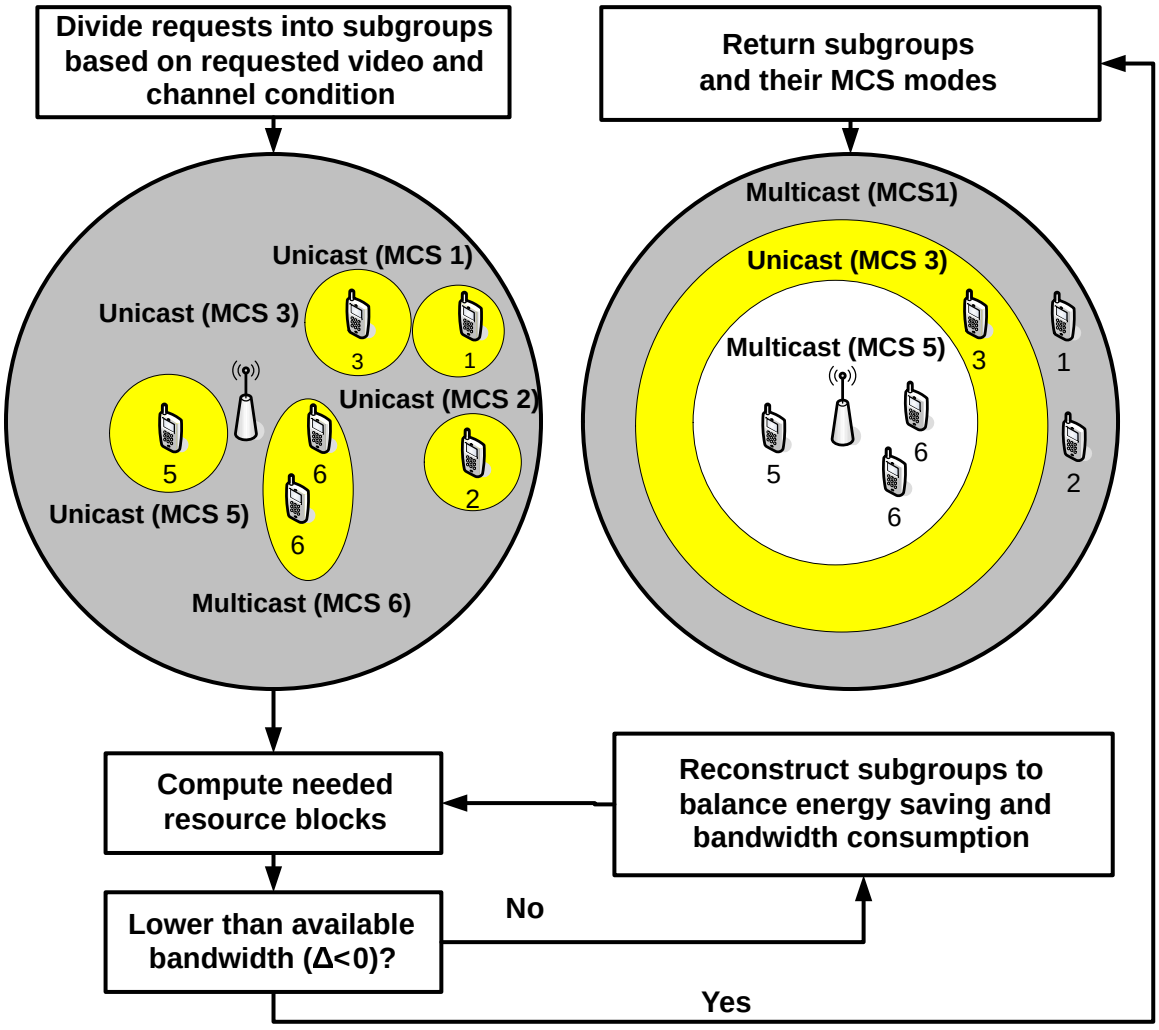


Figure 3.2: An illustrative example for the SCG algorithm.

3.5 Hybrid Streaming over Multi-Cell Single Frequency Networks

The independent cell networks discussed in the previous section follow the conventional cellular design philosophy to increase the network capacity: neighboring cells adopt different frequency bands to minimize inter-cell interference. Such philosophy, however, is largely driven by the unicast nature of conventional cellular networks. For multicast/broadcast sessions, minimizing inter-cell interference essentially means that each cell operates on its own [36]. This is suboptimal since the same signals are transmitted to many mobile devices within the coverage range. It is possible to allow multiple neighboring cells to simultaneously send the same signals in order to boost the signal strength received by the mobile devices at the edges of cells. That is, by sending the same signals from multiple base stations, a mobile device may receive the aggregate signals from several base stations, which leads to better channel conditions, higher MCS modes, higher energy saving,

and lower network resource consumption. Single frequency networks are popular in broadcast services, from FM/AM radios to digital TV [36]. However, managing SFNs in dynamic mobile networks is not an easy task, because mobile networks have to concurrently support both unicast and multicast, which have contradicting goals: minimizing inter-cell interference versus maximizing inter-cell joint signal strength. In this section, we carefully model hybrid video streaming in SFNs and propose (near-)optimal allocation algorithms to solve it.

3.5.1 Mathematical Formulation

The formulation in Eq. (3.3) considers a single cell. We consider H hexagonal cells that form a *dynamic* single frequency network, where each block can be assigned to an SFN independently. Such an extension requires two major enhancements: (i) expanding the solution space to multiple cells and (ii) modeling SFN gains from neighboring cells. We explain each of the enhancements below.

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

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

Combining these two enhancements, we get the formulation for an SFN in Eq. (3.4). The objective function in Eq. (3.4a) maximizes the average energy saving across all H cells. The constraint in Eq. (3.4b) makes sure that each cell is not overloaded. The constraint in Eq. (3.4c) ensures that every mobile device receives at an MCS mode, which is equal to or smaller than its maximum MCS mode. The constraints in Eqs. (3.4d) and (3.4e) relate variables $y_{v,m,n}^h$ and $x_{v,m}^h$. The constraint in Eq. (3.4f) takes the SFN gains into consideration.

$$\max_{\mathbf{x}} \quad \gamma = 1 - \frac{\Big[\sum_{h'=1}^{H}\sum_{v'=1}^{V}\sum_{m'=1}^{M}\sum_{z'=1}^{Z_{v'}} \hat{w}^{h'}_{v',m',z'} \sum_{n'=1}^{m'} y^{h}_{v',m',n',z'} \lceil \lceil \frac{qTr_{v'}}{c_{n'}} \rceil \Big/ S \rceil \Big]}{\sum_{h'=1}^{H} N^{h'}} \tag{3.4a}$$

$$\text{s.t.} \quad \sum_{v'=1}^{V}\sum_{m'=1}^{M}\sum_{z'=1}^{Z_{v'}} x^{h}_{v',m',z'} \lceil \frac{qTr_{v'}}{c_{m'}} \rceil \leq dTS, \ \ \forall\, h \in [1,H] \tag{3.4b}$$

$$(1 - \sum_{n'=1}^{m} y^{h}_{v,m,n',z}) \hat{w}^{h}_{v,m,z} = 0, \ \ \forall\, h \in [1,H] \tag{3.4c}$$

$$y^{h}_{v,m,n,z} \leq 1 - x^{h}_{v,m',z}, \ \ \forall\, m' \in [n+1, m], \ h \in [1,H] \tag{3.4d}$$

$$y^{h}_{v,m,n,z} \leq x^{h}_{v,n,z}, \ \ \forall\, h \in [1,H] \tag{3.4e}$$

$$\hat{w}^{h}_{v,m,z} = w^{h}_{v,m,z} + \sum_{h' \in [1,H] \setminus \{h\}} x^{h'}_{v,m,z} \delta^{h,h'}_{v,m,z}, \ \ \forall\, h \in [1,H] \tag{3.4f}$$

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

3.5.2 Proposed Algorithms

Similar to Eq. (3.3), Eq. (3.4) is a Binary Integer Programming problem, which can be solved by existing optimization solvers [33, 43]. We use CPLEX [33] to implement an optimal algorithm, called SFNOPT (Single Frequency Network OPTimum). SFNOPT has an exponential complexity, so we propose a heuristic algorithm, called SFNG. Its pseudocode is presented in Figure 3.3. The main idea of SFNG is to start with the best-case scenario where each video is transmitted with as many MCS modes as possible, such that the average energy saving at mobile devices is maximized. Most likely, this requires excessive amount of radio resources, which may not be feasible due to the constraint in Eq. (3.4b) on the available amount of bandwidth for video streaming services. To overcome infeasible solutions, we iteratively reduce the video traffic within the cell $\hat{h} \in [1,H]$ that suffers from the largest excessive network load. To achieve this goal, we reduce the number of unicast/multicast streams in $\hat{h}$ by removing one stream at each iteration. The selection of which video stream to drop is decided based on the profit and cost analysis dictated by α, β, and τ. In each iteration, an MCS mode $m^* \in [1,M]$ of allocation window $z^* \in [1, Z_v]$ and video $v^* \in [1,V]$ is removed so that the network load of $\hat{h}$ is reduced at the expense of lower energy saving. Once a stream is dropped, we reset $x^{\hat{h}}_{v^*,m^*,z^*}$ to 0 and then re-compute the required bandwidth for the current solution to determine its feasibility. The SFNG algorithm terminates as soon as a feasible allocation is derived.

Lemma 3 (Complexity)**.** *The SFNG algorithm terminates in polynomial time:* $O(HV^2M^3Z^2)$.

Proof. Let $Z = \max_{v=1}^{V} Z_v$. The for-loop starts from line 4 has a complexity of $O(HVM^2Z)$. The while-loop starts from line 7 repeats up to $HVMZ$ times, the for-loop starts from line 8 repeats up to VMZ times, and the line 9 updates up to M $y^{\hat{h}}_{v,m,n}$ values. Thus, SFNG's time complexity is $O(HVM^2Z) + O(HV^2M^3Z^2) = O(HV^2M^3Z^2)$. □

Algorithm 2: Hybrid Streaming Algorithm for Single Frequency Networks (SFNG)

Inputs: $\{V^h, Z^h, R^h\} \leftarrow$ A set of requests for video streams and their data rates in a cell h
$W_h \leftarrow$ The bandwidth still available for video services over cell h
$Bandwidth(v_z, m) \leftarrow$ computes the required bandwidth to stream segment v_z using MCS m
Output : $X^h \leftarrow$ The set of video segments to be served during the current allocation window

$W_r^h = 0; \quad X^h = \{\emptyset\};$
for $h \in [1, H]$ **do**
  for each required segment (v, z) **do**
    for $m \in [M_{Min}, M_{Max}]$ **do**
      $n^h_{(v,z,m)} =$ the number of viewers interested to receive this segment using MCS m;
      if $(n^h_{(v,z,m)} > 0)$ **then**
        $x^h_{(v,z,m)} = 1;$ // Create a subgroup for this video segment
        $W_r^h \ += Bandwidth(v_z, m);$ // Update the required bandwidth to serve videos
        $X \ += x^h_{(v,z,m)};$ // Update the set of video segments to be served
        for $n \in [1, m]$ **do**
          Update the intermediate variable $y^h_{(v,z,m,n)}$ using Eqs. (3.3d) and (3.3e);
        end for
      end if
    end for
  end for
end for
Let $\hat{h} = \text{argmax}_{h=1}^{H}(W_r^h - W_h)$ // Find the cell with the highest deficit in resource blocks
while $(W_{\hat{h}} < W_r^{\hat{h}})$ **do**
  for each subgroup $x^{\hat{h}}_{(v,z,m)}$ **do**
    Compute $\alpha^{\hat{h}}_{v,m,z}, \ \beta^{\hat{h}}_{v,m}, \ \tau^{\hat{h}}_{v,m,z};$ // Estimate both profit (Eq. 3.3a) and cost (Eq. 3.3b)
  end for
  Sort $X^{\hat{h}}$ ascendingly based on their ratios (i.e., $\tau^{\hat{h}}_{v,m,z}$);
  $g_{(v,z,m)} \leftarrow X^{\hat{h}}.getHead();$ // Get the subgroup with the minimum ratio
  $W_r^{\hat{h}} \ -= Bandwidth(v_z, m);$ // Update the required bandwidth to serve videos
  $x^{\hat{h}}_{(v,z,m)} = 0;$ //Eliminate this subgroup
  for each subgroup $x_{(v,z,m)}$ **do**
    for $n \in [1, m]$ **do**
      Update the intermediate variable $y^{\hat{h}}_{(v,z,m,n)}$ using Eqs. (3.3d) and (3.3e);
    end for
  end for
  Let $\hat{h} = \text{argmax}_{h=1}^{H}(W_r^h - W_h)$ // Find the cell with the highest deficit in resource blocks
end while
return X

Figure 3.3: SFNG: An efficient algorithm to solve the allocation problem in SFN.

3.6 Evaluation

In this section, we present extensive trace-driven simulation results from a popular packet-level simulator. We demonstrate the near optimality of our algorithms, and we show that they significantly increase the number of served users and reduce the overall energy consumption, while imposing minimal overhead on the cellular network. In addition, we simulate a realistic SFN with 10 base stations in downtown Vancouver, Canada, and we show that our SFN solution further increases the number of served mobile users and saves more energy of mobile devices. We also show that our algorithms outperform the closest three solutions in the literature [19, 67, 76] as well as the energy saving scheme introduced in [51].

3.6.1 Simulation Setup

We have implemented an on-demand video streaming system in OPNET [82]. We have also implemented the proposed SCG, SCOPT, and SFNG using a mixture of C/C++, Matlab, and CPLEX [33] in the simulator. The heuristic SCG algorithm is evaluated against the optimal solutions generated by SCOPT. We do not compare SFNG against SFNOPT, because the latter incurs prohibitively long running time: it may take hours to terminate. Moreover, we have implemented the maximum throughput algorithm [19], proportional fair algorithm [76], combined unicast-multicast algorithm [67], and energy saving algorithm [51], and we refer to them as *MT, PR, COMB, ES*, respectively. In addition to the transmission scheduling algorithms, we customize the simulator to employ a few practical heuristics. For instance, if an incoming request from a mobile user is rejected due to resource scarcity, this mobile user will retry for up to 3 times with an exponential back-off waiting period starting from 2 seconds. After being rejected three times, it stops requesting the desired video. As another example, we incorporate batching in the sense that all requests for videos within the duration of an allocation window are grouped together to be served at the beginning of the next allocation window. These heuristics are likely to be implemented in real video streaming services.

3.6.2 Test Scenarios for Independent Cell Networks

We consider multiple base stations that operate independently, where each cell covers a 10×10 km^2 area. We consider up to 1,000 mobile devices in a cell, and these users join our system following a Poisson process with mean λ, which is set to 20 users per second by default. These mobile devices are randomly deployed in the transmission coverage area, such that more users are located close to base stations as cellular operators typically construct their networks so that more base stations are in the crowded regions. In particular, we assume 90% of users are in 1/3 of the cell radius. Our system does not require any prediction for the user mobility in order to perform its handover operations, make its scheduling decisions, or obtain the density of user distributions within its cells. Therefore, in our simulations, mobile users can either: remain static or follow a random waypoint model, which is chosen for simplicity since it does not depend on any GPS traces of human walks or

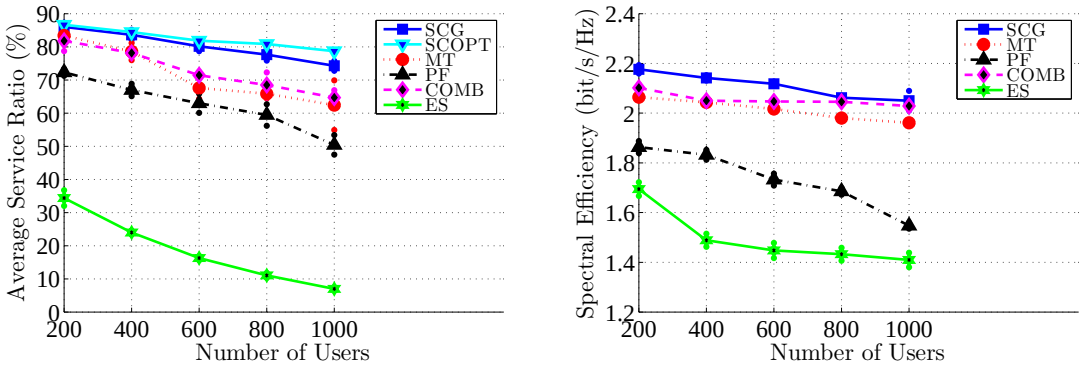
cellphone location tracking. Upon joining the network, a user randomly requests a video and leaves once the video is finished.

3.6.3 Results for Independent Cell Networks

We compare our SCG algorithm versus three multicast-only approaches (i.e., MT [19], PR [76], and COMB [67]) and a unicast-only approach (i.e., ES [51]). The performance metrics are service ratio, spectral efficiency, energy saving, Peak Single-to-Noise Ratio (PSNR), frame loss rate, initial buffering time, and number of re-buffering events. We simulate LTE networks where mobile devices in each cell generate requests from a pool of 1000 possible video streams. We vary the number of users in a cell from 200 to 1000, and report the mean results from 5 simulation runs in Figures 3.4(a)–3.5(b). The variance for each value is included as points in these figures as well. These results indicate that our proposed algorithms not only outperform others with significant margins on achieved service ratio, but also save more energy than multicast-only approaches without causing any violation at the buffer levels nor degraded video quality. Detailed simulation results are discussed below.

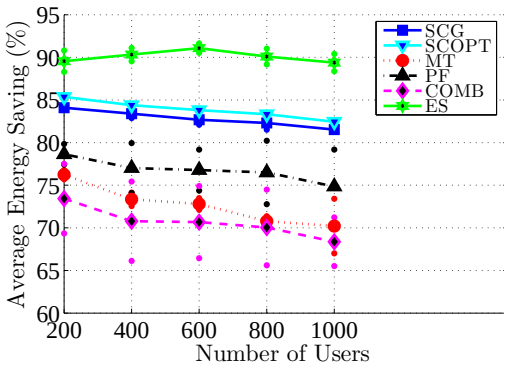
Service ratio. Due to the limited radio resources in cellular networks, it may be impossible to serve all requesting mobile users. Therefore, we compute the service ratio as the fraction of admitted mobile users to the number of received requests. Figure 3.4(a) indicates that our SCG algorithm outperforms other algorithms on achieved average service ratio. For instance, when there are 1000 mobile users in each cell, the proposed algorithm admits 74.5% of users at any given time, while systems employing MT, PF, COMB, and ES algorithms accept only 62.5%, 50.5%, 65%, and 7% of users, respectively. This shows that our SCG algorithm provides a service ratio that is 20%, 47%, 15%, and 965% higher than the MT, PF, COMB, and ES. Last, we note that the results are from a cellular network that dedicates all network resources to the video streaming service. Similar outcomes are observed under different parameters, such as reserved network resources and number of mobile users. Compared against the optimal solution given by SCOPT, our SCG algorithm gives only 0.76% and 5.94% lower average service ratio when the numbers of users in each cell are 200 and 1000, respectively.

Spectral efficiency. Here, the spectral efficiency is defined as the total transmitted data rate (in bits per second) divided by the allocated bandwidth (in Hertz) [18]. As it is shown in Figure 3.4(b), the proposed heuristic algorithm outperforms the other four approaches by providing a spectral efficiency between 2.05 and 2.18 bits/second/Hertz, depending on the number of mobile terminals within the cell. These performance results are at least 28% and 17% higher than the unicast scheme (ES) and the fair proportional multicast policy (PF), respectively. Compared with those multicast approaches applying the multicast subgrouping concept (i.e., MT and COMP), our heuristic algorithm still gives up to 5% increase in its spectral efficiency. Such improvement is achieved by applying the hybrid unicast-multicast approach, in which users with poor channel quality conditions can be removed from a multicast subgroup and served via unicast connections. Then this multicast sub-

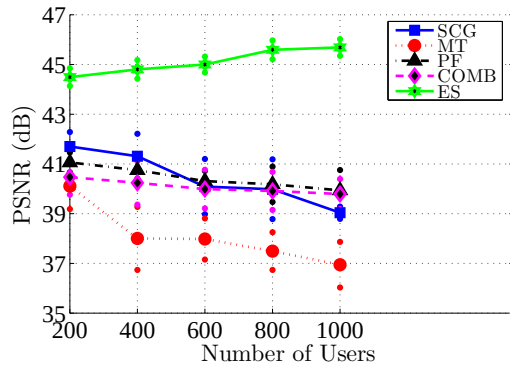


(a) Service Ratio

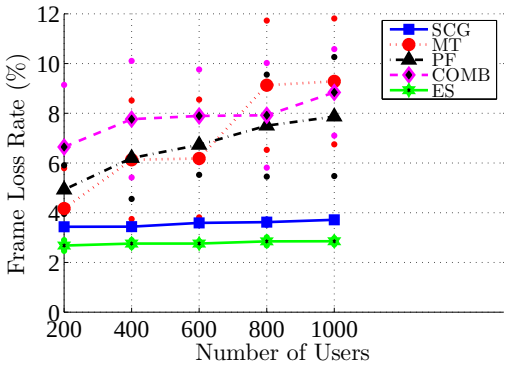
(b) Spectral Efficiency



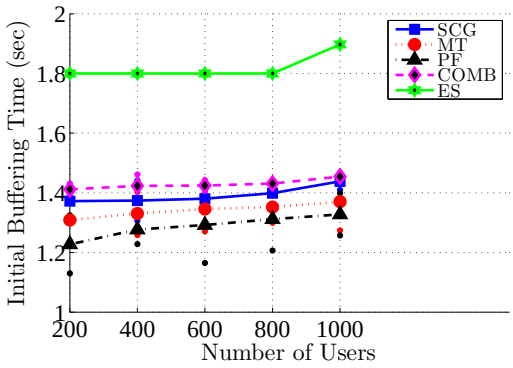
(c) Energy Saving



(d) Video Quality in PSNR



(e) Frame Loss Rate



(f) Initial Buffering Time

Figure 3.4: Comparisons of the achieved performance of the proposed algorithms against the state-of-the-art approaches.

