HYBRID MULTICAST-UNICAST STREAMING OVER MOBILE NETWORKS

by

Md. Mahfuzur Rahman
B.Sc., Bangladesh University of Engineering & Technology, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Spring 2014

APPROVAL

Name: Md. Mahfuzur Rahman

Degree: Master of Science

Title of Thesis: HYBRID MULTICAST-UNICAST STREAMING OVER MOBILE NETWORKS

Examining Committee:

Dr. Kay C. Wiese
Associate Professor, Computing Science
Simon Fraser University
Chair

Dr. Mohamed Hefeeda,
Associate Professor, Computing Science
Simon Fraser University
Senior Supervisor

Dr. Joseph G. Peters,
Professor, Computing Science
Simon Fraser University
Supervisor

Dr. Jiangchuan Liu,
Associate Professor, Computing Science
Simon Fraser University
SFU Examiner

Date Approved: December 9th, 2013

Abstract

Modern cellular networks support unicast and multicast video streaming. Streaming over unicast results in high energy saving for mobile devices, but it can only serve a limited number of users. Whereas multicast streaming can serve large number of mobile users, but it increases the energy consumption for mobile devices. We formulate a resource allocation problem that considers both unicast and multicast simultaneously to maximize the average energy saving across all mobile devices and to serve large number of mobile users. We propose efficient algorithms to solve this problem. Our simulation results show that the proposed algorithms: (i) result in high energy saving close to optimal solutions, (ii) serve large number of mobile users compared to that of unicast-only approach, (iii) are more elastic than unicast- and multicast-only approaches, and (iv) lead to higher energy saving as more network resources are allocated to streaming services.

Keywords: video streaming; dynamic modulation and coding; next generation cellular networks; unicast-multicast hybrid streaming; resource allocation; optimization; mobile energy saving

To all the members of my family!

"Acts of kindness protect one from ruin wrought by evil."

— *Prophet Muhammad (peace be upon him), Fiqh-us-Sunnah*

Acknowledgments

I would like to express my gratitude and thanks to my senior supervisor, Dr. Mohamed Hefeeda for his patient guidance and encouragements throughout the graduate study. His passion about research and insights in my work have been an immense help for me. He has always been there to help smooth continuation of the research. The completion of the thesis would not have been possible without him.

I would like to thank my supervisor, Dr. Joseph G. Peters and my thesis examiner Dr. Y for being on my committee and reviewing the thesis. I would also like to thank Dr. X for taking the time to chair my thesis defense.

I would like to thank my postdoc in the Lab, Abdul Hasib for his guidance and encouragements during his stay in the Lab. I would also like to thank all the members at NSL, particularly Ahmed Hamza and Saleh Almowuena for their cordial love and help in the research work. I would like to take this opportunity to thank Dr. Cheng-Hsin Hsu, a previous NSL student, for his support and guidelines throughout the thesis.

I would also like to thank all of my friends in British Columbia and back home whose support and inspiration have made the journey easier. I would like to express my gratitude especially to the friends in British Columbia whose company has made my graduate life enjoyable.

I would like to take this opportunity to express my gratitude to my parents, younger brother and sister whose affection, love and care have influenced every step of my life in a positive way. I would like extend my special thanks and gratitude to my wife for her support, love and care which have greatly simplified my life.

Contents

List of Tables

List of Figures

Preface

This dissertation is original, unpublished work by the author, Md. Mahfuzur Rahman with close collaboration with senior supervisor, Dr. Mohamed Hefeeda and previous NSL student, Dr. Cheng-Hsin Hsu.

Chapter 1

Introduction

1.1 Introduction

The demand for multimedia streaming over mobile networks has been steadily increasing: worldwide mobile traffic amount reached 885 petabytes per month in 2012 of which 51% carries video while the traffic amount was 520 petabytes per month in 2011 [61]. So, the video traffic should be served carefully and efficiently. Video traffic in cellular networks can be served using unicast or multicast. Unicast is not optimized for streaming the same video to many users as the same content is transmitted repeatedly consuming excessive network resources. On the other hand, multicast is effective in terms of serving many users watching the same video. Although current 3G mobile networks support only unicast, several 4G mobile networks have multicast, e.g., the Multicast and Broadcast Service (MBS) in WiMAX [12] and the Evolved Multimedia Broadcast Multicast Service (eMBMS) in Long Term Evolution(LTE) [59]. Realizing the power of multicast, some U.S. mobile operators plan to launch eMBMS service in their LTE networks [69] to efficiently stream the same video to numerous mobile users. Multicast in cellular networks is useful for live sports events and other popular live events. Mobile users start receiving from the current moment although they may arrive at different times creating a natural case for grouping users into multicast sessions. But many other applications including video on-demand streaming, time-shifted events, and mobile video recorders can all benefit from multicast due to increasingly larger storage space available in modern mobile devices to prefetch some video segments for future consumption. When considering prefetching, popular videos like latest TV episodes and highlights of recent sports events are requested by many users at somewhat close times, e.g., in the evening of the release day. These requests can be grouped into multicast sessions. These applications lead to tremendous loads on the mobile networks which can be efficiently handled by leveraging multicast.

In order to efficiently handle the numerous incoming requests, cellular base stations have to intelligently determine whether to serve each request using unicast or multicast to maximize the average energy saving across all mobile devices and to minimize the network load. This decision is challenging because: (i) user demands and network conditions are diverse and dynamic, and (ii) there exists a tradeoff between the network load and energy saving. This tradeoff comes into play due to two common features of modern mobile networks. First, network interfaces on mobile devices support multiple Modulation and Coding Scheme (MCS) modes to adapt to different channel conditions. For example, mobile devices closer to the base station may use more aggressive, i.e., higher MCS modes due to better channel conditions and can achieve higher transfer rates. Whereas mobile devices at the cell edge can only use lower MCS modes resulting in low transfer rates because of comparatively poor channel conditions. Second, network interfaces on mobile devices may be turned off to save energy during idle periods. For the same amount of data, higher MCS modes lead to more energy saving. This is because, higher MCS modes result in higher data transfer rates allowing corresponding users finish earlier and achieve longer idle periods. If a user is served using unicast, it can use the most aggressive MCS mode possible based on its channel conditions alone to achieve the highest energy saving. But this leads to excessive network load as this may result in duplicate video streams. On the other hand, multicast greatly reduces the network loads. But the users suffer from lower transfer rate as the MCS mode in the multicast session is determined by user with the worst channel condition. Such tradeoff between network load and energy saving motivates the study of a *hybrid* on-demand video streaming system which is the subject of this thesis, leveraging both unicast and multicast. Figure 1.1 demonstrates a sample solution of our resource allocation problem. For example, the mobile device in the mid-circle receives video using *unicast* with MCS mode 3 as there is only one receiver. The two mobile devices at the outer-most circle receive video using *multicast* with MCS mode 1, although one of the two devices is capable of receiving at a higher MCS mode of 2 which would have resulted in more energy saving for that mobile device. But this would have also consumed more resource blocks as multicast would have not been possible in that case. It indicates when a user in a multicast group cannot receive at a specific MCS mode, our solution tries to send at a *lower* MCS mode so that all the mobile devices in the multicast group can receive at that MCS mode.

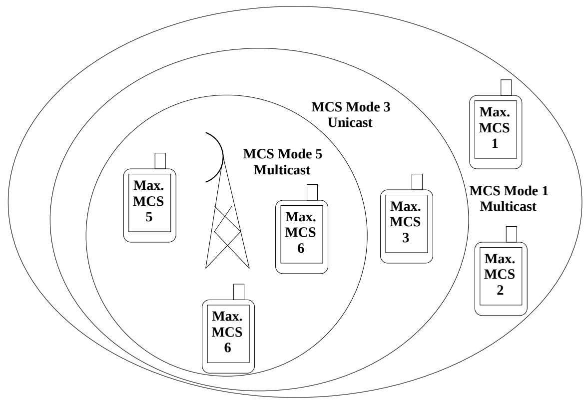

Figure 1.1: An illustrative example showing that mobile devices receive videos using either unicast or multicast.

1.2 Problem Statement

We consider wireless networks with multicast streaming support, such as LTE [59]. Mobile users arrive asynchronously to the system and request different on-demand videos following a Zipf-like distribution. The channel conditions of these users vary in the cellular network. Each cell in the wireless network can support a certain amount of resource blocks based on the downlink transmission bandwidth and a fraction of these resource blocks are reserved for on-demand video services. We define a *resource block* as the minimum allocation unit that can be assigned to a user. A resource block consists of a group of transmission symbols along the frequency domain as well as the time domain. For example, in LTE downlink, a resource block occupies 180KHz in the frequency domain and 0.5ms in the time domain. A request from a user can be satisfied by a unicast or multicast streaming based on availability of resources, channel conditions of the requesting user along with other factors. We are interested in finding out a resource block assignment to serve these users through a mixture of unicast and multicast streaming in order to maximize the average energy saving for the requesting mobile users.

The problem addressed by this thesis can be stated as follows:

Problem 1:Radio Resource Allocation. *Given a number of mobile users with different channel conditions requesting various on-demand videos and a certain fraction of resource blocks reserved to serve these requests. Find the optimal resource allocation considering both unicast and multicast* *streaming such that: (i) the average energy saving across all mobile devices is maximized (ii) the number of resource blocks consumed does not exceed the number of resource blocks reserved and (iii) all mobile devices served with a certain video receive that video at the same rate for smooth playout.*

1.3 Thesis Contributions

The contributions of the thesis can be summarized as follows:

1. We propose a novel resource allocation algorithm that considers a hybrid on-demand streaming system. It consists of both unicast and multicast in mobile networks to maximize the average energy saving across all mobile devices.

2. We propose greedy algorithm to solve the resource allocation problem in polynomial time. The proposed greedy solution runs in real-time and produces close-to optimal results, within 1.8% of the optimal. Thus, our proposed solution can be deployed in real systems.

3. We successfully merge the merits from both multicast and unicast in our solution. It allows the cellular networks to support a large number of mobile devices analogous to multicast-only networks, and achieve almost-optimal energy saving analogous to unicast-only networks. On experiments, compared to 71% energy saving of the multicast-only solution, our proposed solution leads to 89% energy saving, which is almost the same as the unicast-only network.

4. We extend our solution to multicell Single Frequency Networks and it results in even higher energy saving, up to 97% because of the larger optimization room, stronger signals, lower interference, and other merits of Single Frequency Networks.

5. We implement our algorithm in OPNET simulator. The simulation results prove the superiority of the proposed solution over conventional unicast-only and multicast-only approaches.

1.4 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 provides background information needed to understand the concepts discussed in this thesis. It also presents a literature review in order to report the existing works related to the thesis. Chapter 3 explains the proposed solution. In Chapter 4, the proposed algorithms are evaluated through extensive simulations to verify different aspects of the proposed solution. We conclude the thesis in Chapter 5 along with some directions for future work.

Chapter 2

Background and Related Work

In this chapter, we present the background information needed to understand the work properly. As part of that, we also provide an overview of the wireless network used for the simulations and importance of the thesis. Towards the end of the chapter, we cover a literature survey related to our work.

2.1 Background

2.1.1 Wireless Networks Support for Multimedia

Multimedia in wireless networks emerges all over the world due to continuous improvement in cellular networks, development of multimedia-enriched smartphones and various popular multimedia applications. In order to ensure smooth operation of these multimedia applications over wireless networks, some level of performance-oriented quality of service guarantee is required. Such quality of service guarantee is challenging in wireless environment mainly because of limited resources, changing network characteristics and user mobility. There are two quality of service models: (i) integrated service or IntServ model and (ii) differentiated service or DiffServ model [43]. These service models address different requirements from various multimedia applications. IntServ model is strict in ensuring quality of service of the admitted users, while DiffServ model is relatively flexible.

There are two types of services defined under IntServ model: (a) guaranteed service and (b) best-effort service. In guaranteed service, each end-user request is served in accordance with the quality of service requirements like minimum bandwidth, maximum delay. However, with best-effort service there is no such guarantee associated with the service requests. With the guaranteed service, resources are reserved based on the quality of service requirements specified by the multimedia applications. The reserved resource of a traffic flow may vary in response to resource variations in wireless networks which may lead the multimedia applications to specify a range of values as the quality of service requirement instead of a single value. On the other hand, the DiffServ model aggregate the incoming requests into different traffic classes and are served in accordance with the associated class. There are different variations of DiffServ model (e.g. relative DiffServ [14]) to handle the traffic classification and service quality associated with them.

The multimedia applications running over wireless networks need to consider (i) interference and (ii) dynamics of wireless environment like available resources, channel conditions of requesting mobile users to satisfy quality of service requirements. The various designs proposed for wireless networks to satisfy quality of service requirements for multimedia applications use different combinations of bandwidth management, admission control, and scheduling. They also employ different measurement techniques like channel quality monitoring to help the decision making process [43]. For example, Lee [31] proposes a framework for IntServ model where the reserved bandwidth is dynamically adjusted and predictive hand-off is employed. As another example, Qingwen et al. [33] propose a scheduling algorithm at medium access control layer to meet quality of service requirements for various applications.

2.1.2 LTE Overview

LTE is a standard for wireless data communication technology developed by the Third Generation Partnership Project (3GPP). It is a preferred development path for GSM/W-CDMA/HSPA networks currently deployed, and an option for evolution of CDMA networks. The goal of LTE is to provide a framework for increasing data rates and overall system capacity, frequency flexibility and improving spectral efficiency and cell-edge performance. Another goal is to offer redesign and simplification of the network architecture to an IP-based system with significantly reduced transfer latency compared to the 3G architecture. LTE aims to achieve a data rate of 300 Mbps in downlink and 170 Mbps in uplink [67]. It also sets a goal to improve the radio-network delay and achieve a delay of less than 5 milliseconds [5]. LTE has the provision for scalable carrier bandwidths varying from 1.4 MHz to 20 MHz and supports both frequency division duplexing (FDD) and time-division duplexing (TDD). Moreover, LTE has built-in support for multicast and broadcast.

The IP-based network architecture of LTE is designed to support only packet-switched services compared to the circuit-switched model of previous cellular systems. The IP address is assigned when the mobile device is switched on and released when switched off. Evolved Packet Core (EPC), the core network architecture of LTE is designed to replace the GPRS Core Network, and supports seamless handovers for both voice and data to base stations with older network technologies such as GSM, Universal Mobile Telecommunications System (UMTS) and CDMA2000. The Evolved Packet Core is designed to work with other access technologies not developed by 3GPP, like WiMAX and WiFi [67]. LTE aims to provide seamless internet connectivity between the mobile device and the packet data network (PDN) without any disruption during user mobility.

The LTE access network is simply a network of base stations which are called evolvoed NodeB (eNodeB)s resulting in a flat architecture. There is no centralized controller present in the architecture. The eNodeBs are usually inter-connected by the X2-interface among themselves and connected to the core network by the S1-interface. This distributed topology in LTE helps to speed up the connection set-up and reduce the time required for a handover. The connection set-up time for a real time data session, in many cases, is crucial for end-users, especially in on-line gaming [67]. The time required for a handover is particularly essential for real-time services where the end-users tend to terminate calls if the handover takes too long. This simple architecture reduces the operating cost [66], and along with high spectral efficiency enables LTE to reduce the cost per byte, which is expected to decrease by a factor of six compared to HSPA [65].

LTE uses a multicarrier technology, Orthogonal Frequency Division Multiple Access (OFDMA) on the downlink where the subcarriers can be shared among multiple users. It provides opportunity to exploit variations in both frequency and time domains for achieving high peak data rates in high spectral efficiency. This requires fast processors. The adoption of OFDMA in downlink leads to high peak-to-average power ratio which requires expensive power amplifiers with high requirements on linearity, increasing the battery consumption. Although this is not a problem for eNodeB, but crucial for mobile devices due to having limited battery life. Hence LTE employs Single Carrier Frequency Division Multiple Access (SC-FDMA) for the uplink, a technology that provides advantages in power efficiency and results in longer battery life compared to that of a pure OFDMA approach. Moreover, LTE considers multiple-input and multiple-output (MIMO) which refers to a technique that employs multiple transmit and receive antennas, often in combination with multiple radios and parallel data streams. It offers significant increases in data throughput without additional bandwidth or increased transmit power. For example, with a 2×2 MIMO system, a gain of a factor of 2 is expected on the peak throughput [65].

The high data rate and throughput along with low latency enable LTE to offer new and advanced mobile broadband services to end users. One of the major objectives of this network evolution is to provide these services with a quality at least equivalent to current standard. This means that LTE aims to meet Quality of Experience expectations of the end users in terms of real-time services like voice over IP, multi-user gaming over IP, high definition on-demand video and live TV, while continues to improve the quality of service for all legacy applications like email, internet browsing. The reduced latency in LTE greatly improves the user experience with interactive applications like NetMeeting, video conferencing. An increasing need for broadband access at home is observed and the same applies to mobile services for two main reasons: (i) as end users become used to higher speeds at home, they tend to expect the same quality of service even when mobile for a seamless experience and (ii) the possibility of offering higher bandwidth in remote areas where ADSL throughput is no longer sufficient and fibre may not be economically viable compared with LTE [65]. In such areas the LTE infrastructure may offer mobile services as well as broadband access at home.

TeliaSonera launched the first publicly available LTE service in Stockholm and Oslo on December 14, 2009 [74]. *Verizon Wireless* launched the first large-scale LTE network in North America in 2010 [76, 75]. Even in India, *Airtel* launched LTE service in April 2012 [60]. Such wide adoption of LTE along with high spectral efficiency, high data rates and flexible access architecture, it is anticipated to become a success amongst operators as well as customers and become the first truly global wireless standard.

2.1.3 Importance of Saving Energy on Mobile Devices

Microelectronics have experienced significant development over last few decades. This makes powerful mobile device development a reality. Now, mobile devices are equipped with high processing capabilities, improved graphical user interface (GUI), and multiple radio interfaces. The mobile devices with more advanced computing capability and connectivity than feature phones are called smartphones [73]. The hardware enhancements in smartphones lead to resource intensive software development. As a result, smartphones have powerful Operating Systems (OS) like Android OS, iOS, Windows, Blackberry OS with enormous number of attractive features. The fast expanding application markets for different smartphones along with these development improvements, make smartphone one of the most popular and desired devices in the market. Canalys report [72] for smartphone shipment analysis in 2011 shows that smartphones overtook even client PCs.

Due to the relatively small size of the mobile devices, the size of the batteries that can fit into these devices is small. Hence, these batteries have limited capacity leading to have limited battery life. The most powerful battery for mobile devices in the market belongs to Samsung Galaxy S4 having a battery of 2600 mAh [63]. Among other powerful mobile devices, iPhone 5S comes up with a battery of 1560 mAh and HTC One SV is equipped with a battery of 1800 mAh [63]. Because of this limited battery life, energy consumption in mobile devices is a crucial aspect. This realization of energy aspect of mobile devices, makes *mobile device energy saving* an evolving and important research era. Hence, in this work we are concerned with saving energy on mobile devices from the cellular network perspective.

2.1.4 Importance of Maximizing Wireless Interface Off time

According to Cisco's *Global Mobile Data Traffic Forecast* report, there will be 788 million mobile-only Internet users and global mobile data traffic will increase by a factor of 26 by 2015 [62]. So, mobile users are going to be one of the dominating population in the Internet world with a large amount of data consumption. A major fraction of the energy in smartphones is consumed by the wireless network interface controller (WNIC) [37]. A Toshiba 410 CDT mobile computer demonstrates that nearly 18% of the power is consumed by the wireless interface [24]. The base energy consumption of a HTC Tilt 8900 series phone ranges between 155-475mW, depending on the intensity of the backlight while the WiFi radio consumes over 1000mW during transmission [42]. As another example, the base energy consumption of a Nokia N900 is about 12 mW while the WiFi radio consumes 765 mW during receiving and 1000mW during sending [37]. Receivers in cellular networks (e.g. 3G and GSM) consume even more energy than WiFi. For example, consider a download of size 10KBytes. In this case, receivers in 3G consume six times the energy consumed by WiFi while in GSM consume three times energy of that of WiFi, and this ratio dramatically increases with the increase in data sizes [6]. So, it is of utmost importance to minimize the *on time* of the network interface of a mobile device to maximize the battery life without sacrificing the quality of service. Realizing this fact, IEEE 802.11 Access Point supports *Power Saving Mode* [42] where the network interface wakes up periodically to receive data instead of keeping it on constantly. Thus, the main concern of the proposed work is to reduce energy consumption of the mobile devices from network perspective.

2.1.5 Importance of Allocating Resources along subchannel

We consider allocating resource blocks along the subchannel. We first allocate the resource blocks for all subchannels for the first symbol. If we need more resource blocks, then we allocate all subchannels in the second symbol and so on. This approach minimizes the number of required symbols to satisfy any request. This in turns minimizes the energy consumption of the mobile devices because previous studies [50, 27] show that the energy consumption of a mobile device depends on the number of symbols received, and almost independent of the number of subchannels. This can be explained by a simple illustrative example. Lets consider allocating S resource blocks in a time period of T symbols. If we allocate along the symbols, it requires S symbols giving an *idle period* of $T - S$ for the network interface assuming $T \geq S$. But if we allocate along the subchannels, it requires 1 symbol which results in an *idle period* of $T - 1$, much higher than the other approach. Specifically consider a typical setting with $T = 20, S = 16$. The latter approach achieves 4.75 times energy saving of the other approach. This concept is illustrated in Figure 3.3 where data from two different videos, v_1, v_2 is allocated along the subchannels. Such high gain in terms of energy saving motivates us to allocate resource blocks along the subchannel.

