UNIVERSITY OF CALIFORNIA, SAN DIEGO

Green Cellular Networks through QoS Aware Dynamic Base Station - Mobile Device Reconfiguration

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Electrical Engineering (Computer Engineering)

by

Ranjini B. Guruprasad

Committee in charge:

Professor Sujit Dey, Chair
Professor Rajesh Gupta
Professor Bhaskar Rao
Professor Ramesh Rao
Professor Tajana Rosing
Professor Alex Snoeren

2017

The Dissertation of Ranjini B. Guruprasad is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

DEDICATION

To My Parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation is a culmination of many years of effort, which if not had been fully supported and encouraged by my family, mentors, friends and colleagues, would have remained a distant dream.

Starting from the day I stood outside Prof. Dey's class to seek admission in to the PhD program, Prof. Dey has given me complete support and understanding throughout the PhD journey. It is his guidance and patience that have shaped the ability in me to think, research and solve interesting problems. I would also like to thank my committee members for their feedback on the work and for useful discussions.

I would like to acknowledge Dr. Kyuho Son for the long and patient discussions during our collaboration which improved my understanding of various technical aspects. I would like to thank Prof. Santhosh Nagaraj and Prof. Mahasweta Sarkar for their valuable inputs.

Some of the most memorable moments of graduate school were spent with my lab mates Chetan, Yao, Ali, Hasti and Po-Han. Discussion with them regarding our respective PhD journeys and things outside graduate school has helped me grow as an individual. I would like to especially thank Chetan who has always been there whenever I needed a helping hand. Many thanks to Maryam Sukosh who helped me on innumerable occasions to navigate through the set up for TA sessions and for patiently listening and encouraging me to go on when I was experiencing low moments. Chris Shenton from graduate office has been my go-to person for arranging logistics and waivers during my stay away from university. Thanks a lot Chris!!!

I would like to thank my parents who have always been my pillars of support in all my endeavors and have stood rock solid with me during the most of challenging times. My father always believed in my abilities and infused in me positive approach and fighting spirit to take on challenges. My mother, who is my most vocal critic is

also the person without whom this thesis would not have seen the light of day. It is her willingness to stand by me that has enabled me to complete my PhD and fulfill a dear dream. I cannot thank Medha enough for brightening my days with her smile and bear hugs and for her patience when I have been away from her while trying to complete my research and thesis writing. Dhruv, Rashmi, Karthik and Prabha Aunty have been my sound boards and continuously encouraged me to complete my PhD and assured me that things will fall in place. The achievement of completing my PhD degree equally belongs to me and my parents, Medha, Dhruv, Rashmi and Karthik.

Most importantly, my thanks and prayers to The Almighty for all the blessings and allowing me to complete my PhD.

Chapter 2, in part, contains material as it appears in the Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'13). "Rate Adaptation and Base Station Reconfiguration for Battery Efficient Video Download". Ranjini Guruprasad, Sujit Dey. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, contains material as it appears in IEEE Transactions on Multimedia. "Battery Aware Video Delivery Techniques Using Rate Adaptation and Base Station Reconfiguration". Ranjini Guruprasad, Sujit Dey. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, contains material as it appears in Journal of Communications and Networks (JCN): Special issue on Green Communication. "Dynamic Cell Reconfiguration Framework for Energy Conservation in Cellular Wireless Networks". Kyuho Son, Ranjini Guruprasad, Santhosh Nagraj, Mahasweta Sarkar, Sujit Dey. The dissertation author was the primary student investigator and author of this paper.

Chapter 4, in part, contains material as it appears in the Proceedings of the Internation Conference on Communications (ICC'15). "Power-efficient Base Station

Operation through User QoS-aware Adaptive RF Chain Switching Technique”. Ranjini Guruprasad, Kyuho Son, Sujit Dey. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, contains material as it appears in IEEE Transactions on Green Communications and Networking. “User QoS-aware RF Chain Switching for Power Efficient Co-operative Base Stations”. Ranjini Guruprasad, Sujit Dey. The dissertation author was the primary investigator and author of this paper

VITA

2002	Bachelor of Engineering, Electronics and Communications People's Education Society Institute of Technology (PESIT)
2005	Master of technology, Advanced Electronics and Digital Communications National Institute of Technology Karnataka (NITK), Surathkal
2010–2013	Research Assistant, Department of Electrical and Computer Engineering University of California, San Diego
2010 – 2017	Doctor of Philosophy, Electrical Engineering (Computer Engineering) University of California, San Diego

PUBLICATIONS

Journal

R. Guruprasad, S. Dey, "User QoS-aware RF Chain Switching for Power Efficient Cooperative Base Stations" Accepted for publication in IEEE Transactions on Green Communications and Networking

K. Son, **R. Guruprasad**, S. Dey, "Dynamic Cell Reconfiguration Framework for Energy Conservation in Cellular Wireless Networks" In Journal of Communications and Networks, vol. 18 (4), pp 567–579, August 2016

R. Guruprasad, S. Dey, "Battery Aware Video Delivery Techniques Using Rate Adaptation and Base Station Reconfiguration" In IEEE Transactions on Multimedia, vol. 17 (9), pp 1630–1645, September 2015

Conferences

R. Guruprasad, P. Murali, D. Krishnaswamy, S. Kalyanaraman, "Coupling a Small Battery with Data Center for Fast Frequency Regulation", Accepted in IEEE Power and Energy Society (PES), July 2017

P. H. Chiang, **R. Guruprasad**, S. Dey, "Renewable Energy-Aware Video Download in Cellular Networks" In Proc. of IEEE Wireless Communications and Networking

Conference (WCNC), pp 1622–1627, September 2015

R. Guruprasad, K. Son, S. Dey, “Power-Efficient Base Station Operation through User QoS-Aware Adaptive RF Chain Switching Technique” In Proc. of International Conference on Communications (ICC), pp 244–250, June 2015

R. Guruprasad, S. Dey, “Rate Adaptation and Base Station Reconfiguration for Battery Efficient Video Download” In Proc. of IEEE Wireless Communications and Networking Conference (WCNC), pp 339–344, April 2013

ABSTRACT OF THE DISSERTATION

Green Cellular Networks through QoS Aware Dynamic Base Station - Mobile Device Reconfiguration

by

Ranjini B. Guruprasad

Doctor of Philosophy in Electrical Engineering (Computer Engineering)

University of California, San Diego, 2017

Professor Sujit Dey, Chair

Anytime-anywhere connectivity offered by cellular networks and mobile devices with multimedia capabilities have revolutionized important sectors of the society such as health care, education, finance, e-commerce and entertainment. To cater to the resulting explosive growth in mobile data traffic in an economically and environmentally sustainable manner, it is critical to efficiently manage the spectral and energy/power consumption of cellular networks. In this thesis, we identify the key challenges faced by the cellular networks in efficiently managing energy/power consumption and propose solutions to alleviate the same.

Rapid advances in processing capabilities of mobile devices and relatively slower advances in battery capacity capabilities has created a huge gap between power required for processing advanced multimedia applications and the available battery capacity. Data and compute intensive mobile video is the leading multimedia application and leads to quick drain in the mobile battery level. In the first part of the thesis, we address the above challenge by developing battery aware mobile video download techniques that increase the battery available time while maintaining the required user experience levels. Extensive experiments have demonstrated the feasibility and efficacy of our approach.

Base stations are the dominant contributors to power consumption of cellular networks. To ensure that quality of service requirements is always met, base stations are over provisioned to handle maximum load and are always switched on. This is leads to wasteful expenditure of electricity when load is less than maximum. To address this, we develop techniques that adapt the coverage area of base stations depending on load to reduce base station power consumption. Simulation experiments have demonstrated the significant power savings is possible using the proposed techniques.

Multi-input, multi-output technologies which require multiple Radio Frequency (RF) chains are being adopted to increase the data rates and coverage capabilities of base stations. This implies that the already dominant contribution of RF chains to power consumption of base stations will significantly increase. We conclude the thesis by developing techniques that switch off RF chains depending on load to reduce base station power consumption. Simulation experiments demonstrate the power savings possible using proposed techniques compared to existing techniques.

Chapter 1

Introduction

In the past three decades, cellular networks and mobile devices have spurred a tremendous growth in connectivity and information availability across the world. The connectivity and information availability has played a critical role in successful digitization of sectors like banking, finance, education, health care, transportation and hospitality. Further, innovations using advanced technologies such as augmented reality, virtual reality, Internet of things (IoT) have led to real-time and interactive gaming, e-commerce, social connectivity platforms that have become an integral part of urban lifestyle. This is evident by the estimates that the total number of mobile subscriptions is expected to increase from 7.5 billion in 2016 to 9 billion by 2022 with a compound annual growth rate of 3% [1] and exceeds the estimated global population [2]. The growth in connectivity is accompanied by significant increase in worldwide total mobile data volume from 8.8EB to an estimated 71 EB (with a CAGR of 42%) with mobile video contributing to more than 75% of the mobile data [1]. To cater to the explosive growth in connectivity and mobile data volume in an economically and environmentally sustainable manner, it is crucial to efficiently manage the limited spectral and energy resources available. In this thesis, we will address the challenges in efficiently managing the energy resources of mobile devices and power consumption of cellular networks and propose solutions to alleviate the same.

1.1 Power/Energy Needs of Mobile Devices and Cellular Networks

To cater to the explosive growth in mobile data subscriptions and traffic, it is estimated that the total number of base stations (BSs) in cellular networks all over the world will grow to 11.2 million by 2020 [3], a 47% increase compared to the number of BSs deployed in 2014. Further, deployment of massive number of antennas at BSs is seen as a promising paradigm to increase data rates [4]. This is expected to increase the electricity consumption and thereby, decrease the energy efficiency of cellular networks [4]. Specifically, the electricity consumption of BSs which constitutes 80% of electricity consumption of cellular networks is estimated to increase from 84TWh to 109TWh by 2020 (38% increase since 2014) if measures are not taken to reduce the power consumption of BSs. The increasing electricity consumption has two effects - (a) the carbon equivalent emissions is estimated to increase to 235 Mto CO2e by 2020 (a 37% increase since 2014) [3] and (b) the electricity bill which currently contributes to 10-15% of the operating expenses in developed markets and about 50% [5] in developing markets. Hence, it is critical to decrease the power consumption of BSs to enable the cellular networks operate in an economically and environmentally sustainable manner.

The explosive growth in data traffic in mainly due to data intensive multimedia applications such as web browsing, mobile video, gaming, augmented reality and virtual reality. The mobile devices are continuously evolving with increasingly complex processing architectures to support such data and compute intensive multimedia applications. To power the data and compute intensive applications, mobile device manufacturers are equipping the mobile devices with increasingly higher energy density batteries. However, there is a limit to increasing the battery energy density as it proportional to the thickness of the battery which in turn leads to bulky handsets. This has resulted in a gap between the power required and available battery energy density. The energy density gap has to be minimized to realize the promise of many advanced techniques such as augmented reality, virtual reality, Internet of Things and applications of the same.

1.2 Contributions and Overview

In this dissertation, we focus on several approaches to reduce the power/energy consumption of cellular networks and mobile devices and demonstrate the efficacy of the proposed approaches through experiments. The first key contribution of the dissertation addresses the energy density gap of mobile device battery by focusing on reducing the battery consumption due mobile video download. The second key contribution addresses the reduction of power consumption of the BSs at the system level enabling economically and environmentally sustainable operation of cellular networks. The third key contribution focuses on the power consumption of the BSs at the component level, thus enabling a fine grained control on the power consumption of the BSs. We first provide an overview of our contributions, followed by individual chapters that give detailed treatment to each of the contributions, including the related literature and how our work relates to and differentiates from the existing literature. Our contributions are summarized below.

1.2.1 Battery Aware Video Download Techniques

Broadly, there has been a two pronged approach to address the gap in power required and available battery energy density of mobile devices. One approach is to increase the battery density by using materials and packing architectures that lend themselves to increased energy density with a small form factor [6]. The other approach has been to develop and implement battery aware application processing techniques on the mobile device such as [7]. The battery aware techniques proposed in this thesis belong to the second category that addresses the battery energy density gap. In particular, we focus on developing battery aware video download techniques because mobile video contributes to over 75% of mobile data traffic [1].

Downloading and viewing mobile video on mobile devices has been on a steady increase from 5% in 2010 to 20% in 2016 [8]. However, there still exists barriers to wider adoption to mobile video download and viewing such as limited battery capacity and connectivity . While advanced 4G and 5G networks are being designed to improve the connectivity, there is still a void in techniques that address the challenge of limited battery capacity. Commercially available video download and streaming clients such as YouTube, Apple's HTTP Live streaming (HLS) and Microsoft's Smooth Streaming focus on optimizing the user experience but does not take in to account the effect of video download on battery consumption.