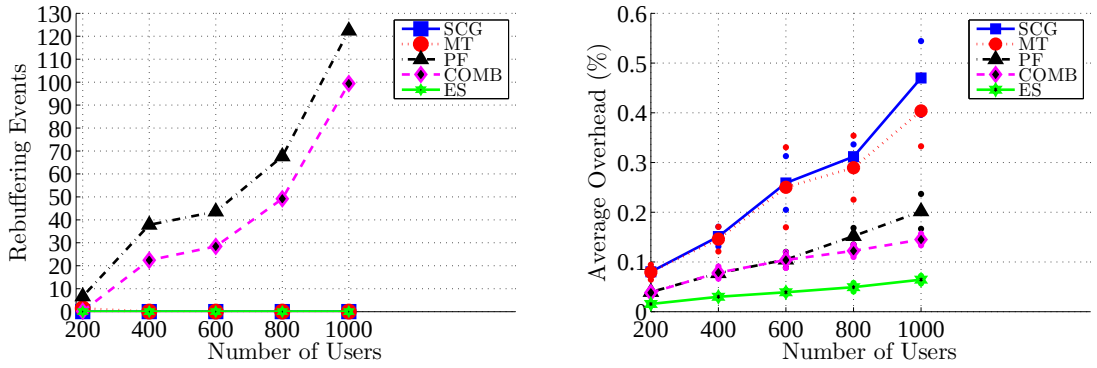
group would be sent using a higher modulation and coding scheme, thereby increasing the achieved spectral efficiency of the mobile system.

Energy saving. We define the energy saving as the percentage of time a served mobile device can turn off its network interface, to reduce its energy consumption. In 4G/5G cellular networks, the time required to switch the network interface between active and idle is small. According to Huang et al. [56], switching an LTE interface on contributes 1.2% to the total power consumption when a single packet is transmitted. In our system, the number of packets transmitted to each user during the active period is larger than one packet since each burst transmission represents a two-second video segment. Thus, to compute the energy saving, it is sufficient to account for the time duration when a network interface is off. Unicast-only approaches achieve the maximum energy saving possible in independent-cell networks since individual mobile users are served according to their best MCS modes. Figure 3.4(c) illustrates that our SCG algorithm leads to 6.5% and 9.5% lower saving than the ES algorithm when there are 200 and 1000 users in a cell, respectively. However, compared to multicast-only approaches (i.e., MT, PF, and COMB), our proposed algorithms outperform them by 9–20% in energy saving. Comparing the results achieved by our SCG algorithm versus those computed by the optimal SCOPT algorithm, we notice that the energy saving obtained in our SCG algorithm is close to the optimal with a small gap of 1.3% on average.

Video quality. Figures 3.4(d) and 3.4(e) 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 unicast-only approaches (ES) achieves the highest PSNR and the lowest frame loss rate. This is because it only admits very few mobile users at a time, making it less commercially viable. In contrast, with 200 mobile users in each cell, our proposed SCG algorithm yields an average of 41.7 dB in PSNR and 3.43% in frame loss rate. Even when the number of mobile users is increased from 200 to 1000, the SCG algorithm still achieves 39.04 dB in PSNR and 3.7% in frame loss rate.

Allocation window size. In video streaming systems, a playback starts after an initial buffering time and continues while the video stream is being downloaded. The initial buffering time in our algorithm depends mainly on the transmission scheduling window size. Intuitively, longer scheduling windows provide more chances for expanding the multicast groups, thereby result in higher service ratios. Yet, larger scheduling windows increase the initial buffering time. Given that our heuristic algorithms terminate in less than 1 millisecond, we recommend short scheduling windows for short initial buffering time. In our simulations, the window size is set to be 2 seconds by default, which is equivalent to the size of video chunks produced by adaptive video streaming solutions, such as Microsoft Silverlight. At this window size, the initial buffering time is shown in Figure 3.4(f), which shows that our algorithms outperform unicast-only approaches in initial buffering time, and scale well with many more mobile users.

Number of re-buffering events. We instrument our simulator to keep track of the buffer status of each mobile device. 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 that our proposed algorithms never lead to buffer overflow events. Then, we calculate the number of re-buffering



(a) Rate of Re-buffering Events

(b) Channel Quality Overhead

Figure 3.5: Comparisons of the achieved performance of the proposed algorithms against the state-of-the-art approaches.

events of different algorithms, and report the numbers in Figure 3.5(a). This figure shows that our SCG algorithm results in no re-buffering event.

Feedback Overhead. Mobile devices in our algorithms and other state-of-the-art algorithms [19, 51, 67, 76] 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 MCS mode at which the block error rate is low, e.g., $< 5\%$. In LTE Release 12 [2], two different reports can be obtained from mobile devices: sub-band and wide-band feedback. Sub-band reports give channel state information for each sub-band, whereas wide-band reports give 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 (DRX) for energy saving, not all users utilize the dedicated upload control channels all the time. Instead, the wide-band reports are sent by mobile devices only when they receive videos. 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 3.5(b) shows the overhead occurred in the five algorithms when the number of users within each cell is varied. Although the SCG algorithm admits more users than other works, the feedback overhead in our algorithm is still less than 0.08% and 0.47% in the cases where the number of users are 200 and 1000, respectively.

Support for scalable videos. Even though the previous results are obtained using non-scalable videos, our proposed algorithms can be easily generalized to support scalable video coding. To do so, each video-segment is divided into layers. We can then include an additional constraint to consider the dependency among layers in scalable videos as follows:

$$x_{v,z,l,m} = 1 \text{ if } x_{v,z,l',m'} = 1, l < l', m \leq m'. \tag{3.5}$$

This condition assures that for each video-segment (v,z), no higher layer (l') is transmitted unless its base and lower layers are already scheduled. Our algorithms, analysis, and implementations still work after this augmentation. To study the impact of scalable video coding on the proposed algorithms, we used the freely licensed animation video sequence *Big Buck Bunny*, whose traces are available from the Video Trace Library [21]. *Big Buck Bunny* consists of $14,315$ frames in the HD 1920×1080 pixels format with a frame rate of 24 frames/sec and average bit rates between 36.393 Kbit/sec and 1.094 Mb/sec. More details about this H.264 sequence can be found in [87]. During our simulation, more than 1000 mobile terminals are deployed around a base station. These mobile terminals are experiencing different channel quality conditions, and they are requesting the scalable video stream at the same time. From the obtained results, our algorithm achieves an energy saving equals to 92.58% and a PSNR value equals to 41.12 dB. On the other hand, the conventional multicast in [76] as an example achieves an energy saving equals to 84.25% and a PSNR value equals to 20.39 dB. That means our algorithm outperforms the conventional multicast by providing almost 10% and 102% improvement in both energy saving and PSNR value, respectively. Such improvements are achieved at the cost of a slightly increased consumption of radio resources (i.e., $< 4.75\%$), which can be acceptable especially during non-rush hours.

3.6.4 Test Scenarios for Multi-Cell SFNs

We construct a multi-cell SFN using the actual base station locations of a Canadian cellular operator in Vancouver, Canada, which are obtained from the published information [26]. In particular, we consider 10 base stations around the West Georgia Street in downtown, Vancouver as shown in Figure 3.6. These 10 cells are assumed to be in the same Multicast Broadcast Single Frequency Network (MBSFN) area. 9 base stations at the east side have a maximum transmission power of 0.3 Watt, whereas the left-most base station has a maximum power of 0.5 Watt. Mobile users arrive to the video streaming service following a Poisson process with a mean arrival rate of 30 users per second in each cell. The initial locations of mobile users are uniformly distributed within each cell.

We consider two test scenarios: static and mobile. In the mobile scenario, users move randomly in either east-west or north-south directions to mimic mobile users commuting along urban streets. A mobile device represents a pedestrian who walks at 4.5 km/hour or a driver who drives at 50 km/hour. Mobile users never leave the multi-cell SFN throughout simulations. We consider the following algorithms: SCG, MT, PF, COMB, and ES, and our heuristic algorithm for multi-cell SFN (SFNG). Each simulation scenario run lasts for 20 minutes.

3.6.5 Results for Multi-Cell SFNs

Performance in the static scenario: Table 3.2 gives the average performance results across mobile devices in the static scenario. We notice that both SCG and SFNG clearly outperform multicast-only approaches [19]- [67] in term of energy saving and the unicast approach [51] in term of service ratio. In fact, ES [51] results in a fairly low service ratio of 25.24%, which may drive users away

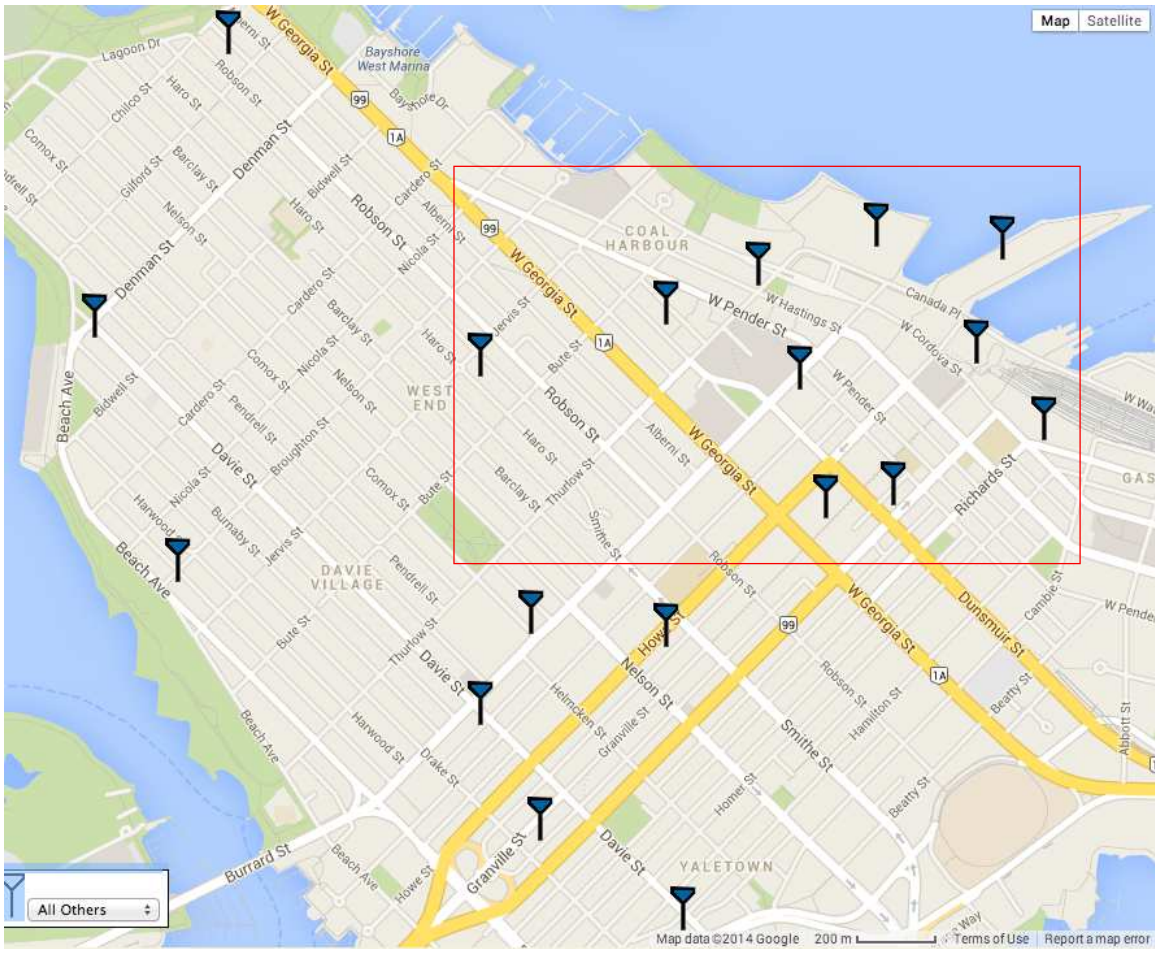


Figure 3.6: Locations of base stations of a leading Canadian cellular operator in downtown Vancouver, British Columbia.

from the video streaming service. On the other hand, the proposed SFNG algorithm achieves a service ratio of 91.66%, which is up to 65.75% higher than those service ratios delivered by the state-of-the-art multicast-only approaches [19]- [67]. We also observe that our SFNG in multi-cells SFNs outperforms our SCG in independent cell networks by up to: (i) 5.73% in energy saving, and (ii) 12.55% in service ratio. This reveals that our SFNG algorithm indeed capitalizes the advantage of SFNs.

Performance in the mobile scenario: Table 3.3 presents the average service ratio and energy saving when the mobility model is applied. These performance results are inline with our earlier findings in Table 3.2. They also confirm that our algorithms outperform other multicast-only and unicast-only approaches under diverse user distributions. However, our SFNG algorithm is superior in its achieved results than SCG by up to 6.57% in service ratio and up to 3.10% in energy saving. This can be attributed to the fact that, in SFNs, video streams are transmitted simultaneously over the air from multiple synchronized base stations, which allow mobile devices in the same SFN area

Table 3.2: Performance Results in Static Scenario.

Metric	MT	PF	COMB	ES	SCG	SFNG
Energy Saving (%)	72.60 ±0.68	77.51 ±2.18	70.81 ±4.32	90.24 ±0.86	84.43 ±0.43	89.27 ±0.36
Service Ratio (%)	68.43 ±2.47	55.30 ±1.94	70.94 ±1.68	25.24 ±1.21	81.44 ±1.43	91.66 ±1.40

Table 3.3: Performance Results in Mobile Scenario.

Metric	MT	PF	COMB	ES	SCG	SFNG
Energy Saving (%)	74.81 ±1.49	80.68 ±3.48	74.13 ±3.84	90.52 ±0.83	86.88 ±0.54	89.58 ±0.69
Service Ratio (%)	69.05 ±4.07	62.33 ±3.03	71.81 ±2.34	33.35 ±0.66	89.48 ±1.58	95.36 ±1.62

to receive stronger signals. For example, mobile devices treat (leverage) the signals from different base stations as multipath components. Hence, mobile users enjoy higher SNR levels, and can survive more aggressive MCS modes for higher service ratio and energy saving.

3.7 Summary

We studied the transmission scheduling problem for large-scale video streaming over cellular networks and proposed novel algorithms to utilize both unicast and multicast. Our main goal is to support more mobile users with less consumed energy on mobile devices. Next generation cellular networks enable two multicast schemes: (i) independent cells in which each base station initiates multicast sessions only to those users within its transmission coverage, and (ii) multi-cell single frequency network (SFN) in which multiple cells collaborate to deliver synchronized video streams using identical radio frequency bands. We formulated optimization problems for the hybrid video streaming service in these two schemes. Then we developed two optimal algorithms (SCOPT and SFNOPT) to solve the two allocation problems and two heuristic algorithms (SCG and SFNG) for faster and near-optimal results, even in cases with highly dense user distributions.

We considered an LTE network as an example to assess the performance of our algorithms with respect to the service ratio, spectral efficiency, energy saving, video quality, frame loss rate, initial buffering time, and number of re-buffering events. We implemented the proposed algorithms and the closest and most recent four solutions in the literature in OPNET. Our detailed simulation results indicate that: (i) our algorithms for independent cell networks admit more users, consume less energy, and provide lower frame loss rate without causing any buffer violation or degraded video quality compared to the multicast-capable algorithms, (ii) our algorithms achieve energy saving close to unicast approaches, while supporting almost 11 times more users, and (iii) our extended algorithms for SFNs perform better than algorithms that do not leverage the features of SFN.

Chapter 4

Dynamic Configuration of Single Frequency Networks

In this chapter, we present a novel dynamic configuration of single frequency networks in order to increase the number of admitted mobile terminals within cells. We formulate the transmission scheduling problem in such systems, and we provide optimal and heuristic algorithms to achieve the optimization goal. We evaluate the proposed algorithms in the cases of single cell configuration, static single frequency networks, and dynamic single frequency networks. We then discuss the impact of control overhead on the proposed algorithms as well as state-of-the-art approaches. We also show the performance of the proposed algorithms under different user behavior models.

4.1 Introduction

Although the capacity of cellular networks has increased with recent generations, the growth in demand of wireless bandwidth has outpaced this increase in capacity. Not only more users are relying on wireless networks, but also the demand from each user has substantially increased. For example, it has become common for mobile users to stream full TV episodes, sports events, and movies while on the go. Further, as the capabilities of mobile devices improve, the demand for higher quality and even 3D videos will escalate, which will strain cellular networks. Within 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 1.9 TB [98]. 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, e.g., in the evening of the

release day. Because these videos are not immediately played back, their requests can be grouped into multicast sessions [42]. These and similar applications offer optimization opportunities to save the radio resources of mobile networks.

As defined in recent 4G standards, e.g., [36], 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.

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 [83, 95], and the size of an SFN and its impact on packet delivery [11], 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 user distribution change with time; which is the problem we address.

The rest of this chapter is organized as follows. Section 4.2 summarizes the related works in the literature. Section 4.3 describes the system model used in this chapter, and Section 4.4 states and formulates the considered problem. Sections 4.5 and 4.6 present the proposed optimal and heuristic algorithms, receptively. Section 4.7 presents our simulation results to assess the performance of our algorithms and compare them against others. Section 4.8 concludes the chapter.

4.2 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. [83] and Talarico and Valenti [95] 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. [97] extend this assessment and provide a comprehensive evaluation of SFN performance under more realistic conditions. Alexiou et al. [11] 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 [11, 83, 95, 97] 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. [28] optimize the unicast multimedia connections over Dynamic Adaptive Streaming over HTTP (DASH) with respect to fairness, stability, and efficiency. Elsherif et al. [40] propose a transmission scheduling 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, Zhixue et al. [71] and Liang et al. [69] 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 transmission scheduling within the network.

The concept of multicast over mobile network has been explored by a number of research works [6, 18, 19, 61, 101, 103]. For instance, Araniti et al. [19] and Won et al. [101] address the issue of transmission scheduling over OFDMA, which is mainly related to the different data rate and quality requirements of users in the same multicast group. Keller et al. [61] 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. [103] 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. [76] and Lee et al. [66] 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, Deng et al. [38] utilize the hybrid approach to guarantee a certain level for the quality of service.

Our proposed transmission scheduling algorithm in this chapter aims at increasing the bandwidth utilization of a cellular network, but it is different 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 [38, 66, 76]. 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 changed. Therefore, we in-

Table 4.1: Symbols Used in This Chapter.

Symbol	Description
α	The skewness factor of the Zipf distribution
λ	The arrival rate of the Poisson distribution
$n_{v,z}$	No. mobile devices watching the segment z of a video v
$n_{v,m,z}$	No. mobile devices watching segment z of video v with m
$x_{v,m,z}$	Whether segment z of video v is sent MCS mode m
$g_{v,z}$	Multicast group in which segment z of video v is transmitted
$g_{v,z,m}$	Subgroup interested in watching segment z of video v with MCS mode m
A	Set of active SFN areas in which cell c can join
A_c	Set of active SFN areas in which cell c is enrolled
A_L	Upper limit of SFN areas in which a cell c can enroll
$y_{a,c}$	Whether cell c is enrolled into SFN area a or not

crease 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 [19, 67, 76], 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.