2.1.6 Single Frequency Network

A Single Frequency Network (SFN) is a broadcast network of multiple base stations transmitting the same signal over the same frequency channel. This approach varies from a traditional Multi-Frequency Network (MFN) where each base station operates on a separate frequency channel. The aim of an SFN is efficient utilization of the radio spectrum compared to that of an MFN. Since the base stations forming an SFN operate in the same frequency, reception of more delayed signals from multiple base stations can be utilized even for improvement of the power efficiency of the base stations [41]. Power contributed from the individual base stations operating in a single frequency adds up. That is why a Single Frequency Network shows so-called SFN gain [41]. In LTE, a particular set of frequency resource blocks is reserved for an SFN such that the overall performance remains consistent even if a mobile device moves from one cell to another. This principle often eliminates interference and allows mobile devices to receive the combined signal which may result in better reception [13]. In an LTE network, the network system is usually divided into a number of areas called eMBMS Single Frequency Networks (MBSFNs). In each area, an on-demand video stream can be simultaneously multicast to multiple mobile devices. It is possible to define up to 256 MBSFN areas at the same time where each of the base stations is capable of joining a maximum of 8 areas.

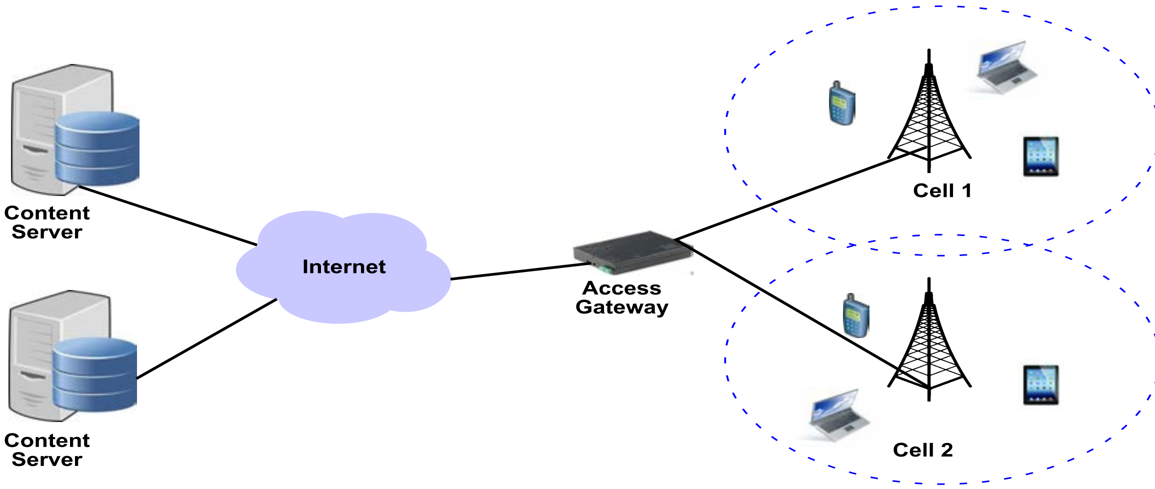

Figure 2.1: Architecture of an on-demand video system over mobile networks.

2.1.7 On-Demand Video Service

A typical framework for an on-demand video service over mobile networks is shown in Figure 2.1. It consists of four main components: (i) content servers, (ii) access gateways, (iii) cellular base stations, and (iv) groups of interested mobile users. Content servers are managed by providers, which may be, e.g., television providers, online video rental stores, multimedia e-learning centers, and other entertainment websites. When these content servers are located at remote places, their videos are assumed to be delivered to a designated cellular base station through high-capacity internet connections as shown in Figure 2.1. These physical links are over-provisioned and capable of carrying the desired video streams from the content provider to the access gateway. The second component in the on-demand video service is the access gateway which is directly connected to the cellular base stations using high-capacity wireless channels or optical fiber cables. The third component, cellular base stations are installed at fixed locations such that every receiving mobile user within their transmission coverage is continuously served. Finally, interested mobile users are the users requesting different video services.

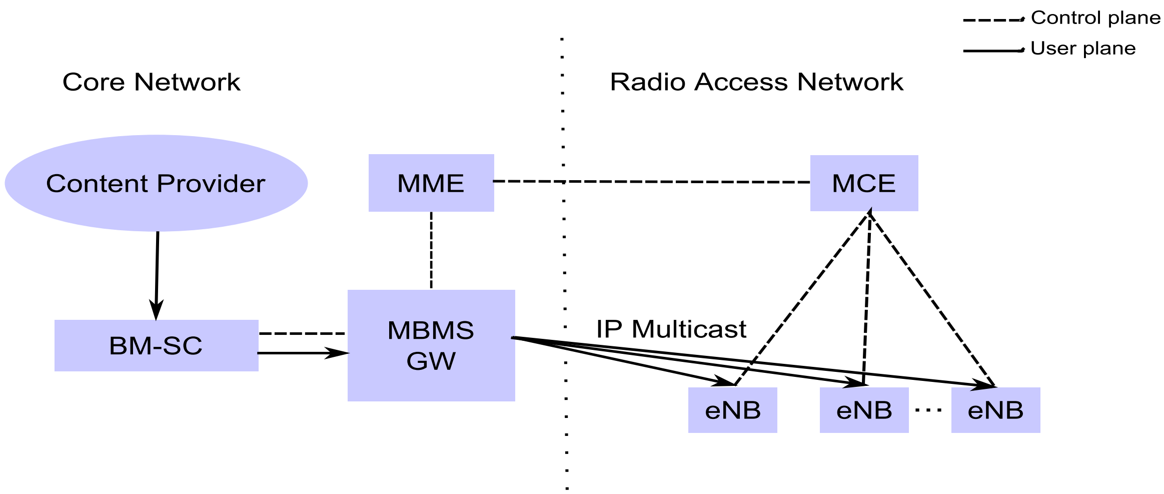

Figure 2.2: LTE eMBMS architecture.

2.1.8 Reasons behind Choosing LTE as an Example Network

Two of the key features required by our algorithms from the network perspective are: (i) multicast support and (ii) Single Frequency Network formation support. The multicast support is provided in LTE through eMBMS. eMBMS has been initially introduced in the 3GPP Release 8. The performance of eMBMS has been improved in two aspects: (i) minimizing the energy consumption of mobile devices and (ii) providing a good coverage within the network. To increase the energy saving on mobile devices, eMBMS usually follows a scheduling algorithm in which videos are transmitted as short bursts with high bit rates which is also a desired feature of our algorithm to ensure sufficient idle periods for mobile devices. The coverage in a mobile network is often restricted by the link quality of the worst-case mobile device in the transmission range. To enhance the reception of interested clients, eMBMS normally operates on a single frequency across a group of cells, known as a Single Frequency Network. Hence, LTE supports both of the key features demanded by our algorithms along with some desired features.

As shown in Figure 2.2, eMBMS consists of four network elements: (i) Broadcast/Multicast Service Centre, (ii) MBMS Gateway (MBMS-GW), (iii) Mobility Management Entity (MME), and (iv) Multicell/multicast Coordination Entity (MCE) [13]. Broadcast/Multicast Service Centre (BMSC), 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 flows through the core network. MBMS-GW is considered as a logical entity that helps in multicasting any IP packet generated from Broadcast/Multicast Service Centre to the entire LTE base stations located in a certain MBSFN area. Moreover, MBMS-GW handles further session control signaling via the MME, which is a main entity for the LTE access network since it plays an important role in performing a number of controlling procedures such as: mobile device's tracking, paging, and bearer activating. Finally, MCE ensures the full functionality of an MBSFN 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.

2.2 Related Work

2.2.1 On-demand Streaming Services

There are several techniques to enable on-demand streaming services in *wired* networks in the literature [1, 2, 25, 23, 46, 21, 18, 7]. These techniques can be broadly classified into three main categories: *piggyback designs*, *periodic broadcasting techniques*, and *patching methods*. Piggyback designs adjust the playout frame rates of multiple unicast sessions of the same video stream, in order to gradually align their streaming time and merge them into a multicast group [1, 2]. The merging process takes place while the streams are in progress without incurring any additional latency for the users [17]. Thus, piggyback designs can effectively improve resource utilization without increasing the latency. Periodic broadcasting techniques divide each video into a number of segments where every segment is repeatedly broadcast over a single communication channel [25, 23, 46]. Although these techniques make effective use of the bandwidth, they can cause noticeable start-up delay [22]. Patching methods allow receivers to immediately join an existing multicast session, while the missed portions are transmitted via temporary unicast connections [21, 18, 7]. It requires that the receiver receives video from both unicast and multicast sessions concurrently and buffers the video from the multicast session [21]. Since, the receiver can start playback of the video immediately from the unicast session, it does not cause any additional start-up delay. These approaches are designed for wired networks, and do not take network load of a shared air medium and energy saving of mobile devices into consideration. Our proposed solution is more general, as our formulation allows mobile devices to request chunks of videos at different times. Hence, piggyback, periodic broadcasting, and patching are all special cases of our optimization problem.

There have been some studies considering on-demand streaming services over mobile networks, but in a more restricted sense. For example, Hillestad et al. [20] propose an adaptive algorithm for streaming scalable on-demand videos over fixed WiMAX networks. However, the authors do not exploit multicast to reduce the network load. Majumdar et al. [35] propose a multimedia streaming approach leveraging both Forward Error Correction (FEC) and Automatic Repeat ReQuest (ARQ). The appropriate parameters of source and channel coding are determined so that the overall transfer rate is maximized. They also present an algorithm for multicasting scenarios, which employs FEC only as ARQ is less applicable in multicast. Paul et al. [32] also use error control schemes combining ARQ and FEC to conserve power. Hlavacs and Buchinger [22] propose a patching system for mobile networks. Their system may suffer from low energy saving as eventually all videos are multicast. Yoon et al. [49] concentrate on the implementation details of a multicast video service in LTE networks. The aforementioned studies [20, 35, 29, 22, 49] do not consider energy conservation, which is crucial to battery-powered mobile devices with stringent energy constraints.

2.2.2 Energy Saving on Mobile Devices

Tremendous research efforts are devoted to save energy on mobile devices. A number of methods of extending battery lifetime within the operating system and middleware layer exists in the literature. One of the earliest works in this area is called STPM [3] which proposes a self-tuning operating system module that adapts itself to the network access patterns and intent of applications. STPM allows the applications to disclose hints about their intent in using the wireless network. This allows STPM to enable power management only when appropriate, and to decide if the network interface can be disabled for periods longer than the beacon period [37]. *Beacon messages* are periodically generated messages used for power saving modes and the interval between two successive beacon messages is called the *beacon period* or *beacon interval*. If applications can specify the maximum delay that it can tolerate on incoming packets, STPM can make power management decisions based on these constraints to ensure that the applications do not suffer. STPM also allows applications to disclose the start and end periods of each transfer. Another approach, SAPSM [39] tags applications with a priority. This allows the network interface to switch to constant awake mode which is normal mode of operation for mobile devices before the introduction of power saving mode, only when a high priority application has a network activity, and staying at power saving mode in the presence of a low priority application. E-Mili [52] observes that a majority of energy is just spent idle listening and hence adaptively downclocks the network radio to reduce the amount of energy consumed during idle listening. These approaches of power management is done via changes to the operating system and applications on the mobile devices, whereas our approach involves scheduling in the wireless networks. Hence, these approaches can work in combination with our proposed solution.

Some approaches manipulate the medium access control (MAC) layer to save energy for the mobile devices. As an example, the primary design goal of Energy Conserving- Medium Access Control (EC-MAC) protocol [9, 45] is to achieve energy efficiency. The EC-MAC protocol considers an infrastructure network with a single base station serving a number of mobile devices within its coverage area. The base station broadcasts a schedule message containing slot permissions for the subsequent data phase with the intent to avoid collision. The gain in energy saving is mainly achieved by reducing the number of retransmission through collision avoidance over wireless channel. Moreover, the base station may optimize the transmission schedule to allow individual mobile devices transmit and receive within contiguous transmission slots.

Zhu and Cao [56] present a novel service model for streaming applications over wireless networks in order to conserve power. A new scheduling algorithm, called rate-based bulk scheduling (RBS) is designed for the base station to determine data flows to be served at different times. The concept of proxy server is employed to buffer data for the mobile devices so that the wireless network interface can sleep for a long time period to save power. Moreover, a novel adaptive technique is presented in order to adjust the sleep time of the wireless network interface according to the channel condition of the mobile device. The authors analytically prove that it is more power efficient than other rate-based fair queuing algorithms, and experimentally show that it significantly reduces the power consumption.

Transmission power management can affect battery life in mobile devices. For example, increasing the transmission power can result in higher throughput. However, this may cause increased interference for other users, and may turn out to be inefficient use of the available battery energy. Luna et al. [34] propose an approach in order to transmit a video frame using the minimal required transmission energy under delay and quality constraints. The authors consider selection of source coding parameters jointly with transmitter power and rate adaptation to reduce power consumption.

The power consumption in cellular radios depends on the signal strength. If the signal is weak, it consumes more power and suffers reduced data rate. According to Bartendr [44], the communication energy per bit can be as much as 6x higher when the signal is weak compared to that of when it is strong. Hence in order to save energy, applications should communicate when the signal is strong, either by deferring non-urgent communication or by advancing anticipated communication to coincide with periods of strong signal. Such scheduling for applications, requires predicting signal strength, in order to anticipate opportunities for energy-efficient communication. Moreover, the prediction mechanism should cost little energy. The realization of this concept leads to Bartendr [44], a practical approach to energy-aware scheduling algorithm. Bartendr establishes the relationship between signal strength and power consumption using measurements. Zorzi and Rao [57] also suggest to avoid transmission when the channel condition is poor from the link layer perspective to save energy. These approaches try to reduce energy consumption of mobile devices from different perspectives, and are different from the approach proposed in this work.

2.2.3 Multicast and Hybrid Unicast/Multicast Transmissions

The concept of *multicast* in wireless network is explored by a number of research works mostly due to the benefit of efficient utilization of the wireless link. C. Chiang [11] addresses multicast in his thesis, but he considers *multihop* wireless network instead of the conventional *single hop* cellular networks. However, he does not point out the power saving issue of the mobile devices. Hartung et al. [19] explore the mobile broadcast services from the end user perspective and the way broadcast services are delivered in 3G networks. The authors also address the broadcast service in the new wireless standard like LTE, and suggest a hybrid unicast/broadcast transport layer integration to improve the end user performance. But they do not address the power saving aspect of the mobile devices. GLBM [53] considers multicast communication in wireless mesh network and proposes a multicast algorithm which improves multicast performance in terms of delay, jitter and throughput. Nguyen et al. [36] present network coding based scheduling policy in order to increase the bandwidth efficiency of multicast and unicast sessions in a wireless network. At the same time, they propose Markov Decision Process based scheduling algorithm to maximize the quality of multimedia applications. But energy saving of the mobile devices is missing from the performance metrics.

Research considering a mixture of unicast and multicast services is still nascent. Only a few research efforts have been dedicated towards this approach. For example, Wang et el. [47] consider mixed unicast-multicast services in LTE networks and attempt to minimize the radio resource demands for the overloaded cell. Single Frequency Network (SFN) for multicast service improves the quality of service at the cost of more radio resource consumption than simple point-to-multipoint (PTM) multicast without SFN. Hence, for each multicast service in the cell, the authors select between SFN and PTM for the multicast service to reduce radio resource demands keeping quality of service as good as possible. For minimizing the resource, the problem is trivial: select SFN for multicast services common in all cells in the SFN area and PTM for the remaining. So, they manipulate the set of multicast services that are not common in all cells (hence, specific to the overloaded cell) in the SFN area to decide whether to choose SFN or PTM for each of them to minimize resource demands and improve quality of service. They consider the resource demands for all unicast services in the cell to be constant whereas in our case, the resource distribution between unicast and multicast is done dynamically. Moreover, our optimization goal is to minimize energy consumption for the mobile devices which is also different from them. Lee et al. [29] describe a scheme that uses both unicast and multicast communications to reduce service blocking probability and bandwidth consumption. Different from our work, they do not take energy consumption into consideration.

2.2.4 Cooperative Wireless Networks

Cooperative wireless network is defined as the wireless network where the mobile users may improve their effective quality of service in a cooperative manner and each user is capable of transmitting data as a cooperative agent for another user [38]. Khalek et al. [26] address hybrid unicast/multicast streaming to minimize the energy consumption for the mobile devices. But the authors consider multihop cooperative wireless network where each mobile user is capable of transmitting and relaying video streams along with receiving, instead of traditional single hop cellular networks. Fitzek et al. [15] discuss the energy saving strategies along with security in a cooperative wireless networks and suggests the use of cooperation in wireless communication. Cooperative retransmission strategy [51] is proposed in a P2P wireless network to reduce energy consumption of the mobile devices. The reduced energy consumption is achieved by employing effective error/loss recovery scheme for reliable multicast where a mobile device can recover error by cooperative retransmission from another peer in its proximity. Ramadan et al. [40] describe a power saving approach for the mobile devices in a cooperative wireless network where the wireless access point distributes the video packets among the mobile devices instead of sending the stream independently to each mobile device, and then mobile devices exchange the received packets among themselves over short-range wireless links using bluetooth technology. Zhu and Mutka [55] and Ananthanarayanan et al. [4] also consider the P2P wireless network in a cooperative manner to reduce energy consumption for the mobile devices.

In summary, our work leverages both unicast and multicast in mobile networks to maximize the average energy saving across all mobile devices under a given bandwidth constraint. To the best of our knowledge, this problem has not been rigorously investigated in the literature.

Chapter 3

Problem Statement and Proposed Solution

In this chapter, we describe the system model considered for our problem. Then we state the problem and find out the hardness with proof. Finally, we mathematically formulate the problem and propose near-optimal greedy algorithm to efficiently solve the problem.

3.1 System Model

Our on-demand video streaming system consists of base stations, mobile devices and resource allocators as illustrated in Figure 3.1. Each resource allocator is responsible for making the resource allocation decisions for a single base station or a set of adjacent base stations forming a Single Frequency Network. Mobile devices arrive into the system asynchronously. The video streaming requests from the mobile devices are forwarded to the responsible resource allocator. The requests can be driven by instantaneous demands of the mobile users or by any prefeching system running inside the mobile device to prefetch video segments most likely to be played out in near future [54, 48]. We assume the requests are initiated by mobile users. So user inputs like delay, VCR functionalities like fast forward, rewind or jumping to a random video position can be supported which enables our system to be applicable for diverse applications including on-demand video streaming, live or time-shifted sports or other popular events, and mobile personal video recorders. Based on the incoming video streaming requests from the mobile devices, the resource allocator periodically solves the resource allocation optimization problem leveraging both unicast and multicast to: (i) maximize

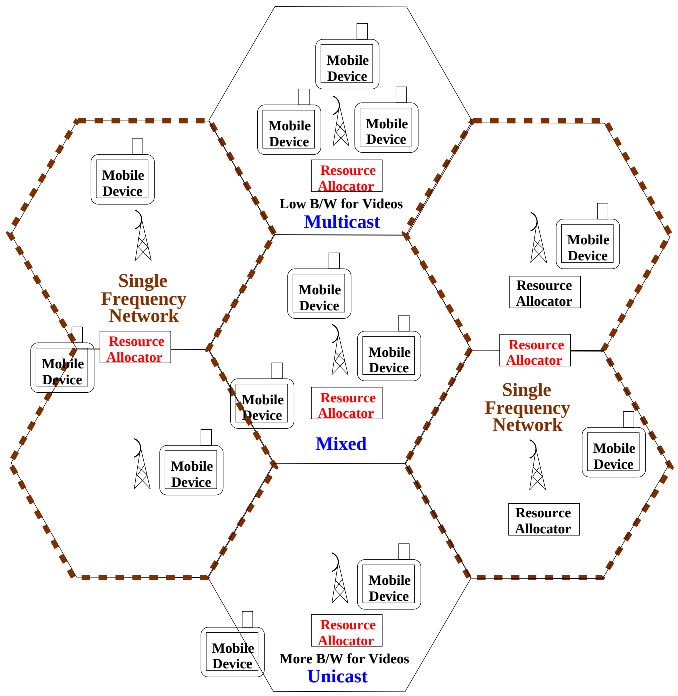

Figure 3.1: The considered resource allocation problem in hybrid cellular networks.

the average energy saving across all mobile devices, (ii) consume only the network resource allotted for video streaming, and (iii) ensure smooth playout on all mobile devices. The resource allocators send the solution of the optimization problem to the base stations to stream videos accordingly.

For simplicity of discussion, we assume the mobile devices within the same cell in Figure 3.1 can be grouped together into a multicast session. We also assume that the channel condition depends on the distance between the mobile device and base station which is realistic. So, the mobile devices closer to the base station have better channel conditions and vice versa. We use Figure 3.1 to demonstrate the generality of the considered problem in two dimensions. First, we consider both normal wireless network and Single Frequency Network. In normal network, the resource allocator solves the optimization problem for a single base station while for a Single Frequency Network, it solves the optimization problem for the set of base stations forming the Single Frequency Network. It is essential for a Single Frequency Network that the resource allocation decisions are made by one allocator considering all the composing base stations to find out the optimal allocations. Further

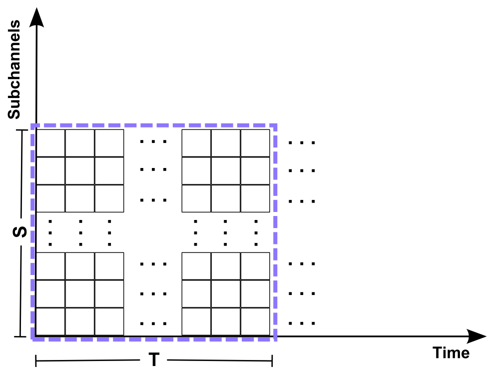

Figure 3.2: An allocation window with T symbols and S subchannels represented by the blue dotted square.