As power consumption due to video download over the cellular wireless connection exceeds that due to video playback [9], we focus on selecting the optimal physical layer parameters involved in video download to minimize the battery consumption during video download. We achieve this via reconfiguration of base station physical layer components such as number of radio frequency (RF) chains, multi-input multi-output (MIMO) transmission scheme, modulation order, coding rate, and download rate and video bit rate adaptation to minimize the battery consumption of the mobile device. The base station reconfiguration and rate adaptation is carried out in a manner that there is no degradation in user experience. We discuss and evaluate our framework for video download as well as streaming scenarios. Through experiments using real world channel conditions and power consumption models based on actual implemented hardware, we demonstrate that the proposed battery aware techniques result in significant savings in battery lifetime and no degradation in user experience compared to non-battery aware video download and streaming techniques.

1.2.2 Dynamic Cell Reconfiguration Framework

The last decade has seen significant research and commercial deployment of energy efficient BSs, including energy efficient power amplifiers and baseband processing [10], [11], [12]. Energy efficient wireless protocols and network techniques have been also proposed that take advantage of variable traffic loads and QoS requirements. The second key contribution of this thesis proposes energy efficient cellular network techniques that minimize the power consumption of BSs at the system level while satisfying the QoS of associated users.

The various components of the BS can be grouped in to two categories. The first category of components contribute the static power consumption of the BS and is a constant irrespective of the load. The second category contribute to the dynamic power consumption of the BS and is dependent on the BS load. Current base stations are designed to handle worst case load. Further, in order to ensure that there is no loss in coverage, BSs are always maintained on. This implies that when the load is low or there is no load, there is unnecessary expenditure of energy due to the static power consumption.

Depending on the load, BSs can be switched off by transferring users to neighboring active (on) BSs resulting in significant static power savings. Further, the transmit power budget can be adapted depending on the load. The second key contribution of the thesis is an integrated dynamic cell reconfiguration framework that dynamically switches on/off base stations and adapts the user association and transmit power budget of BSs depending on the load. We discuss and evaluate the framework under static and dynamic BS load conditions. Using measurements from actual BS power consumption and real world BS traffic traces, we demonstrate that the proposed dynamic cell reconfiguration techniques result in higher power savings compared to techniques that always maintain

BSs on.

1.2.3 QoS Aware RF Chain Switching

The dynamic cell reconfiguration techniques result in significant savings in BS power consumption as both static and dynamic components of BS power consumption is reduced to zero. However, switching off BS has the major limitation of creating coverage holes which can result in degradation of user Quality of Service (QoS). Further, BS switch off requires tens of minutes and hence termed as long time scale techniques. Such long time scale techniques cannot exploit the fine time scale variations of BS load.

Taking cognizance of the above limitations of the system level BS power minimization techniques and identifying further opportunity to minimize power at finer time scales, the third key contribution of this thesis are techniques that increase the power efficiency of BSs at the component level at times scales of seconds to minutes. The power amplifier in the radio frequency (RF) chain is the dominant contributor to the BS power consumption. The final contribution of this thesis is the RF chain switching technique which minimizes the power consumption of cluster of BSs in a manner that the QoS requirements of all the cluster users and BS utilization bounds of individual BSs in the cluster are satisfied.

The adaptive RF chain switching technique achieves the above by jointly adapting the number of RF chains, time slots and frequency blocks of individual BSs and user association of cluster users to minimize the number of RF chains in the cluster and thereby, power consumption of the cluster of BSs. The short time scale technique allows finer control on BS power consumption and does not result in coverage holes. Using measurements from actual BS power consumption and real world BS traffic traces, we demonstrate that the proposed adaptive RF chain switching techniques result in higher power savings compared to techniques that always maintain RF chains on.

Chapter 2

Battery Aware Video Download techniques using Rate Adaptation and Base Station Reconfiguration

2.1 Introduction

By 2022, mobile video is expected to contribute to about more than 75% of the total mobile data traffic [1], making it the leading multimedia application on mobile devices. As mobile video is a data and compute intensive application, it places significant demands on processing and battery capabilities of mobile devices. While the processing capabilities of mobile devices continue to increase significantly, the incremental improvements in battery technologies will lead to frustratingly lower battery lifetime. Consequently, it is critical to develop techniques that can lower mobile video battery consumption. It has been shown that RF and baseband components used for video download are major contributors to battery consumption in addition to decoder and display used for playback [13]. With the adoption of MIMO technologies that use multiple antennas with power consuming baseband processing, power due to radio frequency (RF) and baseband components will dominate the power consumption for high bit rate mobile video applications. Hence, this chapter focuses on reducing battery demand imposed by MIMO RF and baseband components while downloading video.

We first consider the widely adopted Progressive Download video delivery approach, which attempts to download video at a rate higher than the video bit rate and hence the video playback rate, thereby buffering video at the mobile device while it is simultaneously being played back [14]. The higher download rate and hence buffering is done to avoid buffer underflow (stalling) in case of bad network conditions during the video session, but there is no consideration about the effect of video download on the mobile device battery. In contrast, we propose a new battery efficient video download approach that utilizes elasticity of the video buffer to dynamically adapt the video download rate, sometimes even stopping video download, enabling reconfiguration or idling of the base station RF and baseband components in a manner that reduces or eliminates battery demand of the mobile device RF and baseband components. While adapting the download rate, the proposed approach also tries to avoid buffer underflow, and since the video bit rate is never adapted, user experience is not compromised while enhancing battery lifetime.

To further enhance battery lifetime, we next consider adapting the video bit rate in addition to adapting the video download rate as the former can further reduce the amount of data to be downloaded and hence the battery load. However, adapting the video bit rate will compromise video quality, leading to a possible tradeo-off between enhanced longevity of video experience and video quality. Adaptive Bit Rate (ABR) streaming techniques [15] are gaining popularity, but they primarily address minimizing stalling of video under challenging network conditions. In contrast, we propose battery aware adaptive bit rate streaming techniques which adapt video bit rate, download rate and MIMO RF and baseband configurations, depending on the battery and buffer levels, and network load and channel conditions experienced during video streaming to maximize the longevity of video experience while ensuring a desired level of quality of video experience. We extend the conventional notion of video user experience to include the longevity of video watching (which can be limited by battery lifetime) by introducing the Video Experience Longevity (VEL) metric. We use the VEL metric to quantify and compare the performance of the proposed battery aware ABR techniques with other ABR techniques. As dynamic streaming over HTTP (DASH) is a widely accepted standard for ABR streaming, we will henceforth refer to ABR as DASH.

2.1.1 Related Work

In this section, we will briefly describe past work related to base station and mobile device MIMO reconfiguration, video bit rate adaptation and battery efficient video delivery. As we will discuss, either these techniques do not address maximizing battery lifetime, or the ones that address do not consider using rate adaptation and transceiver reconfiguration whose effectiveness we will demonstrate in this chapter.

Base station reconfiguration techniques have been developed for cognitive radios for dynamic spectrum management [16], which is not the focus of the work presented in this chapter. The focus in [17] was on choosing optimal MIMO parameter set to minimize overall link energy while satisfying bit error rate and throughput. While the above technique does not consider video delivery, [18] proposed to use Space Time Multiplexing (STM) and Space Time Block Coding (STBC) to reduce video distortion due to wireless video delivery; however, the latter does not address energy consumption. In [19], rate adaptation and corresponding switching between Single Input Multi Output (SIMO) and MIMO is proposed to save uplink RF transmission energy when mobile device is transmitting files. However, [19] does not aim to reduce downlink RF and baseband processing battery consumption when mobile device is downloading video, which is the objective of this work. The energy efficient rate adaptation (EERA) technique proposed in [20] achieves energy efficiency at the client by selecting RF and MIMO baseband components at the access point and client Wireless Network Interface Card

(WNIC) in a manner that reduces per bit energy while maintaining the minimum required goodput determined by the video bit rate and channel condition. However the energy efficient rate adaptation technique proposed in [20] does not utilize the elasticity of the video buffer to dynamically adapt the video download rate, including stopping transmission, to avoid stalling and reduce battery load, which constitutes an important part of our proposed approach. Also, mode selection in [20] requires base station to allocate maximum number of antennas to each user which places high demand on base station resources whereas our techniques have no such requirement.

Recently, there has been significant research done on developing video bit rate adaptation techniques [21], [22], including several commercial HTTP based Adaptive Bit Rate video streaming solutions like Apple HTTP Live Streaming [23], Microsoft Smooth Streaming [24] and Adobe Open Source Media Framework (OSMF)[1]. Unlike the above adaptive HTTP streaming clients and techniques which to the best of our knowledge (based on available public information at the time of writing this manuscript, including the Adobe OSMF source files) focus on ensuring user experience in a non-battery aware manner, our proposed techniques focus on maximizing battery lifetime while also ensuring desired level of video experience.

Techniques have also been developed to address energy and battery life of mobile devices during video delivery. In [25], a base station scheduling technique is proposed which utilizes the Variable Bit Rate (VBR) encoding of multiple broadcast streams in a manner that does not under/overflow the client buffers and allows transmission of video streams in bursts, the latter allowing switching off the client WNIC in between bursts to reduce energy consumption on mobile devices. However, the above approach cannot be applied to on-demand unicast video delivery (like YouTube) which is the target of the work detailed in this chapter. Authors in [26] propose battery aware video streaming by

[1] ”Open source media framework,” [Online]. Available: http://www.osmf.org

changing video encoding parameters such as bit rate, frames/second in real time using a proxy server and switching off the client WNIC after bulk download. Our proposed approach does not require computationally intensive real time transcoding and utilizes different bit rate representations of a given video available on the server for video bit rate adaptation. Battery and stream aware adaptive multimedia (BaSe-AMy) streaming techniques proposed in [27] adapt video bit rate depending on battery level, packet loss and remaining video stream duration. However, these techniques do not adapt download rate and transceiver configuration which increases the battery savings achieved by our proposed techniques.

To the best of our knowledge, this is the first work which proposes to (a) jointly adapt video download rate and MIMO transceiver components to maximize battery lifetime and ensure user experience during video download and (b) additionally adapt video bit rate to maximize video experience longevity while maintaining desired level of video experience during adaptive bit rate streaming; (c) quantify the performance of adaptive bit rate streaming techniques in terms of both video viewing time and user experience. In Section 2.2, we provide an overview of our battery aware video delivery approach. In Section 2.3, we formulate the download rate and transceiver configuration selection as an optimization problem and provide a solution. In Section 2.4, we present the simulation framework developed for video download and experimental results obtained using different video download techniques. In Section 2.5, we formulate bit rate, download rate and transceiver configuration selection as an optimization problem and offer a solution which guarantees minimum desired video quality, and subsequently extend the solution with a heuristic to achieve higher video quality when possible. We conclude the section with formulation of the “Video Experience Longevity” metric. In Section 2.6, we present the simulation framework developed for DASH streaming and experimental results obtained using different DASH based streaming techniques. We conclude in Section 2.7.

2.2 Battery Aware Video Delivery - Overview

In this section, we will describe our overall approach towards battery aware video delivery. We will then discuss in detail different ways video bit rate and download rate can be selected and base station and mobile device can be reconfigured, to reduce battery load and the effect on user experience.

2.2.1 Overall Approach

Our overall approach towards video bit rate and download rate adaptation and corresponding transceiver reconfiguration for battery aware video delivery consists of two main objectives namely, maximization of battery lifetime and ensuring user experience.

Our approach towards prolonging battery life [28] is based on the following factors: (1) minimizing battery load (current drawn from battery), and duration of load, and (2) idling the battery allowing it to recover charge, and increasing the duration of idling. Our proposed approach affects the above two factors in the following three ways. (a) Varying video download rate: A required video download rate is determined by the video bit rate (rate at which video is encoded by the encoder and decoded by the decoder), amount of data buffered at the mobile, and channel conditions. The required download rate is achieved by the base station with suitable configuration of its RF and baseband components, with corresponding mobile device configurations, the latter affecting battery load. Hence, for a given video bit rate, by utilizing the elasticity that the video buffer offers, the download rate can be varied and the base station reconfigured in a way that reduces the battery load imposed by the mobile device RF and baseband processing. (b) Stopping download: If for certain periods of time, video download and hence related processing on mobile device can be stopped, the battery load can be reduced to just playback load which is much lower than load due to downloading. Due to significant difference in consecutive loads (download + playback followed by playback only load), effect on battery is similar to that of idling thereby enabling battery to recover charge [28], [29] [We show this later in Section 2.4.5]. We term this as "download idle". Note that extensive analysis of charge recovery due to idling is presented in [28] using the analytical Rakhmatov Vrudhula (RV) rechargeable lithium ion battery model and authors in [29] have shown the ability of battery to recover charge due to idling using measurements on a commercially available lithium ion battery. (c) Varying video bit rate: As bit rate determines the amount of data that needs to be downloaded, bit rate can be varied in a manner that minimizes amount of data to be downloaded. This offers the opportunity to either further reduce the duration of download and hence introduce download idle periods, or choose lower download rates and less power intensive modes reducing the load imposed on the battery.