4.3 System Model

We list all symbols used in this chapter and their definitions in Table 4.1. We consider a wireless cellular network with base stations, mobile devices, and multi-cell coordination entities as illustrated in Figure 4.1. 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

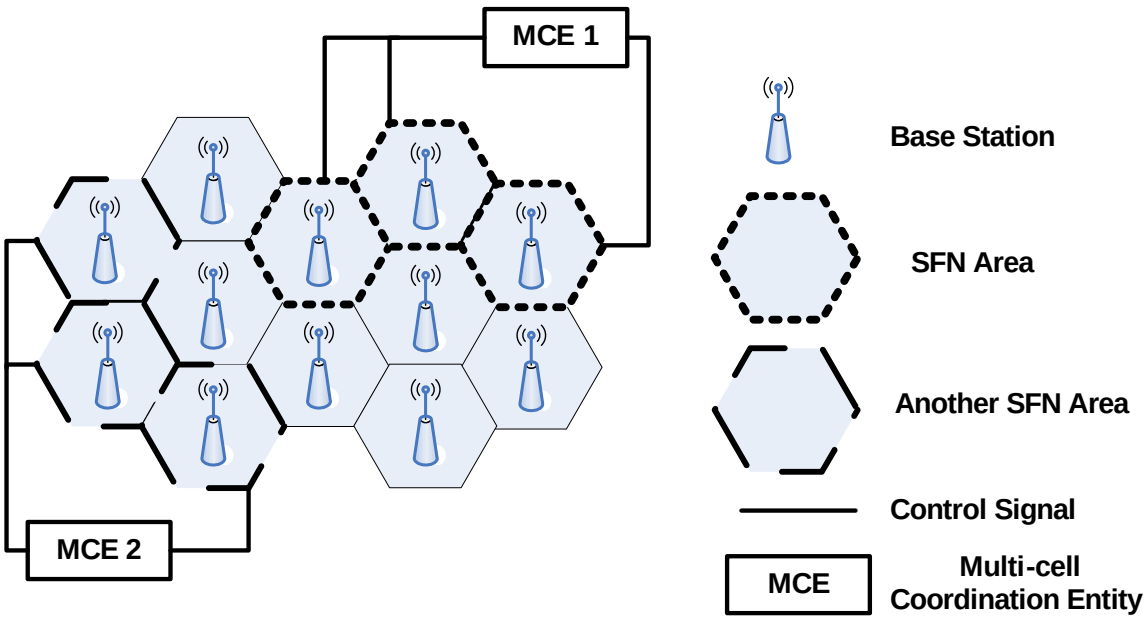


Figure 4.1: The considered model for a mobile network.

compared to multicasting within a single-cell [36]. Since the coordinated cells are sending using identical radio signals, receivers at cell edges get multiple copies of the same data but from different base stations. Hence, the 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 4.1 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, i.e., dTS resource blocks, for video services. The mobile terminal 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 [62]. Hence, the sub-carrier power allocation is P/S.

In this chapter, 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

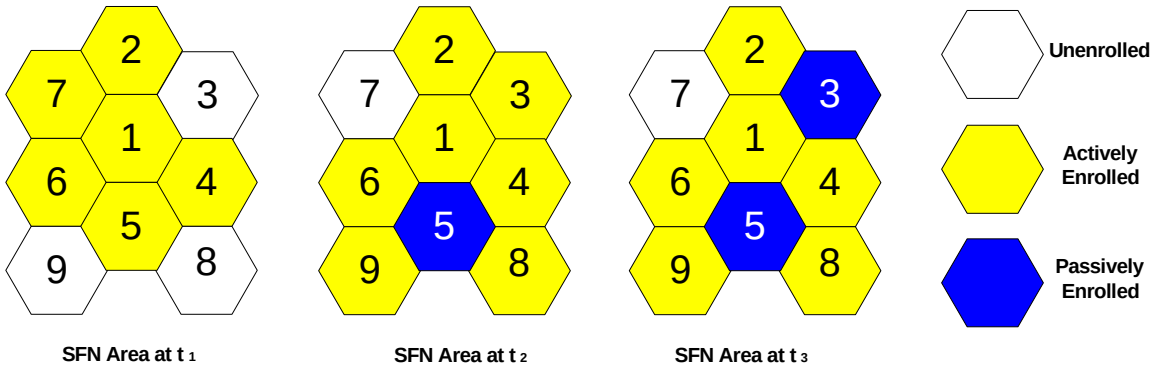


Figure 4.2: Three possible types of cells within an SFN area.

that there are three possible types of cells within an SFN area, as shown in Figure 4.2. 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 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 4.5 and 4.6.1.

4.4 Problem Definition and Complexity

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.

4.4.1 Problem Statement

Multiple prior works consider static configurations for single frequency networks, e.g., [36]. 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 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 chapter can be stated as follows:

Sub-Problem 2 (SFN Configuration)**.** *For a given cell c, select the optimal subset of SFNs to join so that 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 approaches adopt a conservative approach, which restricts the rate of the multicast session to the user with the worst channel condition. These approaches introduce 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 that they do not slow down all users in the multicast session.

Thus, our second problem in this chapter 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 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 3 (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 that the average service ratio across cells is maximized.*

$$\max_{\mathbf{X}} \quad O = \sum_{v=1}^{V} \sum_{z=1}^{Z_v} \sum_{m=1}^{M} x_{vzm} n_{vzm} \tag{4.1a}$$

$$s.t. \qquad B = \sum_{v=1}^{V} \sum_{z=1}^{Z_v} \sum_{m=1}^{M} x_{vzm} \frac{r_{vz}\Gamma}{c_m} \leq dTS \tag{4.1b}$$

$$\sum_{m=1}^{M} x_{vzm} \leq 1 \tag{4.1c}$$

$$SFN_i \left(\frac{O}{B}\right) > SFN_{i-1} \left(\frac{O}{B}\right) \tag{4.1d}$$

$$\sum_{a=1}^{A} y_{a,c}^{t_i} \leq A_L \tag{4.1e}$$

$$x_{vzlm} \in \{0,1\}, \forall\, v \in [1,V], z \in [1,Z_v], m \in [1,M] \tag{4.1f}$$

4.4.2 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 Eq. (4.1), in which we consider the trade-off between the number of served users at a certain modulation and coding scheme, i.e, n_{vzm} and their resource requirements, i.e., $r_{vz}\Gamma/c_m$. The objective function of Eq. (4.1a) is to maximize the total number of served users within a cell. Eq. (4.1b) implements the resource constraint. Eq. (4.1c) ensures that at most one MCS mode is selected for each video segment. Eq. (4.1d) 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, Eq. (4.1e) 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, keeping 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 or not. Solving this optimization problem, we can find the decision matrix $\mathbf{X}$ which contains the set of video segments selected for transmission and their corresponding MCS modes.

4.4.3 Problem Complexity

The formulation of Eq. (4.1) 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 i^{th} subchannel and j^{th}

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 following: 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 i^{th} subchannel and j^{th} 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.

4.5 Proposed Optimal Algorithm

The goal of our proposed optimal algorithm is to maximize the service ratio during a transmission scheduling window. The proposed algorithm is shown in Figure 4.3. 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

Algorithm 3: Optimal Service Ratio Algorithm

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: $n_{(v,z)} \mathrel{+}= n_{(v,z,m)}$;
9: $g_{(v,z,m)} = n_{(v,z)}$; // Calculate the gain in service ratio resulted by serving this subgroup
10: $g_{(v,z)} \mathrel{+}= g_{(v,z,m)}$; // Merge this subgroup into its larger streaming group
11: **end for**
12: $X \mathrel{+}= g_{(v,z)}$; // Update the set of video segments to be served during this window
13: $W_r \mathrel{+}= Bandwidth(g_{(v,z)}, c)$; // Update the required bandwidth to serve video requests
14: **end for**
15: **if** $(W_c < W_r)$ **then**
16: // Find the optimal set of subgroups that maximizes the service ratio as given in Figure 4.4
17: **return** $X \leftarrow$ **OPTSET**(X, W_c);
18: **end if**

Figure 4.3: Proposed transmission scheduling algorithm to maximize the service ratio for a video service over mobile networks.

is 170 [2], assuming an average video bit rate of 300 Kbps [4] and maximum bandwidth of 20 MHz [36]. 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 it will be shown in Section 4.7, both running time and bandwidth overhead are reduced by around 360.7% and up to 48.8%, respectively,

Function OPTSET(X, W_c)

if (($X = \emptyset$) or ($W_c \leq 0$)) **then**
  return $[X = \emptyset, S = 0]$; // Return an empty set with zero gain in service ratio
else
  $g_{(v,z)} \leftarrow X.getHead()$; // Retrieve a streaming group from the given set
  // Calculate the service ratio if this group is chosen to be served
  $[X_s, S_s] \leftarrow$ **OPTSET**$(X, W_c - Bandwidth(g_{(v,z)}, c))$
  // Calculate the service ratio if its subgroup with the lowest MCS is eliminated
  $[X_d, S_d] \leftarrow$ **OPTSET**$(X + (g_{(v,z)} - g_{(v,z,m_{Min})}), W_c)$
  if $((S_s + n_{(v,z)}) > S_d)$ **then**
    return $[(X_s + g_{(v,z)}), (S_s + n_{(v,z)})]$
  else
    return $[X_d, S_d]$
  end if
end if

Figure 4.4: Proposed function to find the optimal set of subgroups that maximizes the service ratio.

while our heuristic algorithm maintains near-optimal results ($< 1.38\%$ on average) with respect to the achieved service ratio.

4.6 Proposed Heuristic Algorithm

We propose a heuristic algorithm to solve the problem defined in Eq. (4.1). 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.

4.6.1 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 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 dynamically help in determining whether an SFN reconfiguration is required. This algorithm is illustrated in Figure 4.5. 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

Algorithm 4: 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

1: Sort M_c ascendingly based on their weights;
2: $v_m \leftarrow M_c.getHead()$; // Get the served video with the minimum weight
3: Sort U_c descendingly based on their weights;
4: $v_u \leftarrow U_c.getHead()$; // Get the unserved video with the maximum weight
5: **while** ($Weight(v_u, c) > Weight(v_m, c)$) or ($W > 0$ and $U_c = \emptyset$) **do**
6: **if** ($W > 0$ and $U_c \neq \emptyset$) **then**
7: // Case 1: expand an SFN area to include cell c as it is given in Figure 4.6
8: $[W, M_c] =$ **Expand(**A, c, v_u, W, M_c**)**;
9: **else if** ($W > 0$ and $U_c = \emptyset$) **then**
10: // Case 2: passively enroll cell c into an SFN area as it is given in Figure 4.7
11: $[W, M_c] =$ **Support(**A, c, W, M_c, λ**)**;
12: **else if** ($W = 0$ and $U_c \neq \emptyset$) **then**
13: // Case 3: replace the video v_m with video v_u as it is given in Figure 4.8
14: $[W, M_c] =$ **Replace(**$A_c, c, v_m, v_u, W, M_c, \alpha$**)**;
15: **end if**
16: $v_m \leftarrow M_c.getHead()$; // Get the served video with the minimum weight
17: $v_u \leftarrow U_c.getHead()$; // Get the unserved video with the maximum weight
18: **end while**

Figure 4.5: Proposed algorithm for reconfiguring an SFN.

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

Function Expand(A, c, v_u, W, M_c)

for $a \in A$ **do**
  // Calculate the gain in service ratio G_a resulted from adding cell c into SFN area a
  $G_a \mathrel{+}= Weight(v_u, x); \quad \forall\, x \in a$
  $G_a \mathrel{+}= Weight(v_u, c);$
  // Calculate the cost of signaling control C_a caused by adding cell c into SFN area a
  $C_a = C_{sys} + C_{init};$
  if $c \notin a$ **then**
    $C_a \mathrel{+}= C_{poll};$
  end if
end for
// Find the SFN area a_{opt} with the maximum ratio of service ratio gain to signaling control cost
$a_{opt} \leftarrow \max_{a \in A}(G_a / C_a);$
Expand cell c **into SFN area** a_{opt};
$W \mathrel{-}= Bandwidth(v_u, c);$ // Update the bandwidth available for multicasting over SFN
$M_c \mathrel{+}= v_u;$ // Add the video v_u into the set of served multicast streams in c
return $[W, M_c]$

Figure 4.6: Proposed function to find an SFN area in which a cell can enroll.

Function Support(A, c, W, M_c, λ)

for $a \in A$ **do**
  $v_a \leftarrow M_a.getHead();$ // Get the served video with the maximum weight in SFN area a
  // Calculate the gain in service ratio G_a resulted from adding cell c into SFN area a
  $G_a \mathrel{+}= Weight(v_a, x); \quad \forall\, x \in a$
  $G_a \mathrel{+}= Weight(v_a, c);$
  // Adjust this gain by considering the probability of having no outstanding traffic in cell c
  $G_a \mathrel{\times}= e^{-\lambda};$
  // Adjust this gain by considering the remaining time duration of video v_a
  $G_a \mathrel{\times}= T(v_a);$
  // Calculate the cost of signaling control C_a caused by adding cell c into SFN area a
  $C_a = C_{sys} + C_{init} + C_{poll};$
end for
// Find the SFN area a_{opt} with the maximum ratio of service ratio gain to signaling control cost
$a_{opt} \leftarrow \max_{a \in A}(G_a / C_a);$
Passively enroll cell c **into SFN area** a_{opt};
$W \mathrel{-}= Bandwidth(v_{a_{opt}}, c);$ // Update the bandwidth available for multicasting over SFN
$M_c \mathrel{+}= v_{a_{opt}};$ // Add the video v_u into the set of served multicast streams in c
return $[W, M_c]$

Figure 4.7: Proposed function to find an SFN area in which a cell can support.

Function Replace(A_c, c, v_m, v_u, W, M_c, α)

1: // Calculate the gain in service ratio G_u resulted from starting the video session v_u
2: $G_u \ += \ Weight(v_u, a); \quad \forall\, a \in A_c$
3: // Adjust this gain by considering the popularity of the video v_u
4: $G_u \ \times= \ 1/(v_u)^{\alpha}$;
5: // Adjust this gain by considering the remaining time duration of video v_u
6: $G_u \ \times= \ T(v_u)$;
7: // Calculate the cost of signaling control C_u caused by starting the video session v_u
8: $C_u \ = \ C_{sys} \ + \ C_{init} + \ C_{poll} + \ C_{stop}$;
9: // Calculate the gain in service ratio G_m resulted from continuing the video session v_m
10: $G_m \ += \ Weight(v_m, a) \quad \forall\, a \in A_c$;
11: // Adjust this gain by considering the popularity of the video v_m
12: $G_m \ \times= \ 1/(v_m)^{\alpha}$;
13: // Adjust this gain by considering the remaining time duration of video v_m
14: $G_m \ \times= \ T(v_m)$;
15: **if** $((G_u/C_u) > (G_M))$ **then**
16: **Replace video stream** v_m **with** v_u;
17: // Update the bandwidth still available for multicasting over SFN
18: $W \ += \ Bandwidth(v_m, c) - \ Bandwidth(v_u, c)$;
19: // Update the set of served multicast streams in c
20: $M_c \ += \ v_u - \ v_m$;
21: **else if** $((G_u/C_{stop}) > (G_M))$ **then**
22: **Shrink cell** c **from SFN area** a **in which** v_m **is multicasted**;
23: // Update the bandwidth still available for multicasting over SFN
24: $W \ += \ Bandwidth(v_m, c) - \ Bandwidth(v_u, c)$;
25: // Update the set of served multicast streams in c
26: $M_c \ += \ v_u - \ v_m$;
27: **end if**
28: **return** $[W, M_c]$

Figure 4.8: Proposed function to replace a video session with another video.

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 Eq. (4.1a).

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 4.2. 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 Eq. (4.1a). 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, 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, cell c is going to discard v_u and look for another outstanding stream.

4.6.2 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 2 seconds as in Microsoft Live Smooth Streaming [75]).

The proposed transmission scheduling algorithm is presented in Figure 4.9. 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 which 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 which 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 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 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, the process of removing a subgroup is repeated until a solution is found.

Algorithm 5: 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: $n_{(v,z)} \mathrel{+}= n_{(v,z,m)}$;
9: $g_{(v,z,m)} = n_{(v,z,m)} \times Weight(v_{(z,m)}, c)$; // Calculate the weight of this subgroup
10: $g_{(v,z)} \mathrel{+}= g_{(v,z,m)}$; // Merge this subgroup into its larger streaming group
11: **end for**
12: $X \mathrel{+}= g_{(v,z)}$; // Update the set of video segments to be served during this window
13: $W_r \mathrel{+}= Bandwidth(g_{(v,z)}, c)$; // 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: $W_r \mathrel{-}= Bandwidth(g_{(v,z)}, c)$; // Update the required bandwidth to serve video requests
19: $g_{(v,z)} \mathrel{-}= g_{(v,z,m_{Min})}$; // Eliminate the subgroup with lowest MCS
20: $X \mathrel{+}= g_{(v,z)}$; // Update the set of video segments to be served during this window
21: $W_r \mathrel{+}= Bandwidth(g_{(v,z)}, c)$; // Update the required bandwidth to serve video requests
22: **end while**
23: **return** X

Figure 4.9: 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 4.9 ensures the feasibility of the produced solution by satisfying the constraint in Eq. (4.1b). Moreover, in each iteration it removes the least profitable streaming ensuring that the algorithm is not trapped into 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 $O(N^2Log(N))$.

4.7 Evaluation

In this section, we present an extensive 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 [19, 67, 76] as well as the energy saving scheme introduced in [51]. 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.

4.7.1 Simulation Setup

Simulator and Algorithms: We have implemented simulator for mobile video streaming systems using OPNET modeler and its associated LTE Specialized module. To evaluate the proposed algorithms, we have also implemented the maximum throughput algorithm [19], proportional fair algorithm [76], combined unicast-multicast algorithm [67], and energy saving algorithm [51], 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.

Wireless Network Configuration: We use the LTE Release-12 standard to evaluate the performance of the proposed algorithms [81]. In Table 4.2, 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 [81, 106]. 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 [82]. 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 [2] 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, i.e., MCS 4, 8, 14, and 22, to support all possible channel qualities [2]. 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

Table 4.2: LTE Network Configurations.

Parameter	**Value**
Physical Profile	LTE 20 MHz FDD
Maximal Transmission Power	0.01 Watt
eNodeB Antenna Gain (dBi)	15 dBi
User Equipment Antenna Gain (dBi)	-1 dBi
Common Subframe Allocation (CSA) Period	8 frames
eMBMS Subframe Allocation per Frame	6 subframes (Max.)
Maximum Downlink Bit Rate	1736 Kbps
Modulation and Coding Scheme (MCS)	4, 8, 14, 22
Evolved Packet System Bearer for Uplink	Best Effort
Propagation Model	Urban Macrocell
Scheduling Mode	Link Adaptation
Mobility Model	Random Waypoint

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 transmission scheduling algorithm once every allocation window of 2 seconds. We set the cell size to be around 10 Km by 10 Km 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 (EPC) 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 which 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 72 km/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 (CQI) report to the associated base stations every 100 ms, 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 1/3 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 API, while the bit rate information is embedded in the video meta-data. If the bit rate is not embedded, we use the video length and size to calculate its average bit rate, in a way similar to the dataset in [29]. The video format is MPEG-4, and these videos are categorized in

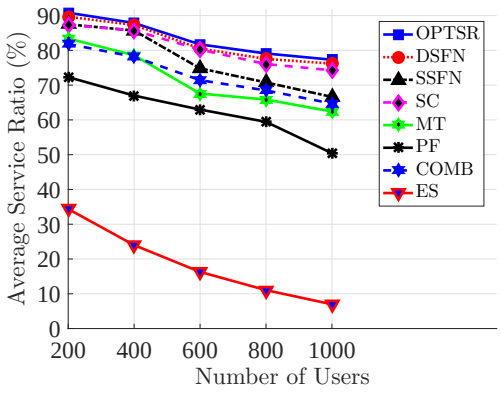
four resolution classes: 240p, 360p, 480p, and 720p (250 videos for each class), where each video belongs to a single resolution class (i.e., a non-adaptive video). We rank these videos based on the view count, and then we employ the Zipf distribution with a skewness factor α to assign synthetic popularity 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. Different 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.

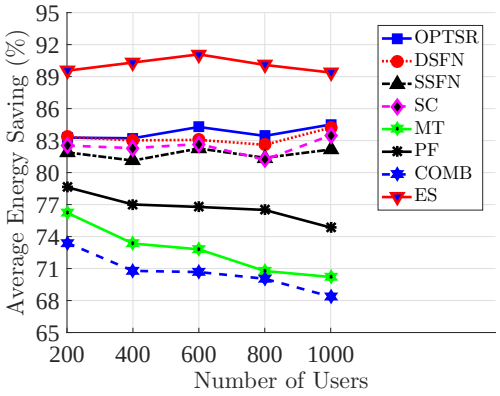
4.7.2 Comparison Against Current Algorithms

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 5 simulation runs in Figure 4.10. 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.