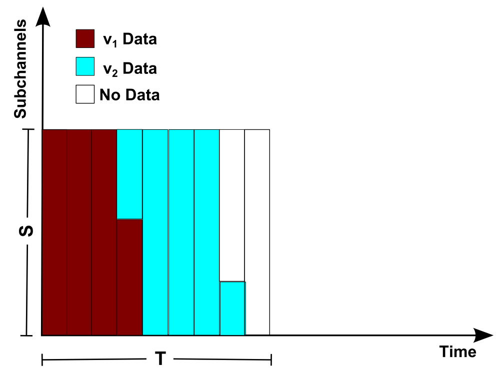

Figure 3.3: Resource blocks are allocated along the subchannels first as shown for an allocation window with T symbols and S subchannels.

details about the generalization of the problem for Single Frequency Networks are described in Section 3.5. Here we consider the simple case where the allocator is responsible for one base station. Second, based on: (i) channel conditions of individual mobile devices and (ii) reserved bandwidth for video streaming services, resource allocators may stream videos over multicast, unicast, or a mixture of both. For example, the top cell in Figure 3.1 contains mobile devices with analogous channel conditions and the reserved bandwidth for video streaming services is low. This situation dictates that the resource allocator may transmit video over a single multicast stream in this cell. On the other hand, the bottom cell has mobile devices with diverse channel conditions and has sufficient bandwidth reserved for on-demand video streaming services. This condition indicates that resource allocator may stream video to these mobile devices over unicast to maximize the average energy saving. However, the middle cell contains two mobile devices close to the base station with analogous channel conditions and the one towards the cell boundary having different channel condition. It has reasonable amount of resources reserved for on-demand video streaming services and this scenario refers that the allocator may decide to stream video to these mobile devices using a mix of both unicast and multicast.

Next generation wireless networks adopt OFDMA modulation scheme where radio resources are divided along both time and frequency domains [50]. We consider an *allocation window* with T columns of *symbols* and S rows of *subchannels* as shown in Figure 3.2. A column $t \in [1, T]$

indicates radio resources along the time domain whereas a row $s \in [1, S]$ indicates resources along the frequency domain. A *resource block* [1], uniquely identified by the pair (t, s) is the minimum allocation unit in the network and is represented by the small squares in Figure 3.2. We assume a fraction d of network resources is reserved for video streaming services. d is dynamic and adjustable based on network loads. For example, if the cell experiences more voice calls it can reduce d and vice versa. Hence, the resource allocator distributes dTS blocks to optimally serve the incoming video requests in each allocation window. The solution computed at the beginning of one allocation window can be reused for several consecutive allocation windows unless the channel conditions of the mobile devices change, or the population of mobile devices change due to new arrivals into the system or departures from the current system. The length of the allocation window is determined by the parameter T and affects the allocation decisions. Larger value of T indicates longer allocation window which provides more flexibility and more opportunity to improve. But it may cause longer *service delay* for mobile devices arrive towards the beginning of the allocation windows as allocation decisions are made at the beginning of an allocation window, and users have to wait for services until beginning of the next allocation window. On the other hand, smaller value of T ensures shorter service delay, but provides less flexibility to the resource allocator. In case of true on-demand video streaming with real time constraint on the service delay, a *patching* solution [21, 18, 7] can be adopted. A threshold can be used for a new request to join an on-going multicast session of video and at the same time a temporary unicast session is created to serve the earlier segments of the video to the requesting mobile user. The new request is considered at the beginning of next allocation window and potentially be merged with other requests to form a multicast session.

We assume there are V videos in the system and the network supports M different MCS modes. We also make the assumption that a video, v is divided into Z_v consecutive parts in the length of the allocation window. For a particular video v, the amount of data needs to be transmitted for smooth playout in each allocation window is qTr_v where q represents the symbol time and r_v represents the encoding rate of the video. However, the capacity of each resource block depends on the MCS mode of the requesting mobile device. If the channel quality of the mobile device is good, more aggressive MCS mode can be used which results in high capacity and vice versa. Let $m \in [1, M]$ is the maximum MCS mode that can be used for the mobile device and c_m denotes the corresponding capacity. Then the number of resource blocks required in an allocation window for this mobile user is $\lceil qTr_v/c_m \rceil$. The required number of blocks along with how the blocks are allocated in each

[1] We interchangeably use resource block and block to refer to the resource allocation unit throughout the thesis.

allocation window determines the time a mobile device can turn off its network interface and thus the *energy saving* for the mobile device. We assume that the base station allocates the resource blocks along the subchannel first, and when all the subchannels in a column finishes, only then moves to the next column to maximize the energy saving for the mobile device.

3.2 Problem Statement and Hardness

3.2.1 Problem Statement

Each cell in the wireless network can support a certain amount of resource blocks based on the downlink transmission bandwidth and a fraction of these resource blocks are reserved for on-demand video services. A request from a user can be satisfied by a unicast or multicast streaming based on availability of resources, channel conditions of the requesting user along with other factors. We are interested in finding out a resource block assignment to serve these users through a mixture of unicast and multicast streaming in order to maximize the average energy saving for the requesting mobile users.

Problem 1:Radio Resource Allocation. *Given a number of mobile users with different channel conditions requesting various on-demand videos and a certain fraction of resource blocks reserved to serve these requests. Find the optimal resource allocation considering both unicast and multicast streaming such that: (i) the average energy saving across all mobile devices is maximized (ii) the number of resource blocks consumed does not exceed the number of resource blocks reserved and (iii) all mobile devices served with a certain video receive that video at the same rate for smooth playout.*

3.2.2 Hardness of the Problem

Lemma. *The resource allocation problem in Problem 1 is NP-Complete.*

Proof. Let us define the set of decision variables $a_{i,j,k}$ for all $i \in [1, S], j \in [1, T], k \in [1, V]$ for resource block allocation such that $a_{i,j,k} = 1$ if the resource block in i^{th} subchannel and j^{th} symbol is allocated to video k, 0 otherwise.

Given an allocation assignment represented by the above variables, we can 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 to check that for any $i \in [1, S], j \in [1, T],\ \sum_{k=1}^{V} a_{i,j,k} \leq 1$. These can be done in polynomial time. So, Problem 1 is in NP.

In order to prove that an NP-Complete instance is reducible to our problem, we will use the 0-1 knapsack problem. The 0-1 knapsack problem is defined as: there are n items x_l such that $l \in [1, n]$ and $x_l = 1$ if the item is chosen, 0 otherwise. The value and weight of item x_l are v_l and w_l respectively. The capacity of the knapsack is W. We assume non-negative values and weights. We have to choose among these n items to maximize the value without violating the capacity of the knapsack. Mathematically:
Maximize $\sum_{l=1}^{n} x_l v_l$ subject to $\sum_{l=1}^{n} x_l w_l \leq W,\ x_l \in \{0, 1\}$

In order to reduce this instance of 0-1 knapsack to an instance of our problem, we set $n = TSV$. We choose $l \in [1, n]$ in a way such that for each combination of (i, j, k) where $i \in [1, S], j \in [1, T], k \in [1, V]$, l has a unique value, and vice versa. We define a new variable $x'_{i,j,k}$ for any $i \in [1, S], j \in [1, T], k \in [1, V]$ such that $x'_{i,j,k} = 1$ if resource block in i^{th} subchannel and j^{th} symbol is allocated to video k, 0 otherwise. For any $l \in [1, n]$ and corresponding combination (i, j, k) where $i \in [1, S], j \in [1, T], k \in [1, V]$, we set $x'_{i,j,k} = x_l$. Hence, if an item x_l is included in the 0-1 knapsack, the associated resource block in i^{th} subchannel and j^{th} symbol is allocated to video k, and vice versa. w_l in 0-1 knapsack is a direct representative of weight in our problem, we set $w_l = 1$ for any $l \in [1, n]$. v_l for any $l \in [1, n]$ represents the energy saving of the mobile devices achieved by allocating resource block in i^{th} subchannel and j^{th} symbol to video k for corresponding (i, j, k). Finally, we set knapsack capacity, $W = dTS$.

This reduction can be done in polynomial time. We claim that the reduced 0-1 knapsack problem will have a solution iff our considered problem has a solution which can easily be verified.

Hence, the resource allocation problem in Problem 1 is NP-Complete.

□

3.3 Mathematical Formulation

Next we concentrate on the mathematical formulation of our problem where we make decision at the resource block level. For each of the resource blocks reserved for on-demand video streaming we determine (i) video segment assignment (ii) whether to use unicast or multicast and (iii) MCS mode to maximize the average power saving across all mobile devices with guarantee of smooth playout. Let us define a boolean decision variable $x_{v,m,z}$ ($v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$) to denote

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

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

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

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

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

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

---

whether the segment z of video v is unicast/multicast using MCS mode m. If segment z of a specific video v is transmitted using MCS mode m, then $x_{v,m,z} = 1$, otherwise $x_{v,m,z} = 0$. If $x_{v,m,z} = 1$, then depending on the number of mobile devices that can watch segment z of video v with maximum MCS mode m, we can decide whether to unicast or multicast. Let $w_{v,m,z}$ ($v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$) denotes the number mobile users watching segment z of video v with a maximum MCS mode m. If $w_{v,m,z} > 1$ we need to multicast, otherwise we use unicast. Note that, we never transmit a segment of a video using a specific MCS mode, if there are no mobile devices requesting that segment in the first place, and in the second place no mobile devices are capable of receiving it with that MCS mode. This leads to the definition of an intermediate boolean variable $y_{v,m,n,z}$ for each $v \in [1, V]$, $m, n \in [1, M]$, $n \leq m$, $z \in [1, Z_v]$. If a mobile device with maximum MCS mode m would receive segment z of video v using MCS mode n with $n \leq m$, then $y_{v,m,n,z} = 1$, otherwise $y_{v,m,n,z} = 0$. The variable $y_{v,m,n,z}$ can be determined by $x_{v,m',z}$, $m' \in [n, m]$ as follows:

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

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

We present our mathematical formulation in Eq. (3.3). $\mathbf{x}$ is the vector representation of the variable, $x_{v,m,z}$ for $v \in [1, V], m \in [1, M], z \in [1, Z_v]$. The objective function in Eq. (3.3a) is to maximize the average energy saving across all mobile devices. The size of a video v that needs to be transmitted in an allocation window is qTr_v, and the minimum number of symbols is $\lceil \frac{\lceil \frac{qTr_v}{c_m} \rceil}{S} \rceil$, where m is the MCS mode. The constraint in Eq. (3.3b) guarantees that the resource allocator does not allocate more than d fraction of network resources reserved for on-demand video streaming services. The constraint in Eq. (3.3c) ensures that every mobile device receives its allocation window with only one feasible MCS mode to smoothly render the video. Finally, the constraints in

Eqs. (3.3d) and (3.3e) are used to guide the computation of the intermediate variables, $y_{v,m,n,z}$.

3.4 Algorithms Proposed for a Single Cell

In this section, we consider the resource allocation in a single cell. Later on, we extend our solution to multicell setup.

3.4.1 SCOPT

The resource allocation algorithms run on the resource allocators responsible to determine how to stream videos in the base stations considering the mobile devices associated with the base stations to maximize the average energy saving of mobile devices as shown in Figure 3.1. The resource allocator is located in the base station. We propose SCOPT (Single-Cell OPTimum) to find out the optimal solution. We use CPLEX [64] and Matlab [68] in the implementation of the optimal algorithm. SCOPT considers all possible combinations of the multicast/unicast streaming to find out the optimal solution. Hence, SCOPT has exponential running time, in the worst-case.

3.4.2 SCG

Realizing the exponential running time of SCOPT, we develop a greedy algorithm, called SCG (Single-Cell Greedy) to find out the near-optimal solution with realistic running time. We start with the ideal case where we try to serve all the mobile devices using individual unicast sessions to *maximize* the average energy saving. But this requires excessive amount of resources and may not be feasible due to the constraint in Eq. (3.3b) on available amount of resources for on-demand video streaming services. To convert this infeasible solution into a feasible one, we iteratively reduce the number of unicast/multicast streams by removing one stream at a time until the solution becomes feasible. The selection of stream for removal at each iteration is decided based on some profit and cost analysis. In order to remove a stream, we reset $x_{v,m,z}$ to 0 and re-compute the required number of resource blocks for the current solution to ensure that the constraint in Eq. (3.3b) is not violated. For example, if we reset $x_{1,3}$ [2] to 0, it reduces the network load for the on-demand streaming service by $\lceil \frac{qTr_1}{c_3} \rceil$ resource blocks. But this leads to a negative impact on average energy saving. The mobile devices receiving video v with MCS mode 3, have to receive at a lower MCS mode which reduces

[2] When we are not interested in the segment number of a video v, we use $x_{v,m}$ instead of $x_{v,m,z}$ throughout the thesis.

the energy saving γ in Eq. (3.3a) for those mobile devices. This illustrative example explains the trade-off between *profit* (Eq. (3.3b)) and *cost* (Eq. (3.3a)) of removing a particular streaming.

Let $\alpha_{v,m,z}$ and $\beta_{v,m}$ be the *offsets* of profit and cost to reset $x_{v,m,z}$ to 0 in an allocation window. Mathematically, we define $\alpha_{v,m,z} = \sum_{m'=m}^{M} w_{v,m',z} y_{v,m',m,z} \lceil \lceil \frac{qTr_v}{c_m} \rceil \Big/ S \rceil$ and $\beta_{v,m} = \lceil \frac{qTr_v}{c_m} \rceil$. Our greedy algorithm iteratively *refines* an infeasible allocation by removing the MCS mode with the minimum ratio of profit reduction to cost reduction. In particular, our algorithm evaluates the ratio $\tau_{v,m,z} = \alpha_{v,m,z}/\beta_{v,m}$ for all $x_{v,m,z} = 1$ and drops the stream for segment z of video v in MCS mode m with the smallest $\tau_{v,m,z}$ value in each iteration. Our algorithm stops once a feasible solution is produced i.e. the constraint in Eq. (3.3b) is satisfied. The pseudocode of SCG is presented in Figure 3.4. The inputs to the algorithm are fraction of resources reserved for on-demand video streaming services (d), length of the allocation window in seconds (T), number of subchannels (S), number of videos (V), number of allowed MCS modes (M), segment information of the videos ($\mathbf{Z}$), number of users watching different videos with various MCS modes ($\mathbf{w}$), the symbol time (q), encoding rates of the videos ($\mathbf{r}$) and data capacities of various MCS modes ($\mathbf{c}$). The algorithm outputs the video segments to be streamed with associated MCS modes ($\mathbf{x}$).

3.4.3 Correctness and Complexity

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

Proof. The while-loop starts from line 6 ensures the feasibility of the produced solution by satisfying constraint in Eq. (3.3b). Moreover, in each iteration it removes the least profitable streaming ensuring that the algorithm is not trapped into an infinite loop. Hence, the algorithm is correct. The dominating computational complexity of the algorithm occurs in lines 6–8: (i) the while-loop starts from line 6 iterates at most VMZ times, (ii) the for-loop starts from line 7 iterates VMZ times in the worst-case, and (iii) line 8 updates at most M $y_{v,m,n,z}$ values. Combining all these computational complexity we can easily find that the time complexity of the SCG algorithm is $O(V^2 M^3 Z^2)$. □

We note that for real networks, V, M, Z are not large numbers and the complexity does not depend on the number of users, which can be large. For example, the maximum number of videos that can be concurrently streamed on the most recent LTE network is 23 [58], assuming average video bit rate of 1736 Kbps [71] and maximum wireless bandwidth of 20 MHz [13]. Similarly, the largest value for M is 28 [58], and for Z is 5 [28] assuming an allocation window of 10 seconds.

Algorithm: SCG
Inputs: $d, T, S, V, M, \mathbf{Z}, \mathbf{w}, q, \mathbf{r}, \mathbf{c}$
Output: $\mathbf{x}$

foreach $v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$
  initialize $x_{v,m,z} = 1$ if $w_{v,m,z} > 0$; $x_{v,m,z} = 0$ otherwise
let $\Delta = dTS - \sum_{v=1}^{V} \sum_{m=1}^{M} \sum_{z=1}^{Z_v} x_{v,m,z} \lceil \frac{qTr_v}{c_m} \rceil$
foreach $v \in [1, V]$, $m \in [1, M]$, $n \in [1, m]$, $z \in [1, Z_v]$
  compute $y_{v,m,n,z}$ using Eqs. (3.3d) and (3.3e)
while $\Delta > 0$
  foreach $v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$,
  where $x_{v,m,z} = 1$
    update $y_{v,m,n,z}$, and **compute** $\alpha_{v,m,z}$, $\beta_{v,m}$
    and $\tau_{v,m,z}$ as described in section 3.4.2
  let v^*, m^* z^* lead to the minimum τ_{v^*,m^*,z^*} considering
  all possible combinations for v, m, z
  let $x_{v^*,m^*,z^*} = 0$
  let $\Delta = \Delta - \alpha_{v^*,m^*,z^*}$
return $\mathbf{x}$

Figure 3.4: The proposed algorithm to solve the single-cell allocation problem.

Moreover, all computations are simple scalar operations. Thus, the algorithm can easily run in real time.

3.5 Supporting Single Frequency Networks

We extend our problem space to consider Single Frequency Networks. A Single Frequency Network is a broadcast network of multiple base stations with many advantages like less interference, improved hand-off. Due to the benefits offered by a Single Frequency Network, in this section, we discuss our problem formulation in the context of a Single Frequency Network.

3.5.1 Extended Formulation

The formulation in Eq. (3.3) assumes a single cell. In real deployments, Single Frequency Networks help to improve the Signal to Interference plus Noise Ratio (SINR) for the receiving mobile devices. This results in better channel conditions for some of the mobile devices enabling them to receive at higher MCS modes than single cell setup. These mobile devices can then turn off their network interfaces for longer period of time which increases the average energy saving. Hence, we extend our problem formulation for Single Frequency Networks. We consider a *dynamic* Single Frequency Network with H hexagonal cells where each block can independently be assigned to a Single Frequency Network. The extension to Single Frequency Networks requires two major enhancements: (i) expanding the solution space to multiple cells and (ii) modeling Single Frequency Network gains from neighboring cells. Both of them are explained below.

Expanding solution space. We concurrently consider H cells, and use a *superscript* h ($h \in [1, H]$) to variables to indicate association with cell h whenever applicable. For example, N_v^h represents the number of mobile devices in cell h ($h \in [1, H]$) watching video v ($v \in [1, V]$). $x^h_{v,m,z}$ ($v \in [1, V]$, $m \in [1, M]$, $z \in [1, Z_v]$, and $h \in [1, H]$) is the decision variable in the extended formulation. The superscript gives us the flexibility to expand the solution space for all H cells.

Modeling Single Frequency Network gains. $w_{v,m,z}$ 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 ensure a certain bit error rate, say $< 5\%$. The actual MCS adaptation algorithm resides in the link layer, and is out of the scope of this thesis. Without loss of generality, we model the Single Frequency Network gain of mobile devices watching segment z ($z \in [1, Z_v]$) of video v ($v \in [1, V]$) with maximum MCS mode m ($m \in [1, M]$), from cell h' ($h' \in [1, H]$) to cell h ($h \in [1, H]$, $h \neq h'$) by $\delta^{h,h'}_{v,m,z}$, which represents the number of more/fewer mobile devices in h that have maximum MCS mode m if cell h' would transmit segment 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 can be written as: $\hat{w}^h_{v,m,z} = w^h_{v,m,z} + \sum_{h' \in [1,H] \setminus \{h\}} x^{h'}_{v,m,z} \delta^{h,h'}_{v,m,z}$.

Adoption to these two enhancements results in a formulation for a Single Frequency Network in Eq. (3.4). $\mathbf{x}$ is the vector representation of the variable, $x^h_{v,m,z}$ for $v \in [1, V], m \in [1, M], z \in [1, Z_v], h \in [1, H]$. The objective function in Eq. (3.4a) is to maximize the average energy saving across all H cells. The constraint in Eq. (3.4b) ensures that each cell is not overloaded. The constraint in Eq. (3.4c) ensures that every mobile device receives at an MCS mode equal to or smaller than its maximum MCS mode. The constraints in Eqs. (3.4d) and (3.4e) relate variables $y^h_{v,m,n,z}$ and $x^h_{v,m,z}$, and can be used as a guideline to compute the intermediate variables $y^h_{v,m,n,z}$. The constraint in Eq. (3.4f) takes the Single Frequency Network gains into consideration.