While maximizing battery lifetime, we need to also ensure user experience. Consequently, our approach needs to ensure that (1) the video download rate variation, including periods of idling, is done in a way that does not lead to buffer overflow or underflow (stalling of video playback), so that user experience is not affected; (2) the base station reconfiguration is done taking into account the wireless channel condition (estimated using Signal to Noise Ratio - SNR) so that a desired bit error rate (BER) (hence PSNR [30], [31], and video quality) is satisfied; and (3) when additional video bit rate adaptation is done, a minimum video quality is satisfied in a way that increases the overall Video Experience Longevity.

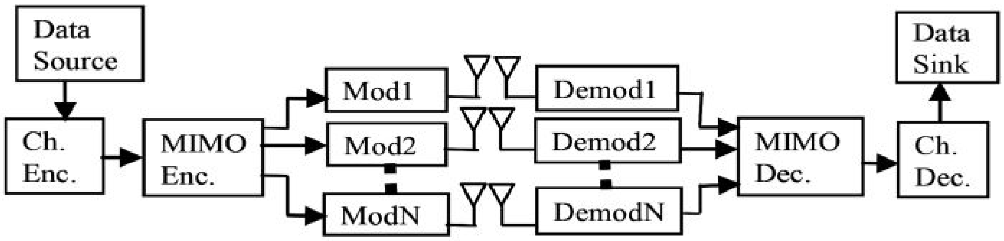


Figure 2.1. MIMO transmitter and receiver

Table 2.1. MIMO transmitter parameters

Channel Coding Rate (CR)	1, 2/3, 1/2, 1/3
MIMO Encoding Rate $MIMO_{Enc}$	STM, STBC
Modulation Schemes (Mod)	Binary Phase Shift Keying (BPSK), Quadrature Amplitude Modulation (QAM) - 4QAM, 16QAM, 64QAM
Number of Antennas (N_T)	1, 2, 3, 4

Table 2.2. MIMO receiver parameters

Number of Antennas (N_R)	1, 2, 3, 4
MIMO Decoding Rate $MIMO_{Dec}$	Zero Forcing (ZF), K-Best
Channel Decoding (Ch_{Dec})	Viterbi Decoding, Turbo Decoding

2.2.2 Download Rate Adaptation and Base Station Reconfiguration

In this section, we will first describe RF and baseband processing components of base station and mobile device, and their effects on power consumed. Subsequently we will discuss ways download rate can be varied and transceiver be reconfigured to reduce battery load. Note, we sometimes refer as baseband components both RF antenna chains and baseband components.

Fig. 2.1 shows a MIMO transmitter and receiver. The transmitter consists of channel encoder, MIMO encoder, and set of antennas each with an associated modulator. The receiver consists of antennas, demodulator, MIMO decoder and channel decoder. Tables 2.1 and 2.2 list some of the possible configuration choices that can be used for

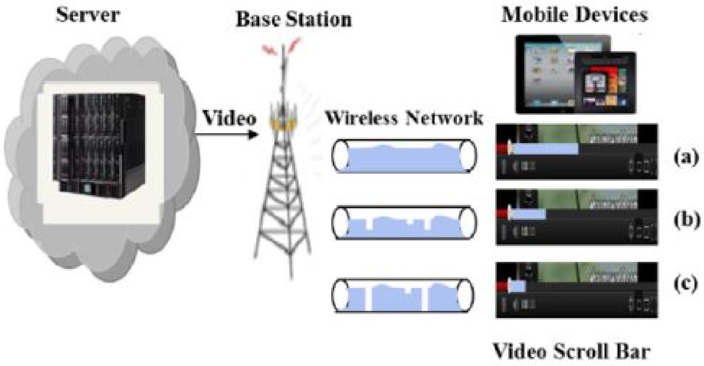


Figure 2.2. Video download using different rates

MIMO transmitter and receiver. The set of all possible combinations of transmitter and receiver baseband components constitutes the configuration spaces of base station and mobile device respectively. Henceforth, we will refer to the combination of transmitter - receiver antennas, channel encoding rate, MIMO encoding, modulation, MIMO and channel decoding algorithms as the transceiver mode selected.

Among all the MIMO receiver baseband components, the antenna RF chain is the most power intensive, and the battery load can increase significantly with increase in number of antennas. We consider two MIMO decoding algorithms, Zero Forcing (ZF) and K-Best, both of whose power consumption depends on the number of antennas and modulation scheme used; however, ZF is more power efficient but provides less BER performance than K-Best. Note the power consumed by demodulation is included in MIMO decoding, as demodulation is performed as part of MIMO decoding. Finally, power consumed by channel decoding depends on the algorithm used. Viterbi decoding consumes less power than Turbo decoding, but also has a lower BER performance than Turbo [32]. The battery load of a receiver configuration can be estimated by adding the power consumptions of the individual receiver components as elaborated in Section 2.3.2.

Fig. 2.2 shows typical video download scenarios from the video server through the

Table 2.3. Examples of modes with different download rates

Mode A	CR:1/2, STM, BPSK,2X2, ZF, Viterbi
Mode B	CR:1/2, STBC, 4QAM, 2X1, ZF, Viterbi
Mode C	CR:1/2, STM, 4QAM, 4X4, K-Best, Viterbi
Mode D	CR; 1/2, STM, 4QAM, 2X2, ZF, Viterbi

base station to the mobile device over the wireless network. The pipes are representative of the wireless network. The height and shape of the contents of the pipe depict the amount and flow of video data. The red portion on the scroll bar indicates the portion of downloaded video that has been viewed and the blue portion indicates the buffered portion. Fig. 2.2a depicts the scenario wherein the video is downloaded as fast as possible (as is the case with HTTP Progressive Download) depicted by the near fullness of the pipe and buffer. This may require the highest download rate possible under the given channel condition and BER value. Multiple transceiver modes may satisfy the required download rate under a given channel condition (SNR) and BER value. Some of these modes may actually increase the power consumption in the base station, but will reduce the mobile battery load. For example, the two modes A and B listed in Table 2.3 result in the same download rate. For the given SNR, mode B increases the power consumed by the base station as it uses 4QAM modulation scheme which consumes more power than BPSK used in mode A. However, mode B will reduce battery load, as only one receiver antenna is used as opposed to two antennas used in mode A. Note that the reduction in battery load due to reduction in receiver antennas far outweighs any increase in battery load due to higher order demodulation. There may also exist certain modes that reduce mobile battery load without increasing power consumption at the base station. For example, if channel condition improves, for the same download rate, it may be possible to reconfigure receiver to use ZF decoding instead of K-Best if BER requirement is met. Hence even when high download rate is required, it may be possible to choose a

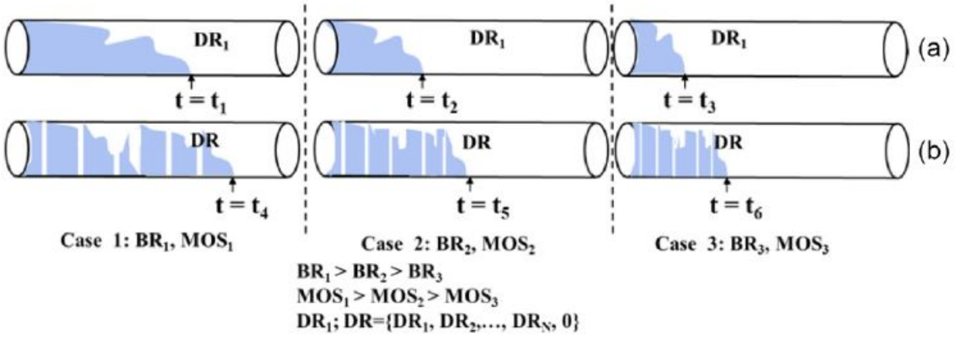


Figure 2.3. Adaptive Bit Rate streaming with different rates

transceiver configuration which reduces battery load.

The opportunities for finding battery efficient modes can be increased if the required download rate can be reduced. As shown in Fig. 2.2b, using the elasticity of the buffer, it is possible to reduce the download rate (depicted by dips in the pipe) which results in lesser buffered data (smaller blue portion), as long as there is no buffer underflow. For instance, consider modes C and D in Table 2.3. If the download rate needed is reduced by half, given the same channel condition and BER requirement, mode D can be used instead of mode C. Reconfiguring to mode D will significantly reduce the battery load, as it uses less number of antennas and less power intensive ZF MIMO decoding.

When buffer levels permit, download rate can be reduced to zero. Download idling reduces the battery load to just the playback load, thereby enabling battery to recover charge. Note that the idling will deplete the buffer (shown in Fig. 2.2c as gaps in the pipe and smallest blue portion on the scroll bar), and hence can be done if no buffer underflow can be ensured.

2.2.3 Video Bit Rate Adaptation

In this section, we will elaborate on how video bit rate adaptation affects battery lifetime and video quality. We pictorially represent adaptive bit rate video streaming in Fig. 2.3. As in Fig. 2.2, the pipes are representation of wireless network; height and shape of contents indicate the amount and flow of video data across time; to conserve space, we have omitted the server, base station and mobile device. Cases 1, 2 and 3 in Fig. 2.3 illustrate the effect of using bit rates, and with associated Mean Opinion Score (MOS) values BR_1, BR_2 and BR_3, while Figs. 2.3(a) and 2.3(b) show the effect of using single download rate, and a set of download rates , on the amount and flow of data in the pipe. Note that the download rates in the set are listed in descending order.

From Fig. 2.3(a), we make the following observations. (1) When highest download rate DR_1 and bit rate BR_1 are used, as in case 1, the battery load is highest because the download duration t_1 is longer than $t_2$and t_3. Using lower bit rate (cases 2 and 3) reduces the amount of data to be downloaded, and hence duration of download ($t_3 < t_2 < t_1$) and battery load. Case 1 of Fig. 2.3(b) illustrates the proposed video download rate (and mode) adaptation techniques which use a combination of high and low download rates including download idle from the set DR. As elaborated in the previous subsection the combination of higher, lower download rates and idling offers the potential to reduce battery load. Additionally, using lower bit rates as in cases 2 and 3 of Fig. 2.3(b) reduces battery load and the reduced download duration ($t_6 < t_5 < t_4$) may allow choosing a more battery efficient combination of download rates (and modes), for instance, introducing more periods of idling. Therefore, we can infer that bit rate adaptation potentially furthers the battery savings due to download rate and mode adaptation but at the expense of video quality.

2.3 Battery Efficient Download Rate and Mode Selection

In this section, we will assume fixed video bit rate, and formulate the optimization problem of adapting video download rate and corresponding transceiver configuration to maximize battery life. We then present an algorithm, MoDS that solves the problem using an optimization solver.

2.3.1 Download Rate and Model Selection Problem Definition

The objective of download rate and mode selection is maximization of battery lifetime during video download subject to download rate and application BER constraints. Video download session consists of several download epochs requiring download rate and mode selection in every download epoch. As battery lifetime is a cumulative result of several such selections and their effect on battery level, we split the optimization problem in to sub-problems and solve it in each download epoch in order to make it tractable. Each sub-problem defined in (2.1) below consists of selecting an optimal mode M for the download epoch under consideration such that battery level Bat_{Lev} (function of mode parameters listed in Tables 2.1 and 2.2) is maximized while download rate DR constraint upper bounded by DR^{Max} and lower bounded by DR^{Min}, and BER constraint upper bounded by BER_{App}, are satisfied. The sub-problems though seem independent, are connected with each other as the download rate selected in current epoch changes the buffer level which in turn affects the download rate selection in the subsequent epoch.

$$\max Bat_{Lev}(\mathbf{M}) \tag{2.1}$$

$$\text{Subject to: } DR^{Min} \leq DR(\mathbf{M}) \leq DR^{Max} \tag{2.2}$$

$$BER(SNR, \mathbf{M}) \leq BER_{App} \tag{2.3}$$

The DR^{Max} and DR^{Min} values, which will be defined later in this section, ensure that buffer does not overflow or underflow respectively. The application BER value BER_{App} ensures that video quality (PSNR) is maintained at desired level. Note that it has been shown in [30] and [31] that BER below $3 \cdot 10^{-5}$ results in PSNR levels greater than 37 dB (corresponding to MOS value of 5 [30]) thereby ensuring high video quality for videos with different space–time characteristics. Hence, choosing BER_{App} value lesser than $3 \cdot 10^{-5}$ will ensure that PSNR of the received videos will be greater than 37 dB.