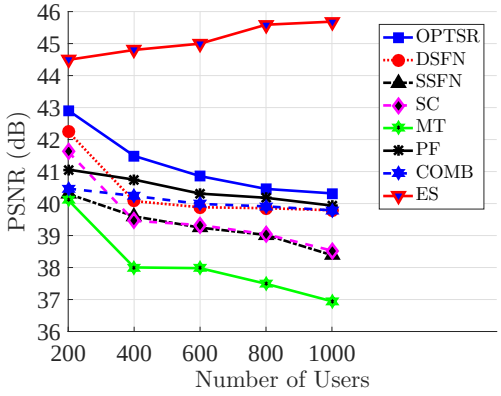
Service Ratio: Due to the limited radio resources in cellular networks, it may not be possible to serve all incoming video requests. For this reason, we estimate the service ratio by computing the fraction of served requests to number of received requests within the system. Figure 4.10(a) indicates that 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



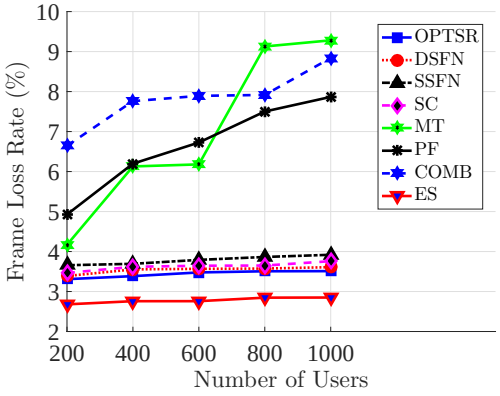
(a) Service Ratio



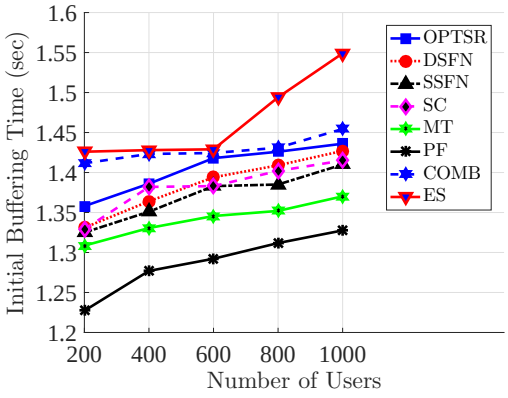
(b) Energy Saving



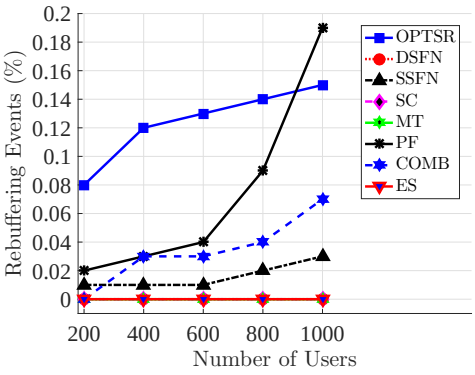
(c) Video Quality in PSNR



(d) Frame Loss Rate



(e) Initial Buffering Time



(f) Rate of Re-buffering Events

Figure 4.10: Comparisons of the achieved performance of the proposed algorithms against the state-of-the-art approaches.

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 4.10(a) that applying the concept of single frequency network improves the achieved service ratio by a significant gain. Figure 4.10(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 service ratio solution given by OPTSR, 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 [105]. 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 4.10(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 static SFN (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 4.10(c) and 4.10(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 42.24 dB 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 39.79 dB 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.

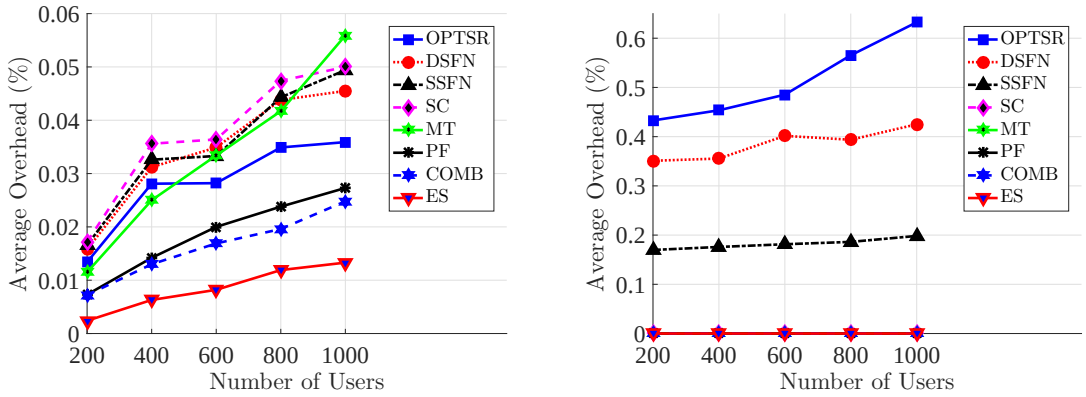
Initial Buffering Time: In video streaming systems, a playback starts after an initial buffering time and continues while the video is being downloaded. The initial buffering time in our algorithm depends mainly on the transmission scheduling window size. Intuitively, longer allocation windows provide more chances for expanding the multicast groups, thereby result in higher service ratios. Yet, larger allocation windows increase the initial buffering time. During our simulations, the window size is set to 2 seconds, which is equal to the size of video chunks produced by video streaming solutions, such as Microsoft Live Smooth Streaming [75]. At this window size, the initial buffering time is shown in Figure 4.10(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 4.10(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 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 4.10(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.

4.7.3 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 [19, 51, 67, 76] 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 [2], 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 (DRX) feature [2] for energy saving, not all users utilize the dedicated upload control channels all the time. Instead, the wide-band reports are sent by mobile terminals only when they receive video streams. We measure the overhead value as the fraction of bandwidth used to



(a) Channel Quality Reports

(b) SFN Control Signals

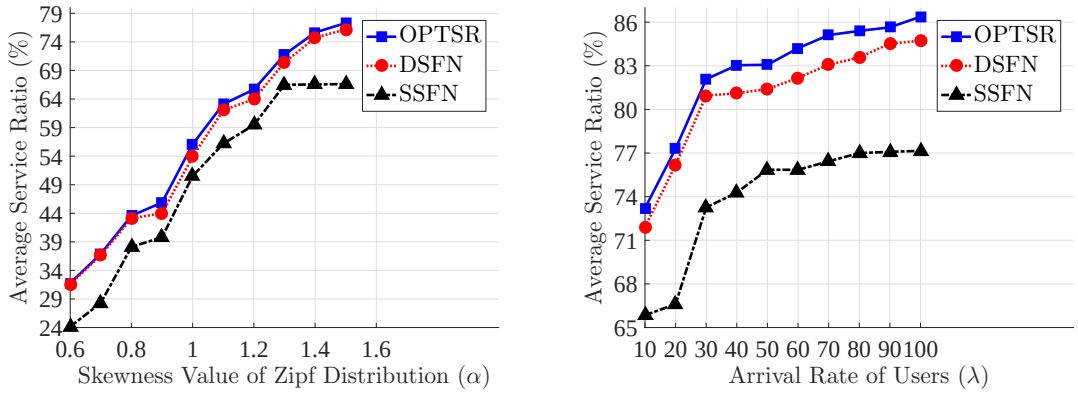
Figure 4.11: Overhead caused by the feedbacks sent to base stations.

send feedback reports to the total bandwidth available for both data and control transmission. Figure 4.11(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 4.11(b) presents the overhead caused by the SFN control signals in our algorithms during two types of configurations: dynamic (DSFN) and static (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 service ratio solution given by OPTSR, 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 OPTSR produces optimal results in terms of service ratios, but it does not consider the overheads during its calculation.

4.7.4 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



(a) Video Popularity Distribution

(b) User Arrival Rate

Figure 4.12: Impact of the user behavior model on the service ratio.

are considered: videos popularity and request arrival. To study the effect of video selection policy on the proposed transmission 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 4.12(a) reports the impact 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 4.12(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 4.12(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 number of mobile devices into multicast groups, which eventually results in higher service ratio. Figure 4.12(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 4.10(a), 4.12(a) and 4.12(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 4.12(a) and 4.12(b) point to a fundamental problem in the static deployment of SFN networks. Typically, the concept of 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 numbers 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 that the best possible bandwidth utilization is reached.

4.8 Summary

Traffic loads on mobile networks have dramatically increased during the recent decade, where a large portion of this traffic can be referred 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 chapter 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. Different from existing works, we do not assume a static configuration of single frequency networks. Instead, we presented optimal and heuristic algorithms which 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, our algorithms exploit 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.

Chapter 5

Adaptive Video Streaming over Heterogeneous Cellular Networks

In this chapter, we discuss video streaming over heterogeneous cellular networks and present how inter-cell interference can be a major challenge in such configurations. We formulate the transmission scheduling problem in these networks. We then propose a self-organized scheduling algorithm, in which each base station independently allocates its radio resources and adjusts its transmission powers such that interference among cells is reduced. To increase the average data rates at mobile terminals, the proposed algorithm also provides bandwidth estimation to terminals in order to help in facilitating their video rate adaptation processes. We evaluate the proposed algorithm using simulations and show its performance compared to the closest related work.

5.1 Introduction

Wireless cellular networks are currently deployed as homogenous networks, in which macrocells are placed based on a preplanned layout. The locations of these macrocells are carefully engineered, and their transmission parameters are configured to achieve target coverage and to control interference among cells. As the demand for higher data rate increases, mobile providers have begun deploying small cells to satisfy this increase in demand, enhance the network coverage at locations with poor signal reception, and offload some traffic from macrocells when possible. By reducing cell size, the spectral efficiency within its transmission coverage is increased through higher frequency reuse, while its transmit power can be minimized such that the power lost during propagation will be lower. Such solution has only been made possible in the recent few years as a result of the advances in hardware miniaturization and the reduction in their production cost. For example, several mobile providers, including Vodafone and Verizon, have started deploying hundreds of small cells, and the number of small cell shipments is forecast to rise from around 4 million to more than 10 million units between 2015 and 2020 [34].

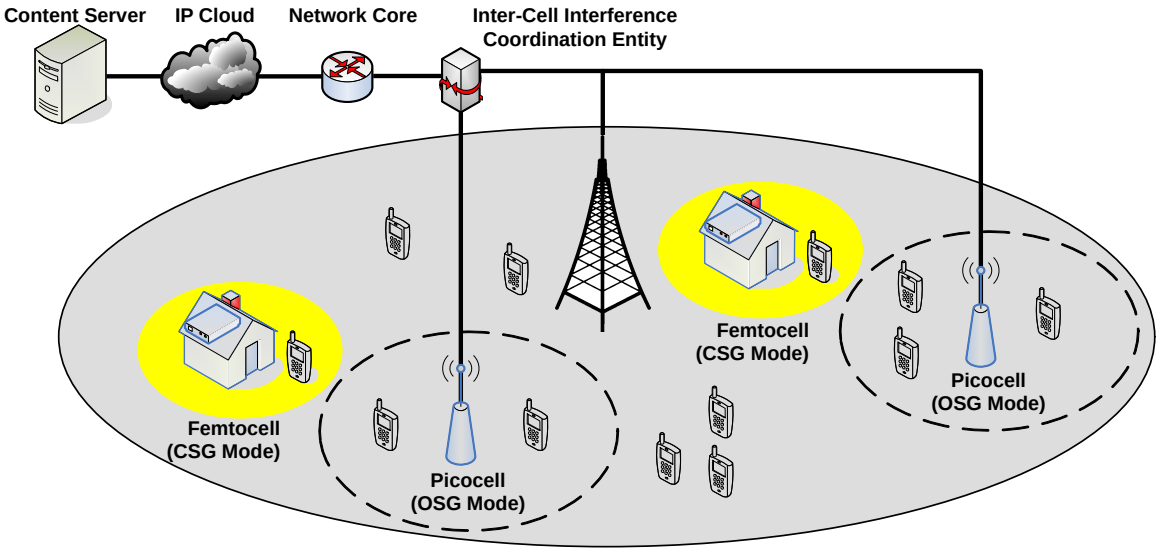


Figure 5.1: Example of heterogeneous mobile networks, in which different cell types share the same frequency and require careful management of inter-cell interference.

Figure 5.1 shows an example of a wireless cellular network consisting of a macrocell with high transmission power (i.e., 5 to 40 W) overlaid with several small cells with lower transmission power (i.e., 10 mW to 2 W) [84]. These small cells have different capabilities. For instance, femtocells are usually utilized in private and enterprise deployments, while picocells are used for wider outdoor deployments and for complementing the coverage of macrocells. This mixture of cells forms what is known as *heterogeneous network*. Since these small cells share the same carrier frequency band with their nearby macrocells, interference management is an important research problem in mobile heterogeneous networks. On the one hand, this type of frequency sharing might cause inter-cell interference between macrocells and small cells. On the other hand, small cells also share the same radio resources among themselves to improve their spectrum efficiency, which might cause another source of interference among small cells. Both types of interference significantly degrade the performance of the wireless network. The high level of inter-cell interference results in high packet drops, increases the block error rate (BLER), reduces the signal-to-noise ratio (SNR), and ultimately decreases the throughput of the whole network.

Small cells operate in three modes: open subscriber group (OSG), closed subscriber group (CSG), and hybrid of the two. The open subscriber group refers to heterogeneous configurations in which base stations are connected directly to the same (or at least cooperative) operator's network. The spectrum allocation here can be coordinated among these cells, and the interference can be managed through some resource partitioning in time or frequency domains. For example, an entity named inter-cell interference coordination (ICIC) has been introduced in LTE-A Release 8 to manage the interference in such OSG systems [2].

The closed subscriber group refers to heterogeneous configurations in which base stations from non-cooperative networks might interfere with each other. In these networks, the number and location of small cells are unknown by the mobile provider, and it will likely have no possible coordination among cells. Hence, cells should organize and manage their radio resources such that the impact of inter-cell interference is minimized. We consider the more general network model in which the hybrid mode of OSG and CSG is supported. We then propose an algorithm to dynamically group incoming video requests into unicast and multicast sessions in order to maximize the quality of served videos, reduce the interference among cells, and increase the number of served mobile terminals. The algorithm adjusts the transmission power for each subcarrier and allocates the available radio resource blocks to mobile terminals. In addition, to enhance the quality of experience for end users, we enable mobile terminals to dynamically and accurately adapt the video quality according to the current network conditions. Specifically, after determining the transmission scheduling decisions, base stations inform terminals about their allocated bandwidth. Therefore, mobile terminals will not need to estimate or guess the available bandwidth, which makes the rate adaptation process more accurate.

We first formulate the video transmission scheduling problem in such heterogeneous networks with the goal of minimizing the inter-cell interference and maximizing the average data rate of served video streams. We prove that this problem is NP-Complete, and we propose a heuristic algorithm to solve it. Through detailed packet-level simulations, we show that the proposed algorithm can achieve substantial improvements in strengthening the received signal at mobile terminals as well as increasing the overall data rate of video sessions. For example, our algorithm can yield at least 25% and up to 252% increase in the average signal strength, compared to the closest work in the literature. Furthermore, the video sessions in our algorithm are transmitted with an average data rate that is ten times higher than the average data rate achieved in the joint unicast-multicast approach [76]. Their average data rate is also up to 37% higher than average data rates in the fair-optimal policies in [27].

The rest of this chapter is organized as follows. Section 5.2 summarizes the related work in the literature, where Section 5.3 describes the considered system model. Section 5.4 states and formulates our problem, and Section 5.5 presents the proposed interference-aware group construction and adaptive bit rate allocation algorithms. Section 5.6 presents our simulation results and comparisons against other works in the literature. Finally, Section 5.7 concludes the chapter.

5.2 Related Work

The presence of inter-cell interference in heterogeneous networks poses an additional challenge for mobile terminals interested in receiving video streaming services. For instance, these terminals have a higher sensitivity toward the degradation in their channel conditions since it negatively impacts their achievable data rates as well as their quality of experiences. Thus, efficient interference management needs to be designed to mitigate the inter-cell interference problem without causing

an under-utilization of the available bandwidth. For this reason, an enhanced inter-cell interference coordination (eICIC) entity has been introduced to manage the interference among macrocells and small cells in mobile heterogeneous networks with open group subscribers [2]. Lopez-Perez et al. [70] evaluate both time and frequency domain techniques developed for eICIC in LTE Release 10. Deb et al. [37] improve the performance of eICIC and present an efficient solution to maximize the network utility and share the downlink radio resources between macrocells and picocells. In this solution, the authors utilize a time domain approach called Almost Blank Subframes (ABS) in which macrocells mute their signals and allow picocells to transmit their data and experience reduced inter-cell interference. The authors in [37] also let the association of mobile terminals be biased toward picocells using a Cell Selection Bias (CSB) technique. Singh et al. [91] address the same joint optimization of time domain resource partitioning and user association but with an objective of fair data rate allocation at mobile terminals.

Elsherif et al. [40] propose a radio resource allocation algorithm in heterogeneous networks to minimize inter-cell interference and then maximize the system throughput. Their algorithm relies on the concept of 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, Zhixue et al. [71] and Liang et al. [69] aim at reaching additional objectives of achieving fairness among mobile terminals and providing adequate quality of services. To achieve these two goals, a graph is constructed to represent the possible interferences between every pair of base stations, and then a vertex coloring approach is utilized to solve the problem of radio resources allocation within the network. Clustering is a partially decentralized approach in which base stations are grouped into a set of clusters. Each cluster operates independently on its own, so inter-cell interference can be significantly reduced. This approach is scalable to larger size of networks, whereas the complexity of its implementation remains reasonable. Hatoum et al. [47] introduce a frequency-domain resource allocation algorithm based on this concept. On the other hand, Argyriou et al. [20] and Zhou et al. [108] employ a time-domain resource partitioning to minimize the interference in a heterogeneous network in a way such that video quality is maximized.

Regarding the problem of adaptive streaming over cellular networks, Chen et al. [28] present a flow management framework for adaptive video delivery and demonstrate its efficiency regarding fairness, stability and resource utilization. Vleeschauwer et al. [100] and Joseph et al. [58] propose similar techniques to maximize the network throughput and quality of experience, respectively. Xie et al. [102] present a video adaptation algorithm to statistically estimate the available bandwidth based on some physical-layer sensing and monitoring approaches and then accordingly select the bit rates for the following video segments. Petrangeli et al. [78] propose an in-network system of coordinated proxies in order to facilitate a fair resource sharing among mobile terminals. On the other hand, Go et al. [44] and Abou-zeid et al. [3] investigate how predicting the bit rates of users can be exploited to offer an energy-efficient video streaming service. Based on such predictions, they develop an optimization framework for network-driven decisions to minimize the transmis-

sion period needed to deliver smooth video streaming experience, reduce the power consumption of base stations, and find a trade-off between both video quality and energy saving. Hamza and Hefeeda [46] present a two-step adaptation streaming approach in order to transmit interactive free-viewpoint videos to a set of heterogeneous clients, with the objective of decreasing the latency of view switching and increasing the quality of rendered virtual views. However, these works address the issue of adaptive video streaming over mobile networks without any consideration for the cell heterogeneity.

Although several algorithms have been proposed in the literature to solve the multicast scheduling problem, very few research efforts consider the more general hybrid approaches with both unicast and multicast [13, 15, 27, 76]. The closest works to our proposed algorithm can be found in [27, 76] since they employ a mixture of multicast and unicast. Monserrat et al. [76] utilize a joint unicast-multicast streaming approach, in which mobile terminals are served using unicast or multicast connections. If a multicast session is initiated, the base station will set its transmission power based on the worst channel condition to ensure accommodating terminals at the cell edge. Chen et al. [27] also apply a joint unicast-multicast streaming approach to maximize the average data rate at mobile terminals; however, they exploit the concept of multicast subgrouping to provide fair and optimal transmission scheduling decisions. Different from prior works, our proposed algorithm is designed to be implemented in cells with hybrid access mode (i.e., a combination of open and closed subscriber groups), which is more challenging because we do not assume full cooperation among cells. The allocation of radio resources and the transmission power of each subcarrier are dynamically adjusted to help cells organizing themselves and reduce the impact of inter-cell interference. The proposed algorithm also utilizes an adaptive video streaming approach such that the best quality representation is chosen for each video segment, without exceeding the available radio resource blocks.