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

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

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

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

$$y^{h}_{v,m,n,z} \leq x^{h}_{v,n,z}, \tag{3.4e}$$

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

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

3.5.2 Algorithms Proposed for a Single Frequency Network

We propose SFNOPT (Single Frequency Network OPTimum) to find out the optimal resource allocation for a Single Frequency Network. The combination of CPLEX [64] and Matlab [68] can be employed in the same way as SCOPT to find out the optimal solution. Since SFNOPT has an exponential running time, we also propose an efficient greedy algorithm, called SFNG (Single Frequency Network Greedy) algorithm to find out a near-optimal solution. We start with the best case where we try to serve all the mobile devices using individual unicast sessions to maximize the average energy saving. This requires excessive amount of resources which may not be feasible due to the constraint in Eq. (3.4b) on available amount of resources for on-demand video streaming services. In order to transform this infeasible solution into a feasible one, we iteratively reduce the network load of the cell $\hat{h} \in [1, H]$ that suffers from the largest excessive network load. We reduce the number of unicast/multicast streams by removing one stream at a time until the solution becomes feasible. The selection of a stream for removal at each iteration is decided based on the profit and cost analysis dictated by α, β and τ. In order to remove a stream, we reset $x^{\hat{h}}_{v^*,m^*,z^*}$ to 0 and re-compute the required number of resource blocks for the current solution to determine its feasibility. The SFNG algorithm terminates once a feasible allocation is produced. Figure 3.5 gives the pseudocode of SFNG. The inputs to the algorithm are fraction of resources reserved for on-demand video streaming services (d), length of the allocation window in seconds (T), number of subchannels (S), number of cells in the Single Frequency Network (H), number of videos (V), number of allowed MCS modes (M),

Algorithm: SFNG
Inputs: $d, T, S, H, V, M, \mathbf{Z}, \mathbf{w}, q, \mathbf{r}, \mathbf{c}$
Output: $\mathbf{x}$

1. **foreach** $h \in [1, H], v \in [1, V], m \in [1, M], z \in [1, Z_v]$
2. **initialize** $x^h_{v,m,z} = 1$ is $w^h_{v,m,z} > 0$; $x^h_{v,m,z} = 0$ otherwise
3. **let** $\Delta^h = dTS - \sum_{v=1}^{V} \sum_{m=1}^{M} \sum_{z=1}^{Z_v} x^h_{v,m,z} \lceil \frac{qTr_v}{c^h_m} \rceil, \quad \forall h \in [1, H]$
4. **foreach** $h \in [1, H], v \in [1, V], m \in [1, M]$,
4. $n \in [1, m]$, and $z \in [1, Z_v]$
5. **compute** $y^h_{v,m,n,z}$
6. **let** $\hat{h} = \text{argmax}_{h=1}^{H} \Delta^h$
7. **while** $\Delta^{\hat{h}} > 0$
8. **foreach** $v \in [1, V], m \in [1, M], z \in [1, Z_v]$,
8. $x^{\hat{h}}_{v,m,z} = 1$
9. **update** $y^{\hat{h}}_{v,m,n,z}$, and **compute** $\alpha^{\hat{h}}_{v,m,z}$, $\beta^{\hat{h}}_{v,m}$ and
9. $\tau^{\hat{h}}_{v,m,z}$ as described in section 3.4.2
10. **let** v^*, m^*, z^* lead to the minimum $\tau^{\hat{h}}_{v^*,m^*,z^*}$
10. considering all possible combinations for v, m, z
11. **let** $x^{\hat{h}}_{v^*,m^*,z^*} = 0$
12. **let** $\Delta^{\hat{h}} = \Delta^{\hat{h}} - \alpha^{\hat{h}}_{v^*,m^*,z^*}$
13. **let** $\hat{h} = \text{argmax}_{h=1}^{H} \Delta^h$
14. **return** $\mathbf{x}$

Figure 3.5: The proposed algorithm to solve the Single Frequency Network allocation problem.

segment information of the videos ($\mathbf{Z}$), number of users watching different videos with various MCS modes ($\mathbf{w}$), the symbol time (q), encoding rates of the videos ($\mathbf{r}$) and data capacities of various MCS modes ($\mathbf{c}$). The algorithm outputs the video segments to be streamed with associated MCS modes for each of the cells in the considered Single Frequency Network ($\mathbf{x}$).

3.5.3 Correctness and Complexity

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

Proof. The correctness comes from lines 6, 7, and 13, which ensures the algorithm only terminates when Eq. (3.4b) is satisfied for all the cells. The for-loop starts from line 4 has a computational complexity of $O(HVM^2Z)$. The while-loop starts from line 7 repeats at most $HVMZ$ times, the for-loop starts from line 8 iterates at most VMZ times, and the line 9 updates at most M $y^{\hat{h}}_{v,m,n}$ values. Thus, the total time complexity of SFNG is $O(HVM^2Z) + O(HV^2M^3Z^2) = O(HV^2M^3Z^2)$. □

3.6 Summary

We considered the problem of resource allocation for wireless networks following a hybrid multicast/unicast approach. We described the system model considered for our problem. Then, we modeled the problem to maximize average energy saving across all mobile devices and mathematically formulated the problem considering a single cell. Since the algorithm to find out the optimal solution is computationally expensive, we proposed an efficient greedy algorithm to find out a near-optimal solution. Then we extended our formulation for Single Frequency Networks because of the benefits offered by such networks. Finally, considering the computational complexity of the optimal solution, we proposed an efficient greedy algorithm to produce a near-optimal solution for such a network.

Chapter 4

Evaluation

In this chapter, we evaluate our proposed solution through extensive simulations. We describe the simulation tools used for the evaluation process along with the simulation setup for our experiments. Then we discuss different performance metrics employed to analyze different aspects of the solution. Finally, we present the simulation results and analyze them.

4.1 Setup

Simulator and Algorithms. We simulated an on-demand video streaming system using OPNET modeler and its associated OPNET LTE Specialized model [70]. OPNET consists of a suite of protocols and technologies in order to facilitate the test and demonstration of technology designs in realistic scenarios before production. OPNET provides models for all network types and technologies to analyze realistic simulated networks in order to compare the impacts of different technology designs on end-to-end behavior. We also use a mixture of C, MATLAB [68] and CPLEX [64] to compute the optimal solutions. Since SFNOPT runs too slowly for on-demand video streaming, we do not consider it in the simulation results. In order to evaluate the proposed solutions, we implemented unicast-only and multicast-only policies adopted in current systems. We consider the unicast-only policy where the users are served in order of arrivals. In case of multicast-only policy, we give priority to multicast groups formed earlier. We compare the results obtained by our solutions with them from different perspectives. We name the current unicast-only and multicast-only policies as $\mathrm{CUR_u}$ and $\mathrm{CUR_m}$ respectively throughout our simulations. We also employ some heuristics to make our simulation setup more realistic. For example, when a mobile device cannot be admitted due to resource limitations, it does not leave the system immediately. Rather it retries for the intended video upto three times with an exponential backoff starting at 2 seconds. After being rejected all these times, it finally stops requesting for the video. As an example of another heuristic, batching of requests is used such that all requests for on-demand videos within the duration of an *allocation window* are grouped together to be served at the beginning of the next *allocation window*. These heuristics improve the performance of the proposed solution from various aspects and are likely to be incorporated in real on-demand video streaming systems.

Wireless Network Configurations. Although the proposed algorithms are applicable to any wireless networks with multicast support, we use LTE Release-9 3GPP standard to evaluate the performance of the proposed algorithms. In Table 4.1, we list some LTE configuration parameters used for the simulations unless otherwise specified. The other parameters are set to the default values of the OPNET LTE module. We configure the LTE downlink with Evolved Packet System (EPS) bearers for the simulations. We define an EPS bearer as a transmission path of defined quality, capacity, delay, and so on [70]. LTE EPS bearer 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. 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. 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 or multiple bearers. For the assumed bearer configurations and MCS modes, depending on the channel conditions of the mobile devices: (i) MCS 4 to MCS 7 are served by the bearer of MCS 4, (ii) MCS 8 to MCS 13 are served by the bearer of MCS 8, (iii) MCS 14 to MCS 21 are served by the bearer of MCS 14, and (iv) MCS 22 to MCS 28 are served by the bearer of MCS 22. The simulator runs the resource allocation algorithm once every allocation window of 10 seconds. We report average results with 95% confidence intervals whenever applicable.

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. For the Single Frequency Network, we assume a service area of three cells with 20% overlapping at the cell boundary. 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. For most of our simulation scenarios, we assume that all mobile devices are quasi-static. However to understand the effect of mobility on the results, we also simulate with mobile devices

Table 4.1: 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
Modulation and Coding Scheme (MCS)	4, 8, 14, 22
Evolved Packet System Bearer for Uplink	Best effort
Propagation Model	Free space, Walfisch-Ikegami line of sight
Scheduling Mode	Link Adaptation Only

following different mobility models. 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 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.

We assume a population of 1000 Mobile *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. Mobile users are randomly distributed within the service area of each cell such that more users, about 90% of the total number of users, 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. We crawl YouTube to retrieve information of 1,000 videos. We use YouTube Dataset [10] and YouTube Data API [78] to serve our purpose. Since the videos are 240p, we scale up the bit rates by a factor of 9 to emulate videos of 720p which are popular in modern smartphones. The popularity, bit rate and length of these videos drive our simulations. We sort the videos in terms of their popularity. In order to imitate realistic video requesting behavior, mobile users request videos from this list following the Zipf distribution with a skewness parameter α. We set $\alpha = 1.5$ for our simulations. But this can be tuned to any desired value to cover a wide range of user behaviors.

Multicast Session Switch Time. We define the *multicast session switch time* for a mobile user currently participating in a multicast session, as the time difference between the time the user sends a join request to another multicast session in order to switch to that multicast session and the time when the corresponding mobile user receives the first data packet from the new multicast group. For example, if a multicast mobile user sends a join request to switch to a new multicast session at time 280 seconds and receives the first data packet from that new multicast session at time 280.5 seconds, then the *multicast session switch time* is 0.5 second. In order to find out the value for this entity for different videos, we use four different videos of bit rates 512 Kbps, 1024 Kbps, 2048 Kbps and 4096 Kbps. We compute multicast session switch time for switching among these different videos over 100 simulations and take the average of the computed results.

4.2 Performance Metrics

The performance metrics used in the evaluation are described below.

- **Energy Saving.** We define *energy saving* as the fraction of time a mobile device can turn off its network interface to save energy. We assume, the required time to turn on/off the network interface is negligible. This suffices that we can use the network interface turn off duration as a direct representative of energy saving.

- **Service Ratio.** Due to the limited resources in wireless networks, it may not be possible to serve all the incoming video requests from mobile devices. The *Service ratio* is defined as the fraction of incoming video requests that are served by the wireless networks at any point of time.

- **Service Delay.** According to the proposed solution, the incoming requests are served at the beginning of a *scheduling window*. Thus, some requests may experience some delay. We define the *service delay* as the time difference between an on-demand video request is placed and the request is actually served, i.e., the requesting mobile device receives the first data packet.

- **Percentage Switching Time.** We define the *percentage switching time* for a user as the percentage of time the user spent in switching among different multicast streams when watching a video. This happens because the proposed greedy algorithm may establish multiple multicast sessions for the same video with different MCS modes to *maximize the energy saving* and the user may switch among these different sessions based on its current channel conditions and algorithmic decision. The overall switching time is the aggregated *multicast session switch time* for all such switches. The time spent in such switching is recorded for each user,

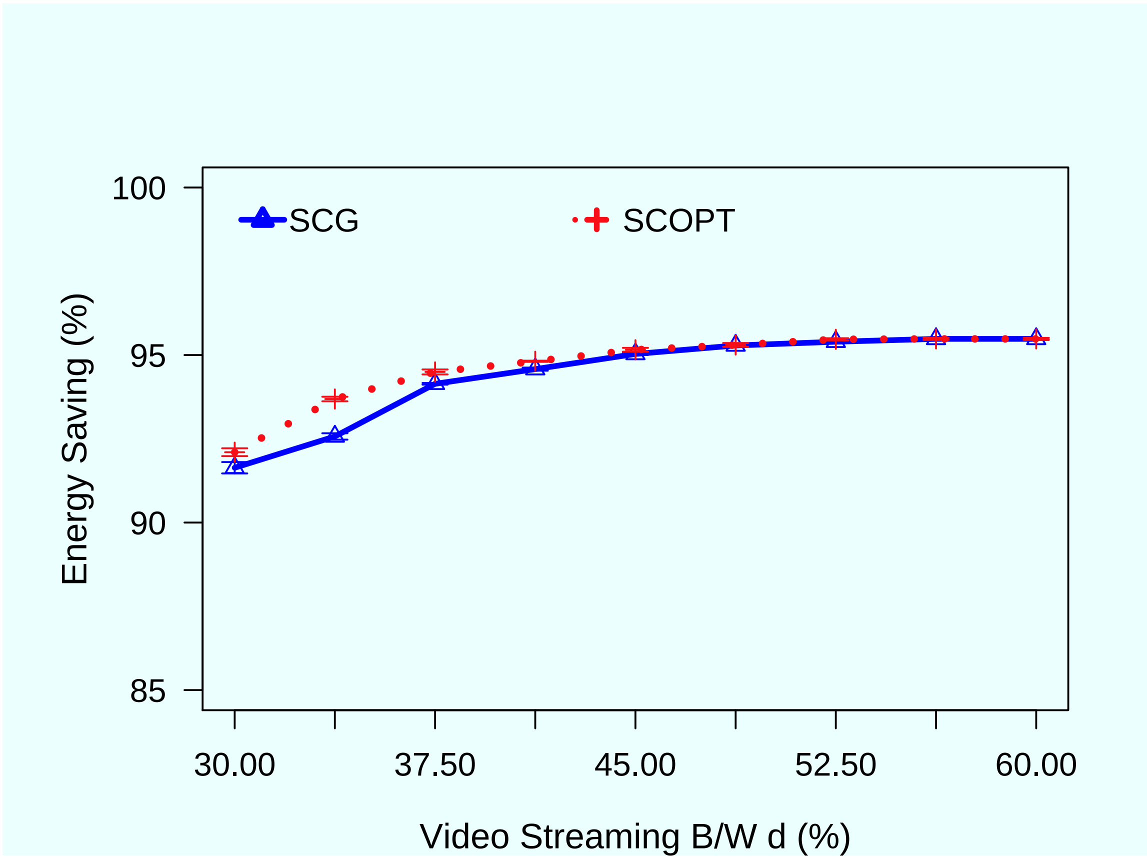

Figure 4.1: Near optimality of our SCG algorithm in LTE 10 MHz eMBMS system.

and the percentage is calculated using the total video watching time of the user as the baseline time. For example, if a user is watching a video for l seconds, and the time spent in switching is l' seconds. Then, the percentage switching time for that user is $\frac{l'}{l} \times 100\%$. *Percentage switching time* over all users is computed as the *average* of all the percentage switching times across all users.

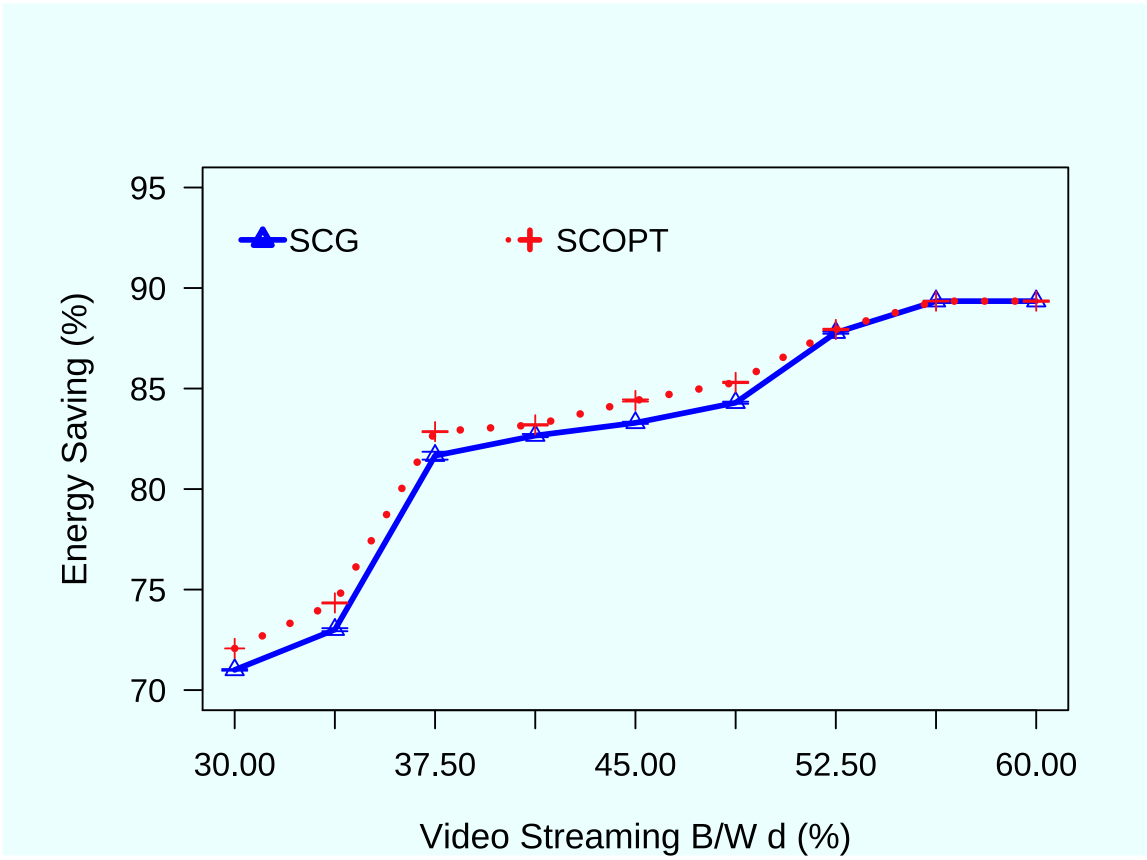

Figure 4.2: Near optimality of our SCG algorithm in LTE 3 MHz eMBMS system.

4.3 Results

4.3.1 Near optimality

We compare the results achieved by our algorithm (SCG) versus those computed by the optimal algorithm (SCOPT). We consider two LTE networks: 10 MHz (Figure 4.1) and 3 MHz (Figure 4.2). We vary the fraction of resources d from 30% to 60% in each case. We configure three 256 Kbps videos for 10 MHz eMBMS network and a single 256 Kbps video for 3 MHz eMBMS network. We simulate a small number of mobile users (36 users) in order to be able to compute the optimal results. In the service area, mobile devices are uniformly distributed to cover wide range of channel conditions. In both of these figures, the energy saving achieved by SCG is fairly close (more than

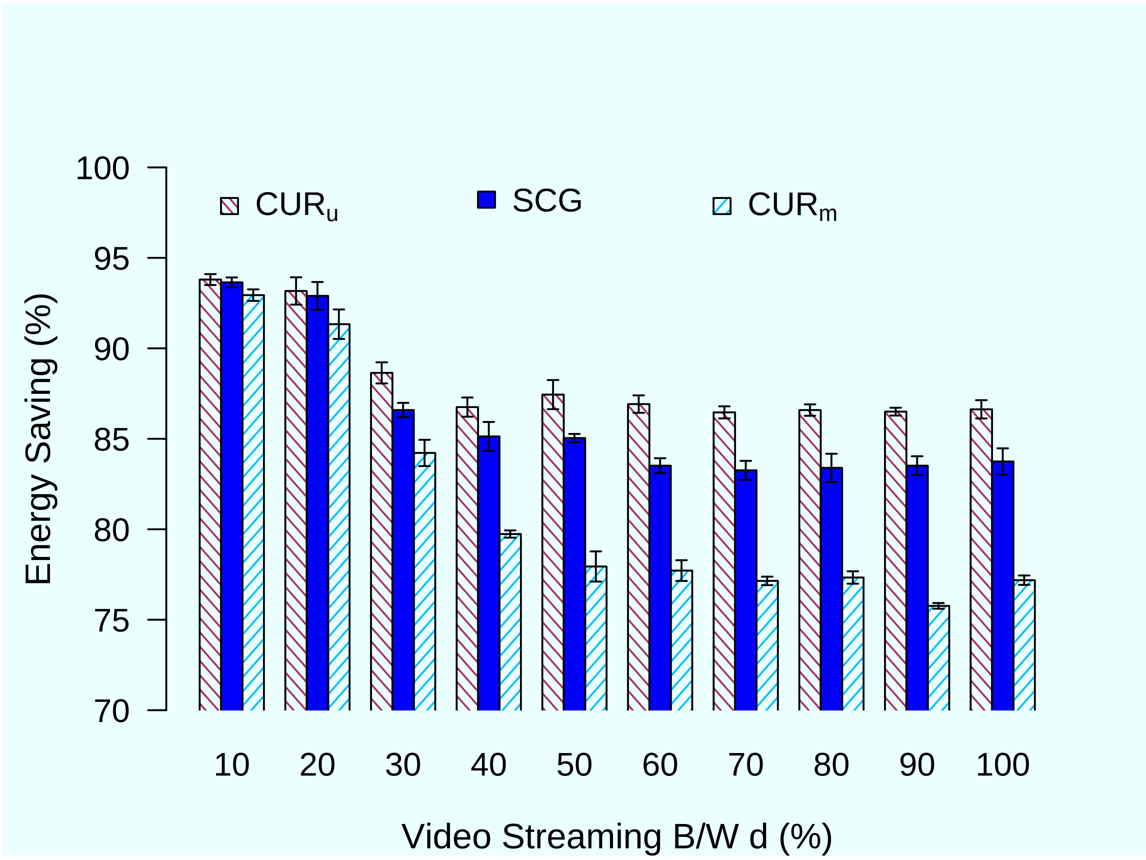

Figure 4.3: Comparing our SCG algorithm versus current unicast algorithm (CUR_u) and multicast algorithm (CUR_m).

98.2% in all cases) to the optimal solution, SCOPT. Moreover, SCG always terminates in < 1 ms, while SCOPT may take as long as 200 ms only for 3 videos and this computation time greatly increases with the increase in number of videos.

In summary, our proposed SCG algorithm achieves energy saving close to the optimal, and the same as optimal as the reserved bandwidth for on-demand video services increases. SCG runs realistically faster compared to SCOPT without sacrificing the optimality too much. Hence, for the rest of the simulations, we no longer report results from SCOPT.