It should be noted that (2.1) may not have a feasible solution always. When no mode satisfies BER_{App}, then download idle ($DR(\mathbf{M}) = 0$) is chosen. This may be at the expense of buffer underflow if DR^{Min} is greater than zero. In case the DR^{Min} constraint is violated, mode which gives highest download rate (lower than DR^{Min}) and satisfies BER_{App} is chosen leading to buffer underflow. On the other hand when DR^{Max} is violated, download idle is chosen to avoid buffer overflow. Having defined the download rate and mode selection problem, we will next discuss the objective and constraint functions.

2.3.2 Modeling of Objective and Constraint Functions

Each download epoch involves video download and simultaneous playback. The RV lithium ion battery model [28], [33] used to estimate the battery level given in (2.4) is characterized by two parameters, namely, which is the battery capacity and β , a function of ion diffusion coefficient, is the measure of battery nonlinearity. The second term in (2.4) represents the ratio of total charge consumed in time T or equivalently in E download epochs due to variable load I and the total charge present in the fully charged battery. The charge consumed in each download epoch i is the sum of the linear term (first term in summation over E) and the summation of nonlinear terms (second term in summation over E) with summation index m. The summation of nonlinear terms is a function of β and accounts for the nonlinearity in diffusion and hence charge recovery when $I_i < I_{i-1}$. Note that our proposed techniques are not battery model specific and can be used with any model that gives an estimate of battery level in response to battery load.

$$Bat_{Lev} = 1 - \frac{1}{\alpha}\sum_{i=1}^{E} I_{i-1}[(t_i - t_{i-1}) + 2\sum_{m=1}^{m=10}\frac{e^{-\beta^2 m^2 (T-t_i)} - e^{-\beta^2 m^2 (T-t_{i-1})}}{\beta^2 m^2}] \qquad (2.4)$$

Maximization of Bat_{Lev} is equivalent to minimizing numerator of second term in (2.4) which represents the battery charge consumed due to battery load I in time T. Further, as charge consumed is estimated in each download epoch of duration D_{Period} , which we assume is a constant, maximization of Bat_{Lev} is equivalent to minimizing battery load I in each download epoch. As each download epoch involves simultaneous download and playback, I is given by

$$I = I_{Download} + I_{Playback} \qquad (2.5)$$

$I_{Playback}$ is the battery load due to video decoder and display used for playback. While the playback load may vary depending on the resolution of the video, for download epochs of the same video session, it is fair to treat it as constant. Hence maximization of Bat_{Lev} is equivalent to minimizing battery load $I_{Download}$ imposed by the mode $\mathbf{M}$ during download subject to the download rate and BER constraints in (2.6). $I_{Download}$ is given by (2.9).

$$\min I_{Download}(\mathbf{M}) \qquad (2.6)$$

$$\text{Subject to: } DR^{Min} \le DR(\mathbf{M}) \le DR^{Max} \qquad (2.7)$$

$$BER(SNR, \mathbf{M}) \le BER_{App} \qquad (2.8)$$

$$I_{Download} = \frac{P_{Download}}{V_{Bat}} \tag{2.9}$$

where V_{Bat} is the battery voltage; we assume that it is constant during discharge. Download power $P_{Download}$ given by (2.10) consists of four components, namely power due to RF chain ($P_{RF-Chain}$), MIMO decoding ($P_{MIMO-Dec}$), channel decoding (P_{Ch-Dec}) and baseband processing ($P_{Baseband}$).

$$P_{Download} = P_{RF-Chain} + P_{MIMO-Dec} + P_{Ch-Dec} + P_{Baseband} \tag{2.10}$$

$P_{RF-Chain}$ depends on N_R and system bandwidth BW. It is determined using (2.11) obtained from relations in [17][34].

$$P_{RF-Chain} = (1.8 \cdot 10^{-8} BW + 0.0061)N_R + 0.1 \tag{2.11}$$

$P_{MIMO-Dec}$ depends on MIMO encoding rate $MIMO_{Enc}$, number of antennas, algorithm chosen (ZF or K-Best) and modulation scheme used $MIMO_{Enc}$ given by (2.12) and (2.13), is dependent on the type of MIMO encoding (STM or STBC) used. $P_{MIMO-Dec}$ is estimated using (2.14)–(2.17), by calculating number of search steps [17] required to decode a symbol and determining number of parallel search engines [35] required to execute the steps. We consider only Viterbi channel decoding algorithm in this work; P_{Ch-Dec} estimate is obtained from [36]. $P_{Baseband}$ is given by (2.18) [17].

$$MIMO_{Enc-STM} = N_R \tag{2.12}$$

$$MIMO_{Enc-STBC} = \begin{cases} N_R, (\frac{N_T}{N_R}) \geq N_T \\ (N_R - 1)\frac{N_T}{N_R}, (\frac{N_T}{N_R}) < N_T \end{cases} \tag{2.13}$$

$$\begin{aligned}P^{STM}_{MIMO-Dec-K-Best} = &10^{-4}[MIMO_{Enc-STM}(0.5N_T^2+1.5N_T)+3.1N_T^2Mod^{2.5}\\ &+0.8N_TMod^{3.5}+1.5N_TMod]\end{aligned} \tag{2.14}$$

$$P^{STM}_{MIMO-Dec-ZF} = 10^{-4}[MIMO_{Enc-STM}(0.3N_T^2+N_T)+0.13N_T^2+0.06N_T^3] \tag{2.15}$$

$$\begin{aligned}P^{STBC}_{MIMO-Dec-K-Best} = &10^{-4}[3.1N_TN_R+4.1N_RMIMO^2_{Enc-STBC}+\\ &N_TMod(1.5+0.8Mod^{2.5}+6.2Mod^{1.5}MIMO_{Enc-STBC})]\end{aligned} \tag{2.16}$$

$$\begin{aligned}P^{STBC}_{MIMO-Dec-ZF} = &10^{-4}[1.9N_TN_R+0.25N_RMIMO_{Enc-STBC}+\\ &MIMO^2_{Enc-STBC}(0.5+2.3N_R+0.5MIMO_{Enc-STBC})]\end{aligned} \tag{2.17}$$

$$P_{Baseband} = 1.62\cdot 10^{-9}N_RBW \tag{2.18}$$

The download rate DR given by (2.19) forms the first constraint function and is calculated using the specifications in 3GPP LTE standard [37].

$$DR = RB\cdot SUB_C\cdot TS\cdot OFDM_{Sym}\cdot Mod\cdot CR\cdot MIMO_{Enc}\cdot T^{-1}_{Frame} \tag{2.19}$$

where RB represents the number of resource blocks associated with BW. SUB_C is the number of subcarriers used in each resource block. TS is the number of slots used to transmit $OFDM_{Sym}$ number of Orthogonal Frequency Division Multiplexing (OFDM) symbols. T_{Frame} is the duration of 3GPP LTE frame.

The upper bound on download rate us given by (2.20). It is calculated using video buffer size Buf_{Size}, amount of data buffered Buf_{Avail} and duration of download epoch D_{Period}.

$$DR^{Max} = \frac{Buf_{Size} - Buf_{Avail}}{D_{Period}} \tag{2.20}$$

Playback time available PBT is calculated using Buf_{Avail} and video bit rate V_{BR} as shown in (2.21).

$$PBT = \frac{Buf_{Avail}}{V_{BR}} \tag{2.21}$$

The lower bound on DR, DR^{Min} given by (2.22) is calculated using PBT, V_{BR} and minimum buffer value Buf_{Min}, chosen to avoid stalling. It should be noted that the lower bound for Buf_{Min} is D_{Period} in which case the PBT will at least be D_{Period}. However, this might stall video when channel conditions do not permit minimum download rate DR^{Min}, hence Buf_{Min} greater than D_{Period} will increase PBT and allow idling while avoiding stalling.

$$DR^{Min} = \begin{cases} 0, PBT > Buf_{Min} \\ V_{BR} + \frac{V_{BR}(\lfloor PBT \rfloor - PBT + Buf_{Min})}{D_{Period}}, PBT \leq Buf_{Min} \end{cases} \tag{2.22}$$

The second constraint in terms BER_{App} of ensures that mode selected does not lead to unacceptable BER and hence adversely impact video quality. We use a BER-SNR look up table (LUT) (Section 2.4.1) in lieu of the BER constraint function in the optimization framework. The BER–SNR LUT lists the BER values for different transceiver configurations under different channel (SNR) conditions.

From (2.4) to (2.22), it is evident that the objective and constraint functions are nonlinear making mode and download rate selection a nonlinear constrained optimization (minimization) problem. In the next subsection we will present a solution to this problem.

2.3.3 Mode and Download Rate Selection (MoDS) Algorithm

In this section, we will describe in detail the MoDS algorithm developed to search the transceiver configuration space (Tables 2.1 and 2.2) for the mode that minimizes the battery load I subject to download rate and BER constraints.

As power calculation functions for MIMO decoding given by (2.14) to (2.17) are different for different MIMO encoding schemes and MIMO decoding algorithms, mode selection in each download epoch needs to be carried out separately for each MIMO encoding scheme and decoding algorithm. This implies that $MIMO_{Enc}$ and $MIMO_{Dec}$ parameters cannot be part of the transceiver mode search space. On the same line of reasoning, Ch_{Dec} cannot be used as an optimization parameter. Hence, we split the transceiver configuration space CS in to two spaces as shown in Fig. 2.4: the outer space OS consisting of parameters $MIMO_{Enc}$, $MIMO_{Dec}$ and Ch_{Dec}, and inner space IS consisting of parameters CR, Mod, N_T and N_R. The BER–SNR LUT used instead of BER constraint function requires the BER constraint to be evaluated for each mode outside the optimization framework. Having made the above two modifications to the problem stated in (2.6), the basic working principle of MoDS algorithm is pictorially shown in Fig. 2.4. For a given point in outer space OS_j, MoDS searches the inner space for the mode (IS_i, OS_j) that minimizes battery load and satisfies download rate. Subsequently the BER constraint is evaluated as shown in Fig. 2.4. This process is repeated till the entire OS is explored resulting in battery efficient mode that satisfies the

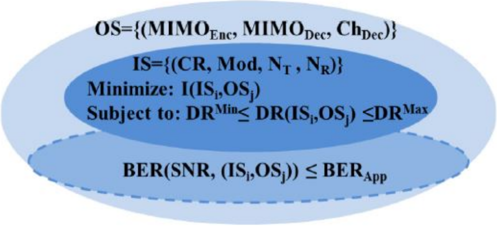

$OS=\{(MIMO_{Enc}, MIMO_{Dec}, Ch_{Dec})\}$
$IS=\{(CR, Mod, N_T, N_R)\}$
Minimize: $I(IS_i,OS_j)$
Subject to: $DR^{Min} \leq DR(IS_i,OS_j) \leq DR^{Max}$
$BER(SNR, (IS_i,OS_j)) \leq BER_{App}$

Figure 2.4. Splitting of configuration space and optimization problem

constraints in (2.6).