5.3 System Model

We consider an OFDMA-based heterogeneous network of B cell types (based on the size of each cell) and N number of mobile terminals as shown in Figure 5.1. Symbols used in the chapter are listed in Table 5.1. We assume that this heterogeneous cellular network either has an independent video server or cooperates closely with a content delivery network. Thus, the network provider manages the video distribution process. We also introduce an adaptive bit rate allocation process, in which the base station is responsible for determining the assigned data rate of each mobile terminal. This process will then help mobile terminals to refine their selection of the available video quality representations, such that the overall average data rate is maximized. In adaptive streaming, a video stream is divided into a sequence of small segments. Video segments are pre-encoded in multiple quality representations, each with a particular video bit rate. We denote the segment quality level by q, where $q = 1, 2, \ldots, q_{max}$, and q_{max} is the maximum quality level. The function $r(v, q)$ maps

Table 5.1: Symbols used in this paper.

Symbol	Description
V	No. videos available at the content server
q	The quality representation level for a certain video
$r(v,q)$	The data rate of a video v using the quality level q
$n_{v,m}$	No. mobile terminals watching video v with m
p_{mi}	Power needed to transmit a certain data rate to
	the terminal i using MCS m
ξ_{is}	The gained signal-to-noise ratio at terminal i
	during the transmission over sub-carrier s
$x_{v,m,q}$	Whether video v is sent using MCS mode m
	and video representations q

the quality level of a video segment to the corresponding bit rate. Higher segment qualities require higher bit rates for successful reception, and therefore $r(v,q)$ is an increasing function of q.

The heterogeneous network in Figure 5.1 contains multiple macrocells, picocells and femtocells. These cells operate in two modes: open subscriber group and closed subscriber group. The control signals generated from the inter-cell interference coordination can be exchanged between cells in the open subscriber group. These control signals help cells in their power adjustment, user association and carrier sharing procedures. There is no coordination among cells when they operate in the closed subscriber group. We assume a general model, where the two modes can exist (i.e., the subscriber group is a hybrid combination of both open and closed modes). Therefore, in our proposed solution each base station self-organizes its radio resource allocation and independently runs its transmission scheduling process.

We consider a general video streaming model, which can be utilized for live and on-demand streaming services. Live streaming is useful in various scenarios, including streaming sports events, live concerts, political debates, and popular TV episodes. Live streaming is also suitable for multicast services as mobile terminals are mostly synchronized: they are watching at the same moment in the video, and functions that may disrupt this synchrony, e.g., pause and fast forward, are not applicable. Furthermore, live streaming of popular events typically attracts a large number of concurrent terminals, which can put a high traffic load on the cellular network. In contrast, mobile terminals in on-demand streaming services arrive asynchronously to the system. That is, terminals might 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 hard to achieve using pure multicast as a few terminals can form a multicast session, especially if they are requesting unpopular videos. 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.

We propose a transmission scheduling algorithm that runs on each base station. The outcome of the scheduling algorithm is the assignment of radio resource blocks among admitted mobile termi-

nals, along with the transmission power for each subcarrier. The transmission scheduling is updated on every new arrival/departure of mobile terminals and on any change in channel state information. The transmission scheduler periodically receives incoming video requests, divides mobile terminals into groups, determines the data rate for each group, and performs interference-aware radio resource allocation. In other words, the proposed transmission scheduling algorithm solves an optimization problem for unicast-multicast video streams to: (i) maximize the average assigned data rate across all mobile devices, (ii) reduce the interference among heterogeneous cells, (iii) minimize the network resources consumed by video streaming, and (iv) ensure smooth playback on all mobile devices. Upon the optimization problem is solved, the transmission scheduling determines which users belong to each sub-group and specifies the allocated bandwidth and the modulation and coding scheme for each video session.

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 $t \in [1, T]$. While each allocation window consists of TS resource blocks, we use d fraction of bandwidth, i.e., dTS resource blocks, for video services. At any instant, a mobile terminal $i \in [1, N]$ is assumed to be within the service range of $b \geq 1$ number of candidate cells. The mobile terminal is informed about which sub-frames and sub-carriers are assigned to its video stream via a control channel broadcasted from its associated base station, and the allocation decision can be changed dynamically at specified intervals.

5.4 Problem Definition

The problem we address is efficiently transmitting video streams over heterogeneous networks, where inter-cell interference may exist. Inter-cell interference is an essential challenge especially with the deployment of small cells, where wireless operators have minor or no control on the locations of these small cells. The concurrent operation of small and traditional cells produces irregularly shaped cell sizes, and thus causing destructive interference. Therefore, we need to design algorithms to dynamically allocate radio resources to mobile terminals in a way that reduces the impact of such interference. Our video transmission scheduling over heterogeneous network problem can be stated as follows:

Sub-Problem 4 (Video Transmission over Heterogeneous Cell Networks)**.** *Given an allocation window of T sub-frames and S sub-carriers in a heterogeneous network, determine the optimal transmission scheduling for each base station that assigns the available resource blocks to hybrid unicast-multicast multimedia sessions, decides the number of multicast groups, and determines the data rate for each mobile terminal such that the average video quality-level for all mobile terminals is maximized.*

$$\max_{\mathbf{X}} \quad \gamma = \frac{1}{N} \sum_{v=1}^{V} \sum_{m=1}^{M} n_{vm} \sum_{s=1}^{S} \sum_{t=1}^{T} x_{vmst} r_m \tag{5.2a}$$

$$\text{s.t.} \quad \sum_{v=1}^{V} \sum_{m=1}^{M} \sum_{s=1}^{S} \sum_{t=1}^{T} x_{vmst} \lceil \frac{\Gamma r_m}{c_m} \rceil \leq dTS \tag{5.2b}$$

$$\sum_{v=1}^{V} \sum_{m=1}^{M} x_{vmst} \leq 1 \tag{5.2c}$$

$$\sum_{s=1}^{S} \sum_{t=1}^{T} p_{st} \leq P \tag{5.2d}$$

$$p_{st} \leq P_s \tag{5.2e}$$

$$x_{vmst} \in \{0, 1\}, \ \forall \ v \in [1, V], m \in [1, M], s \in [1, S], t \in [1, T].$$

We formulate the transmission scheduling problem, whose objective is maximizing the average data rate at mobile terminals and reducing the inter-cell interference among cells. We use the Boolean decision variable x_{vmst} ($v \in [1, V]$, $m \in [1, M]$, $s \in [1, S]$, $t \in [1, T]$) to denote whether the resource block (s, t) of the base station is allocated to transmit video segment v using the MCS mode m in the current scheduling window. That is, $x_{vmst} = 1$ if video v is transmitted with the MCS mode m, and $x_{vmst} = 0$ otherwise. Recall that n_{vm} denotes the number of mobile terminals receiving video segment v using MCS mode m. Therefore, when $n_{vm} = 1$, the base stations stream video v using unicast; and when $n_{vm} > 1$, the base stations stream video v using multicast. When $x_{vmst} = 0$, mobile terminals with maximum MCS mode m receive v with the next *lower* MCS mode $m \in [1, M]$ that is available in the solution.

We define r_m to be the data rate transmitted over (s, t) resource block using one of the m modulation and coding schemes. Given the required power p_{mi} to transmit to mobile terminal i, the data rate r_m in the resource block (s, t) can be given as [60]:

$$r_m = ln(1 + \delta p_{mi} \xi_{ist}), \tag{5.1}$$

where ξ_{ist} is the signal-to-noise ratio (SNR) at the mobile terminal i during the transmission over resource block (s, t), assuming the transmission power to be unity and $\delta = \frac{1.5}{-ln(5 \times BER)}$ is a constant based on the bit error rate (BER) needed to achieve a target quality level. The noise measured in an SNR value is a combination of all unwanted signal sources, including both interfering radio frequencies and signal distortion sources. We also assume two constraints for the power allocation, P and P_s, which represent the maximum power for a base station and the maximum power for an individual sub-carrier s, respectively.

We present the formulation in Eq. (5.2). The objective function in Eq. (5.2a) is to choose the set X of unicast-multicast video sessions that maximize the average data rate within a cell. The total data rate of video v in a scheduling window is r_m, and the number of resource blocks needed is $\lceil \frac{\Gamma r_m}{c_m} \rceil$, where m is the chosen modulation and coding scheme. The first two summations iterate

through all videos and MCS modes, respectively, while the last two summations iterate through the radio resources of a cell. The constraint in Eq. (5.2b) ensures that the video streaming service only consumes up to d fraction of network resources. Eq. (5.2c) assures that each resource block of a base station is assigned to a single video stream, whereas the conditions in Eq. (5.2d) and Eq. (5.2e) implement the power allocation constraints. The following lemma states the hardness of our transmission scheduling problem for video streaming services over heterogeneous networks.

Lemma 4 (Hardness)**.** *The video transmission scheduling over heterogeneous network problem is NP-Complete.*

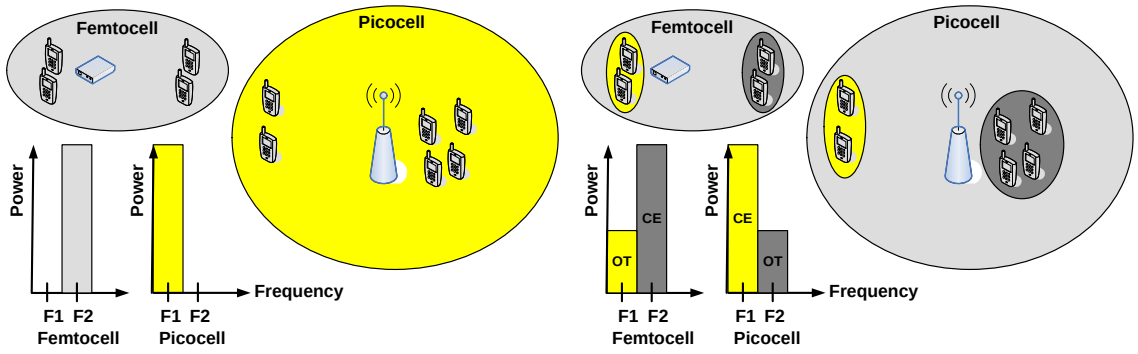
Proof. We reduce the 0-1 knapsack problem to Problem 4. In the 0-1 knapsack problem, we consider O objects, where object o ($1 \leq o \leq O$) has a weight θ_o and a value ϕ_o. The problem is to select a subset of objects for maximizing the total value without exceeding the weight limit $\hat{\theta}$. Given a 0-1 knapsack problem, we generate a corresponding problem instance as follows. For each object o, we create a new video session, and we: (i) add ϕ_o mobile terminals to that video session, and (ii) set the per-block capacity to be proportional to the weight θ_o. Last, we set the dTS value based on the weight limit $\hat{\theta}$. This results in a proper instance of Problem 4 in polynomial time. In addition, a solution to Problem 4 can easily be verified in polynomial time. Therefore, Problem 4 is NP-Complete. □

5.5 Proposed Algorithm

The proposed transmission scheduling algorithm runs in each base station to independently organize its radio resources, adjust the transmission power for every subcarrier, and determine the data rate for each video session. The formulation in Eq. (5.2) is a *binary integer programming* problem, which can be solved by optimization solvers, such as CPLEX [33] and GLPK [43]. Although the optimal solution gives optimum resource allocation, the running time of its worst cases is exponential. Therefore, we develop a heuristic algorithm consisting of two steps: interference-aware group construction and adaptive bit rate allocation. We call the proposed algorithm as adaptive video streaming over heterogeneous cellular networks, and we refer to it shortly as ASH.

5.5.1 Interference-aware Group Construction

A simple transmission scheduling procedure is to serve all members in a multicast group with the same data rate. This data rate is set based on the transmission power required by the terminal with the weakest link, which may lead to increasing the inter-cell interference and downgrading the overall spectral efficiency. To address such drawback, we propose a grouping approach in which mobile terminals are split into multiple unicast-multicast subgroups. Each subgroup is then served by its associated base station using a transmission power and a modulation and coding scheme, with which its worst mobile terminal can decode.



(a) Different Frequency Subcarrier Approach (b) Interference-aware Group Construction Approach

Figure 5.2: Examples of various radio resource allocation approaches.

Figure 5.2 illustrates the grouping concept and how it can assist heterogeneous cells to decrease the impact of inter-cell interferences. In this example, picocell and femtocell utilize the feedback measurements from the associated mobile terminals about their channel quality conditions. Based on these feedback measurements, the two cells independently divide their users into unicast-multicast subgroups, choose the suitable resource blocks for each subgroup, and adjust the transmission power of each subcarrier. This independent organization helps in reducing the negative impact of interfering signals among the two cells. For instance, each of the picocell and femtocell in Figure 5.2 have two subcarriers to serve their mobile terminals. Each cell can initiate a multicast session on one subcarrier, say F_1 for femtocell and F_2 for picocell, as shown in Figure 5.2(a). While this approach helps in avoiding inter-cell interference with the other cell, both cells would have to transmit using a coding and modulation scheme and transmission power suitable for their mobile terminals with the worst channel quality conditions. Thus, following this approach is inefficient since terminals in the same multicast group are likely to have different data rate and quality requirements. The proposed solution overcomes such inefficiency, as shown in Figure 5.2(b), by allowing the two cells to split their multicast sessions into two subgroups: one subgroup for terminals at the cell edge denoted by CE, and another subgroup for the other mobile terminals denoted by OT. These two subgroups are then served by various data rates with the objective of increasing the average data rate at all mobile terminals. Our solution also mitigates the inter-cell interfence because the two cells choose different subcarriers for the subgroups at the cell edge since they are delivered using higher transmission powers.

Utilizing the concept of grouping improves the system throughput without necessitating any additional control overhead on the network. There might be a slight increase in the bandwidth usage since a multicast session can be concurrently transmitted multiple times; however, this slight increase can be justified for the sake of improving both achieved average data rate and spectral efficiency, as illustrated in Figure 5.2(b). The main issue is how such grouping can be constructed efficiently. To tackle this issue, we propose an interference-aware grouping algorithm that can run

Algorithm 6: Interference-aware Group Construction

Inputs: $\{V, R\} \leftarrow$ Set of requested videos and their data rates
$M \leftarrow$ Set of available modulation and coding schemes
$W \leftarrow$ Bandwidth still available for video services
Output : $X \leftarrow$ Set of groups to be served in this scheduling window, along with their transmission power and chosen subcarriers

$W_r = 0$; // Initialize the required bandwidth to serve incoming video requests
$X = \{\emptyset\}$; // Initialize the set of groups to be served during this allocation window
$\tau = 0$; // Initialize the total utility of the set of groups to be served
// Perform the minimum grouping stage to maximize the number of served terminals
for each video v **do**
  $n_v =$ number of terminals interested in receiving this video;
  $x_v = 1$; // Create the group of users interested in this video
  $s_{x_v} =$ **Subcarrier**(x_v); // Select the most suitable subcarrier to serve group x_v
  $m_w =$ **Power**(s_v, x_v); // Adjust the transmission power needed to transmit group x_v
  $\tau_{vm_w} = \alpha_{vm_w}/\beta_{vm_w}$; // Calculate the utility of this group based on Eqs. (5.4) and (5.5)
  $\tau \mathrel{+}= \tau_{vm_w}$; // Update the total utility of the set of groups to be served
  $W_r \mathrel{+}=$ **Bandwidth**(v, m_w); // Update the required bandwidth to serve video requests
  $X = X \cup [x_v, s_{x_v}, m_w]$; // Update the set of video groups to be served during this window
end for
// Perform the quality-aware grouping stage to increase the achievable average data rate
while $(W_r < W)$ **do**
  for each video subgroup $x_{vm} \in X$ **do**
    $[m_p, \tau_{vm_p}] \leftarrow$ **Partition**(v, m); // Divide to 2 subgroups if it increases group utility
  end for
  $\tau_{\hat{v}\hat{m}_p} \leftarrow$ **FindSubgroup**(X, τ_{vm_p}); // Select the subgroup which maximizes the total utility
  $x_{\hat{v}\hat{m}_p} = 1$; // Create a new subgroup to transmit $\hat{v}$ using MCS $\hat{m}_p$
  $s_{x_{\hat{v}\hat{m}_p}} =$ **Subcarrier**$(x_{\hat{v}\hat{m}_p})$; // Select the most suitable subcarrier to serve group $x_{\hat{v}\hat{m}_p}$
  $\hat{m}_p =$ **Power**$(s_{x_{\hat{v}\hat{m}_p}}, x_{\hat{v}\hat{m}_p})$; // Adjust the transmission power to transmit group $x_{\hat{v}\hat{m}_p}$
  $\tau \mathrel{+}= \tau_{\hat{v}\hat{m}_p}$; // Update the total utility of the set of groups to be served
  $W_r \mathrel{+}=$ **Bandwidth**$(\hat{v}, \hat{m}_p)$; // Update the required bandwidth to serve video requests
  $X = X \cup [x_{\hat{v}\hat{m}_p}, s_{x_{\hat{v}\hat{m}_p}}, \hat{m}_p]$; // Update the set of groups to be served during this window
end while
return X

Figure 5.3: Proposed group construction algorithm.

in real time. In every scheduling window, our algorithm divides incoming video requests into subgroups, assigns radio resources to each subgroup, and adjusts transmission power in each subcarrier.

Figure 5.3 shows the pseudo code of our algorithm. After initializing various variables, the algorithm starts with a *minimum grouping stage*, which utilizes a minimum number of unicast-multicast groups in order to maximize the number of served terminals in the system. The algorithm then

performs a *quality-aware grouping stage*. At every iteration of this grouping stage, the algorithm increases the number of served groups, and it searches for the most suitable subgroup configuration that strengthens the received signal at terminals, thereby increasing the chances of achieving higher average data rate. These iterations terminate when there is no further improvement in the objective function given by Eq. (5.2a).

During the minimum grouping stage (i.e., Lines 2 to 20), we start our solution from an intuitive decision in which the number of groups is set to the number of requested video segments. This can be accomplished by setting up a unicast/multicast channel to each video segment with a transmission power suitable for its terminal with worst channel condition. For each video request v, our algorithm utilizes the function **Subcarrier** to allocate this group to its most suitable subcarriers (i.e., the SNR values ξ_v on these subcarriers are higher than other available subcarriers). Mathematically, we can define this process as:

$$\xi_v = \max_{i \in n_v} \min_{s \in S} \xi_{si} \tag{5.3}$$

Following such approach will mitigate the inter-cell interference among cells and increase the average data rate of transmitted video streams. When the subcarriers for video v are chosen, our algorithm uses the function **Power** to determine the transmission power needed to serve its mobile terminals. The power adjustment here is set based on the basis of the assigned modulation and coding rate and according to the predetermined link adaptation curve [2]. Our algorithm also utilizes the function **Bandwidth** to calculate the number of radio resource blocks required to serve each group. This can be done by calculating the capacity of each chosen subcarrier at the assigned modulation and coding scheme. Once the minimum grouping stage is performed, our algorithm calculates the value of objective function denoted by γ_1 that is related to the current group construction. We indicate this group construction with X_1.

In the quality-aware grouping stage (i.e., Lines 21 to 36), our algorithm evaluates if there exists a group configuration formed by two subgroups and whether it increases the value of the objective function compared to the previous iteration. An important issue here is related to the group partitioning process and distribution of the available resource blocks among the enabled groups. Techniques such as random group partitioning initiate several inefficiencies. For this reason, our algorithm utilizes a function called **Partition** to divide each group $x \in X_1$ into two subgroups in a way that increases the average data rate within the cell. The function **Partition** define α_{vm} and β_{vm} to be the expected gain in the received signal strength using the chosen subcarriers and the consumed amount of the available bandwidth, respectively, after dividing a group into two subgroups. These two parameters are used to balance between the gain in average data rate and the cost in

consumed radio resources. Mathematically, we write:

$$\alpha_{vm} = n_{vm} \frac{|\xi_m - \xi_{min}|}{|\xi_{max} - \xi_{min}|}, \tag{5.4}$$

$$\beta_{vm} = \lceil \frac{\Gamma r_m}{c_m} \rceil, \tag{5.5}$$

where ξ_m represents the average SNR at the subcarriers chosen for the new subgroup, ξ_{max} defines the upper threshold at which an increase in the SNR value would lead to no gain in the throughput, and ξ_{min} specifies the lower threshold at which a decline in the SNR value would provide intolerable block error rate.

Each candidate group $x \in X$ is examined in Lines 24 to 27 to observe the gain on the objective function when another subgroup, associated with the j^{th} MCS mode, is enabled in addition to the current group x. The additional enabled subgroup should be suitable to an MCS mode among those feasible ones not already enabled for group x. Since $j \in M$, the function **Partition** examines at most $(\|M\|-1)$ configurations for each group. Our algorithm then tries to maximize the utility of its transmission scheduling by finding the subgroup with maximum possible gain (objective function) and the minimum bandwidth consumption (constraint). In particular, our algorithm in Line 29 uses the function **FindSubgroup** to evaluate the utility $\tau_{vm} = \alpha_{vm}/\beta_{vm}$ of all $x_{vm} = 1$ and divides in each iteration the subgroup x with the highest τ value after partitioning. The algorithm repeats the grouping process in the next iterations, and it stops once the constraint in Eq. (5.2b) is violated.