4.3.2 Comparison Versus Current Algorithms

We compare the energy savings of our SCG algorithm versus current multicast ($\mathrm{CUR_m}$) and unicast ($\mathrm{CUR_u}$) algorithms. We simulate large networks of 1000 mobile users with 1000 videos. We use 20 MHz LTE system and vary the fraction of resources, d from 10% to 100%. The results are shown in Figure 4.3 as bar chart. We also plot the error bars where the length of the bars are determined by the standard deviation of 10 simulation results obtained from 10 different seeds. The standard deviation is added with the mean to compute the upper limit of each error bar, and subtracted from the mean to calculate the lower limit of the bar. It shows that our proposed SCG algorithm consistently saves more energy than conventional $\mathrm{CUR_m}$ and the energy saving difference becomes noticeable as more bandwidth is reserved for on-demand video streaming services. In particular, the largest energy saving gap between proposed SCG and $\mathrm{CUR_m}$ is experienced when 90% of the bandwidth is reserved for on-demand video streaming services. This situation occurs when the mobile users of a particular multicast group are divided into two sets: (i) mobile users with poor channel conditions which can be served using relatively lower MCS mode and (ii) mobile users with good channel conditions which can be served with relatively far aggressive MCS mode. In such a situation, $\mathrm{CUR_m}$ creates a single stream for all the users with lower MCS mode and as a result mobile users with relatively good channel conditions end up saving far less energy than possible. However, our proposed SCG algorithm, capitalizes this situation nicely creating two multicast video streams: (i) one multicast stream with less aggressive MCS mode for mobile users with poor channel conditions and (ii) the other multicast stream with far more aggressive MCS mode for mobile users with good channel conditions. Thus, SCG algorithm attains far more energy saving for those mobile users with good channel conditions consuming a little extra bandwidth as aggressive MCS mode requires far less radio resources compared to that of lower MCS mode. Sometimes the proposed SCG algorithm can outperform $\mathrm{CUR_m}$ by a large margin. $\mathrm{CUR_u}$ represents the maximum energy saving because of individual unicast connections to achieve the highest data rates possible based on the individual channel conditions. The SCG algorithm achieves closer energy saving to $\mathrm{CUR_u}$ as bandwidth reservation for on-demand video streaming services increases after 50%. When d <60%, the number of users served by $\mathrm{CUR_u}$ as shown in Figure 4.4 is small enough to represent the overall channel conditions of average users.

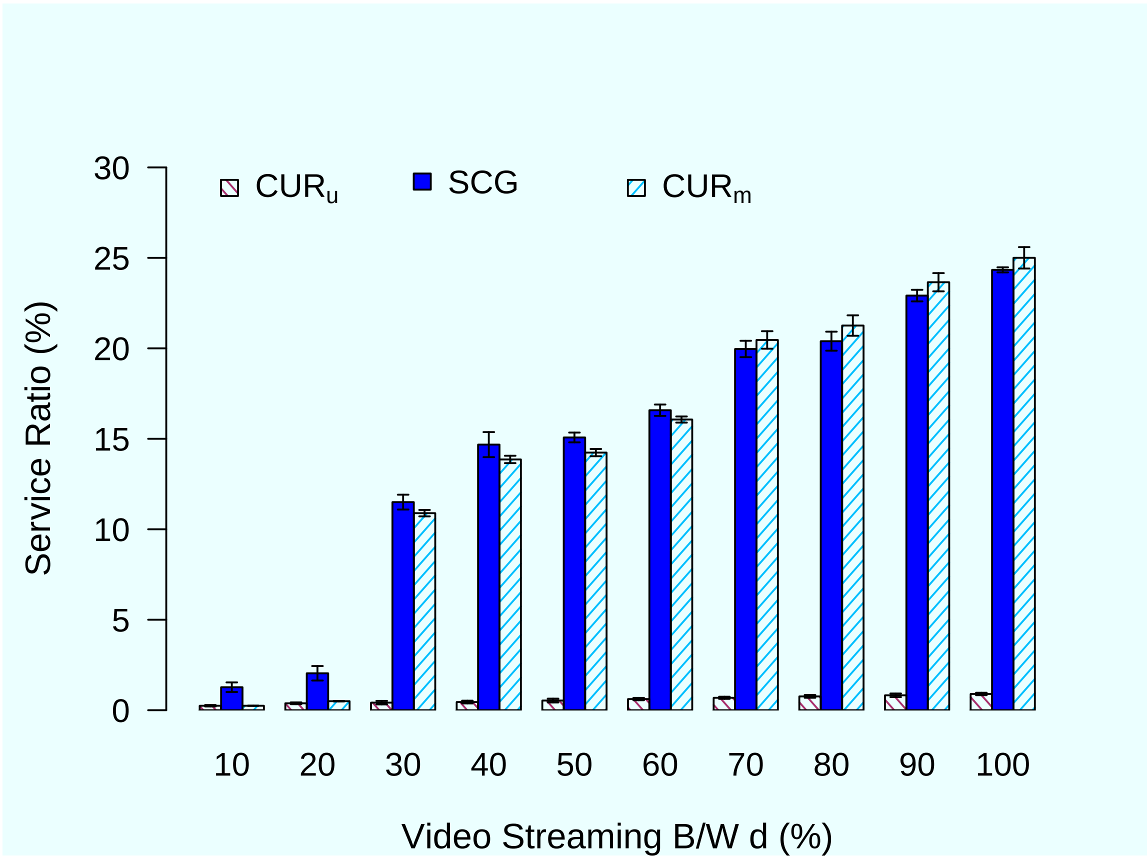

Figure 4.4: Service ratio in large-scale simulations of 1000 mobile devices under different bandwidth constraints.

4.3.3 Scalability to serve more users

In this experiment, we show that our hybrid SCG algorithm can serve more users than current unicast algorithm ($\mathrm{CUR_u}$) and can serve number of users close to current multicast approach ($\mathrm{CUR_m}$). We compare the service ratio of our SCG versus $\mathrm{CUR_u}$ and $\mathrm{CUR_m}$ for different d values. The results are shown in Figure 4.4. This figure shows that $\mathrm{CUR_u}$ serves less than 10 mobile users even when the whole bandwidth is reserved for on-demand video services. On the other hand, the proposed SCG algorithm can serve more than 200 mobile users under the same configuration which is above 20 times improvement of $\mathrm{CUR_u}$ in terms of scalability and fairly close to the service ratio achieved by

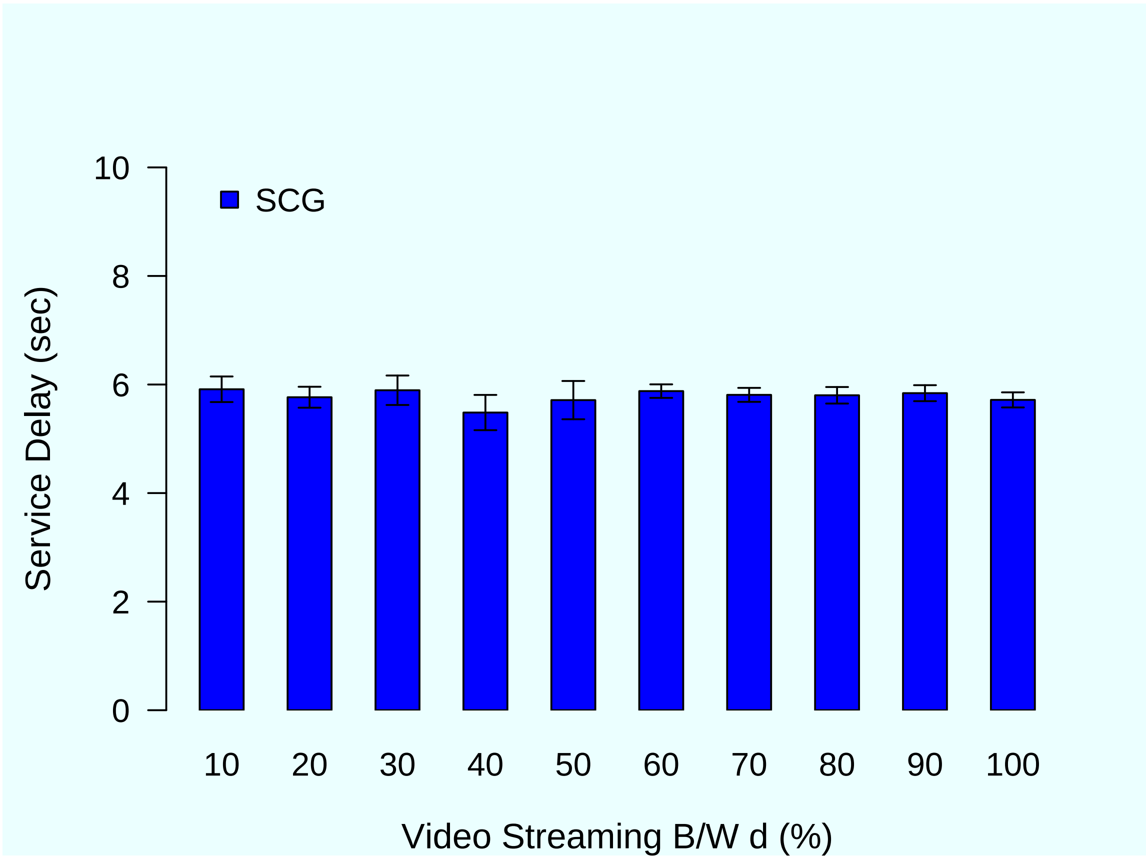

Figure 4.5: Service delay in large-scale simulations of 1000 mobile devices under different bandwidth constraints.

CUR_m. When small amount of bandwidth is available for on-demand video services, in our configuration, less than 70%, our proposed SCG algorithm even outperforms CUR_m in terms of scalability. This is because the SCG algorithm intelligently allocates radio resources to the requesting mobile users to maximize average energy saving across all users while CUR_m distributes radio resources among the requesting multicast users on a first-come first-serve basis. While our SCG algorithm achieves service ratio close to the multicast approach, the SCG achieves higher energy saving (upto 7.75%) as described in Section 4.3.2 and shown in Figure 4.3.

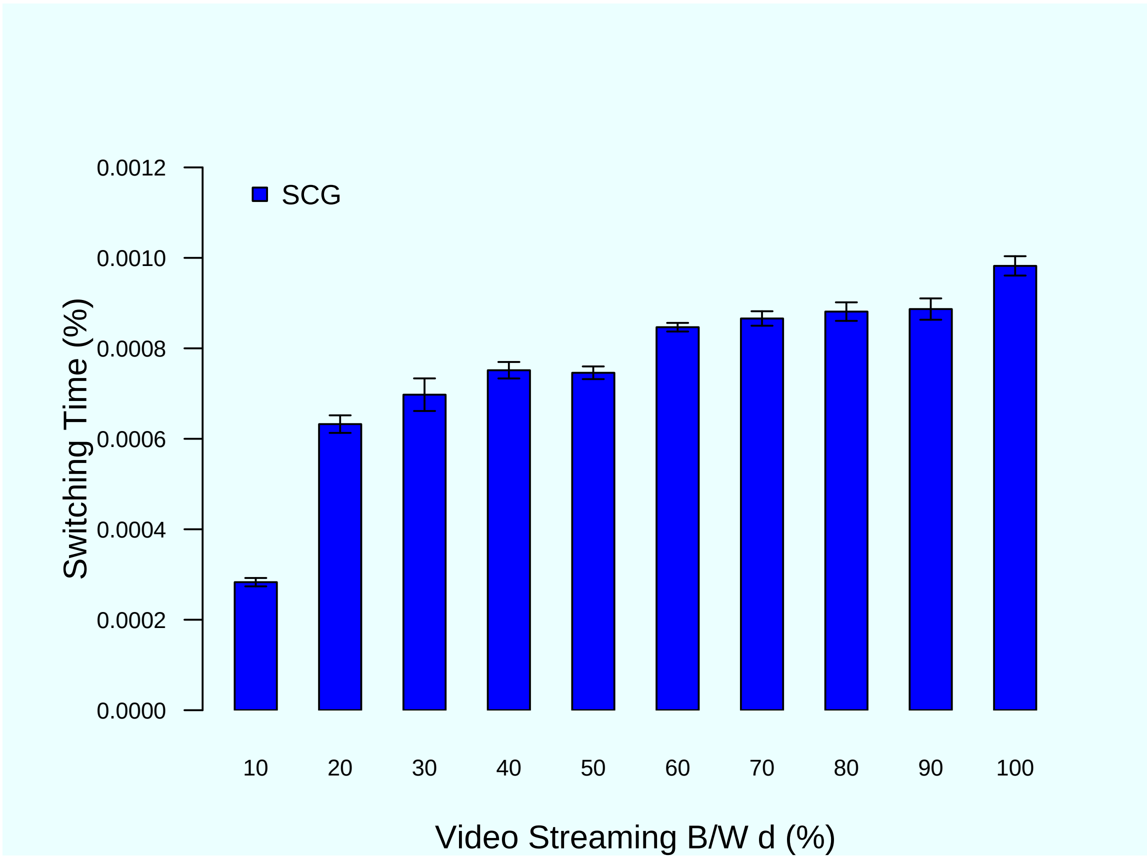

Figure 4.6: Percentage switching time in large-scale simulations of 1000 mobile devices under different bandwidth constraints.

4.3.4 Service Delay

In this experiment, we show the service delay experienced by the users in our SCG algorithm. The results are presented in Figure 4.5 for various d values ranging from 10 to 100. The figure shows that the experienced service delay for the admitted mobile users on average is less than 6 seconds which is slightly over half of the length of the allocation window. If a shorter delay is required for any mobile users, a patching technique [21, 18, 7] may be employed for them to achieve the targeted delay. In patching, the system allows incoming users to immediately join ongoing multicast sessions, and the missed portions of the video are transmitted using temporary unicast connections.

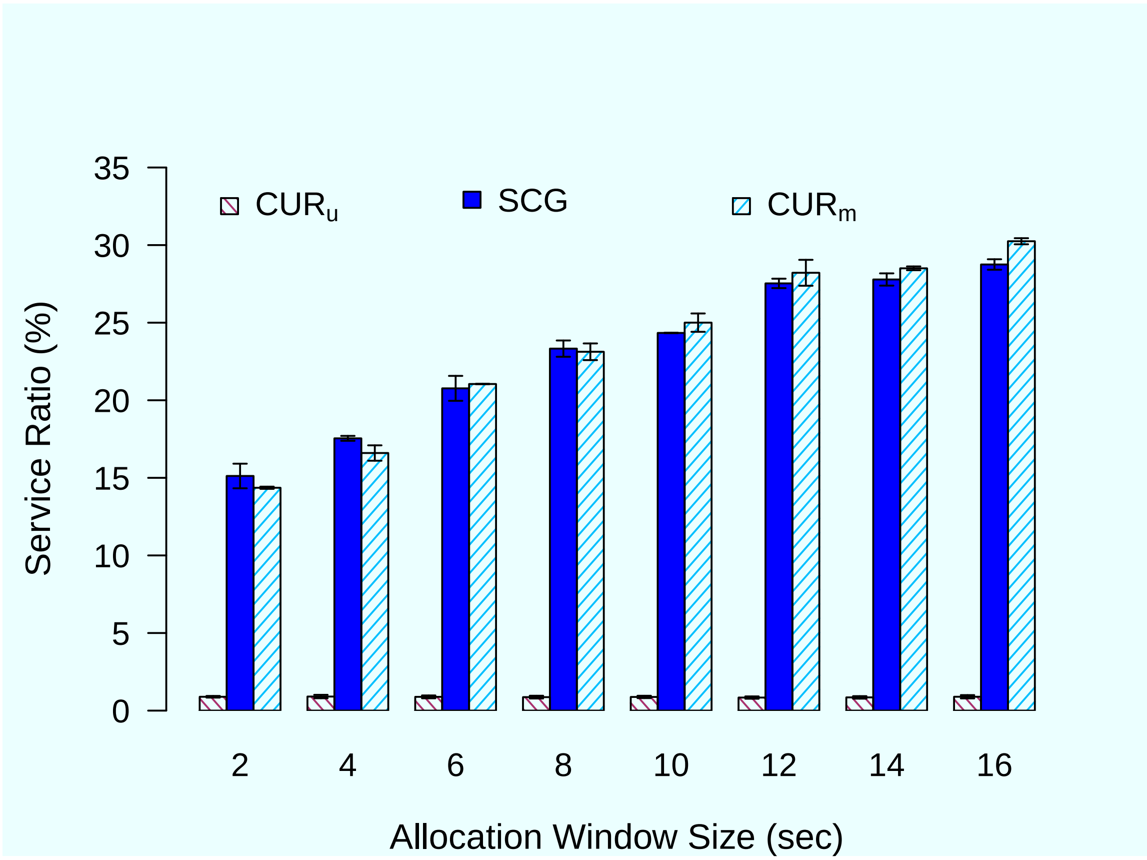

Figure 4.7: Impacts of allocation window size on service ratio.

4.3.5 Percentage Switching Time

In this experiment, we show the average fraction of time a user spends in switching among different multicast sessions for the same video in SCG algorithm is *negligible*. For this experiment, we vary d from 10 to 100. The simulation results are shown in Figure 4.6. The figure shows that the percentage switching time increases as more bandwidth is reserved for on-demand video services. This is because more bandwidth allows the proposed algorithm to allocate more multicast streams which in turns increases the provisions for switching. Note that the percentage switching time increase is slow with the increased bandwidth and is only about 0.001% even with 100% bandwidth reservation for on-demand video services. This indicates that the switching time experienced by users is negligible and implies high stability of the proposed SCG algorithm.

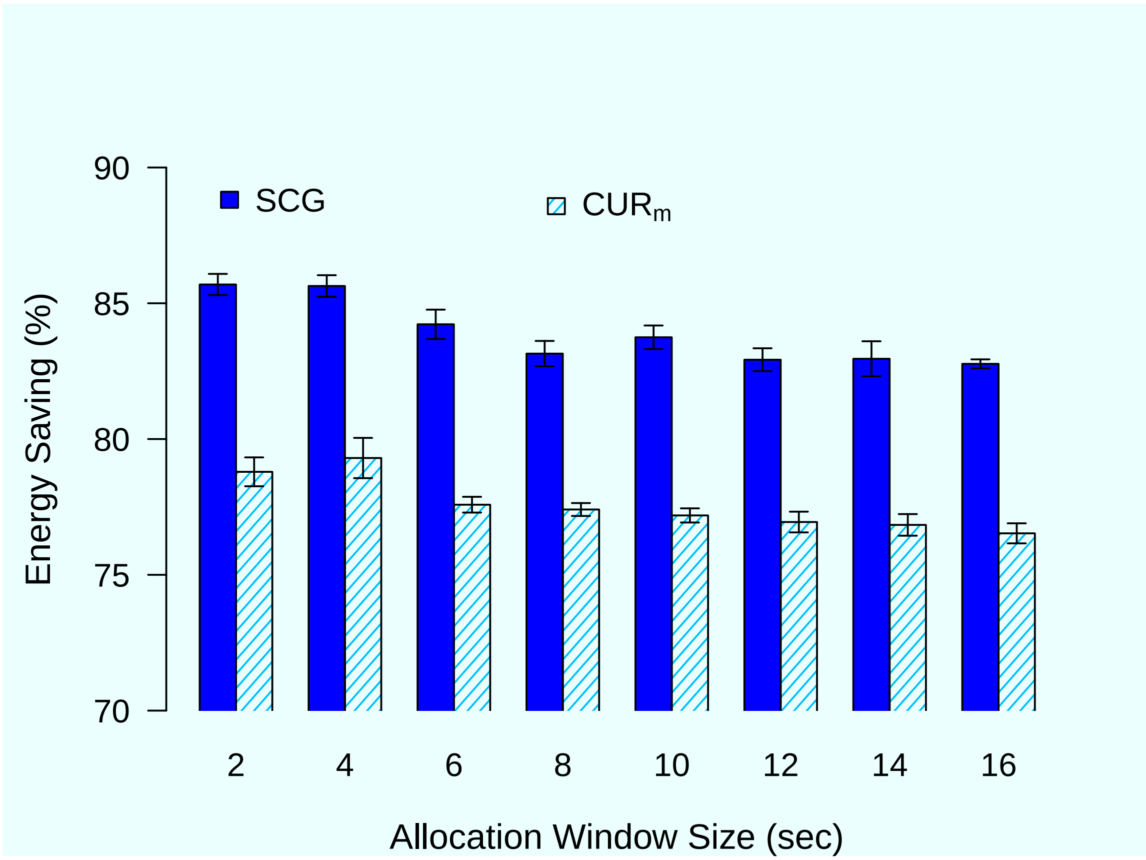

Figure 4.8: Impacts of allocation window size on energy saving.