The outer space OS, the inner space IS, upper bound and lower bound LB representing the maximum and minimum values possible for the elements of inner space, and set of valid inner space points IS^{Valid} form the inputs to the MoDS algorithm shown in Fig. 2.5. Given an outer space point, the nonlinear optimization solver, 'nlopt'[2] is used to determine the mode that minimizes the battery load I and satisfies the download rate constraints. It should be noted that when DR^{Min} is zero, download idling ($DR(\mathbf{M}) = 0$) is chosen as this minimizes the battery load I. If the mode does not belong to IS^{Valid}, it is rounded off to the nearest valid mode by adding such that the resulting mode does not violate the download constraint. The BER value of mode is obtained from the BER-SNR LUT. As pictorially shown, if the BER value of mode v lies to the left of BER_{App}, then the mode is added to the set $Feasible_{Mode}$ as the battery efficient mode for the chosen point OS_j of outer configuration space. When BER value lies to the right of BER_{App}, the inner space is constrained to IS' by lowering and increasing upper and lower bounds respectively; thereby eliminating modes that do not meet the BER requirement. The upper bound is shifted to lower points by first gradually reducing CR and then Mod to lower values. Lowering CR and Mod values constricts the configuration space to modes with lower CR, Mod and BER values, thereby increasing the chances of finding mode that satisfies the BER requirement. If BER requirement is not met even at the lowest

[2] The NLopt nonlinear-optimization package," [Online]. Available: http:// ab-initio.mit.edu/wiki/index.php/NLopt

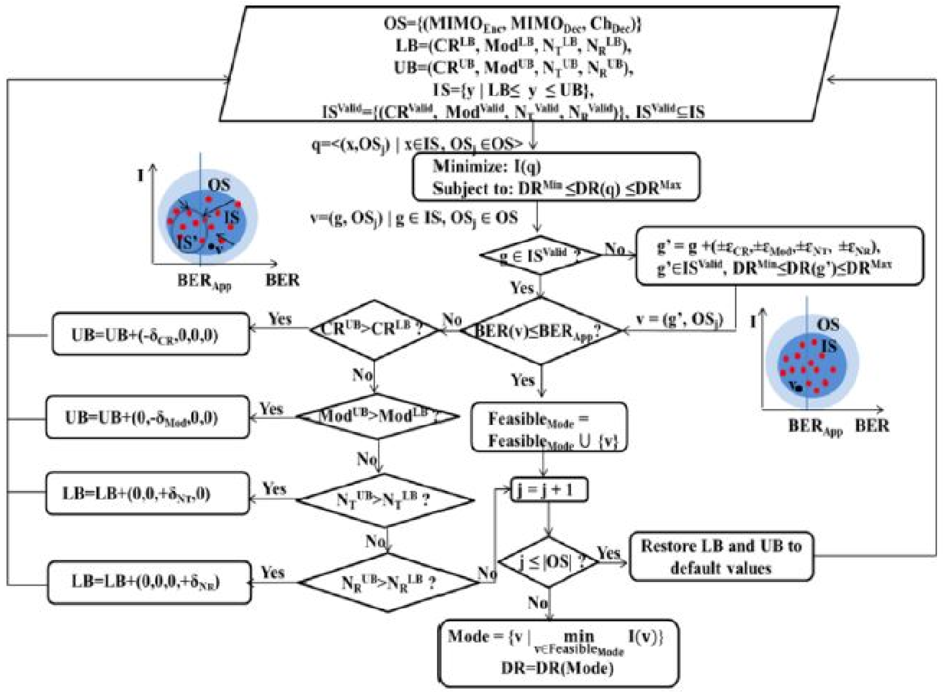

OS={(MIMO_Enc, MIMO_Dec, Ch_Dec)}
LB=(CR^LB, Mod^LB, N_T^LB, N_R^LB),
UB=(CR^UB, Mod^UB, N_T^UB, N_R^UB),
IS={y | LB≤ y ≤ UB},
IS^Valid={(CR^Valid, Mod^Valid, N_T^Valid, N_R^Valid)}, IS^Valid⊆IS
q=<(x,OS_j) | x∈IS, OS_j ∈OS>
Minimize: I(q)
Subject to: DR^Min ≤DR(q) ≤DR^Max
v=(g, OS_j) | g ∈ IS, OS_j ∈ OS
g ∈ IS^Valid ?
No
g' = g +(±ε_CR,±ε_Mod,±ε_NT, ±ε_NR),
g'∈IS^Valid, DR^Min≤DR(g')≤DR^Max
v = (g', OS_j)
Yes
BER(v)≤BER_App?
No
Yes
CR^UB>CR^LB ?
Yes
UB=UB+(-δ_CR,0,0,0)
No
Mod^UB>Mod^LB ?
Yes
UB=UB+(0,-δ_Mod,0,0)
No
N_T^UB>N_T^LB ?
Yes
LB=LB+(0,0,+δ_NT,0)
No
N_R^UB>N_R^LB ?
Yes
LB=LB+(0,0,0,+δ_NR)
No
Feasible_Mode =
Feasible_Mode ∪ {v}
j = j + 1
j ≤ |OS| ?
Yes
Restore LB and UB to default values
No
Mode = {v | min_{v∈Feasible_Mode} I(v)}
DR=DR(Mode)
I
OS
IS
IS'
v
BER_App
BER
I
OS
IS
v
BER_App
BER

Figure 2.5. Mode and Download Rate Selection (MoDS) algorithm

value of CR and Mod, in the final iteration, the lower bound is shifted to higher points by gradually increasing the number of antennas N_T, and N_R. As increasing N_T and N_R values will lead to selection of power intensive modes, it is done in the final iteration. The mode selected in the final iteration is the battery efficient mode corresponding to the chosen OS point OS_j and is added to the set of feasible modes $Feasible_{Mode}$. This process is repeated till the outer space is completely explored and then the most battery efficient mode M is chosen from $Feasible_{Mode}$. The corresponding download rate $DR(\mathbf{M})$ is the chosen rate for the ensuing download epoch. The computational complexity of MoDS algorithm which iteratively searches the IS and OS for battery efficient mode is presented in Appendix A.1.

The overall framework for information and control data exchange between base station and mobile device, mode selection and reconfiguration during battery efficient video download is described in detail in the Appendix A.2. As elaborated in Appendix A.2, additional data transmitted for conveying buffer levels to base station is nominal–a byte resulting in 1.14 mW of power consumption [38]. On the other hand, receiving information from the BS about mode selected requires 8 bytes, and results in about 2.22μW of power consumption when the mode $(1X1, BPSK, CR = 1, ZF)$ is used. In addition to the power consumed due to information exchange, during mode reconfiguration at the mobile device, a change in the number of antennas used in the previous mode to the current mode results in RF component switching power of 100mW per antenna and switching time of 5μs [34].

2.4 Simulation Framework and Results

In this section, we describe the simulation framework developed and experimental results obtained by using our proposed battery aware video download technique MoDS, and compare with results obtained using conventional HTTP Progressive Download

(HTTP-PD) as well as the EERA technique [20] discussed earlier in Section 2.1.1.

We have developed a very modular and flexible MATLAB based simulation framework to estimate battery consumption and assess user experience during video download and playback. The simulation framework consists of power, battery, BER and user experience models, and allows us to implement and assess different video download techniques to download video sequences under varying channel conditions and video quality requirements. We briefly describe the models followed by discussion of the framework integrated with the models and various video download techniques.

2.4.1 Power and Battery Models

The power model is used to estimate the power consumed in the mobile device due to video download and playback. As elaborated in Section 2.3.2, download power consists of four components namely, $P_{RF-Chain}$, $P_{MIMO-Dec}$, P_{Ch-Dec}, and and $P_{Baseband}$ is modeled using (2.10) to (2.18) which are in turn based on measurements made on ASIC implementations of the respective blocks. Similarly, playback power is estimated using measurements from video decoder[3] and mobile device display [39]. Note that since the overall device power is the sum of the power consumed by the different components of download and playback power, the power model can be adapted to a different device by modeling and substituting for the components that are different. For example, if the new device uses a different implementation of say, the baseband, then (2.18) will need to be updated with the appropriate model for the new baseband implementation.

Next we will discuss the RV rechargeable lithium ion battery model [28] which takes the output of power model to estimate the battery level. As elaborated in Section 2.3.2, (2.4) is used to estimate the battery level given the magnitude and duration of battery load which is obtained using (2.9). It should be noted that the RV model can be

[3] [Online]. Available: http://www.privateline.com/imode/MPEG_ 4_ CODEC. pdf

Table 2.4. BER model simulation parameters

Channel Model	Spatial Channel Model (SCM) - CaseII, Vehicular A
Channel Bandwidth	5MHz
SNR(dB)	0-40
FFT Size	512 points
Channel coding	$1, 2/3, 1/2, 1/3$
Modulation Schemes	BPSK, 4QAM, 16QAM, 64QAM
Antenna Configuration	STM: 1X1, 2X2, 3X3, 4X4, STBC: 2X1, 2X2
MIMO Decoding	Zero Forcing, K-Best
Channel Decoding	Viterbi Decoding

used to estimate battery levels of rechargeable lithium ion batteries with different battery voltage and capacities (battery specific parameters required by the battery model are obtained by running discharge tests with constant battery load [28], [33]). This implies that the proposed video download techniques can be evaluated on mobile devices of varying form factors and battery capacities. Moreover, the proposed techniques are not battery model specific and can be used with any model that gives the battery level in response to the battery load.

2.4.2 BER Model

As elaborated in Section 2.3.1, given the channel condition, BER values of transceiver modes are required by MoDS to ensure that desired BER is maintained. We have developed a BER model by using MATLAB to simulate different modes under different channel (SNR) conditions and obtain BER values which are stored in the BER-SNR look up table (LUT). The simulation parameters used to generate the BER–SNR LUT are listed in Table 2.4, including the modulation schemes, antenna configurations, MIMO decoding and channel decoding algorithms. Using the specified channel model, carrier bandwidth and FFT size, the SNR is varied to obtain the BER values of all the modes constituting the reconfiguration space.

2.4.3 User Experience Model

User experience for video download is primarily determined by the video quality and any stalling in video playback. Video quality of received video is affected by adaptation of video characteristics such as video resolution, bit rate, and frame rate, and any packet losses that may occur due to BER during transmission. Since the original video resolution, bit rate, and frame rate are not changed by HTTP-PD, EERA or MoDS, the video quality is not affected. Further, by choosing a very low application BER, BER_{App} ($< 3 \cdot 10^{-5}$, [30], [31]) and carrying out mode selection so as to meet the desired application BER requirements (10^{-6} in our experiments), no loss in PSNR and thereby video quality due to packet loss is ensured. Hence, the only user experience impairment in the case of video download techniques to be compared here is stalling. Consequently, the user experience model uses the stalling–MOS relationship developed in [40] to map the number and duration of stalling events recorded (by the simulation framework developed) to MOS scores.

2.4.4 Simulation Framework

The simulation framework for video download techniques consists of power, battery and BER models along with the video download algorithm/technique and simulation time counter. When video download is initiated, the simulation time counter is started. The simulation step is equal to the download epoch duration D_{Period}, and in our experiments it is fixed at 2s, though it can be made longer or shorter. In case of the proposed battery aware video download technique, the MoDS algorithm determines the battery efficient mode and download rate depending on the current buffer level and channel condition (SNR) for each simulation step.

For the energy efficient rate adaptation (EERA) technique [20], the EERA algorithm determines the energy efficient mode and the download rate depending on the video bit rate and channel condition. While simulating the conventional HTTP-PD technique implemented using the download mechanism (consisting of initial phase and throttle phase) in [41], we fix the desired download rate to maximum value determined using (2.20) in the initial phase and to that which will allow a constant average rate of 1.25 times the video encoding rate when data is sent in bursts of 64KB in the throttle phase. We select the mode that satisfies the download rate and BER requirement and if no such mode exists, then the mode that gives highest download rate (lower than the desired rate) at the given SNR and BER value is chosen. For all the aforementioned techniques, the BER–SNR LUT (Section 2.4.2) is used to ascertain whether the BER of the selected mode satisfies the BER_{App}. The download power and playback power are calculated using the power model (Section 2.4.1) and the resulting battery load is input to the battery model (Section 2.4.1) to estimate the battery level. It should be noted that for MoDS, the power consumed due to information (1.14mW, see Section 2.3.3) and control data exchange (2.2μW, see Section 2.3.3) and RF component switching power (100mW for 5us, see Section 2.3.3) is also added to the download power and playback power before determining battery load. The simulation framework also records the number and duration of stalls (buffer underflow/overflow) if any and uses the user experience model (Section 2.4.3) to quantify the user experience in terms of MOS value.

If the viewer switches to a new video or current video is completely downloaded, new video download begins. This continues till battery is completely drained. The simulation counter at this instant gives the battery lifetime for downloading and watching the chosen video sequence under simulated channel conditions and quality requirements. It should be noted that while battery lifetime is a cumulative result of multiple video download and viewing sessions, user experience is assessed for each session.

Table 2.5 lists the simulation parameters used in our experiments. Video characteristics specify the video bit rate used to encode the video and the sequence of videos

Table 2.5. Video download simulation parameters

Video Characteristics	Video Bit Rate $V_{BR} = 4.12 Mb/s$
	Video Sequence 1: $\{$184s, 226s, 195s, 197s, 226s, 257s, 274s, 231s, 200s, 224s, 298s, 235s, 285s, 198s, 233s, 291s, 298s, 236s, 221s, 205s$\}$
Client Characteristics	Video Buffer Size $Buf_{Size} = 300s$
	Playback Load (Decoder + Display) $I_{Playback}$ = 34mA
	RF Component Switching Power = 100mW
	RF Component Switching Time = 5μs
User Characteristics	Constant SR=1
	Variable SR: $\{$0.5, 0.1, 0.97, 0.43, 0.27, 0.93, 0.22, 0.19, 0.28, 0.67, 0.6, 0.39, 0.93, 0.82, 0.05, 0.82, 0.38, 0.45, 0.01, 0.28$\}$
Algorithm Parameters	Minimum Buffer Level $Buf_{Min} = 10s$
SNR(dB)	High:40, Low:9, Variable: In the range 0-40
BER	Application BER $BER_{App} = 10^{-6}$

watched. Client characteristics enumerate buffer size, playback current, switching power and switching time specifications of RF components (antennas). In our experiments, we also consider the increasingly prevalent "video snacking" user viewing pattern wherein the user begins to watch a video and then switches to a new video without finishing the current video. This pattern is modeled by randomly generated values of snacking ratio (SR) which is the ratio of the duration of the video viewed by the user to the actual duration of the video. In other words, each snacking ratio value specified in user characteristics indicates how much of the corresponding video in the video sequence the user will watch. The value of the algorithm parameter, minimum buffer level Buf_{Min} used to determine DR^{Min} in (2.22) is also listed in Table 2.5. Table 2.5 lists the channel conditions based on measurements of cellular network (high, low and variable) and the application BER requirement that ensures high quality (Section 2.3.1, [30], [31]). Note the resolution of temporal variation in channel condition is assumed to be comparable to the simulation step.