The proposed algorithm has an overall complexity of $O(NM^2)$, where M is the maximum number of MCS modes, and N is the number of mobile terminals in the cell. We note that in practice, M is ≤ 28 and N is in the order of hundreds [2]. Thus, our algorithm is suitable for real implementations. It is worth underlining that the computational cost of the proposed algorithm, unlike the optimal solution, does not depend on the number of resource blocks reserved for the video services dTS, which is pseudo-polynomial and potentially exponential in the length of the input (i.e., the number of bits required to represent dTS).

5.5.2 Adaptive Bit Rate Allocation

In adaptive streaming over HTTP, each video is divided into segments with short lengths (usually between 2 and 20 seconds [64]). Every video segment is then encoded at multiple bit rates. These bit rates are affected by the video resolution, quantization level, and frame rate. Bit rates are also impacted by the video content itself such as its motion and structure. Prior to the beginning of a streaming process, the client downloads a manifest file for its requested video, in which the versions of qualities for each segment are specified. Once the manifest file is retrieved, the video application at the client requests segments according to the available bandwidth.

However, it has been demonstrated that enabling mobile terminals to be fully responsible of requesting video segments may lead to inefficient usage utilization of the network resources, unfair allocation of the wireless bandwidth, and large variations in the received video quality. [9,10,73]. In

Algorithm 7: Adaptive Bit Rate Allocation

Inputs: $X \leftarrow$ Set of groups to be served in the current window
$q_v \leftarrow$ Set of available quality representations for video v
$W \leftarrow$ Bandwidth available for video services
Output : $\Phi \leftarrow$ Set of assigned bit rates for video segments

$W_r = 0$; // Initialize the required bandwidth to serve video groups
$\Phi = \{\emptyset\}$; // Initialize the set of assigned bit rates for video groups
// Prioritizing the video groups based on their per-bit efficiency rates
for each served group $x_{vm} \in X$ **do**
  $n_{vm} =$ number of terminals receiving v using MCS m;
  $q_{vm} = q_l$; // x_{vm} is enabled to receive the lowest quality representation
  $\Phi = \Phi \cup q_{vm}$; // Update the set of assigned bit rates for video groups
  W_r += **Bandwidth**(q_{vm}, m); // Update the required bandwidth to serve video groups
  $u_{vm} = n_{vm}/$**Bandwidth**(q_{vm}, m); // Calculate the per-bit efficiency of this group
end for
// Increasing the assigned bit rate for video groups
while $(W_r < W)$ **do**
  Sort X descendingly based on their per-bit efficiency ratios u_{vm};
  $x_{vm} \leftarrow X$.**getHead**(); // Get the group with the maximum per-bit efficiency
  W_r $-= Bandwidth(q_{vm}, m)$; // Update the required bandwidth to serve videos
  $\Phi[q_{vm}]$ $+= 1$; // x_{vm} is enabled to receive a higher quality representation
  W_r $+= Bandwidth(q_{vm}, m)$; // Update the required bandwidth to serve video groups
  if $(q_{vm} \geq q_h)$ **then**
    $u_{vm} = 0$; // It has been enabled to receive the highest representation
  else
    $u_{vm} = n_{vm}/$**Bandwidth**(q_{vm}, m); // Calculate the per-bit efficiency of this subgroup
  end if
end while
return Φ

Figure 5.4: Proposed adaptive bit rate allocation algorithm.

addition, when multiple mobile terminals request video segments around the same time, it is likely that they incorrectly estimate the available bandwidth within a cell [9, 10, 73]. Wrong estimation might also occur because of the continuous variation in the channel quality conditions of these terminals. Then it subsequently affects the process of data rate selection and negatively impacts the overall quality of experience. To overcome this problem, we propose a novel rate adaptation mechanism to improve the delivered video quality to mobile terminals. This goal is reached with the aid of their associated base station, which is in charge of collecting feedback measurements on the channel quality conditions. Once the base station determines the assigned data bit rate for each mobile terminal, this information is utilized by the terminals in order to refine their selection of the video quality representations.

The proposed adaptive bit rate allocation algorithm is presented in Figure 5.4. Each group is given a per-bit efficiency rate which is determined by two parameters: 1) the number of terminals that receive a particular video at a certain transmission power, and 2) the number of resource blocks needed to transmit this video segment. Dividing the former by the latter parameter gives an efficiency rate which is used in prioritizing the video groups and then determining which set of them are chosen to get higher quality representations during the current scheduling window. The number of available resource blocks is usually limited, so the algorithm aims at selecting the set of groups that increases the average data bit rate, and thereby enhancing the overall quality of experience. To achieve such objective, our proposed algorithm selects the group with the highest efficiency rate and then increases its assigned bandwidth to match the bit rate of the next video quality representation level. Once the assigned bit rate of this group is increased, its efficiency rate should also be recalculated. The scheduler tries again to increase the average bit rate of the groups in hand. If the bandwidth is still enough, the process of increasing the assigned bit rates is repeated until a final solution is found.

The proposed adaptive bit rate allocation algorithm terminates in polynomial time: $O(q|X|^2 log(|X|))$, where $|X|$ is the number of video groups to be served in the cell, and q is the number of quality representation levels available at the content server. The while loop in Figure 5.4 ensures the feasibility of the produced solution by satisfying the constraint in Eq. (5.2b). Moreover, in each iteration, it increases the estimated bit rate of the video group with the highest per-bit efficiency rate, thereby ensuring that the algorithm is not trapped into an infinite loop. The dominating computational complexity of the algorithm occurs in the while loop: (i) the while-loop iterates at most $q|X|$ times, and (ii) sorting the per-bit efficiency rate queue consumes $|X|log(|X|)$. Combining these two computational complexities, the time complexity of the proposed algorithm is $O(q|X|^2 log(|X|))$.

5.6 Evaluation

In this section, we present an extensive simulation performed in OPNET to evaluate the proposed transmission scheduling algorithm (ASH). To compare its performance, we have also implemented the closest algorithms in the literature, which are [27, 76]. The algorithm introduced in [76] utilizes a joint unicast-multicast streaming approach, and we refer to it as JUM. The algorithm in [27] also applies a joint unicast-multicast streaming approach, but it exploits the concept of multicast subgrouping to provide fair and optimal transmission scheduling decisions. We refer to this algorithm as FO. Since the authors of [27] suggested two weighting functions for the multicast sessions, we have implemented both of them: linear and constant weighting functions (FO-Constant and FO-Linear, respectively). We compare the results obtained by our algorithm with these three methods from different perspectives, and we demonstrate that our algorithm significantly outperforms them.

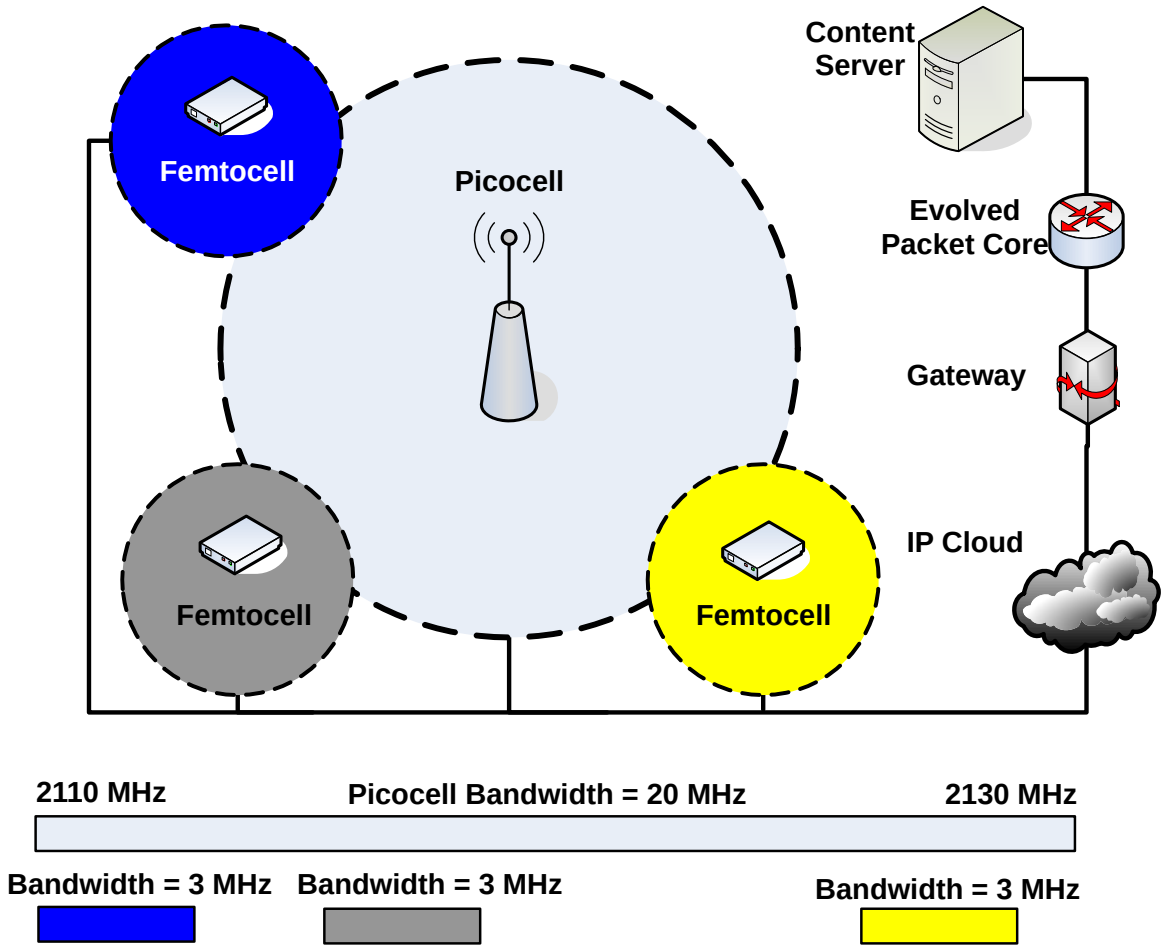


Figure 5.5: The simulation setup for a heterogeneous network.

5.6.1 Wireless Network Configuration

We simulate an LTE network in which a picocell base station is surrounded by three femtocell base stations as shown in Figure 5.5. The picocell is configured with 20 MHz FDD physical profile, which is based on OFDMA in downlink access and SC-FDMA in uplink access. Then we modify the default transmission scheduling algorithm of this profile and replace it with our proposed algorithm. The base frequency of the picocell downlink channel is set to 2110 MHz. The three femtocells are configured with 3 MHz FDD physical profile, and their default transmission scheduling algorithms are modified to broadcast using their entire resource blocks during the simulation time. To act as interfering signal jammers with the picocell eNodeB, the three femtocell base frequency channels are set to be 2112 MHz, 2116 MHz, and 2126 MHz, respectively.

We configure the LTE downlink with Evolved Packet System (EPS) bearers. An EPS bearer is defined as a transmission path of specific quality, capacity, and delay [82]. The simulator runs the proposed data transmission scheduling algorithm once every allocation window of 2 seconds. We

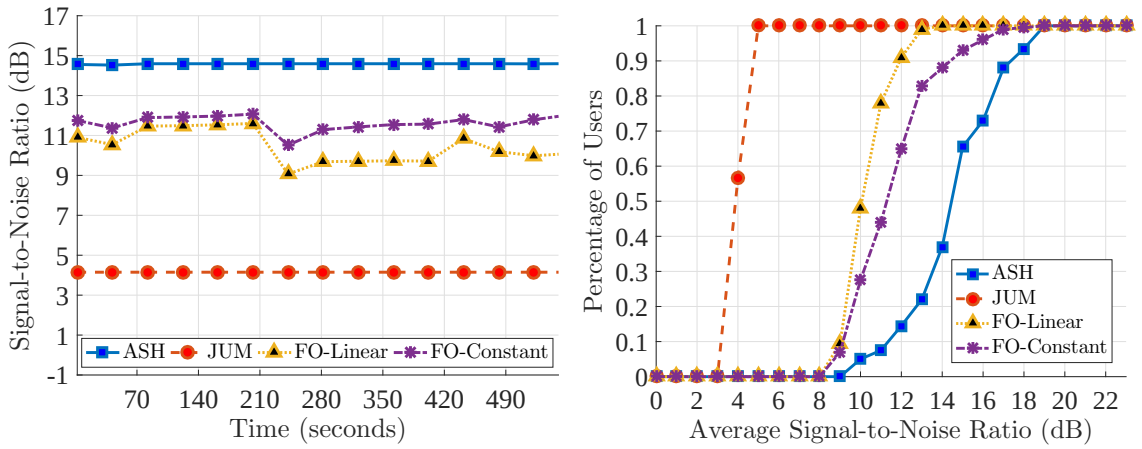
set the radius of the picocell eNodeB to be around 4 Km, whereas it is set to 800 m for the femtocell cells by controlling the power of their base stations.

We consider a set of 10 videos. The quality representations of each video are categorized in five data rate classes: 0.400, 0.750, 1.0, 2.5, and 4.5 Mbps. We assume a population of 200 users joining the system following a Poisson process with mean arrival rate λ. λ is a simulation parameter which we set to 20 users per second by default for our simulations. We chose 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 1.8 km/hr. This mobility model stresses our algorithms since it is difficult to predict the path of receivers and plan ahead of time. We configure the mobile devices to send a channel quality indicator (CQI) report to the associated base stations every 100 ms, which allows the base stations to determine the MCS mode depending on the channel condition. We chose 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 populated within $1/3$ 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.

5.6.2 Comparison Against Current Algorithms

We compare our proposed algorithm versus three approaches (i.e., JUM, FO with constant weight allocation, and FO with linear weight allocation). Five performance metrics are utilized in this comparison: 1) signal-to-noise ratio of the received signals at mobile terminals, 2) spectral efficiency gained from the transmission scheduling decisions, 3) average bit rate allocated for video streams, 4) number of quality switches between video representations during the simulation time, and 5) average packet loss rate in the system. We report the mean results from 5 simulation runs in Figures 5.6 – 5.9. Collectively, these results show that our proposed algorithm not only outperforms others with significant margins on the achieved signal strength at mobile terminals but also increases the average data bit rate without causing any significant quality switching among video representations. The simulation results are discussed below.

Signal-to-Noise Ratio: The signal-to-noise ratio indicates the signal quality at a mobile terminal by dividing the strength of the received signal by the strength of any interference. It is measured in dB, and its value range depends on the cellular technology being used. In LTE networks, an SNR value $\xi \geq 13$ is considered as a good signal, while an SNR value $\xi \leq 0$ is deemed to be a weak signal [96]. Figure 5.6(a) shows the average signal-to-noise ratio over time of the proposed algorithm. As shown in the figure, our algorithm remains consistent and gives an average of 14.6 dB in the strength gain of its transmitted signal, while JUM, FO-Linear, and FO-Constant present an approximate average of 4.2 dB, 10.43 dB, and 11.65 dB, respectively. In other words, our algorithm outperforms its closest works by yielding at least 25% and up to 252% increase in its average achieved SNR value. We plot the cumulative distribution function (CDF) in Figure 5.6(b), which



(a) Average Over Time

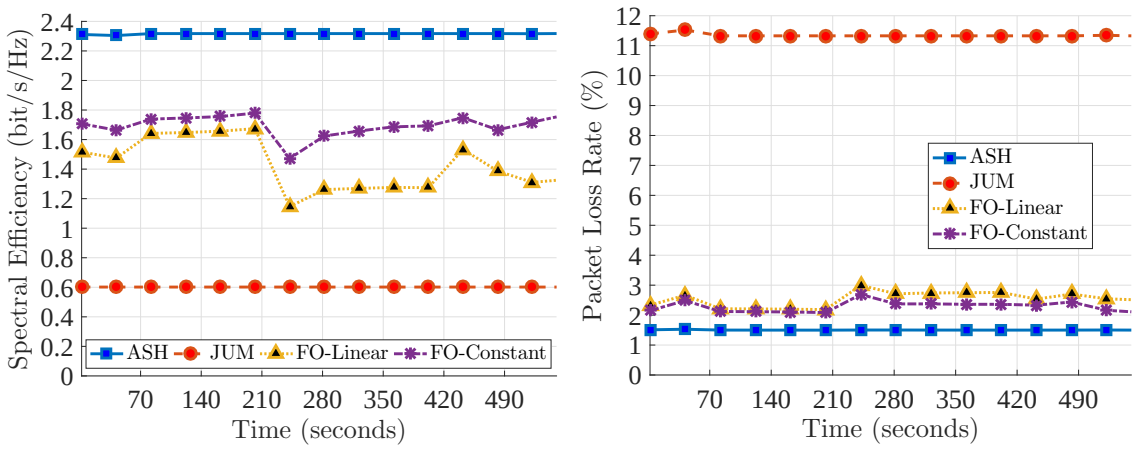
(b) Cumulative Distribution

Figure 5.6: Comparisons of the achieved signal-to-noise ratio of the proposed algorithm against the closest related work.

shows the average SNR value on the x-axis and the cumulative percentage of mobile terminals that achieved that SNR value on the y-axis. For example, about 78% of mobile terminals achieved 13 dB or higher (i.e., good signal [96]) using our algorithm, while about 17%, 1%, 0% of mobile terminals achieved that SNR value using FO-Constant, FO-Linear, and JUM algorithms, respectively.

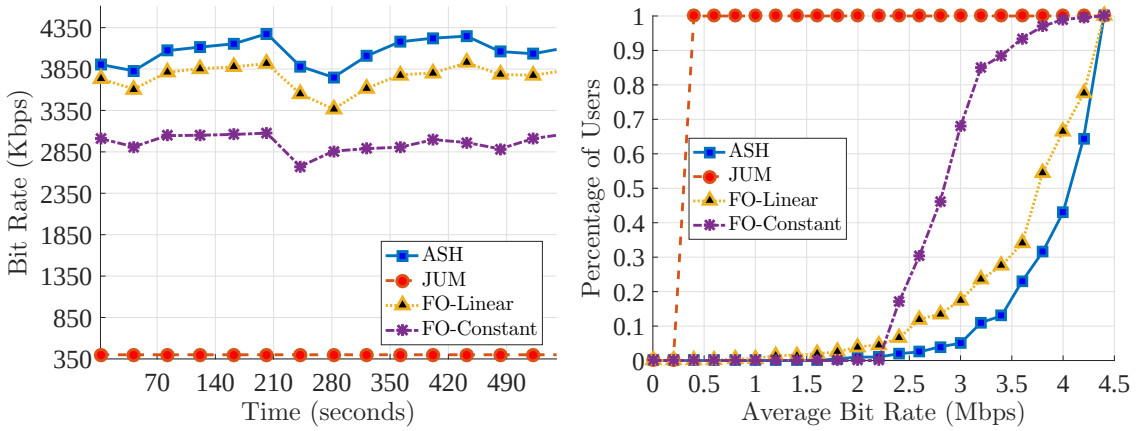
Spectral Efficiency: The spectral efficiency is defined as the transmitted data rate (in bits per second) divided by the allocated bandwidth (in Hertz) [18]. As shown in Figure 5.7(a), the proposed algorithm outperforms the other three approaches by providing a spectral efficiency around 2.3 bits/second/Hertz, which is a three fold increase in spectral efficiency compared to JUM. Our algorithm is also 63% and 36% higher than the obtained values of both FO-Linear and FO-Constant, respectively. This improvement is achieved by applying the subcarrier selection and group construction techniques, in which mobile terminals at locations with high interference can be severed in a separate subgroup with low transmission power. Such procedure confines the interference impact on a limited number of mobile terminals and then helps in increasing the overall spectral efficiency of the mobile system.

Packet Loss Rate: The impact of inter-cell interference in heterogeneous networks can also be measured in terms of the average packet loss rate at mobile terminals. High rates of packet losses substantially reduce the bandwidth utilization within a cell and negatively affect the quality of experience at its end users. Figure 5.7(b) presents the achieved average packet loss rate when the proposed ASH algorithm is applied. Compared to both FO approaches, we observe that the proposed algorithm yields the lowest packet loss rate with an average of 1.5%, whereas FO-Constant and FO-Linear algorithms provide higher loss rates with averages of 2.27% and 2.53%, respectively. The figure also shows that the JUM algorithm gives the worst packet loss rate with an average of 11.3%. This significant packet loss rate can be attributed to the basic adjustment process of transmission powers in JUM, in which there is no consideration for the possible interfering signals at cell edges.