4.3.6 Impact of Allocation Window Size

We compare service ratio of our SCG algorithm versus current multicast (CUR_m) and unicast (CUR_u) approaches for various sizes of allocation window. The results are shown in Figure 4.7. We also compare the energy saving of our SCG algorithm versus CUR_m algorithm for various allocation window sizes as shown in Figure 4.8. Here we do not mention current unicast algorithm because in CUR_u, users are served with individual unicast connections independent of the allocation window size. Figure 4.9 and Figure 4.10 report the effects of allocation window size variation on percentage switching time and service delay respectively. We vary the allocation window size between 2 and 16 seconds inclusively to observe the effects of different window sizes. Figure 4.7 shows an increasing trend of service ratio with the increase in allocation window size. This is because larger allocation

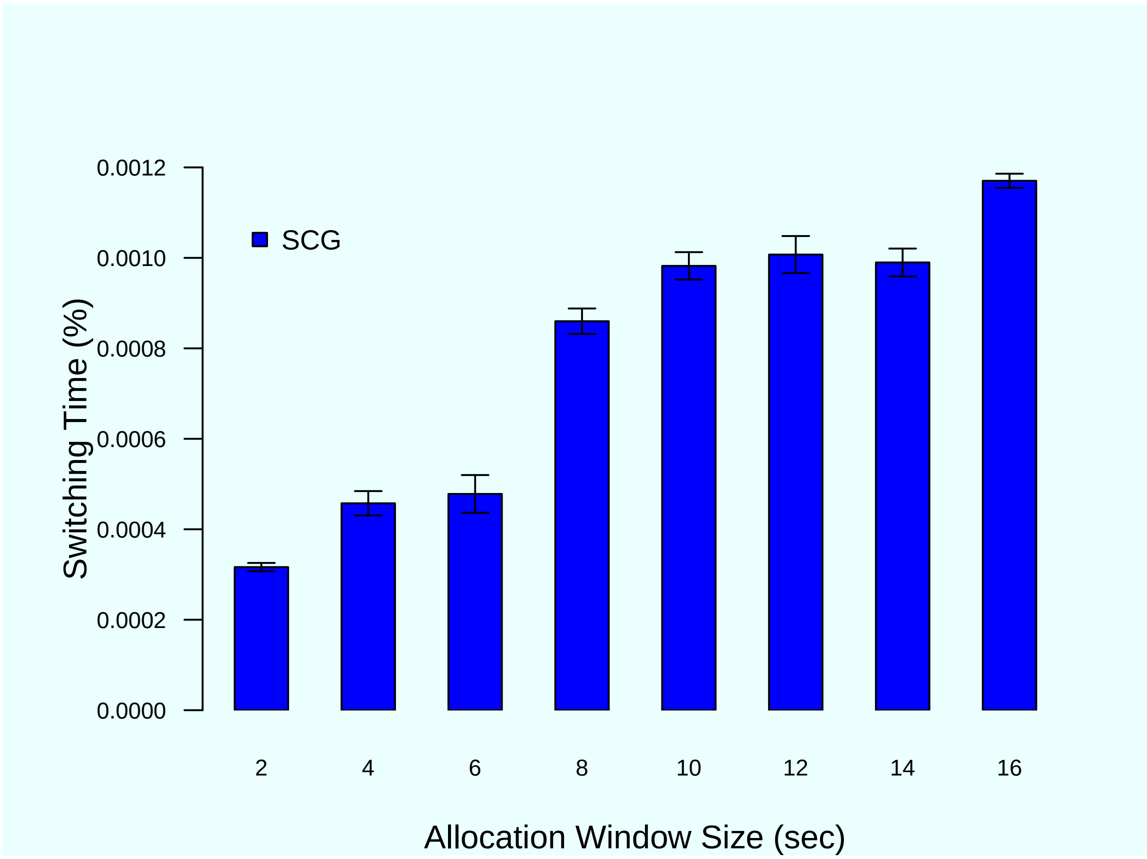

Figure 4.9: Impacts of allocation window size on percentage of time spent in switching.

window size offers more opportunity to batch more requests together to form multicast groups and provides larger optimization scope. With smaller window size of less than 6 seconds, when there is comparatively less number of multicast groups, SCG even outperforms $\mathrm{CUR_m}$ in terms of service ratio. This is because of SCG's algorithmic choice of multicast groups for transmission to maximize average energy saving across all users which prefers large multicast groups with relatively better channel conditions, instead of $\mathrm{CUR_m}$'s first-come first serve approach. Even when the window size is comparatively larger, SCG achieves service ratio close to that of $\mathrm{CUR_m}$. Moreover, as the window size becomes larger, the improvement achieved in service ratio slows down. This is logical because after certain window size, even if we increase the window size, the opportunity to create more multicast groups does not increase proportionately. Figure 4.8 shows that SCG outperforms $\mathrm{CUR_m}$ by

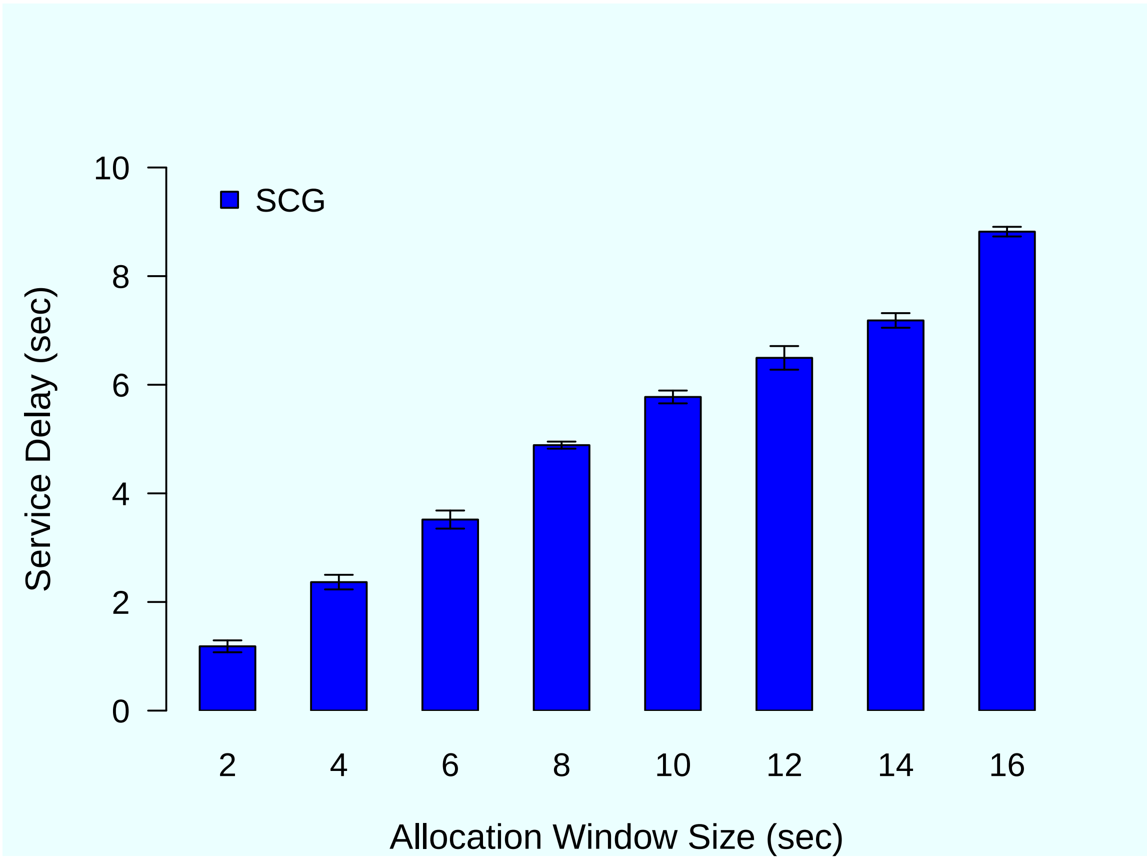

Figure 4.10: Impacts of allocation window size on service delay.

reasonable margin irrespective of the window size which encourages adoption of SCG in wireless networks. However, smaller window size shows better margin of improvement in energy saving. In particular, SCG outperforms $\mathrm{CUR_m}$ in terms of energy saving by the largest margin when the window size is set to 2 seconds. At the same time, larger window size achieves margins fairly close to the largest margin. From the observation of both Figure 4.7 and Figure 4.8, we can conclude that any window size results in reasonable improvement in energy saving of SCG compared to that of $\mathrm{CUR_m}$ without sacrificing the service ratio that much.

Figure 4.9 exhibits an increasing trend of percentage switching time as the window size increases. This is because the larger window size gives provisions for more multicast users through batching, which in turns leads to more multicast sessions creating scopes for more switching among multicast sessions. However, when the window size becomes larger than 8 seconds, the percentage switching time becomes quite stable. Even with the largest window size of 16 seconds, the switching time is about 0.0012% which confirms that the proposed SCG algorithm is free from the *ping-pong effect* in terms of switching. We define *ping-pong effect* as the wild switching of mobile devices among different multicast sessions such that the mobile devices do not spend sufficient time continuously in any of these multicast sessions. Figure 4.10 confirms that the average service delay increases gradually with the increase in allocation window size. This happens because the on-demand video requests arrive after the start of any allocation window are served at the beginning of next allocation window. In order to achieve shorter service delay *patching* technique can always be employed. Combining the observations from the figures, we can conclude that the benefit gained in service ratio by adapting larger allocation window is superseded by the incurred average delay in service. Hence, we would suggest the adaptation of reasonably short allocation window with the fact that the proposed SCG algorithm terminates in less than a second.

4.3.7 Impact of Video Popularity Distribution

In this experiment, we show the superiority of our SCG algorithm over current multicast approach (CUR_m) independent of the video selection policy. We emulate Zipf distribution to select videos from the pool of 1000 videos. The skewness parameter α guides the selection of the videos: higher values of α assign higher probability to most popular videos to be selected and vice versa. We vary α from 0.5 to 1.5 to emulate different video selection policies. Figure 4.11, Figure 4.12 and Figure 4.13 report the effects of variation of skewness α on service ratio, energy saving and percentage switching time respectively. Figure 4.11 shows that the service ratio gradually increases with the increase in skewness α. The higher value of skewness guides more users to select from top ranked videos resulting in more multicast users which provides larger opportunities to serve more users through multicast. That is why, an increase in skewness results in an increase in service ratio. Importantly, the proposed SCG algorithm achieves service ratio close to that of CUR_m irrespective of the skewness α. Figure 4.12 confirms that proposed SCG algorithm always outperforms CUR_m in terms of energy saving irrespective of the chosen skewness value. However, higher values of skewness lead to higher energy saving for SCG compared to that of CUR_m as higher values result in more multicast users providing more improvement opportunity. In particular, the largest margin of improvement in terms of energy saving is achieved when α is set to 1.5, the highest value for our experiments. Figure 4.13 shows a non-decreasing trend of percentage switching time for increased

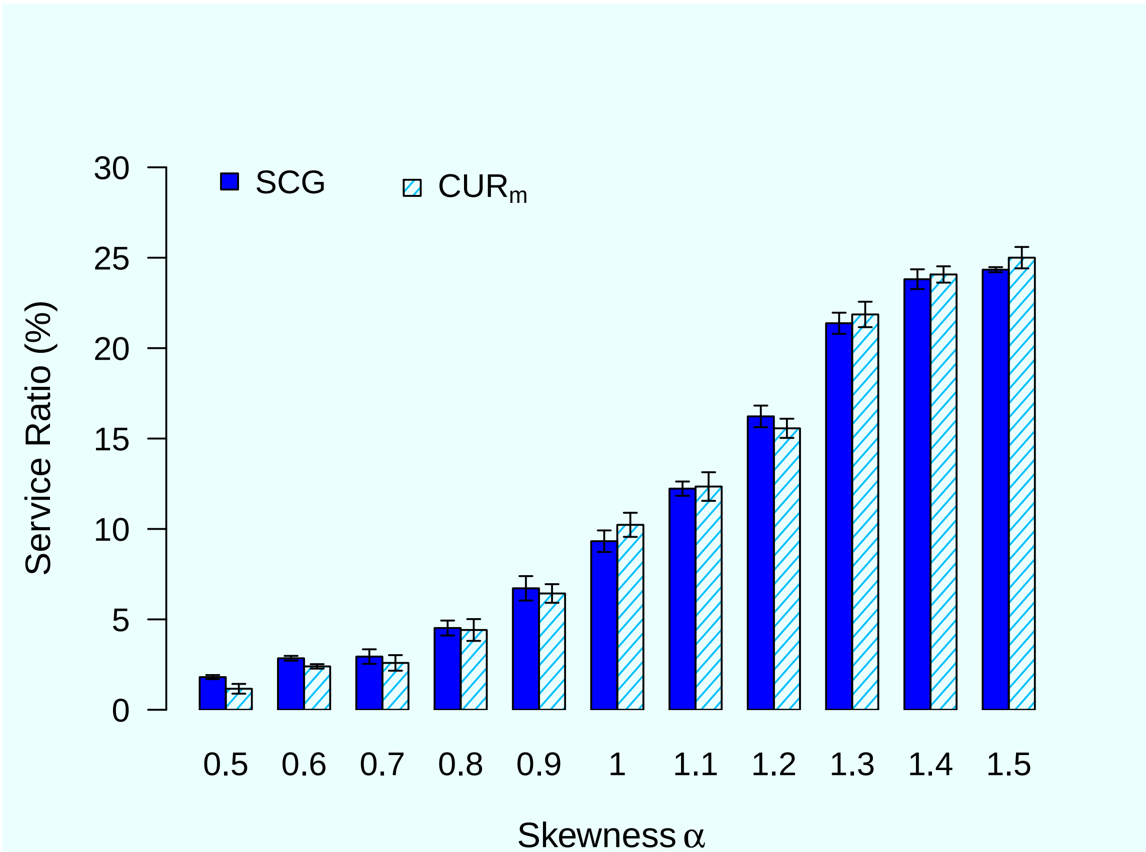

Figure 4.11: Impacts of Skewness α of the chosen Zipfian distribution on service ratio.

values of skewness due to the fact that larger skewness values usually result in more multicast groups giving more flexibility to switch among different multicast groups. In particular, percentage switching time shows an increase when skewness value changes to 0.6, remains stable for a while and shows other increases when changing to 1.1 and 1.5 respectively. This pattern shows that the percentage switching time is affected by a noticeable variation in number of multicast group users. When the increase in skewness value noticeably changes the number of multicast group users, the percentage switching time increases. Hence, percentage switching time shows intermittent increases with the increase in skewness value, and remains stable for the rest of the simulation. However, even when the skewness is set to 1.5, it results in about 0.001% switching time only, which confirms the stability of the proposed SCG algorithm against *ping-pong effect* in terms of switching.

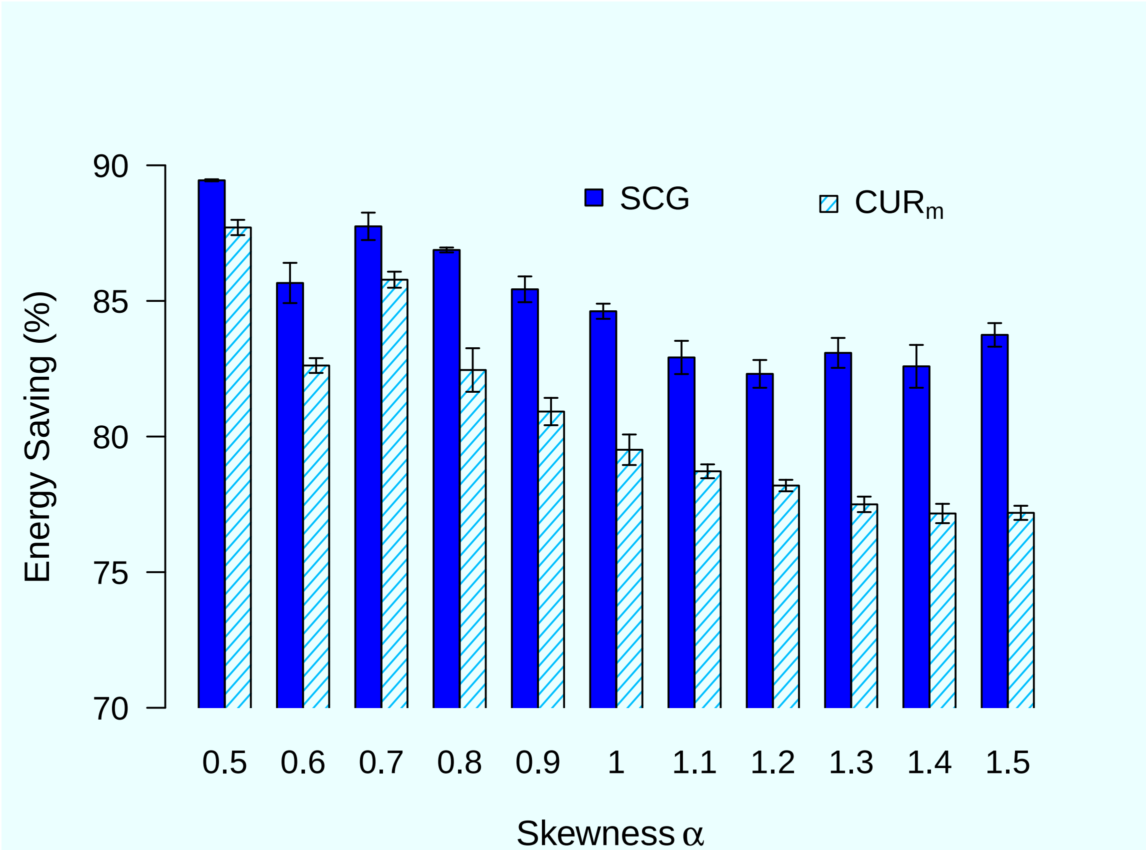

Figure 4.12: Impacts of Skewness α of the chosen Zipfian distribution on energy saving.

4.3.8 Impact of Number of Videos

In this experiment, we compare our SCG algorithm versus current multicast (CUR_m) and unicast (CUR_u) algorithms for different number of videos. We use small number of videos and small number of users (36 users) to understand the behavior. In the service area, mobile devices are uniformly distributed to cover wide range of channel conditions. We use 256 Kbps videos and vary the number of videos from 1 to 6 for this experiment. We use 10 MHz LTE system. The energy saving results are shown in Figure 4.14. For the SCG algorithm, with the increase in number of videos, the energy saving decreases because the videos require more bandwidth to support multiple MCS mode bearers per video. But we have fixed the bandwidth reservation to 60% of the total bandwidth. For upto three videos, the SCG algorithm has equal energy saving as that of CUR_u because the available bandwidth

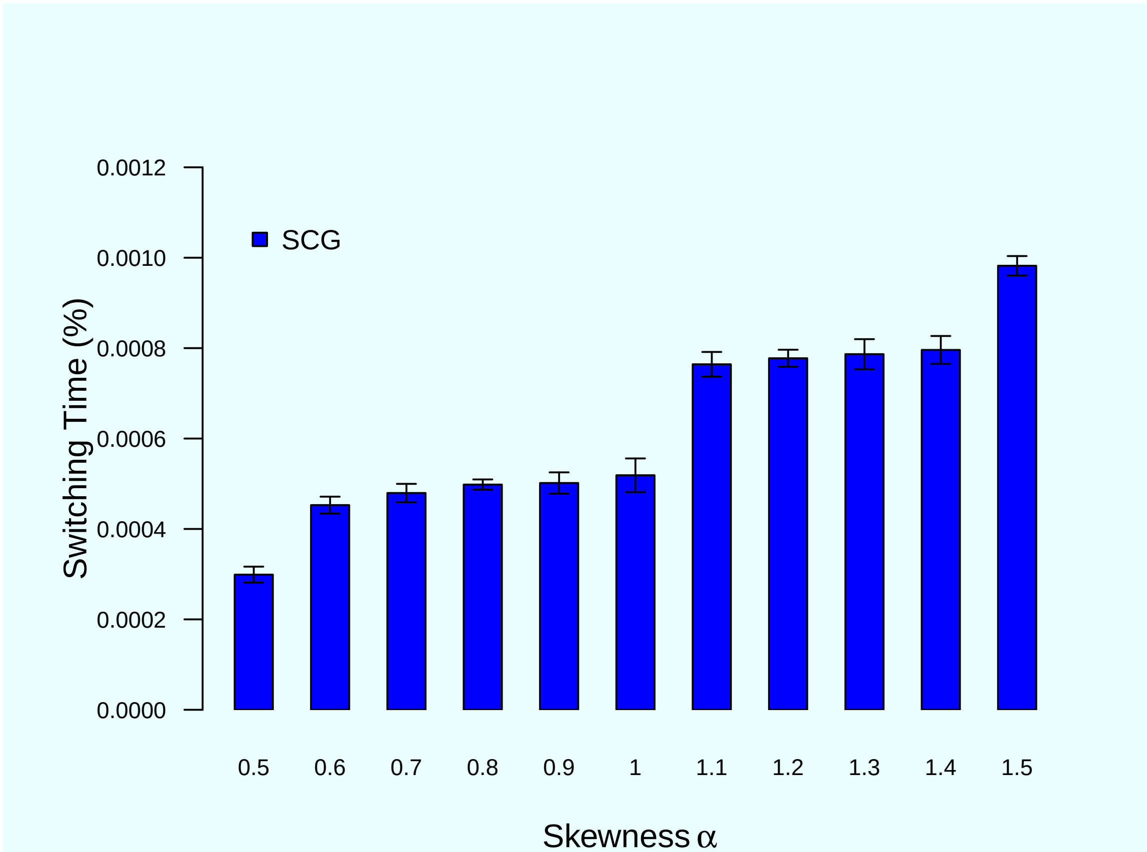

Figure 4.13: Impacts of Skewness α of the chosen Zipfian distribution on percentage of time spent in switching.

allows the SCG algorithm to transmit all four MCS mode bearers per video. The SCG algorithm performs better than $\mathrm{CUR_m}$, even with six videos, because of allowing multiple MCS mode bearers for the SCG algorithm whereas $\mathrm{CUR_m}$ allows a single MCS mode bearer per video. For $\mathrm{CUR_m}$, as the number of videos increases, the energy saving improves slightly. Because the MCS mode is guided by user with the worst channel condition in a multicast stream, and we distribute the same number of users among different videos resulting less number of users per video with increase in number of videos. $\mathrm{CUR_u}$ achieves even higher energy saving than SCOPT, and is not affected by the number of videos because each mobile device is served with its maximum MCS mode. In fact, the optimal solution of our hybrid streaming solves a different problem of serving more users along with maximizing average energy saving, while $\mathrm{CUR_u}$ can only maximize energy saving with poor

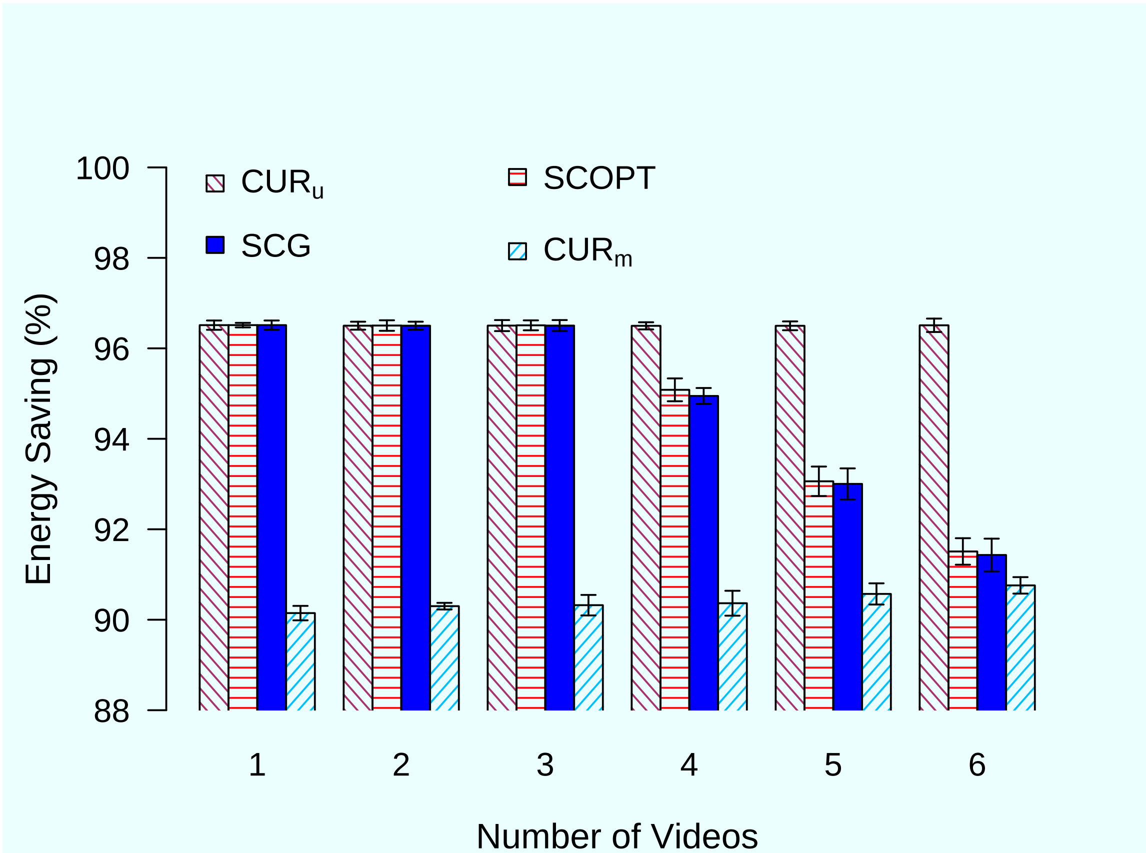

Figure 4.14: Energy saving for different number of videos.

scalability as we have already shown in section 4.3.3.