2.4.5 Experimental Results

Next, we present results obtained by simulating video download under different channel conditions and snacking ratios (and low BER/high video quality requirement) shown in Table 2.5. Figs. 2.6 and 2.7 show the effects on download rate, battery load, level, and lifetime while using HTTP Progressive Download (HTTP-PD, shown as red dot-dash line/red bar) [41], the energy efficient rate adaptation technique (EERA, shown as blue dashed line/blue bar) [20] and our proposed battery aware download technique (MoDS, shown as green solid line/green bar).

We will first describe the download characteristics of each of the techniques and then discuss their impact on battery consumption under different snacking ratio values and SNR conditions. HTTP-PD delivers video at maximum download rate in the initial phase followed by constant average download rate in the throttle phase [41] without attempting to choose battery efficient modes in both phases resulting in maximum battery drain during video download. However, the above factors contribute to reduced download duration and extended playback only period after download during which significant charge recovery takes place as battery load is reduced to only playback load. EERA reduces battery drain by selecting energy efficient modes; however it downloads at the video encoding bit rate (4.12 Mb/s) which not only extends the download duration but also does not fill the buffer and thereby can neither vary download rate nor download idle to achieve additional battery savings. On the other hand, MoDS selects modes that maximize battery level and battery load is further reduced by selecting download idling whenever playback time available is greater than Buf_{Min} [as in (2.22)].

We will first examine the scenario when the mobile device is experiencing good network condition (high SNR). The download rates selected while downloading and viewing a single 184s video with $SR = 1$ (video viewed completely), and the resulting

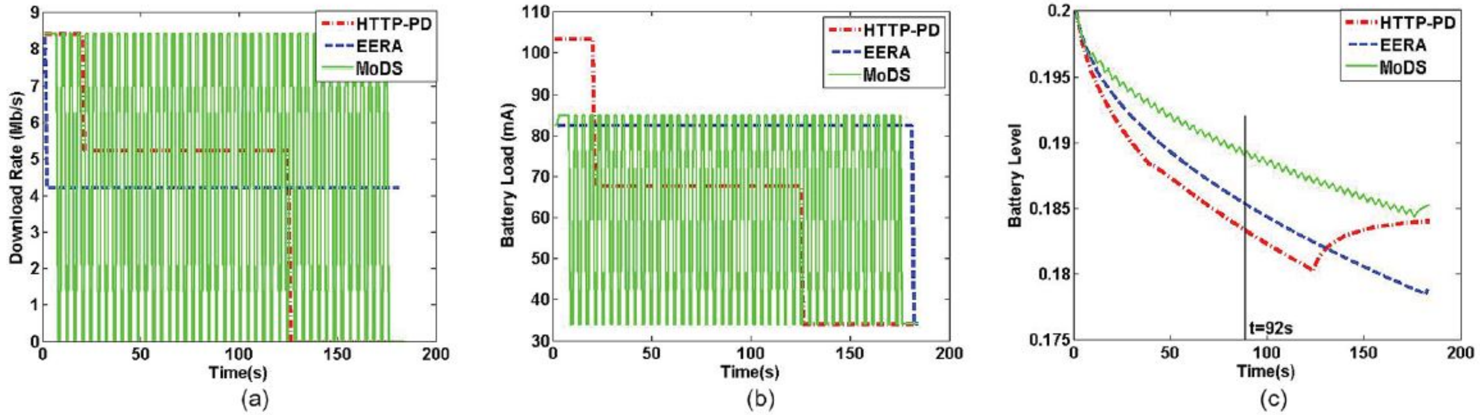


Figure 2.6. Effect of downloading and viewing a single 184s video under high SNR conditions on (a) download rate, (b) battery load, and (c) battery level

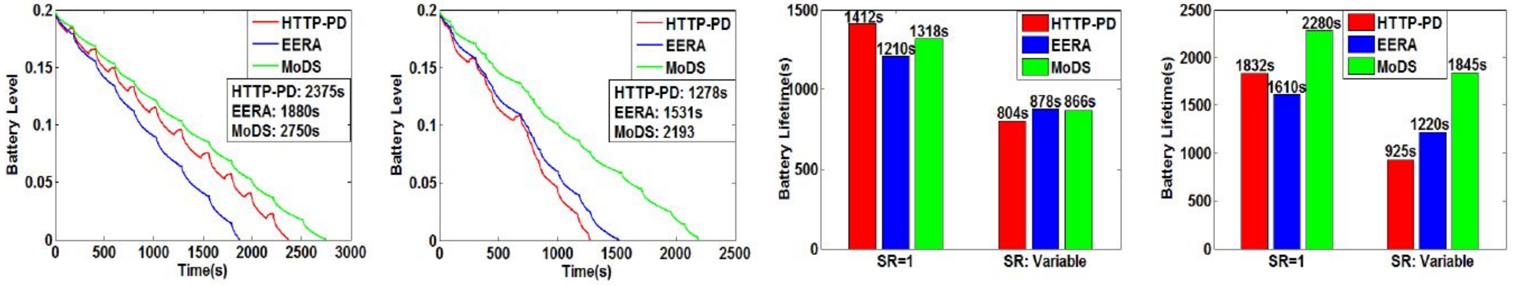


Figure 2.7. Effect of downloading and viewing video sequence 1 on battery lifetime under (a) high SNR with SR=1, (b) high SNR with variable SR, (c) low SNR with SR=1 and variable SR and (d) variable SNR with SR=1 and variable SR

battery load, are shown in Figs. 2.6a and 2.6b respectively. The green solid line shows the effect of MoDS performing download idling. Fig. 2.6c shows the effect on battery level, when the simulation is started with battery level of 0.2. Note that for MoDS, download idle followed by transmission results in alternate fall and rise in load with corresponding rise and fall in battery level clearly indicating that battery recovers charge as a result of idling. HTTP-PD results in maximum battery drain (as explained above) till video download is complete at $t = 125s$. Subsequently, it recovers significant charge during the playback only period lasting about 59s as shown in Fig. 2.6c. This explains how HTTP-PD reduces most of the gains achieved by MoDS through selection of battery efficient modes and idling. At the end of video duration (184s), we can see that EERA causes maximum decrease in battery level, followed by HTTP-PD and finally by MoDS, the latter reducing degradation in battery level significantly compared to EERA but only marginally compared to HTTP-PD. On the other hand, if we consider $SR = 0.5$, then video download and playback will stop at $t = 92s$ indicated by vertical line in Fig. 2.6c. In this case, HTTP-PD cannot utilize the playback only period to recover charge, hence it causes maximum battery drain (about 3.6%) followed by EERA and then MoDS.

Figs. 2.7a and 2.7b show the impact on battery lifetime when video sequence 1 is seen with $SR = 1$ and variable SR respectively and with a starting battery level of 0.2. For $SR = 1$ [Fig. 2.7a], even though HTTP-PD recovers charge at the end of single video as elaborated above, subsequently, as download progresses, the high download load depletes the battery more during download than that can be recovered during playback period. This widens the gap in battery levels between HTTP-PD and MoDS with the lower download load and idling for MoDS further extending the battery lifetime to result in overall gain of 16%. For EERA, the maximum battery drain for single video continues for subsequent videos resulting in 46% lower battery lifetime compared to MoDS. For variable SR [Fig. 2.7b], HTTP-PD cannot recover charge in the playback only period unless SR is comparable to 1. On the other hand, as EERA extends download time and does not vary download rate or idle, variable SR has negligible effect on its performance. MoDS which reduces battery load right from the outset, gains about 71.5% and 43% in battery lifetime over HTTP-PD and EERA respectively. As no stalling is recorded for either of the techniques with either $SR = 1$ or $SR = 0.5$ or variable SR, user experience is same for HTTP-PD, EERA and MoDS and MOS is 5 [40]. It should be noted that the MOS values are the average of the MOS values of videos downloaded and viewed.

Next we will discuss the scenario when mobile device experiences bad channel condition (low SNR). Low SNR condition does not allow filling up the buffer as fast resulting in shorter playback period for HTTP-PD. On the other hand, the selection of modes that minimize battery load by MoDS under low SNR conditions results in reduced download rates (higher download rates require power intensive modes to maintain BER) that not only extend duration of download but also do not allow idling. As can be seen in Fig. 2.7c, for MoDS, this results in loss of about 6.6% over HTTP-PD when $SR = 1$. However, it gains by about 9% in battery lifetime compared to EERA which stalls for 380s (31%) as it attempts to download at video bit rate by selecting battery efficient modes which under low SNR conditions does not allow buffer to fill and avoid stalling. HTTP-PD and MoDS do not result in stalling, hence result in the same user experience (average $MOS = 5$ [40]). On the other hand, EERA on an average (1210s battery lifetime corresponds to approximately 6 videos of video sequence 1 and 380s stalling corresponds to about 63s of stalling per video) results in MOS score below 2 [40]. From Fig. 2.7c, one can see that for variable SR, MoDS gains by 7.7% over HTTP-PD. With no video stalling, $MOS = 5$ for both HTTP-PD and MoDS. Though EERA gains by 1.4% over MoDS, it stalls for 258s (878s, 9 videos, 29s of stalling per video) resulting in MOS below 2. Under variable SNR conditions, gains under high SNR offset the loss under low SNR to result in net gain in battery lifetime for MoDS. In this case, the combination of power intensive modes under low SNR and battery inefficient modes under high SNR reduces the gain due to charge recovery for HTTP-PD. It can be seen from Fig. 2.7d that when $SR = 1$, a gain of 24% over HTTP-PD and 41% over EERA is possible when MoDS is used. HTTP-PD and MoDS achieve $MOS = 5$ with no stalling whereas EERA results in an average of 6s of stalling per video (1610s, 7 videos, 45s stalling) resulting in a MOS score of about 2 [40]. For variable SR [Fig. 2.7d], the above gains for MoDS are extended to 99% and 51% over HTTP-PD and EERA respectively. As with $SR = 1$, no stalling results in $MOS = 5$ for both HTTP-PD and MoDS whereas EERA results in an average of 2.1s stalling per video (1220s, 11 videos, 24s of stalling) resulting in MOS value of 3.5.

Studies conducted in [42] and [43] show that the average video completion rate is as low as 15% on smartphones and slightly higher on Tablets and that 80% of YouTube sessions are less than half of the video duration indicating that video snacking is highly prevalent among users. With reference to these statistics, variable SR values less than 1 is more realistic than constant SR equal to 1. From the above results, it can be seen that MoDS significantly increases battery lifetime in the realistic scenario of variable SR. With respect to SNR, the statistics presented in [33] for signal strength (SNR) experienced by users shows that variable SNR conditions are most prevalent and also that low SNR conditions throughout video download are less likely to occur. Again, the above results show that MoDS performs best under the most prevalent case of variable SNR conditions while the loss or nominal gains under low SNR conditions are less likely to occur.

In the next section, we will present battery aware techniques for DASH video that add to the battery savings achieved by download rate and mode reconfiguration while ensuring minimum desired video quality.

2.5 Bit Rate, Download Rate, and Mode Selection

As explained in Section 2.2.3, adapting the video bit rate may offer the opportunity for further battery savings beyond download rate and transceiver mode adaptation. However, bit rate adaptation may also affect video quality. When the mobile device is battery constrained limiting the longevity of watching video, the overall user experience may be enhanced by considering bit rate adaptation to elongate the battery lifetime and hence the viewing experience even with some acceptable degradation in video quality. In this section, we explore the potential additional benefit of video bit rate adaptation, along with download rate and mode adaptation, to increase the battery lifetime and thereby the video viewing experience, while ensuring an acceptable video quality. We first formulate the optimization problem formally, and then present algorithm developed namely BR-MoDS which uses optimization solver to solve the optimization problem. We then extend the formulation to consider battery level while selecting bit rate and present the B^{2}R-MoDS algorithm that solves the extended optimization problem. We conclude the section by defining the new Video Experience Longevity metric which quantifies the performance of DASH based techniques in terms of battery lifetime (longevity of video experience) and quality of video experience.