(a) Spectral Efficiency (b) Packet Loss Rate

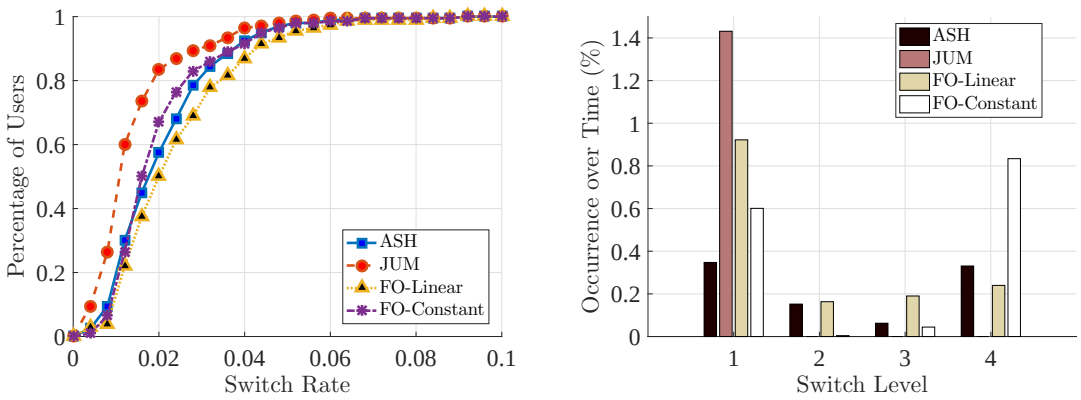
Figure 5.7: Comparisons of the proposed algorithm against the closest related work with respect to spectral efficiency and packet loss.



(a) Average Over Time (b) Cumulative Distribution

Figure 5.8: Comparisons of the achieved average bit rate of the proposed algorithm against the closest related work.

Assigned Bit Rate: Due to the limited radio resources and the variance of channel conditions in cellular networks, mobile terminals are likely to be served with different quality representations. Figure 5.8 indicates that our proposed ASH algorithm outperforms other algorithms on the achieved average bit rate. For instance, when there are 200 mobile users within the cell, Figure 5.8(a) shows that the proposed algorithm provides an average data rate of 4.2 Mbps. This average data rate is suitable for 720p resolution displays, which are popular on modern smartphones and tablets. On the other hand, the obtained average data rate in the other three approaches varies between 410 Kbps and 3.8 Mbps. Also, it is shown in Figure 5.8(a) that our ASH algorithm gives an average data rate almost ten times higher than that average obtained in JUM. This significant improvement demonstrates that our proposed algorithm overcomes the spectral inefficiency found in the conventional multicast policies, in which mobile terminals in multicast groups are bounded by the terminal



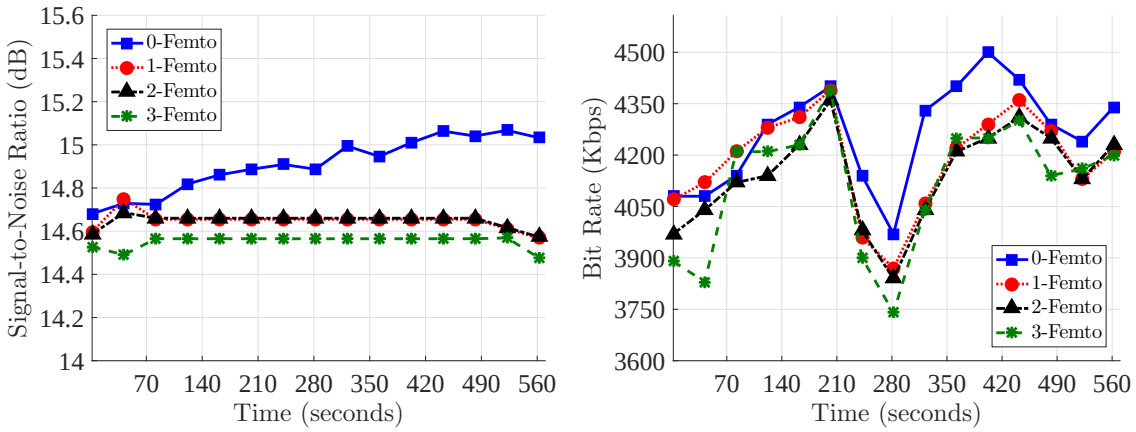
(a) Cumulative Distribution

(b) Switch Level Magnitude

Figure 5.9: Comparisons of the video quality switching of the proposed algorithms against the closest related work.

with worst channel condition. When it is compared against both FO-Constant and FO-Linear, our algorithm enhances the average data rate by 37% and 8%, respectively. Regarding the cumulative distribution function of the average bit rate, the proposed algorithm, as shown in Figure 5.8(b), allocates average bit rates higher than 3 Mbps to almost 95% of the mobile terminals in the system. In contrast, systems employing JUM, FO-Constant, and FO-Linear algorithms guarantee 3 Mbps to less than 0%, 32%, and 82% of the mobile terminals, respectively.

Quality Switching: From a user's perspective, the quality of experience of a video stream is greatly impacted by the average data rate and the frequency of switching among video quality representation levels. Figure 5.9 presents the stability of the video streaming system, in which the number of switching for each mobile terminal is measured in terms of rate (i.e., the frequency of switching among video representation levels per second) and magnitude (i.e., the difference between two consecutive quality representation levels). The video content server offers five different quality representation levels for each video stream (0.400, 0.750, 1.0, 2.5, and 4.5 Mbps). From the obtained results in Figure 5.9(a), it is demonstrated that 91% of mobile terminals experience at most a single switch once every 25 seconds (i.e., 0.04 switches per second), when the proposed algorithm is employed. In addition, when a quality switching occurs, mobile terminals in the proposed ASH algorithm observe a slight switch with 1-level up or down for less than 0.35% of their streaming time (i.e., it occurs once every 288 seconds on average) as shown in Figure 5.9(b). Mobile terminals are experiencing the same magnitude in JUM, FO-Linear, and FO-Constant for almost 1.4%, 0.9%, and 0.6% of the streaming time, respectively. That is, the same magnitude occurs once every 70 seconds, 108 seconds, and 166 seconds in JUM, FO-Linear, and FO-Constant, respectively. On the other hand, mobile terminals in the ASH algorithm observe a switch with 4-level up or down once every 302 seconds on average, while terminals in FO-Constant and FO-Linear are experiencing the same magnitude once every 120 seconds and 417 seconds, respectively.



(a) Average Signal-to-Noise Ratio

(b) Average Bit Rate

Figure 5.10: The impact of varying the number of interfering femtocells on the proposed algorithm.

Impact of Femtocells: We show the effectiveness of the proposed ASH algorithm in handling the inter-cell interference by demonstrating the impact of femtocells on its performance with respect to both gained signal-to-noise ratio and achieved average bit rate. To accomplish such objective, we run the considered simulation scenario in Figure 5.5 four times as following: 1) the three femtocells in the figure are disabled to have *No Interference Case*, 2) only the upper-left femtocell is enabled to have *Low Interference Case*, 3) both femtocells at the left-side are enabled to have *Intermediate Interference Case*, and 4) the three femtocells are enabled to have *High Interference Case*. Figure 5.10(a) reports the impact of varying the number of interfering femtocells on the achieved average signal-to-noise ratio over time. From the obtained results, it can be observed that the proposed algorithm succeeds in minimizing the impact of inter-cell interference since the received signals at mobile terminals experience a minor decline in its strength. For instance, the average signal-to-noise ratio is dropped by only 1.8%, 1.9%, and 2.5% when the number of interfering femtocells is set to 1, 2, and 3, respectively. As a consequence of such downgrading in the signal strength, the average data rate would eventually decrease as it is shown in Figure 5.10(b). As an example, the average data rate in the video streaming service is reduced by 1.8%, 3%, and 3.4% in the cases where a single, two, and three femtocells are deployed, respectively.

5.7 Summary

Inter-cell interference is one of the most important research problems in heterogeneous networks, especially when multiple base stations of different types share the same frequency channels to cover their transmission regions. A high level of inter-cell interference results in downgrading the signal strength at mobile terminals and reducing the overall average data rate. This chapter proposes a novel video transmission scheduling algorithm in which both interference avoidance and adaptive streaming are taken into consideration. Different from existing works, we do not rely on the control of subframe muting or the limitation of frequency reuse. Instead, we present an approach in

which terminals are dynamically arranged in unicast and multicast groups in a manner that maximizes received signal strength at mobile terminals. Once a proper group configuration is reached, the available radio resources are allocated with the goal of increasing the average data rate of video sessions. To accomplish such objective, each base station helped its associated mobile terminals in estimating their available bandwidth and facilitating their video rate adaptation processes. We demonstrate through simulations that compared to the closest works in the literature, the proposed algorithm yields a significant gain in the average data rate and reduces the number of switching in video quality representations. We also compare the performance of the proposed algorithm against three existing multicast approaches with respect to several performance metrics, including the average signal-to-noise ratio, spectral efficiency, and frame loss rate.

According to obtained simulation results, the proposed ASH algorithm outperforms both prior joint unicast-multicast and fair-optimal grouping approaches. Mobile terminals in the proposed algorithm have an average of 14.6 dB in their signal-to-noise ratios, whereas around 60% and 90% of mobile terminals in the closest related work are experiencing less than 12 dB and 14 dB in their signal-to-noise ratios, respectively. Regarding the achieved average bit rate, the proposed algorithm allocates a data rate higher than 3.5 Mbps to almost 84% of the mobile terminals. On the other hand, systems employing FO-Linear and FO-Constant algorithms assign 3.5 Mbps to only 10% and 70% of their mobile terminals, respectively. Also, it is shown that mobile terminals in the proposed algorithm enjoy smooth video streaming experiences since those terminals sense a slight switch with a 1-level magnitude once every 288 seconds, compared to a 1-level change at least once every 70 seconds and up to once every 167 seconds in the closest works. Moreover, we have shown that the proposed algorithm mitigates the effect of interfering femtocells. The results show that the SNR value is decreased by small percentages (i.e., by 1.8% and 2.5% in the cases when the number of femtocells is one and three, respectively). This decrease resulted in a slight reduction in the achieved average data rate (i.e., by 1.8% and 3.4% in the cases where one and three femtocells are deployed, respectively).

Chapter 6

Conclusions and Future Work

In this chapter, we first summarize the contributions of this thesis and show how they can help in supporting the substantial increase in demand for bandwidth capacity, providing good quality of experience for end-users, and minimizing the power consumption of mobile terminals. We then outline possible future research directions for the work presented in this thesis.

6.1 Conclusions

Multimedia streaming over wireless cellular networks encounters several challenges due to the diverse nature of its underlying infrastructure and the various capabilities of its mobile receivers. To address these challenges, we have designed in this thesis novel algorithms in order to enable current and future cellular networks to increase their capacity, optimize the quality of video delivered to users, and extend the lifetime of the batteries in mobile terminals.

There exists a tradeoff between the transmission schemes: unicast connections allow the creation of short transmission bursts that help in lowering energy consumption of mobile terminals, whereas multicast sessions restrict the data rate of each group according to users with worst channel conditions to increase the number of served users within every cell. To balance such tradeoff, we proposed in Chapter 3 transmission scheduling algorithms that concurrently utilized a mixture of unicast and multicast connections in order to construct sets of transmission bursts that result in increasing the overall energy saving at mobile terminals and the total number of served users. The proposed algorithms decided which chunks of videos should be sent, when they must be delivered, and which modulation and coding schemes should be chosen. We also demonstrated that the allocation of radio resources directly affected the load within a cellular network as well as the battery life of mobile terminals. For example, transmitting at higher modulation and coding schemes allowed mobile devices to receive at higher rates and then finish earlier, which in turn resulted in an increased energy saving because mobile terminals might turn off their wireless interfaces for longer time periods.

Current wireless cellular networks support theoretical peak data rates of 300 Mbps downlink, and future mobile networks aim at achieving a theoretical peak data rate of 3 Gbps downlink. A single active terminal could get up to almost 75 Mbps when it experiences an excellent channel quality condition, and this rate might be dropped to about 1 Mbps for a cell-edge user, who usually receives a weak signal and suffers from strong interference [2, 36]. For this reason, major events and sports games like the Super Bowl, where tens of thousands fans gather for hours, impose a significant burden on any mobile network. In such cases, a mobile provider could exploit the concept of multicast over single frequency networks within the surrounding area of an event to deliver related content, such as instant replays and highlight videos. In addition to increasing the average service ratio within the system, multicast over single frequency network improves the network coverage and minimizes inter-cell interference. Different from unicast streams, multicast sessions are preceded by an extended cyclic prefix to accommodate a longer guard time, thus enabling more SFN signals from distant base stations to positively contribute into the useful signal energy. The use of multicast over single frequency networks is expected to accelerate in the near future since it does not require any changes in the architecture of its underlying network or the terminal of mobile receivers.

In Chapter 4, we introduced the idea of dynamically configuring cells in wireless networks to collaboratively form single frequency networks based on the multimedia traffic demands from users in each cell. We formulated the transmission scheduling problem in such complex systems with the goal of maximizing the number of served users in each cell. We proved that this problem is NP-Complete, and we proposed a heuristic algorithm to solve it. Different from existing works, the proposed algorithm did not need dedicated radio resources for single frequency networks; instead, the spectrum could be allocated to multicast sessions at peak times and within certain regions of interest whenever it was likely to be fully utilized. Once a certain event was over, these frequencies could be easily aggregated into the radio resources used for unicast connections. To accomplish such objective, different decisions were performed in the proposed transmission scheduling algorithm. For instance, it could expand the number of areas in which a cell was enrolled to accommodate additional multicast services, replace an existing video stream with another multicast session, or shrink the number of areas in which a cell was enrolled. Through detailed packet-level simulations, we showed that the proposed algorithm can achieve orders of magnitudes increase in the service ratio compared to the current state-of-the-art approaches in 4G networks.

To cope with the substantial bandwidth requirements for video streaming over wireless cellular networks, a mobile network provider could also deploy additional small cells to improve the network coverage and offload some data traffic from the main macrocells. These small cells share the same frequency carrier with existing macrocells; therefore, interference is likely to evolve among cells. To address this problem, we proposed a quality-aware transmission scheduling algorithm in Chapter 5, in which interference-aware multicast grouping and adaptive video streaming approaches were presented. In the proposed algorithm, each base station allocated its radio resources and adjusted its transmission powers such that interference among cells is reduced. To increase the average data rates at mobile terminals, the proposed algorithm also provided bandwidth estimation to termi-

nals in order to help in facilitating their video rate adaptation processes. The proposed transmission scheduling algorithm ran independently and self-organized the radio resources in each cell; consequently, it could operate in a hybrid mode of both open and closed subscriber groups. Compared against the closest state-of-the-art algorithms, our proposed algorithm outperformed them by significant margins with respect to the average signal-to-noise ratio at mobile terminals, the spectral efficiency of each video session, and the rate of frame loss in the network. It also provided higher average data rates without causing frequent switching among video quality representations.

6.2 Future Work

The problems studied in this thesis can be extended in several directions. For instance, the benefits of dynamic single frequency networks can be further amplified with the expected new features in 5G wireless cellular networks, which include the deployment of massive Multiple-Input Multiple-Output (MIMO) antennas and the virtualization of base stations. Massive MIMO enables each base station to have between tens and hundreds of antennas, compared to only a few in the current deployment of base stations. Such enhancement will introduce more opportunities for optimizations. For example, a base station can join multiple single frequency networks using different antennas, while it can at the same time have other antennas to serve local incoming video requests. However, several research challenges need to be addressed before the incorporation of massive MIMO into future wireless cellular networks. Base stations with Massive MIMO eventually have to apply new techniques of beamforming, and these techniques might need high volume of feedback about the channel quality conditions. Hence, efficient beamforming approaches and channel estimation methods should be designed.

Recently, there is an increasing interest to apply the concept of software-defined networking in wireless cellular systems. The primary incentive here is that the principle of software-defined networking will further increase the ability of a cell to optimally configure its radio resources, meet the dynamic demand of mobile terminals, and reduce the cost of both deployment and operation. Yet, utilizing software-defined networking in mobile systems still encounters several outstanding research issues. As an example, there is a somewhat lack of standardization about this emerging technology, and there is no unified programmable interface to include those software-based controllers within the wireless cellular infrastructure. In other words, there is no clear approach on how to abstract the low-level functionality of a mobile network into virtual services, how to maintain a global and logical view of the radio resource within a mobile network, and how a programmable switch can provide the optimal trade-off between performance and flexibility.

In the proposed multimedia streaming model, video content servers are considered to be a part of telecommunication operators. Since video streams are being delivered simultaneously within the same network, having content caches closer to base stations helps in increasing the chances of creating multicast streams. It also helps in overcoming the probability of facing congestion issues over the delivery path, which usually occurs in traditional content delivery networks. Due to the het-

erogeneity of mobile terminals and their channel conditions, video streams can be requested with different quality resolutions and bitrate representations. Therefore, a computationally-efficient and real-time transcoding technique is needed to help mobile providers in giving on-the-fly adaption to the variation in user requests. Another approach would be relying on the massive data centers operating under the hood of public cloud service providers and take advantage of their computational capabilities. Since these data centers might be located many hops away from base stations, mobile providers should perform some countermeasures for the expected network latency and carefully consider effective video caching strategies to minimize the usage of these cloud services. For instance, these caching strategies can be designed based on some prediction probabilities for both video popularity and user distribution.

In this thesis, we have also applied the concept of patching to present a delay mitigation mechanism and provide true-on-demand video services without imposing long waiting times on users. In the current implementation, mobile terminals can join any existing multicast group and receive the missing leading portion of this video session over a unicast connection. Instead of creating unicast connections and then increasing the traffic load on base stations, it is possible to introduce a novel delay mitigation mechanism in which a terminal-to-terminal communication is exploited to offload the traffic from base stations. A terminal-to-terminal communication can occur in the cases where many mobile terminals are interested in a certain video stream and have an overlapped transmission coverage. These terminals can take advantage of the cellular in-band or the unlicensed out-band spectrum to initiate the communication, without the necessity of any support from their associated base stations. In such scenarios, such cooperative effort can significantly improve the total throughput, minimize the energy consumption, decrease the initial buffering time, and maintain reasonable fairness among end-users. However, several issues are still needed to be addressed, such as both power control and interference management between terminal and cellular connections. To illustrate this problem, interference is likely to be generated when a terminal-to-terminal connection is utilizing radio frequencies similar to those used by cellular networks. If an unlicensed out-band spectrum is used to cooperative video streaming services, such interference would not be caused. Nonetheless, mobile terminals cooperating together are going to encounter a different type of interference, whose level is much complicated to control in the unlicensed spectrum. As a consequence, guaranteeing an adequate level of quality of services would be another challenging problem. Another challenge is also to provide sufficient incentives for cooperative users since they utilize their power and bandwidth resource to relay the leading portion of a video stream to other mobile terminals.

Bibliography

[1] 3GPP. Improved video support for packet switched streaming (PSS) and multimedia broadcast/multicast service services (3GPP TR 26.903 V9.0.0), March 2010.

[2] 3GPP. Evolved universal terrestrial radio access and evolved universal terrestrial radio access network: Overall description (3GPP TS 36.300 V12.2.0), September 2014.

[3] Hatem Abou-zeid, Hossam Hassanein, and Stefan Valentin. Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks. *IEEE Transactions on Vehicular Technology*, 63(5):2013–2026, Jun 2014.

[4] Adobe: Bit rates for live streaming, January 2009. Available: `http://tiny.cc/Adobe`.

[5] Adobe Systems: HTTP Dynamic Streaming (HDS), February 2015. Available: `http://tiny.cc/HDS`.

[6] Richard Afolabi, Aresh Dadlani, and Kiseon Kim. Multicast scheduling and resource allocation algorithms for OFDMA-Based systems: A survey. *IEEE Communications Surveys and Tutorials*, 15(1):240–254, January 2013.

[7] Patrick Agyapong, Mikio Iwamura, Dirk Staehle, Wolfgang Kiess, and Anass Benjebbour. Design considerations for a 5G network architecture. *IEEE Communications Magazine*, 52(11):65–75, November 2014.

[8] Akamai Press Releases: Swisscom and Akamai enter into a strategic partnership, March 2013. Available: `http://tiny.cc/Akamai`.

[9] Saamer Akhshabi, Lakshmi Anantakrishnan, Ali C. Begen, and Constantine Dovrolis. What happens when HTTP adaptive streaming players compete for bandwidth? In *Proc. of ACM Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'12)*, pages 9–14, Toronto, Canada, 2012.

[10] Saamer Akhshabi, Ali C. Begen, and Constantine Dovrolis. An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP. In *Proc. of the ACM Conference on Multimedia Systems (MMSys'11)*, pages 157–168, San Jose, CA, 2011.

[11] Antonios Alexiou, Christos Bouras, Vasileios Kokkinos, and George Tsichritzis. Performance evaluation of LTE for MBSFN transmissions. *Wireless Networks*, 18(3):227–240, April 2012.