In summary, our proposed SCG algorithm achieves energy saving close to the optimal even with multiple videos, which is also close to $\mathrm{CUR_u}$ when videos have enough bandwidth for sufficient number of multicast streams, and much higher than using $\mathrm{CUR_m}$ for all mobile devices.

4.3.9 Impact of User Arrival Distribution

In this experiment, we compare our SCG algorithm versus current multicast algorithm ($\mathrm{CUR_m}$) with respect to user arrival distribution. We use Poisson process and vary the mean of the distribution to emulate variations in the user arrivals. The arrival rate λ indicates number of user arrivals per second with higher values offering more multicast opportunity. Figure 4.15, Figure 4.16 and Figure 4.17

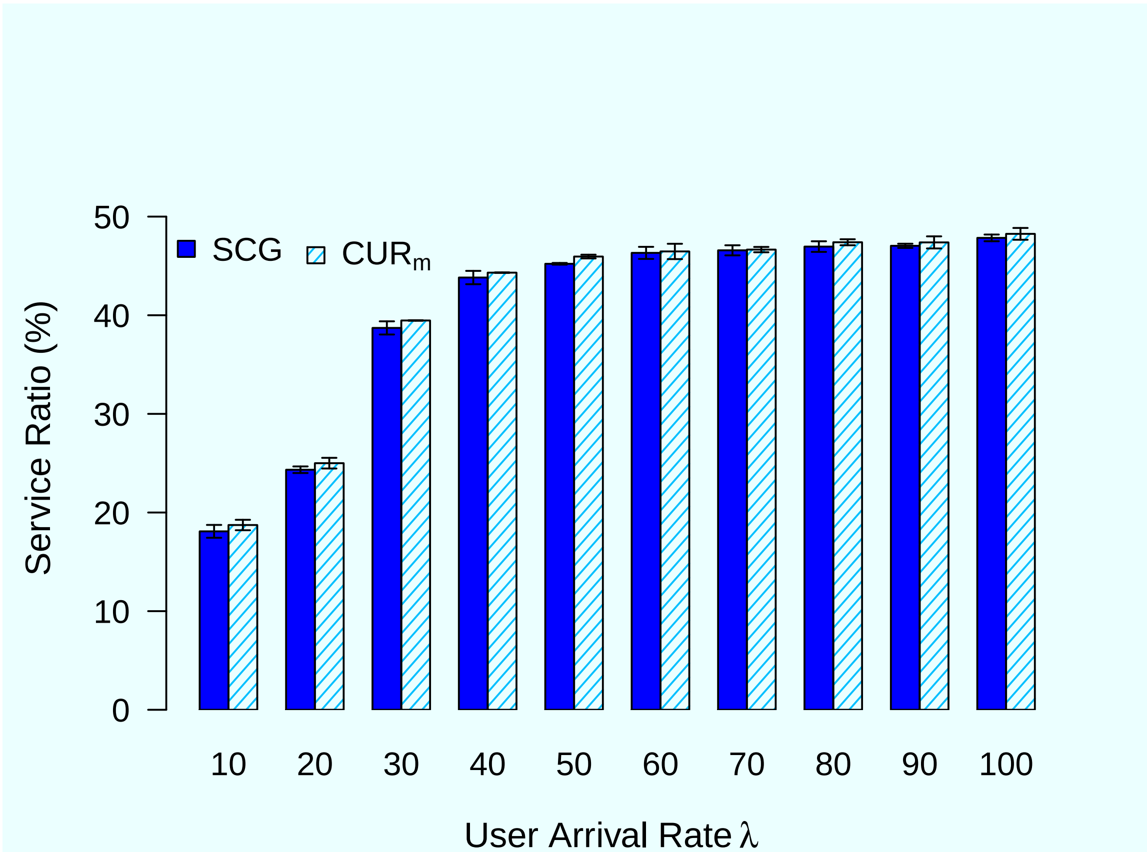

Figure 4.15: Impact of User Arrival Rate λ of the Poisson Process on service ratio.

report the effects of variation of arrival rate of the Poisson process on service ratio, energy saving and percentage switching time respectively. We vary the value of arrival rate from 10 to 100 to observe the effects of different arrival rates. Figure 4.15 shows that the service ratio increases for both SCG and CUR_m with an increase in the arrival rate. Moreover, the proposed SCG algorithm achieves service ratio close to CUR_m irrespective of the arrival rate. High value of the arrival rate ensures large number of user arrivals per second with the same selections of videos as the skewness parameter is kept unchanged for this experiment. This gives opportunity to merge larger number of mobile users to each multicast group which results in higher service ratio. This indicates the effectiveness of the proposed SCG under high load conditions. However, the service ratio increases fast with arrival rate increment until it reaches 40. After that the increment speed slows down and

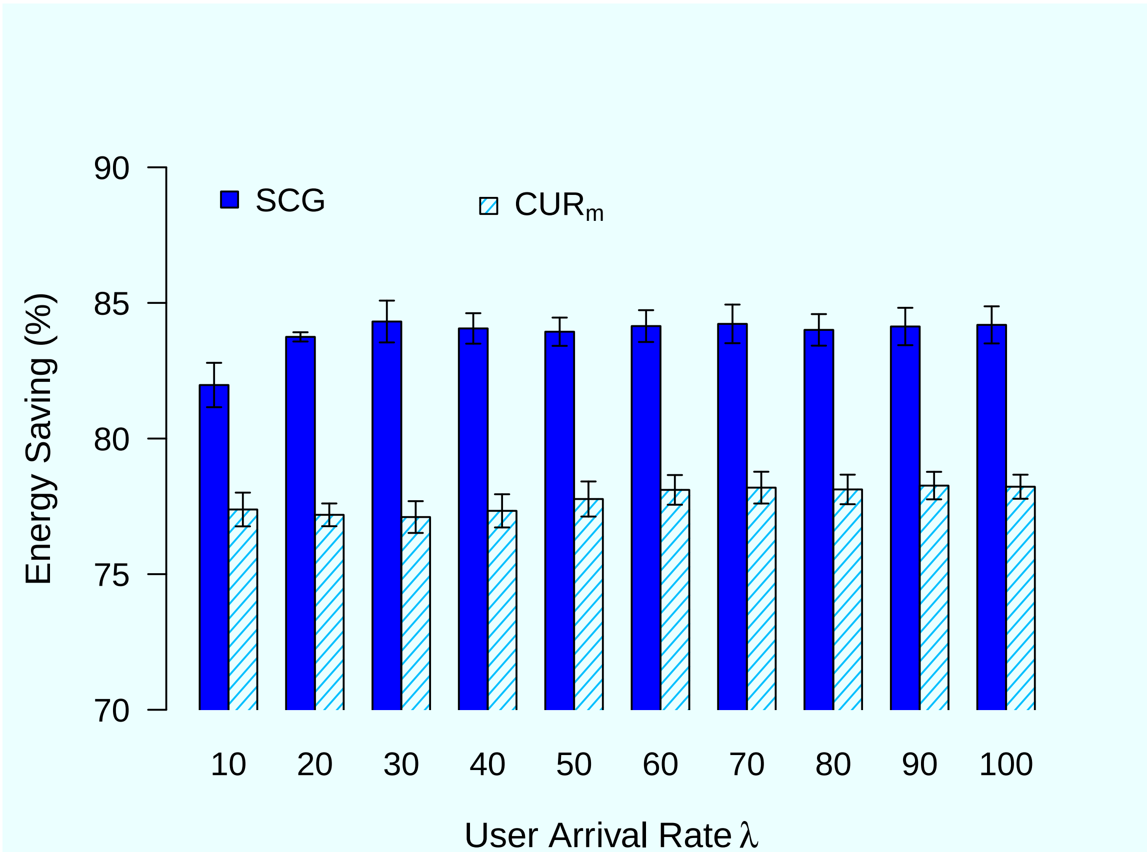

Figure 4.16: Impact of User Arrival Rate λ of the Poisson Process on energy saving.

becomes quite stable. This is because the arrival rate beyond 40 does not increase the opportunity to create larger multicast groups proportionately. Figure 4.16 shows that SCG outperforms CUR$_m$ by a noticeable margin in terms of energy saving irrespective of the arrival rate. In particular, SCG algorithm outperforms CUR$_m$ by the highest margin in terms of energy saving when the arrival rate is 30 for our experiment. Moreover, Figure 4.17 shows that the percentage switching time with the variation of user arrival rate is always less than 0.001% which is an indication of high stability of the proposed SCG algorithm against ping-pong effect in terms of multicast group switching. Combining all these observations for different arrival rates, we can conclude that the proposed SCG algorithm outperforms CUR$_m$ by a noticeable margin in terms of energy saving without sacrificing the service ratio. The simulation results thus confirm the superiority of the proposed SCG over current multicast

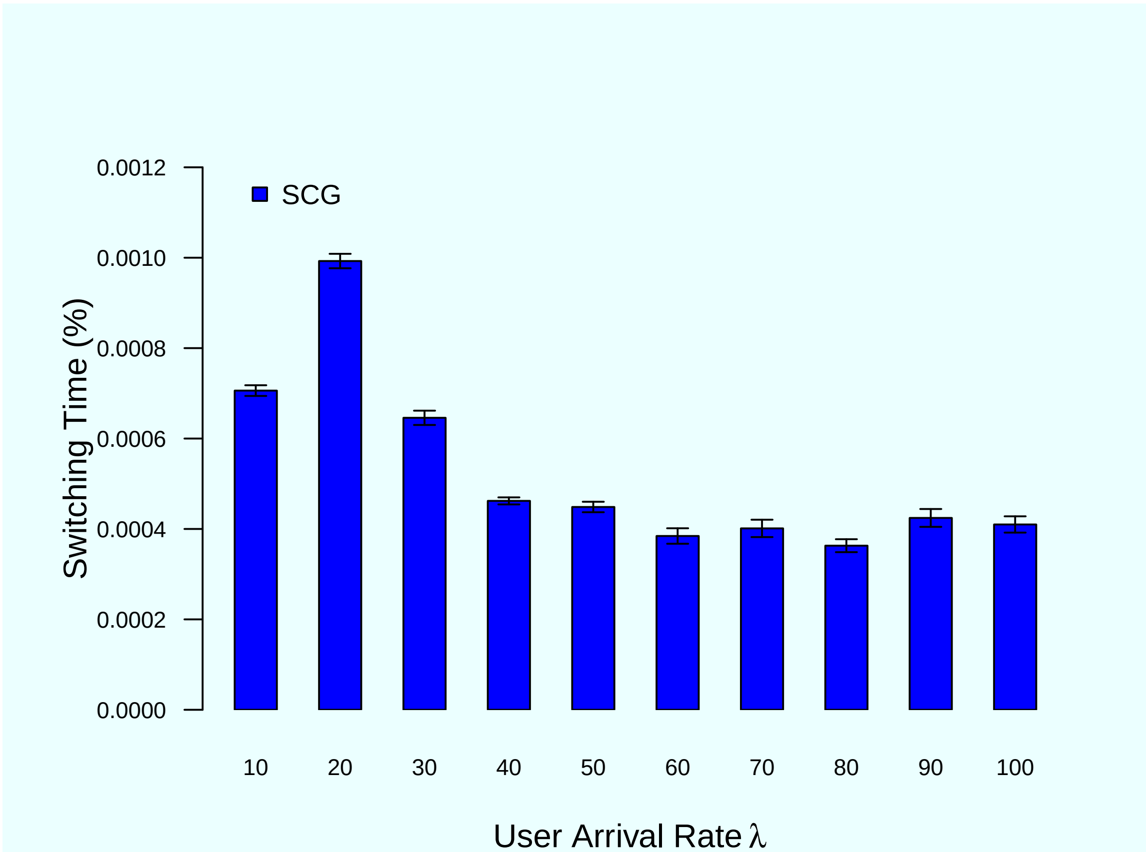

Figure 4.17: Impact of User Arrival Rate λ of the Poisson Process on percentage of time spent in switching.

approach in all load conditions.

4.3.10 Impact of User Mobility

In this experiment, we evaluate our SCG algorithm with different mobility models for the users. We configure the simulation scenario with Random Waypoint mobility model and Walfisch-Ikegami line of sight [30] propagation model. However, due to higher attenuation of Walfisch-Ikegami model, for these results, we set the cell size as 4 Km by 4 Km. The base station and mobile devices' antenna heights are chosen to be 30 meters and 1 meter, respectively. For fair comparison, we conduct our simulations with *no mobility model*, *vehicular mobility model* and *walking mobility model* for SCG,

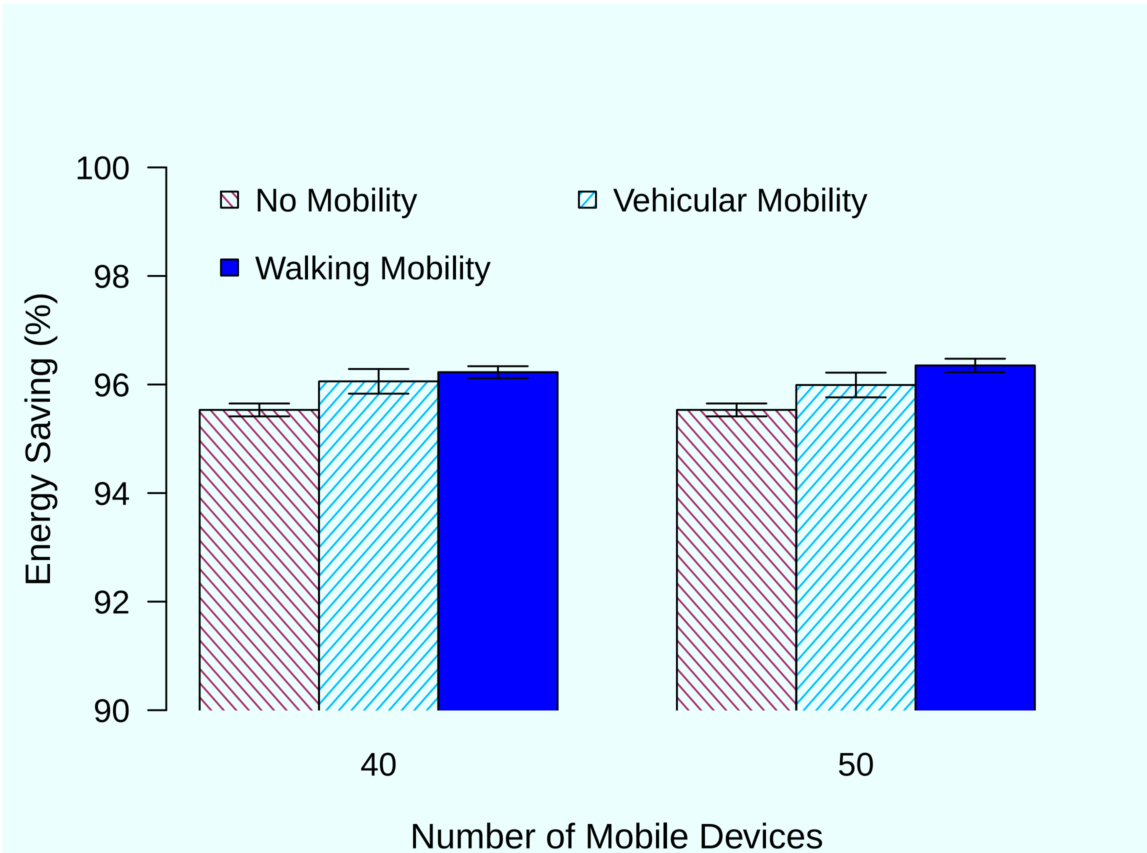

Figure 4.18: Comparison of energy saving of SCG for different Mobility Models.

and report the results in Figure 4.18. We configure three 256 Kbps videos and 10 MHz eMBMS system for this experiment. In order to simulate the vehicular and walking speeds, mobility speeds of 0–72 Km/hour and 4.51–5.43 Km/hour [77] are applied respectively. We use two different number of users, 40 and 50 respectively to verify the results. There is no significant changes brought by the introduction of mobility models into the scenario. The *vehicular* and *walking* mobility models achieve analogous energy savings with that of when there is no mobility. This is because SCG mainly depends on the channel condition distribution of the mobile devices and number of mobile devices belong to a particular channel condition.

In summary, the results do not vary significantly with the mobility models as long as the mobile devices follow the same distribution with respect to channel conditions in the service area. That

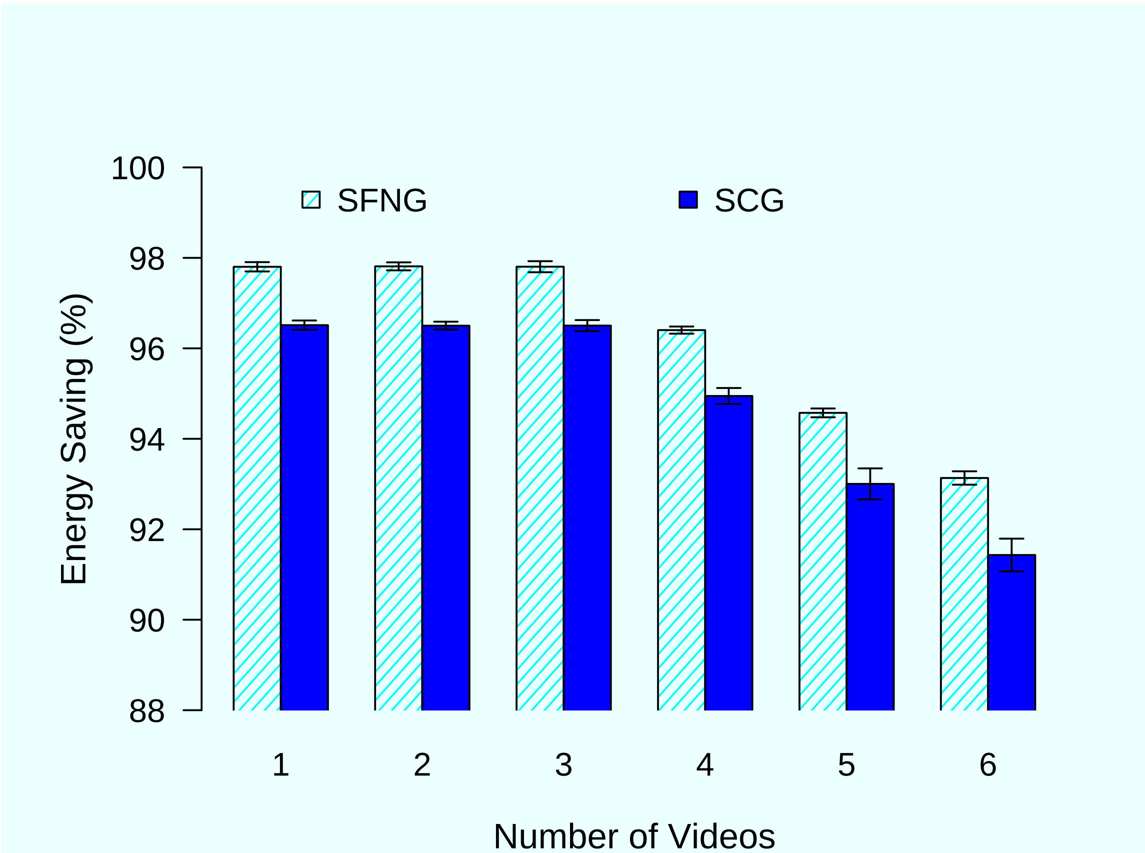

Figure 4.19: Energy saving achieved by SFNG.

is, the SGC algorithm is adaptable to mobility with respect to energy saving even with different mobility models.