2.5.1 Maximization of Battery Lifetime with Acceptable Quality

The objective of video bit rate, download rate and mode selection is maximization of battery lifetime during adaptive video streaming subject to bit rate, download rate and user experience constraints. In adaptive bit rate streaming, the video is fragmented in to equally sized segments, each segment encoded using the set of discrete bit rates available [15]. A segment download can be viewed as a two tiered process wherein first the bit rate for the segment and subsequently, download rate and mode is selected. It should be noted that the download rate and mode selection may need to be done multiple times during a segment download; in other words each segment may consist of one or more ‘download epochs’ during which download rate and mode selection is carried out. This implies that battery lifetime maximization is achieved at two levels, namely at the segment level and at the download epoch level and hence, we will adopt a two tiered approach towards selecting a battery efficient combination of bit rate, download rate and mode. Selections made at either segment or download epoch level maximize battery level and the cumulative result of these selections maximizes battery lifetime. Therefore, henceforth we will refer to maximization of battery level instead of battery lifetime as the objective of bit rate, download rate and mode selection.

First we will formulate the sub-problem that maximizes battery level by choosing bit rate for each segment subject to bit rate and user experience constraints. We consider segments of duration Seg_{Time} encoded using bit rates belonging to set V_{BR-Set}^{Valid} lower bounded by V_{BR-Min} and upper bounded by V_{BR-Max}. The amount of data downloaded Seg_{Data}, for a segment is given by the product of Seg_{Time} and bit rate chosen V_{BR}. Choosing lower bit rates reduces the amount of data downloaded which in turn reduces battery load and/or duration of load thereby maximizing battery level (as elaborated in Section 2.2.3). However, as bit rate selection affects video quality VQ, it has to be done in a manner that the VQ exceeds a certain threshold VQ_{Thr} in order to ensure user experience. In addition to video quality, maintaining user experience also requires that V_{BR} does not exceed the network throughput NW_{TPut} in order to avoid video stalling. Hence bit rate selection to maximize battery level can be viewed as minimizing Seg_{Data} subject to bit rate, video quality and network throughput constraints as shown in (2.23).

$$\min Seg_{Time} * V_{BR} \tag{2.23}$$

$$\text{Subject to: } V_{BR-Min} \leq V_{BR} \leq V_{BR-Max} \tag{2.24}$$

$$VQ_{Thr} \leq VQ \tag{2.25}$$

$$V_{BR} \leq NW_{TPut} \tag{2.26}$$

Video quality VQ is measured in terms of average MOS value, MOS_{Avg}^{Video}. MOS_{Avg}^{Video} defined in (2.27) is the average of MOS values corresponding to bit rates of previously downloaded N segments and the bit rate to be selected using (2.23) for the current $N+1^{th}$ segment. As MOS value corresponding to V_{BR-Max}, $MOS(V_{BR-Max})$ represents maximum video quality, we define the lower bound on video quality VQ_{Thr}, as reduced by the factor VQ_{Red} which specifies the acceptable loss in video quality due to battery aware DASH techniques. VQ_{Thr} is given by (2.28).

$$MOS_{Avg}^{Video} = \frac{MOS_{Seg_1} + .. + MOS_{Seg_N} + MOS_{Seg_{N+1}}(V_{BR})}{N+1} \tag{2.27}$$

$$VQ_{Thr} = VQ_{Red} MOS(V_{BR-Max}), 0 < VQ_{Red} \leq 1 \tag{2.28}$$

Network throughput given by (2.29) is the ratio of Seg_{Data} and segment download duration Seg_{DT}. Seg_{Data} and Seg_{DT} used to estimate NW_{TPut} corresponds to the N^{th} segment, that is the network load conditions experienced during the download of the previous segment influences the selection of bit rate of the current segment. It should be noted that Seg_{DT} may be lesser than, equal to or greater Seg_{Time} than depending on network load and channel conditions.

$$NW_{TPut} = \frac{Seg_{Data}}{Seg_{DT}} \tag{2.29}$$

A feasible solution to (2.23) may not always exist. In case NW_{TPut} is lesser than V_{BR-Min}, then V_{BR-Min} is selected which may lead to video stalling and violation of VQ_{Thr}. When both NW_{TPut} and VQ_{Thr} constraints cannot be satisfied, bit rate which satisfies the NW_{TPut} is selected to avoid video stalling and leads to violation of VQ_{Thr}.

Subsequent to bit rate selection, we will now formulate the download rate and mode selection sub-problem for all the download epochs that constitute the segment download. We use the problem formulation given by (2.6) and elaborated in Section 2.3.2, except that upper bound on DR is the amount of segment data that needs to be downloaded and not amount of data needed to fill the buffer [as in (2.20)] and lower bound ensures that playback time is at least equal to segment time and not minimum buffer level Buf_{Min}, [as in (2.22)]. DR^{Max} corresponding to any download epoch in a segment cannot exceed the difference of total segment data Seg_{Data} and segment data downloaded so far, the latter being the sum of the products of duration of each download epoch D_{Period} and download rate DR chosen for the epoch. On the other hand, DR^{Min} is zero when the playback time available PBT exceeds Seg_{Time}. When PBT available is less than Seg_{Time}, DR^{Min} corresponds to the deficit required to increase PBT to at least Seg_{Time} to ensure that buffer contains enough data to playback the segment without stalling. Hence the bounds on download rate are now defined as shown in (2.30) and (2.31).

$$DR^{Max} = \frac{Seg_{Time}V_{BR} - \sum_{i=1}^{N} D^{i}_{Period}DR^{i}}{D^{N+1}_{Period}} \tag{2.30}$$

$$DR^{Min} = \begin{cases} 0, PBT \geq Seg_{Time} \\ (Seg_{Time} - PBT)V_{BR}, PBT < Seg_{Time} \end{cases} \tag{2.31}$$

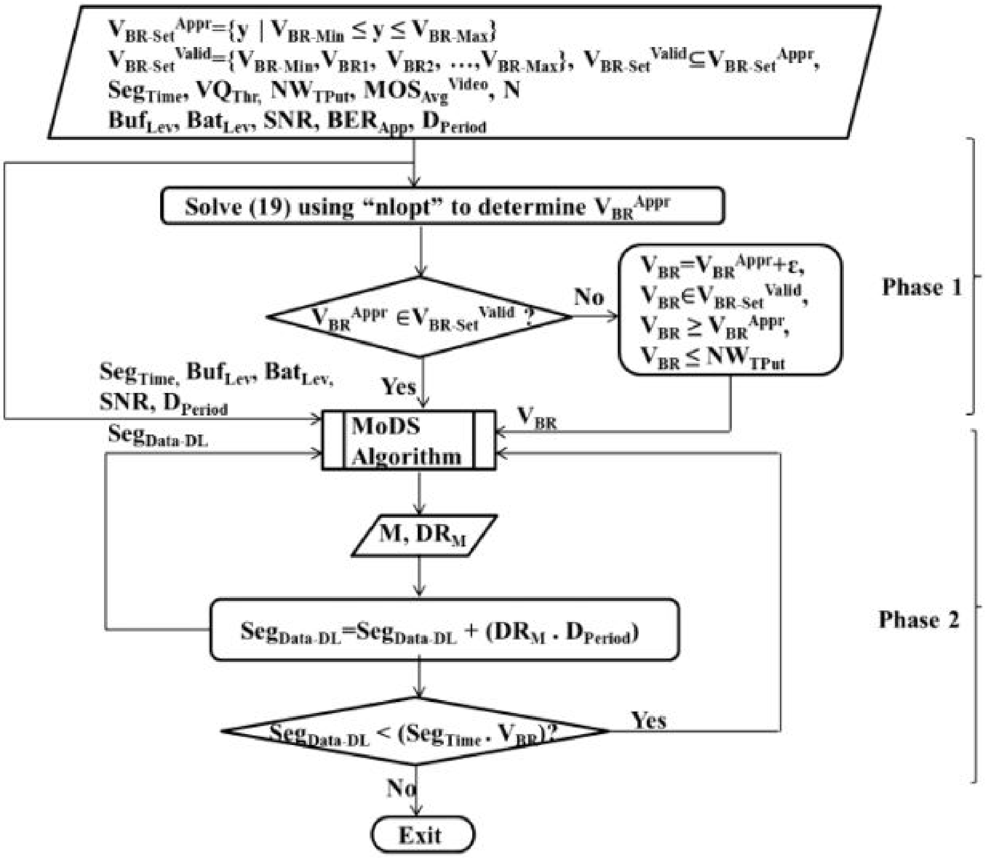

$V_{BR\text{-}Set}^{Appr}=\{y \mid V_{BR\text{-}Min} \le y \le V_{BR\text{-}Max}\}$
$V_{BR\text{-}Set}^{Valid}=\{V_{BR\text{-}Min}, V_{BR1}, V_{BR2}, \ldots, V_{BR\text{-}Max}\}$, $V_{BR\text{-}Set}^{Valid} \subseteq V_{BR\text{-}Set}^{Appr}$,
Seg_{Time}, VQ_{Thr}, NW_{TPut}, MOS_{Avg}^{Video}, N
Buf_{Lev}, Bat_{Lev}, SNR, BER_{App}, D_{Period}
Solve (19) using "nlopt" to determine V_{BR}^{Appr}
$V_{BR}^{Appr} \in V_{BR\text{-}Set}^{Valid}$?
No
$V_{BR}=V_{BR}^{Appr}+\varepsilon$,
$V_{BR} \in V_{BR\text{-}Set}^{Valid}$,
$V_{BR} \ge V_{BR}^{Appr}$,
$V_{BR} \le NW_{TPut}$
Phase 1
Yes
Seg_{Time}, Buf_{Lev}, Bat_{Lev},
SNR, D_{Period}
$Seg_{Data\text{-}DL}$
V_{BR}
MoDS Algorithm
M, DR_M
$Seg_{Data\text{-}DL}=Seg_{Data\text{-}DL} + (DR_M \cdot D_{Period})$
Phase 2
$Seg_{Data\text{-}DL} < (Seg_{Time} \cdot V_{BR})$?
Yes
No
Exit

Figure 2.8. Bit Rate, Mode and Download Rate Selection (BR-MoDS) algorithm

2.5.2 Bit Rate, Mode, and Download Rate Selection (BR-MoDS) Algorithm

In this section we will describe BR-MoDS algorithm that adopts the two tiered problem formulation elaborated in the previous subsection to search the bit rate space and transceiver configuration space (Tables 2.1 and 2.2). Fig. 2.8 shows the inputs, two phases and outputs of BR-MoDS algorithm. As shown in Fig. 2.8, phase 1 involves selecting bit rate V_{BR}^{Appr} that minimizes the Seg_{Data} given the bit rate, video quality and network throughput constraints. If V_{BR}^{Appr} does not belong to V_{BR-Set}^{Valid}, it is rounded off to the nearest higher valid bit rate V_{BR} by adding ϵ such that the resulting bit rate does not violate the network throughput constraints. It should be noted that the rounding off of bit rate does not violate the video quality threshold as a higher bit rate is chosen. The output of phase 1, V_{BR} along with Buf_{Lev}, Bat_{Lev}, SNR and D_{Period} form the inputs to MoDS algorithm (Section 2.3.3, Fig. 2.5). As elaborated in the previous subsection, the bounds on download rate constraint used by the MoDS algorithm are defined by (2.30) to (2.31) instead of (2.20) to (2.22). The MoDS algorithm is iteratively called, with iterations corresponding to download epochs, till the aggregate of the segment data downloaded is equal to Seg_{Data} as shown in phase 2, Fig. 2.8. The output of MoDS is the mode M and download rate used in that epoch.

Having discussed in detail the framework and algorithm developed to maximize battery lifetime during DASH streaming, we next discuss an approach to jointly maximize both battery lifetime and video quality.