[12] Saleh Almowuena and Mohamed Hefeeda. Dynamic configuration of single frequency networks in mobile streaming systems. In *Proc. of ACM Multimedia Systems Conference (MMSys'15)*, pages 153–164, Portland, Oregon, 2015.

[13] Saleh Almowuena and Mohamed Hefeeda. Mobile video streaming over dynamic single-frequency networks. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 13(1):3:1–3:25, October 2016.

[14] Saleh Almowuena and Mohamed Hefeeda. Self-organized transmission scheduling in heterogeneous cellular networks (Under Review). *ACM Multimedia Systems Conference (MMSys'17)*, pages 1–12, 2017.

[15] Saleh Almowuena, Md. Mahfuzur Rahman, Cheng-Hsin Hsu, Ahmad Hassan, and Mohamed Hefeeda. Energy-aware and bandwidth-efficient hybrid video streaming over mobile networks. *IEEE Transactions on Multimedia*, 18(1):102–115, January 2016.

[16] Manish Anand, Edmund Nightingale, and Jason Flinn. Self-tuning wireless network power management. *Wireless Networks*, 11(4):451–469, July 2005.

[17] Apple: HTTP Live Streaming (HLS), February 2015. Available: `http://tiny.cc/HLS2015`.

[18] Giuseppe Araniti, Massimo Condoluci, Antonio Iera, Antonella Molinaro, John Cosmas, and Mohammadreza Behjati. A low-complexity resource allocation algorithm for multicast service delivery in OFDMA networks. *IEEE Transactions on Broadcasting*, 60(2):358–369, June 2014.

[19] Giuseppe Araniti, Massimo Condoluci, Leonardo Militano, and Antonio Iera. Adaptive resource allocation to multicast services in LTE systems. *IEEE Transactions on Broadcasting*, 59(4):658–664, December 2013.

[20] Argyriou Argyriou, Dimitrios Kosmanos, and Leandros Tassiulas. Joint time-domain resource partitioning, rate allocation, and video quality adaptation in heterogeneous cellular networks. *IEEE Transactions on Multimedia*, 17(5):736–745, May 2015.

[21] Arizona State University: Video Trace Library, March 2014. Available: `http://trace.eas.asu.edu`.

[22] Amotz Bar-Noy, Justin Goshi, Richard Ladner, and Kenneth Tam. Comparison of stream merging algorithms for media-on-demand. *Multimedia Systems*, 9(5):411–423, March 2004.

[23] Ali Begen, Tankut Akgul, and Mark Baugher. Watching video over the web: Streaming protocols. *IEEE Internet Computing*, 15(2):54–63, March 2011.

[24] Bell. Crave TV: video-on-demand service, December 2014. Available: `www.cravetv.ca`.

[25] Francesco Capozzi, Giuseppe Piro, Luigi Grieco, Gennaro Boggia, and Pietro Camarda. Downlink packet scheduling in LTE cellular networks: Key design issues and a survey. *IEEE Communications Surveys and Tutorials*, 15(2):678–700, July 2013.

[26] Cellular Coverage Maps, November 2014. Available: `www.cellumap.com`.

[27] Jiasi Chen, Mung Chiang, Jeffrey Erman, Guangzhi Li, Kadangode Ramakrishnan, and Rakesh Sinha. Fair and optimal resource allocation for LTE multicast (eMBMS): Group partitioning and dynamics. In *Proc. of IEEE INFOCOM'15*, pages 1266–1274, Hong Kong, P.R. China, April 2015.

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

[29] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and social network of YouTube videos. In *Proc. of IEEE Workshop on Quality of Service (IWQoS'08)*, pages 229–238, Enschede, Netherlands, June 2008.

[30] Seong Chuah, Zhenzhong Chen, and Yap Tan. Energy-efficient resource allocation and scheduling for multicast of scalable video over wireless networks. *IEEE Transactions on Multimedia*, 14(4):1324–1336, August 2012.

[31] Claudio Cicconetti, Luciano Lenzini, Enzo Mingozzi, and Carl Eklund. Quality of service support in IEEE 802.16 networks. *IEEE Network Magazine*, 20(2):50–55, March 2006.

[32] Cisco Visual Networking Index: Global mobile data traffic forecast update 2014-2019, February 2015. Available: `http://tiny.cc/Cisco14`.

[33] CPLEX: IBM ILOG Optimizer, July 2009. Available: `http://tiny.cc/CPLEX`.

[34] Crossing the Chasm: Small Cells Industry, November 2015. Available: `http://tiny.cc/SmallCell`.

[35] Erik Dahlman, Gunnar Mildh, Stefan Parkvall, Janne Peisa, Joachim Sachs, Yngve Selen, and Johan Skold. 5G wireless access: requirements and realization. *IEEE Communications Magazine*, 52(12):42–47, December 2014.

[36] Erik Dahlman, Stefan Parkvall, and Johan Skold. *4G: LTE/ LTE-advanced for mobile broadband*. Academic Press, Waltham, Massachusetts, December 2013.

[37] Supratim Deb, Pantelis Monogioudis, Jerzy Miernik, and James Seymour. Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets. *IEEE/ACM Transactions on Networking*, 22(1):137–150, Feb 2014.

[38] Hui Deng, Xiaoming Tao, and Jianhua Lu. QoS-aware Resource Allocation for Mixed Multicast and Unicast Traffic in OFDMA Networks. *EURASIP Journal on Wireless Communications and Networking*, 2012(1):1–10, June 2012.

[39] Derek Eager, Mary Vernon, and John Zahorjan. Minimizing bandwidth requirements for on-demand data delivery. *IEEE Transactions on Knowledge and Data Engineering*, 13(5):742–757, September 2001.

[40] Ahmed Elsherif, Zhi Ding, Xin Liu, and Jyri Hamalainen. Resource allocation in two-tier heterogeneous networks through enhanced shadow chasing. *IEEE Transactions on Wireless Communications*, 12(12):6439–6453, December 2013.

[41] Jeffrey Erman and Kadangode Ramakrishnan. Understanding the super-sized traffic of the super bowl. In *Proc. of ACM Conference on Internet Measurement Conference (IMC'13)*, pages 353–360, Barcelona, Spain, 2013.

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

[43] GLPK: GNU Linear Programming Kit, June. 2012. `http://tiny.cc/GLPK`.

[44] Yunmin Go, Oh Chan Kwon, and Hwangjun Song. An energy-efficient HTTP adaptive video streaming with networking cost constraint over heterogeneous wireless networks. *IEEE Transactions on Multimedia*, 17(9):1646–1657, September 2015.

[45] Ehsan Haghani, Shyam Parekh, Doru Calin, Eunyoung Kim, and Nirwan Ansari. A quality-driven cross-layer solution for MPEG video streaming over WiMAX networks. *IEEE Transactions on Multimedia*, 11(6):1140–1147, October 2009.

[46] Ahmed Hamza and Mohamed Hefeeda. Adaptive streaming of interactive free viewpoint videos to heterogeneous clients. In *Proc. of ACM Multimedia Systems Conference (MMSys'16)*, pages 1–12, Klagenfurt, Austria, 2016.

[47] Abbas Hatoum, Rami Langar, Nadjib Aitsaadi, Raouf Boutaba, and Guy Pujolle. Cluster-based resource management in OFDMA femtocell networks with QoS guarantees. *IEEE Transactions on Vehicular Technology*, 63(5):2378–2391, June 2014.

[48] Mohamed Hefeeda and Cheng-Hsin Hsu. On burst transmission scheduling in mobile TV broadcast networks. *IEEE/ACM Transactions on Networking*, 18(2):610–623, April 2010.

[49] Mohamed Hefeeda, Cheng-Hsin Hsu, and Kianoosh Mokhtarian. Design and evaluation of a proxy cache for peer-to-peer traffic. *IEEE Transactions on Computers*, 60(7):964–977, July 2011.

[50] Helmut Hlavacs and Shelley Buchinger. Hierarchical Video Patching with Optimal Server Bandwidth. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 4(1):8:1–8:23, January 2008.

[51] Mohammad Hoque, Matti Siekkinen, Jukka Nurminen, Sasu Tarkoma, and Mika Aalto. Saving energy in mobile devices for on-demand multimedia streaming – a cross-layer approach. *ACM Transactions on Multimedia Computing, Communications and Applications*, 10(3):1–23, April 2014.

[52] Cheng-Hsin Hsu and Mohamed Hefeeda. Broadcasting video streams encoded with arbitrary bit rates in energy-constrained mobile TV networks. *IEEE/ACM Transactions on Networking*, 18(3):681–694, June 2010.

[53] Cheng-Hsin Hsu and Mohamed Hefeeda. Flexible broadcasting of scalable video streams to heterogeneous mobile devices. *IEEE Transactions on Mobile Computing*, 10(3):406–418, March 2011.

[54] Kien Hua, Ying Cai, and Simon Sheu. Patching: A multicast technique for true video-on-demand services. In *Proc. of ACM Multimedia Conference*, pages 191–200, Bristol, United Kingdom, September 1998.

[55] Sha Hua, Yang Guo, Yong Liu, Hang Liu, and Shivendra Panwar. Scalable video multicast in hybrid 3G/Ad-Hoc networks. *IEEE Transactions on Multimedia*, 13(2):402–413, April 2011.

[56] Junxian Huang, Feng Qian, Alexandre Gerber, Z. Mao, Subhabrata Sen, and Oliver Spatscheck. A close examination of performance and power characteristics of 4G LTE networks. In *Proc. of ACM Conference on Mobile Systems, Applications, and Services (MobiSys'12)*, pages 225–238, Low Wood Bay, United Kingdom, 2012.

[57] Shraboni Jana, Amit Pande, An Chan, and Prasant Mohapatra. Mobile video chat: issues and challenges. *IEEE Communications Magazine*, 51(6):144–151, June 2013.

[58] Vinay Joseph and Gustavo Veciana. NOVA: QoE-driven optimization of DASH-based video delivery in networks. In *Proc. of IEEE INFOCOM'14*, pages 82–90, April 2014.

[59] Kashyap Kambhatla, Sunil Kumar, Seethal Paluri, and Pamela Cosman. Wireless H.264 video quality enhancement through optimal prioritized packet fragmentation. *IEEE Transactions on Multimedia*, 14(5):1480–1495, October 2012.

[60] Mohammad Katoozian, Keivan Navaie, and Halim Yanikomeroglu. Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization. *IEEE Transactions on Wireless Communications*, 8(1):66–71, January 2009.

[61] Lorenzo Keller, Anh Le, Blerim Cici, Hulya Seferoglu, Christina Fragouli, and Athina Markopoulou. MicroCast: Cooperative video streaming on smartphones. In *Proc. of ACM Conference on Mobile Systems, Applications, and Services (MobiSys'12)*, pages 57–70, Low Wood Bay, UK, June 2012.

[62] Hongseok Kim, Gustavo de Veciana, Xiangying Yang, and Muthaiah Venkatachalam. Distributed α-Optimal User Association and Cell Load Balancing in Wireless Networks. *IEEE/ACM Transactions on Networking*, 20(1):177 –190, February 2012.

[63] Juyeop Kim, Taesoo Kwon, and Dong-Ho Cho. Resource allocation scheme for minimizing power consumption in OFDM multicast systems. *IEEE Communications Letters*, 11(6):486–488, June 2007.

[64] Dilip Krishnappa, Michael Zink, and Ramesh Sitaraman. Optimizing the video transcoding workflow in content delivery networks. In *Proc. of ACM Multimedia Systems Conference (MMSys'15)*, pages 37–48, Portland, Oregon, March 2015.

[65] Wen Kuo, Wanjiun Liao, and Tehuang Liu. Adaptive resource allocation for layer-encoded IPTV multicasting in IEEE 802.16 WiMAX wireless networks. *IEEE Transactions on Multimedia*, 13(1):116–124, February 2011.

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

[67] Seung Joon Lee, Yongjoo Tcha, Sang-Yong Seo, and Seong-Choon Lee. Efficient use of multicast and unicast channels for multicast service transmission. *IEEE Transactions on Communications*, 59(5):1264–1267, May 2011.

[68] Baochun Li, Zhi Wang, Jiangchuan Liu, and Wenwu Zhu. Two decades of Internet video streaming: A retrospective view. *ACM Transactions on Multimedia Computing, Communications and Applications*, 9(1s):33:1–33:20, October 2013.

[69] Yu Liang, Wei Chung, Guo Ni, Ing Chen, Hongke Zhang, and Sy Kuo. Resource allocation with interference avoidance in OFDMA femtocell networks. *IEEE Transactions on Vehicular Technology*, 61(5):2243–2255, June 2012.

[70] David Lopez-Perez, Ismail Guvenc, Guillaume de la Roche, Marios Kountouris, Tony Quek, and Jie Zhang. Enhanced intercell interference coordination challenges in heterogeneous networks. *IEEE Wireless Communications*, 18(3):22–30, June 2011.

[71] Zhixue Lu, Tarun Bansal, and Prasun Sinha. Achieving user-level fairness in open-access femtocell-based architecture. *IEEE Transactions on Mobile Computing*, 12(10):1943–1954, October 2013.

[72] Carlos Luna, Yiftach Eisenberg, Randall Berry, Thrasyvoulos Pappas, and Aggelos Katsaggelos. Joint source coding and data rate adaptation for energy efficient wireless video streaming. *IEEE Journal on Selected Areas in Communications*, 21(10):1710–1720, December 2003.

[73] Ahmed Mansy, Mostafa Ammar, Jaideep Chandrashekar, and Anmol Sheth. Characterizing client behavior of commercial mobile video streaming services. In *Proc. of Workshop on Mobile Video Delivery (MoViD'14)*, pages 8:1–8:6, Singapore, Singapore, 2014.

[74] Marco Nicosia. Internet video: new revenue opportunity for telecommunications and cable providers, July 2010. Available: `http://tiny.cc/Cisco2010`.

[75] Microsoft: Live Smooth Streaming, February 2015. Available: `http://tiny.cc/MSSmooth`.

[76] Jose Monserrat, Jorge Calabuig, Ana Fernandez-Aguilella, and David Gomez-Barquero. Joint delivery of unicast and e-MBMS services in LTE networks. *IEEE Transactions on Broadcasting*, 58(2):157–167, June 2012.

[77] Netflix: Letter to Shareholders, October 2016. Available: `http://tiny.cc/Netflix2016`.

[78] Stefano Petrangeli, Jeroen Famaey, Maxim Claeys, Steven Latré, and Filip De Turck. QoE-driven rate adaptation heuristic for fair adaptive video streaming. *ACM Transactions on Multimedia Computing, Communication and Application*, 12(2):28:1–28:24, October 2015.

[79] Andrew Pyles, Xin Qi, Gang Zhou, Matthew Keally, and Xue Liu. SAPSM: smart adaptive 802.11 PSM for smartphones. In *Proc. of ACM Conference on Ubiquitous Computing (UbiComp'12)*, pages 11–20, New York, NY, September 2012.

[80] Bozidar Radunovic, Alexandre Proutiere, Dinan Gunawardena, and Peter Key. Dynamic channel, rate selection and scheduling for white spaces. In *Proc. of ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT'11)*, pages 2:1–2:12, Tokyo, Japan, December 2011.

[81] Riverbed: LTE Model User Guide, July 2012. Available: `http://tiny.cc/OPNETLTE`.

[82] Riverbed: OPNET Modeler Suite, July 2010. Available: `http://tiny.cc/OPNET`.

[83] Letian Rong, Olfa Haddada, and Salah-Eddine Elayoubi. Analytical analysis of the coverage of a MBSFN OFDMA network. In *Proc. of IEEE Global Telecommunications Conference (GLOBECOM'08)*, pages 1–5, New Orleans, Louisiana, November 2008.

[84] Nazmus Saquib, Ekram Hossain, Long Le, and Dong Kim. Interference management in OFDMA femtocell networks: issues and approaches. *IEEE Wireless Communications*, 19(3):86–95, June 2012.

[85] Aaron Schulman, Vishnu Navda, Ramachandran Ramjee, Neil Spring, Pralhad Deshpande, Calvin Grunewald, Kamal Jain, and Venkat Padmanabhan. Bartendr: a practical approach to energy-aware cellular data scheduling. In *Proc. of ACM Conference on Mobile Computing and Networking (MobiCom'10)*, pages 85–96, New York, NY, September 2010.

[86] Patrick Seeling, Frank Fitzek, Gergö Ertli, Akshay Pulipaka, and Martin Reisslein. Video network traffic and quality comparison of VP8 and H.264 SVC. In *Proc. of ACM Workshop on Mobile Video Delivery (MoVid'10)*, pages 33–38, Firenze, Italy, October 2010.

[87] Patrick Seeling and Martin Reisslein. Video transport evaluation with H.264 video traces. *IEEE Communications Surveys and Tutorials*, 14(4):1142–1165, September 2012.

[88] Souvik Sen, Bozidar Radunovic, Jeongkeun Lee, and Kyu-Han Kim. CSpy: Finding the best quality channel without probing. In *Proc. of ACM Conference on Mobile Computing and Networking (MobiCom'13)*, pages 267–278, Miami, FL, USA, October 2013. ACM.

[89] Somsubhra Sharangi, Ramesh Krishnamurti, and Mohamed Hefeeda. Energy-efficient multicasting of scalable video streams over WiMAX networks. *IEEE Transactions on Multimedia*, 13(1):102–115, February 2011.

[90] Shomi: Video-on-demand service, July 2014. Available: `www.shomi.com`.

[91] Sarabjot Singh and Jeffrey Andrews. Joint resource partitioning and offloading in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, 13(2):888–901, February 2014.

[92] Chetna Singhal, Swades De, Ramona Trestian, and Gabriel-Miro Muntean. Joint optimization of user-experience and energy-efficiency in wireless multimedia broadcast. *IEEE Transactions on Mobile Computing*, 13(7):1522–1535, July 2014.

[93] Thomas Stockhammer. Dynamic Adaptive Streaming over HTTP: Standards and design principles. In *Proc. of ACM Multimedia Systems Conference (MMSys'11)*, pages 133–144, San Jose, CA, February 2011.

[94] Gary Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, December 2012.

[95] Salvatore Talarico and Matthew Valenti. An accurate and efficient analysis of a MB-SFN network. In *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*, pages 6994–6998, Florence, Italy, May 2014.

[96] Tony Chun and Andrew Lund. DIGI: Meter Engineers Handbook for Cellular Communication, March 2015. Available: http://tiny.cc/DIGI2015.

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

[98] Verizon Wireless: Customers Use 1.9 Terabytes of Data in Stadium at Super Bowl, July 2014. Available: http://tiny.cc/Verizon2014.

[99] Verizon Wireless: Verizon Delivers LTE Multicast Over Commercial 4G LTE Network in Indy, May 2014. Available: http://tiny.cc/VerizonIndy.

[100] Danny Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, Gang Li, and R. Miller. Optimization of HTTP adaptive streaming over mobile cellular networks. In *Proc. of IEEE INFOCOM'13*, pages 898–997, April 2013.

[101] Hyungsuk Won, Han Cai, Do Eun, Katherine Guo, Arun Netravali, Injong Rhee, and Krishan Sabnani. Multicast scheduling in cellular data networks. *IEEE Transactions on Wireless Communications*, 8(9):4540–4549, September 2009.

[102] Xiufeng Xie, Xinyu Zhang, Swarun Kumar, and Li Erran Li. piStream: Physical layer informed adaptive video streaming over LTE. In *Proc. of ACM Conference on Mobile Computing and Networking (MobiCom'15)*, pages 413–425, Paris, France, 2015.

[103] Jian Xu, Sang Lee, Woo Kang, and Jong Seo. Adaptive resource allocation for MIMO-OFDM based wireless multicast systems. *IEEE Transactions on Broadcasting*, 56(1):98–102, March 2010.

[104] YouTube: Product Statistics, March 2016. Available: http://tiny.cc/YouTube2016.

[105] Ya-Ju Yu, Pi-Cheng Hsiu, and Ai-Chun Pang. Energy-efficient video multicast in 4G wireless systems. *IEEE Transactions on Mobile Computing*, 11(10):1508–1522, October 2012.

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

[107] Xinyu Zhang and Kang Shin. E-MiLi: energy-minimizing idle listening in wireless networks. *IEEE Transactions on Mobile Computing*, 11(9):1441–1454, September 2012.

[108] Hao Zhou, Yusheng Ji, Xiaoyan Wang, and Baohua Zhao. ADMM based algorithm for eICIC configuration in heterogeneous cellular networks. In *Proc. of IEEE INFOCOM'15*, pages 343–351, April 2015.

[109] Hao Zhu and Guohong Cao. On supporting power-efficient streaming applications in wireless environments. *IEEE Transactions on Mobile Computing*, 4(4):391–403, July 2005.