4.3.11 Single Frequency Networks

In this experiment, we show the benefit of our SCG algorithm in Single Frequency Networks. We assume a Single Frequency Network with three cells such that the adjacent cells produce 20% overlapping region at the cell boundary. We consider 36 mobile devices uniformly distributed in each of these three cells to cover wide range of channel conditions. We employ three 256 Kbps videos and 10 MHz eMBMS system. Figure 4.19 presents the energy savings for the proposed SFNG and SCG algorithms. We reserve 60% of the total bandwidth for on-demand video services for this experiment. This figure shows that SFNG further improves the energy saving of SCG.

In summary, our proposed SFNG further enhances the performance of SCG by successfully leveraging the merits of Single Frequency Networks.

4.4 Summary

The simulation results show that our SCG algorithm achieves more than 98.2% energy saving of that of the optimal solution which is very close. The SCG algorithm runs relatively much faster than SCOPT and produces the solution in real time. The SCG algorithm outperforms the conventional multicast algorithm by a significant margin of upto 7.75% in terms of energy saving, yet serves more than 95.9% of the users served by the multicast approach for any amount of resource reservation. The simulation results confirm that our SCG algorithm outperforms the conventional unicast algorithm in terms of scalability by a large margin. The SCG algorithm serves more than 20 times the number of users served by the unicast algorithm while saves average energy close to that of unicast approach.

The average service delay experienced by the users in our SCG algorithm is close to half of the allocation window size, and we suggest to adopt patching technique if shorter delay is required. The percentage switching time among different multicast groups on average is negligible for our SCG algorithm and this result confirms that our SCG algorithm is free from the ping-pong effect. The SCG algorithm saves more energy and achieves service ratio close to current multicast algorithm independent of allocation window size, video selection policy and user arrival patterns. However, our SCG serves more users than current multicast with smaller allocation window size, achieves higher energy saving margin with increase in multicast users as dictated by larger skewness values. The simulation results show that our SCG algorithm is well-suited for all load conditions and various mobility models.

The simulation results show that our SCG algorithm successfully merges the benefits of current unicast ($\mathrm{CUR_u}$) and multcast ($\mathrm{CUR_m}$) algorithms. It successfully achieves energy saving close to $\mathrm{CUR_u}$ algorithm, and at the same time serves large number of users close to $\mathrm{CUR_m}$ algorithm. Moreover, our SFNG algorithm successfully leverages the merits of Single Frequency Networks and further enhances the performance of the SCG algorithm.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

We have studied the resource allocation problem in next generation wireless networks considering *hybrid* unicast/multicast video transmissions. We have mathematically formulated the problem to maximize average energy saving of all the participating mobile devices. Initially we have devised our problem to maximize energy saving for a single cell and then have extended our problem scope to multiple cells participating in a Single Frequency Network. We proved that the resource allocation problem is NP-Complete. We presented two algorithms SCOPT and SFNOPT to solve the problem optimally for a single cell and multiple cells forming a Single Frequency Network respectively. Due to the exponential running time of these algorithms, we developed two efficient greedy algorithms namely SCG and SFNG for near-optimal solutions for a single cell and a Single Frequency Network respectively. We proved that these greedy algorithms terminate in polynomial time. In order to evaluate the proposed algorithms, we chose LTE as an example of the next generation networks as it already has built-in multicast support. We developed our proposed algorithms in OPNET and ran extensive simulations to verify them. The findings of our extensive simulations can be summarized as follows:

1. Our proposed SCG, SFNG algorithms always outperform conventional multicast algorithm ($\mathrm{CUR_m}$) in terms of average energy saving.

2. SCG achieves energy saving close to SCOPT, while it can run real time confirming its deployability in real next generation wireless networks.

3. SCG results in higher energy saving if more bandwidth is reserved for on-demand video services or less number of videos is supported.

4. SCG achieves energy saving close to current unicast approach (CUR_u) if enough bandwidth is reserved for on-demand video services, yet exhibits far higher scalability than conventional CUR_u.

5. SFNG in a Single Frequency Network attains higher energy saving than SCG in a single cell confirming that the proposed SFNG algorithm successfully capitalizes the benefit of potential Single Frequency Networks.

In summary, the simulation results confirm the effectiveness and efficiency of the proposed algorithms.

This work promotes the adoption of unicast/multicast hybrid approach for resource allocation to meet the rapidly growing demand for video services over next generation cellular networks, and at the same time maximizes average energy saving for mobile devices. Since cellular networks have limited resources, the hybrid approach can be a viable option as confirmed by our simulation results. Adoption of such approaches require multicast support in the wireless networks which has been provided by LTE and WiMAX. These technologies are expected to be deployed widely in real networks in the near future. This emphasizes the importance and timeliness of the proposed work in this thesis.

5.2 Future Work

The work in this thesis can be extended in multiple directions. For example, we can incorporate *prefetching* into the proposed solution. The prefetching system will be running in the mobile devices and will interact with the associated base station for intelligent decision making. Two of the key concerns addressed by prefetching from users' QoS perspective are (i) start up time reduction and (ii) seeking delay reduction. Incorporation of prefetching into our proposed system improves the overall performance in various ways. For example, from the base station's point of view, prefetching will generate more video requests providing more opportunities for merging them into multicast sessions, which in turn gives our algorithms more scope to improve the overall system performance, power saving and number of requests served in particular. On the other hand, the prefetching system can take feedback from the associated base station about the on-going multicast sessions and intelligently utilizes this information in making prefetching requests for mobile users. For example, in the prefetching decision, we can assign weights to video segments based on whether the desired segments are on-going or not. If a video segment is on-going, it should be assigned more weight in order to promote its selection to improve the overall system performance.

Another direction of improvement is to incorporate *peer-to-peer (P2P) network* along with *prefetching*. In that case, any mobile device after playing a video segment does not delete it immediately. It rather preserves the segment for a while as a potential provider in a P2P network. We can use the WiFi Direct protocol [8] to create the mobile P2P network. The benefits of the WiFi P2P network using mobile devices are many folds. Base stations in wireless networks have limited resources. Since in P2P networks mobile devices are usually served by another provider mobile device without involving the base station, the P2P reduce the network load on mobile stations and more mobile devices can be served this way. It also reduces data cost as 3G data is usually more expensive compared to that of WiFi. Moreover, data reception over WiFi instead of 3G significantly reduces energy consumption for mobile devices as WiFi consumes less energy than 3G for data transfer [16]. There are challenges associated with this approach as well. Mobile devices that work as providers in P2P network consume more energy, because in addition to normal energy consumption for receiving data from base stations, they have to consume energy while serving other mobile devices in a P2P network. So, the decision should be made whether the gain achieved by serving through P2P network out-weigh the cost incurred by the provider mobile device in terms of energy consumption along with other criteria (e.g. base station load reduction). Moreover, the fast movement of mobile devices make the mobile P2P network more challenging and this by itself is another active research field.

Bibliography

[1] C. Aggarwal, J. Wolf, and P. Yu. On optimal batching policies for video-on-demand storage servers. In *Proc. of IEEE International Conference on Multimedia Computing and Systems (MMCS'96)*, pages 253–258, Hiroshima, Japan, June 1996. 14

[2] C. Aggarwal, J. Wolf, and P. Yu. On optimal piggyback merging policies for video-on-demand systems. In *Proc. of ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'96)*, pages 200–209, New York, NY, USA, May 1996. 14

[3] M. Anand, E.B. Nightingale, and J. Flinn. Self-tuning wireless network power management. *Wireless Networks*, 11(4):451–469, July 2005. 15

[4] G. Ananthanarayanan, V.N. Padmanabhan, L. Ravindranath, and C.A. Thekkath. Combine: leveraging the power of wireless peers through collaborative downloading. In *Proc. of the 5th ACM international conference on Mobile systems, applications and services (MobiSys '07)*, pages 286–298, New York, NY, USA, June 2007. 18

[5] D. Astely, E. Dahlman, A. Furuskar, Y. Jading, M. Lindstrom, and S. Parkvall. LTE: the evolution of mobile broadband. *IEEE Communications Magazine*, 47(4):44–51, April 2009. 7

[6] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proc. of the 9th ACM SIGCOMM conference on Internet Measurement Conference (IMC'09)*, pages 280–293, New York, NY, USA, November 2009. 10

[7] A. Bar-Noy, J. Goshi, R. Ladner, and K. Tam. Comparison of stream merging algorithms for media-on-demand. *Multimedia Systems*, 9(5):411–423, 2004. 14, 23, 44

[8] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano. Device-to-device communications with Wi-Fi Direct: overview and experimentation. *IEEE Wireless Communications*, 20(3):96–104, June 2013. 62

[9] J. Chen, K.M. Sivalingam, and P. Agrawal. Performance comparison of battery power consumption in wireless multiple access protocols. *Wireless Networks*, 5(6):445–460, December 1999. 16

[10] X. Cheng, C. Dale, and J. Liu. Statistics and social network of YouTube videos. In *Proc. of the 16th International Workshop on Quality of Service, 2008 (IWQoS'08)*, pages 229–238, Enschede, Netherlands, June 2008. 36

[11] C. Chiang. *Wireless Network Multicasting*. PhD thesis, University of California Los Angeles, 1998. 17

[12] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund. Quality of service support in IEEE 802.16 networks. *IEEE Network Magazine*, 20(2):50–55, March 2006. 1

[13] E. Dahlman, S. Parkvall, and J. Sköld. *4G LTE/LTE-Advanced for Mobile Broadband*. Elsevier Ltd., March 2011. 11, 13, 28

[14] C. Dovrolis and P. Ramanathan. A case for relative differentiated services and the proportional differentiation model. *IEEE Transactions on Network*, 13(5):26–34, September 1999. 7

[15] F.H.P. Fitzek and M.D. Katz. *Cooperation in Wireless Networks: Principles and Applications*. Springer, July 2006. 18

[16] N. Gautam, H. Petander, and J. Noel. A comparison of the cost and energy efficiency of prefetching and streaming of mobile video. In *Proc. of the 5th ACM Workshop on Mobile Video (MoVid'13)*, pages 7–12, New York, NY, USA, February 2013. 62

[17] L. Golubchik, J. Lui, and R. Muntz. Adaptive piggybacking: a novel technique for data sharing in video-on-demand storage servers. *Multimedia Systems*, 4(3):140–155, June 1996. 14

[18] C. Griwodz, M. Liepert, M. Zink, and R. Steinmetz. Tune to lambda patching. *ACM SIGMETRICS Performance Evaluation Review*, 27(4):20–26, March 2000. 14, 23, 44

[19] F. Hartung, U. Horn, J. Huschke, M. Kampmann, T. Lohmar, and M. Lundevall. Delivery of broadcast services in 3G networks. *IEEE Transactions on Broadcasting*, 53(1):188–199, March 2007. 17

[20] O. Hillestad, A. Perkis, V. Genc, S. Murphy, and J. Murphy. Delivery of on-demand video services in rural areas via IEEE 802.16 broadband wireless access networks. In *Proc. of ACM international workshop on Wireless multimedia networking and performance modeling*, pages 43–52, New York, NY, USA, October 2006. 15

[21] H. Hlavacs and S. Buchinger. Hierarchical video patching with optimal server bandwidth. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP'08)*, 4(1):1–23, January 2008. 14, 23, 44

[22] H. Hlavacs and S. Buchinger. Optimal server bandwidth for mobile video on demand. *Annals of Telecommunications*, 65(1):31–46, February 2010. 14, 15

[23] K. Hua, M. Tantaoui, and W. Tavanapong. Video delivery technologies for large-scale deployment of multimedia applications. *Proc. of the IEEE*, 92(9):1439–1451, September 2004. 14

[24] C.E. Jones, K.M. Sivalingam, P. Agrawal, and J.C. Chen. A survey of energy efficient network protocols for wireless networks. *Wireless Networks*, 7(4):343–358, September 2001. 10

[25] L. Juhn and L. Tseng. Harmonic broadcasting for video-on-demand service. *IEEE Transactions on Broadcasting*, 43(3):268–271, September 1997. 14

[26] A.A. Khalek and Z. Dawy. Energy-efficient cooperative video distribution with statistical QoS provisions over wireless networks. *IEEE Transactions on Mobile Computing*, 11(7):1223–1236, July 2012. 18

[27] J. Kim, T. Kwon, and D. Cho. Resource allocation scheme for minimizing power consumption in OFDM multicast systems. *IEEE Communications Letters*, 11(6):486–488, June 2007. 11

[28] S. Lederer, C. Müller, and C. Timmerer. Dynamic adaptive streaming over HTTP dataset. In *Proc. of the 3rd Multimedia Systems Conference (MMSys '12)*, pages 89–94, New York, NY, USA, February 2012. 28

[29] J. Lee, H. Park, S. Choi, and J. Choi. Adaptive hybrid transmission mechanism for on-demand mobile IPTV over WiMAX. *IEEE Transactions on Broadcasting*, 55(2):468–477, June 2009. 15, 18

[30] J.S. Lee and L.E. Miller. *CDMA Systems Engineering Handbook*. Artech House, Inc., January 1999. 56

[31] K. Lee. Adaptive network support for mobile multimedia. In *Proc. of the 1st ACM international conference on Mobile computing and networking (MobiCom'95)*, pages 62–74, New York, NY, USA, December 1995. 7

[32] P. Lettieri, C. Fragouli, and M.B. Srivastava. Low power error control for wireless links. In *Proc. of the 3rd ACM international conference on Mobile computing and networking (MobiCom '97)*, pages 139–150, New York, NY, USA, September 1997. 15

[33] Q. Liu, X. Wang, and G.B. Giannakis. A cross-layer scheduling algorithm with QoS support in wireless networks. *IEEE Transactions on Vehicular Technology*, 55(3):839–847, May 2006. 7

[34] C.E. Luna, Y. Eisenberg, R. Berry, T.N. Pappas, and A.K. 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. 16

[35] A. Majumdar, D. Sachs, I. Kozintsev, K. Ramchandran, and M. Yeung. Multicast and unicast real-time video streaming over wireless LANs. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(6):524–534, June 2002. 15

[36] D. Nguyen, T. Nguyen, and X. Yang. Multimedia wireless transmission with network coding. In *Packet Video 2007*, pages 326–335, Lausanne, Switzerland, November 2007. 17

[37] D. Ning, A. Pathak, D. Koutsonikolas, C. Shepard, Y.C. Hu, and L. Zhong. Realizing the full potential of PSM using proxying. In *Proc. of INFOCOM*, pages 2821–2825, Orlando, FL, USA, March 2012. 10, 15

[38] A. Nosratinia, T.E. Hunter, and A. Hedayat. Cooperative communication in wireless networks. *IEEE Communications Magazine*, 42(10):74–80, October 2004. 18

[39] A.J. Pyles, X. Qi, G. Zhou, M. Keally, and X. 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, USA, September 2012. 15

[40] M. Ramadan, L.E. Zein, and Z. Dawy. Implementation and evaluation of cooperative video streaming for mobile devices. In *Proc. of the 19th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2008 (PIMRC 2008)*, pages 1–5, Cannes, France, September 2008. 18

[41] V. ŘÍČNÝ. Single Frequency Networks (SFN) in digital terrestrial broadcasting. *Radioengineering*, 16(4):3, December 2007. 11

[42] E. Rozner, V. Navda, R. Ramjee, and S. Rayanchu. NAPman: network-assisted power management for WiFi devices. In *Proc. of the 8th ACM international conference on Mobile systems, applications, and services (MobiSys'10)*, pages 91–106, New York, NY, USA, June 2010. 10

[43] M. Schaar and P. A. Chou. *Multimedia over IP and Wireless Networks: Compression, Networking, and Systems*. Academic Press, March 2007. 6, 7

[44] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain, and V.N. Padmanabhan. Bartendr: a practical approach to energy-aware cellular data scheduling. In *Proc. of the 16th ACM international conference on Mobile computing and networking (MobiCom '10)*, pages 85–96, New York, NY, USA, September 2010. 16, 17

[45] K.M. Sivalingam, J. Chen, P. Agrawal, and M.B. Srivastava. Design and analysis of low-power access protocols for wireless and mobile ATM networks. *Wireless Networks*, 6(1):73–87, February 2000. 16

[46] D. Tran and T. Nguyen. Broadcasting techniques for video-on-demand in wireless networks. In *Handbook of Mobile Broadcasting: DVBH, DMB, ISDB-T, and MediaFLO*, pages 675–696. Auerbach Publications, April 2008. 14

[47] M. Wang, C. Feng, and T. Zhang. Min-max load balancing scheme for mixed multicast and unicast services in LTE networks. *The Journal of China Universities of Posts and Telecommunications*, 19(2):30 – 35, April 2012. 17

[48] S. Wang, T. Lin, Y. Wang, C. Hsu, and X. Liu. Fusing prefetch and delay-tolerant transfer for mobile videos. In *Proc. of the 11th ACM international conference on Mobile systems, applications, and services (MobiSys '13)*, pages 525–526, New York, NY, USA, June 2013. 20

[49] J. Yoon, H. Zhang, S. Banerjee, and S. Rangarajan. MuVi: a multicast video delivery scheme for 4G cellular networks. In *Proc. of ACM International Conference on Mobile Computing and Networking (Mobicom'12)*, pages 209–220, New York, NY, USA, August 2012. 15

[50] Y. Yu, P. Hsiu, and A. Pang. Energy-efficient video multicast in 4G wireless systems. *IEEE Transactions on Mobile Computing*, 11(10):1508–1522, October 2012. 11, 22

[51] Q. Zhang, F.H.P. Fitzek, and V.B. Iversen. Design and performance evaluation of cooperative retransmission scheme for reliable multicast services in cellular controlled P2P networks. In *Proc. of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007 (PIMRC 2007)*, pages 1–5, Athens, Greece, September 2007. 18

[52] X. Zhang and K.G. Shin. E-MiLi: energy-minimizing idle listening in wireless networks. *IEEE Transactions on Mobile Computing*, 11(9):1441–1454, September 2012. 15

[53] L. Zhao, A. Y. Al-Dubai, and G. Min. GLBM: a new QoS aware multicast scheme for wireless mesh networks. *Journal of Systems and Software*, 83(8):1318 – 1326, August 2010. 17

[54] Y. Zhao, N. Do, S. Wang, C. Hsu, and N. Venkatasubramanian. Enabling efficient offline access to online social media and social networks. Technical report, Networking and Multimedia Systems Lab, National Tsing Hua University, May 2013. 20

[55] D. Zhu and M. Mutka. Cooperation among peers in an Ad Hoc network to support an energy efficient IM service. *Pervasive and Mobile Computing*, 4(3):335–359, June 2008. 18

[56] H. Zhu and G. Cao. On supporting power-efficient streaming applications in wireless environments. *IEEE Transactions on Mobile Computing*, 4(4):391–403, July 2005. 16

[57] M. Zorzi and R.R. Rao. Energy-constrained error control for wireless channels. *IEEE Personal Communications*, 4(6):27–33, December 1997. 17

[58] Evolved Universal Terrestrial Radio Access (E-UTRA); physical layer procedures (release 9). Third Generation Partnership Project (3GPP) standard TS 36.213 ver. 9.2.0, 2010. 28

[59] Improved Video Support for Packet Switched Streaming (PSS) and Multimedia Broadcast/Multicast Service (MBMS) Services (Release 9). Third Generation Partnership Project (3GPP) Standard TR 26.903 Ver. 9.0.0, March 2010. 1, 3

[60] Bharti Airtel launches India's first 4G service. http://www.airtel.in/about-bharti/media-centre/bharti-airtel-news/mobile/airtel-launches-india-s-first-4g-service. 9

[61] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2012-2017. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html. 1

[62] Global Mobile Data Traffic Forecast by Cisco. http://arstechnica.com/tech-policy/2011/03/world-mobile-data-traffic-toexplode-by-factor-of-26-by-2015/. 10

[63] GSM Arena. http://www.gsmarena.com/. 10

[64] IBM ILOG CPLEX. http://www.ibm.com/software/integration/optimization/cplex-optimizer. 27, 30, 34

[65] Long Term Evolution (LTE) - Alcatel-Lucent. http://www.unispheresoluciones.com/archivosparadescargar/alcatel_lucent_lte_overview.pdf. 8, 9

[66] Long Term Evolution (LTE) - Motorola. http://www.motorola.com/web/Business/Solutions/Industry Solutions/Service Providers/Wireless Operators/LTE/_Document/Static Files/6833_MotDoc_New.pdf. 8

[67] LTE. http://www.3gpp.org/lte. 7, 8

[68] MATLAB - The Language of Technical Computing. http://www.mathworks.com/ products/matlab/. 27, 30, 34

[69] Mobile TV is back: Ericsson launches broadcast video for 4G. http://www.theregister.co.uk/2013/02/25/broadcast_tv_yet_again/. 1

[70] OPNET Modeler Suite. http://www.opnet.com. 34, 35

[71] Recommended bit rates for live streaming. http://www.adobe.com/devnet/adobe-media-server/articles/dynstream_live/popup.html. 28

[72] Report: smart phones overtake client PCs in 2011. http://www.canalys.com/newsroom/smart-phones-overtake-client-pcs-2011. 9

[73] Smartphone. http://en.wikipedia.org/wiki/Smartphone. 9

[74] TeliaSonera, First in the world with 4G services. http://www.teliasonera.com/en/media/pressreleases/2009/12/teliasonera-first-in-the-world-with-4g-services/. 9

[75] Verizon 4G LTE speed test using Droid Bionic (video). http://www.smartkeitai.com/verizon-4g-lte-speed-testmotorola- droid-bionic-video/. 9

[76] Verizon Wireless rolled out their LTE network in 38 major markets on December 5, 2010, Happy 1st Anniversary, Verizon Wireless 4G LTE! http://news.verizonwireless.com/news/2011/12/pr2011-12-05a.html. 9

[77] Walking. https://www.princeton.edu/ achaney/tmve/wiki100k/docs/Walking.html. 57

[78] YouTube Data API (v3). https://developers.google.com/youtube/v3/. 36