2.5.3 Joint Maximization of Battery Lifetime and Video Quality

The BR-MoDS algorithm described above selects the minimum (optimal) bit rate that satisfies the video quality and network throughput constraints even though battery level and network conditions may allow selection of higher bit rate as it aims to maximize only battery lifetime and not aggregate video quality. On the other hand, aggregate video quality can be potentially enhanced by choosing a higher video quality threshold VQ_{Thr} which will result in choosing higher bit rates, but will decrease battery savings. This implies that joint maximization of battery savings and aggregate video quality is required to balance the battery lifetime–video quality tradeoff achieved by bit rate adaptation. However, while bit rate impacts video quality directly, it has an indirect relationship with battery lifetime. Bit rate determines the amount of data to be downloaded, which in turn (along with battery and buffer levels, channel and network load) determines the mode and download rate and hence battery lifetime. This indirect relationship does not lend itself naturally to a joint battery lifetime–aggregate video quality maximization formulation. Hence in this section, we propose a heuristic approach which uses information about battery level and network conditions during bit rate selection to opportunistically maximize both battery lifetime and aggregate video quality.

One possible way of utilizing battery level information during bit rate selection is to scale bit rate with battery level. The basis for this approach is that when battery level is high, battery can support higher drain due to higher bit rates whereas when battery level is low, lower bit rates have to be chosen because higher bit rates will deplete the battery to a greater extent than when battery level is high. However, though the choice of low bit rates when battery level is low will conserve battery and extend video viewing time, it will also result in consistently low video quality and may not meet the video quality constraint. A better approach will be scaling bit rate with the ratio of battery level Bat_{Lev} to the starting battery level $Bat_{Lev-Init}$. Using the ratio ensures that scaling of bit rate and rate of increase in scaling during a session is lesser when Bat_{Lev} is higher, and much more when Bat_{Lev} is lower. For instance, consider the two cases when battery level reduces by 0.05 and 0.1, the ratio values are 0.95 and 0.9 respectively when $Bat_{Lev-Init}$ is 1 and 0.75 and 0.5 when is 0.2. This results in wider range of bit rates selected during a session when Bat_{Lev} is low with higher bit rates boosting quality and lower bit rates offsetting the drain due to higher bit rates. It should be noted that whenever the bit rate selected exceeds NW_{TPut}, it is set to NW_{TPut} in order to avoid buffer underflow. As the bit rate selection stated in (2.23) selects the minimum bit rate that satisfies the constraints, based on the above observations, we modify the lower bound on bit rate V_{BR-Min} to a battery level dependent bit rate V_{BR}^{BA} given by (2.32).

$$V_{BR}^{BA} = V_{BR-Min} + Bat_{Lev} Bat_{Lev-Init}^{-1} (V_{BR-Max} - V_{BR-Min}) \tag{2.32}$$

This implies that the lower bound on bit rate shifts higher or lower depending on

Bat_{Lev} thereby using battery level information for bit rate selection. The modified bit rate selection problem is same as that stated in (2.23) except that V_{BR-Min} is replaced by V_{BR}^{BA}. The new algorithm termed Battery Level Aware BR-MoDS, B^{2}R-MoDS is same as BR-MoDS except that phase 1 is modified to reflect the above change. The computational complexity of BR-MoDS and B^{2}R-MoDS which use nonlinear optimization solver 'nlopt' to determine the minimum bit rate V_{BR}^{Appr} is presented in the Appendix A.1.

2.5.4 Battery Aware Video Streaming - Framework

In our proposed framework, the execution of BR-MoDS B^{2}R-MoDS algorithms is distributed as the bit rate selection is mobile device driven (like any DASH based technique) and download rate and mode selection carried out by MoDS is base station driven. The framework is the same as that elaborated in Appendix A.2 except that the bit rate is sent by the mobile device prior to each segment download. Also, the initial information conveyed by mobile device at the beginning of video session consists of V_{BR-Max}, V_{BR-Min}, maximum PBT possible and also the segment time Seg_{Time}. Subsequently, for each of the download epoch that constitutes the segment download, the information exchange between base station and mobile device is as explained in Appendix A.2. However, the buffer status update is used to calculate DR^{Max} and DR^{Min} defined in (2.30) and (2.31).

2.5.5 Video Experience Longevity (VEL) Metric

In this section, we develop the Video Experience Longevity (VEL) metric to quantify the performance of the proposed battery aware bit rate adaptation techniques in terms of both the longevity of video experience and the quality of video experience as compared to alternative DASH based techniques. In this chapter, for comparison we consider the non-battery aware rate adaptation algorithm proposed in [22] for DASH [44] (termed RA-DASH) and the battery aware rate adaptation technique (termed BaSe-AMy) proposed in [27]. The VEL metric is developed to compare performances of the different techniques for the most demanding scenario when the mobile device continuously downloads and watches videos till the battery gets exhausted. In this scenario, note that the longevity of video experience Exp_{Time} is the same as battery lifetime $Bat_{Lifetime}$ minus any stalling time $Stall_{Time}$ during the video sessions, as given by (2.33) below. However, even in other user scenarios, a DASH technique with higher VEL score than another technique can be considered more efficient in terms of battery lifetime and/or video experience. While modeling of the quality of video experience VE continues to be an active area of research, in this chapter we model VE as shown in (2.36) as a weighted sum of video spatial quality measured by the MOS_{Avg}^{Tot} defined in (2.34) as the average of MOS_{Avg}^{Video} [defined in (2.27)] of all the K videos streamed till the battery dies, and video temporal quality reflected by a term $NStall_{Norm}$ defined in (2.35), which measures how free the video experience is from stalls/jitter. The weights w_{MOS} and w_{NStall} in (2.36) reflect relative priority for spatial quality versus stall-free video in determining user experience. Note that we normalize $NStall_{Norm}$ to 5 in line with MOS score so we can consider both of them in VE; when there is no stalling, the value is 5, while in the extreme case that no video playback is possible at all due to stalling, the value is 0.

$$Ext_{Time} = Bat_{Lifetime} - Stall_{Time} \tag{2.33}$$

$$MOS_{Avg}^{Total} = \frac{MOS_{Avg}^{Video_1} + MOS_{Avg}^{Video_2} + .. + MOS_{Avg}^{Video_K}}{K} \tag{2.34}$$

$$NStall_{Norm} = \frac{5Exp_{Time}}{Bat_{Lifetime}} \tag{2.35}$$

$$VE = w_{MOS} MOS_{Avg}^{Total} + w_{NStall} NStall_{Norm}, 0 < w_{MOS}, w_{NStall} \leq 1 \tag{2.36}$$

Next we define the VEL metric in (2.37) to quantify the joint gain/loss in experience longevity and quality of video experience achieved by the proposed battery aware DASH techniques over DASH techniques used for comparison. The ratio increases (decreases) when there is gain (loss) in experience longevity relative to gain (loss) in video experience.

$$VEL = \frac{1 + \Delta Exp_{Time}}{1 - \Delta VE} \tag{2.37}$$

ΔExp_{Time} defined in (2.38) and ΔVE defined in (2.39) represent the gain/loss in experience longevity and video experience respectively achieved by the proposed battery aware DASH (BA-DASH) techniques (BR-MoDS and B^{2}R-MoDS) over other DASH techniques (non-battery aware [22] and battery aware [27]).

$$VEL = \frac{Exp_{Time_{BA-DASH}} - Exp_{Time_{DASH}}}{Exp_{Time_{DASH}}} \tag{2.38}$$

$$\Delta VE = \frac{VE_{BA-DASH} - VE_{DASH}}{VE_{DASH}} \tag{2.39}$$

Note that ΔExp_{Time} and ΔVE for technique used for comparison (RA-DASH or BaSe-AMy) are zero, implying VEL value of 1. VEL for the proposed techniques can be greater than or lesser than 1 depending on values of ΔExp_{Time} and ΔVE. If a proposed technique has VEL greater than 1, it is more efficient than the DASH technique used for comparison in terms of experience longevity and/or video experience.

2.6 Simulation Framework and Results

In this section, we describe the simulation framework developed to adaptively stream different video sequences under varying channel and network load conditions and video quality requirements. We will then discuss the experimental results obtained using the proposed battery aware DASH techniques, as compared to using the conventional RA-DASH technique and battery aware rate adaptation technique, BaSe-AMy.

2.6.1 Simulation Framework

In this section, we will elaborate on the modifications made to the simulation framework developed in Section 2.4 to estimate battery consumption during adaptive bit rate streaming and playback. As rate adaptation techniques for DASH adapt video bit rate under challenging channel conditions and network load, we extend the simulation framework developed in Section 2.4 to simulate varying network load (equivalent to varying number of users) by modulating the peak throughput available to a particular user while downloading video. Also in the framework, MoDS algorithm is replaced by BR-MoDS and B^{2}R-MoDS algorithms. When video download is initiated, the simulation time counter is started. As before, in our experiments, simulation step is fixed at 2s. In the simulation step that marks the beginning of segment download, BR-MoDS/B^{2}R-MoDS determines the bit rate of the segment. For all the subsequent simulation steps that download this segment data, MoDS algorithm (Sections 2.3.3 and 2.4.4) determines the mode and corresponding download rate. The simulation counter when the battery is fully drained gives battery lifetime when user downloads and watches chosen video sequence under simulated channel and network load conditions and quality requirements. In order to capture the effect of bit rate adaptation on user experience, we modify the User Experience Model (Section 2.4.3) to include the MOS corresponding to the bit rate

Table 2.6. Simulation parameters for DASH streaming

Video Characteristics	Video Bit Rate V_{BR}^{Valid}(Mb/s):$\{4.5, 3.75, 3.125, 2.6, 2.17, 1.81, 1.51\}$
	MOS Values: $\{4.8, 4.6, 4.49, 4.35, 4.2, 4.02, 3.9\}$
	Segment Time Seg_{Time} = 10s
	Video Sequence 2: $\{$404s, 1942s, 124s, 360s, 526s, 190s, 757s, 738s, 360s, 255s, 232s, 396s, 181s, 219s, 319s, 139s, 348s, 408s$\}$
Client Characteristics	Video Buffer Size Buf_{Size} = 50s
	Playback Load (Decoder + Display) $I_{Playback}$ = 34mA
Video Quality Requirements	Quality Threshold VQ_{Thr}= 4.32 (VQ_{Red}=0.9, 10% degradation from highest MOS value of 4.8)
Network Level	Variable, Peak Throughput = 2.52-8.4Mb/s

selected. We use the bit rate–MOS model [45] to map the bit rate of each segment to a MOS value and calculate the average MOS value for the video streamed using (2.27). Given these MOS values and stalling measurements, the video experience VE of the user is measured using (2.36).

To allow comparison, we use the same framework to simulate the RA-DASH and BaSe-AMy, except that, instead of using BR-MoDS/B^{2}R-MoDS, we use the algorithm implemented in [22] and [27] respectively to determine bit rate. For RA-DASH and BaSe-AMy, the download rate is determined by (2.30), and mode that satisfies the download rate and BER requirement is selected. It should be noted that if download rate determined by BR-MoDS/B^{2}R-MoDS or for RA-DASH/BaSe-AMy exceeds the peak throughput, then the base station limits download rate to the peak throughput rate. The user characteristics, channel conditions and application BER requirements are identical to those in Table 2.5.

Table 2.6 lists the other required simulation parameters used in our DASH streaming experiments. The video characteristics consist of the set of video bit rates available for selection, the corresponding MOS values (derived from the video bit rate–MOS mapping for VGA screen resolution presented in [45]), duration of segments and the video sequence viewed. Each of the videos in video sequence 2 are available in segments of duration Seg_{Time} and each of these segments are encoded using V_{BR-Set}^{Valid}. Client characteristics enumerate buffer size and playback current requirements. The video quality requirements specify the maximum quality reduction acceptable VQ_{Red} and the quality threshold VQ_{Thr} that must be satisfied. Table 2.6 also lists the peak throughput under variable network load conditions.

2.6.2 Experimental Results

In this section, we will present the experimental results obtained by simulating adaptive bit rate streaming of video under variable network load conditions and different channel conditions. In all the experiments reported below, we set the weights w_{MOS} and w_{NStall} in (2.36) to 0.5, giving equal priority to spatial quality and stall-free video.

Fig. 2.9 shows the selection of bit rate (shown as green solid line) and download rate (shown as blue pluses) while streaming a video of 200s duration using RA-DASH, and our proposed BR-MoDS and B^{2}R-MoDS techniques, under variable network load (shown as red dashed line representing the variation in peak throughput) and variable channel conditions. For lack of space, we do not illustrate the same for BaSe-AMy technique. The 200s video has the same bit rate/MOS characteristics shown in Table 2.6. Fig. 2.9a shows that RA-DASH attempts to track the network throughput while selecting bit rates, and downloads at the highest rate possible during each download epoch. From Fig. 2.9b, it can be seen that BR-MoDS chooses the lowest bit rate possible initially, followed by higher bit rates (in order to boost MOS_{Avg} and satisfy the video quality constraint) and also lowest download rates possible. Fig. 2.9c shows that B R-MoDS, as designed, chooses bit rates higher than that selected by BR-MoDS (except

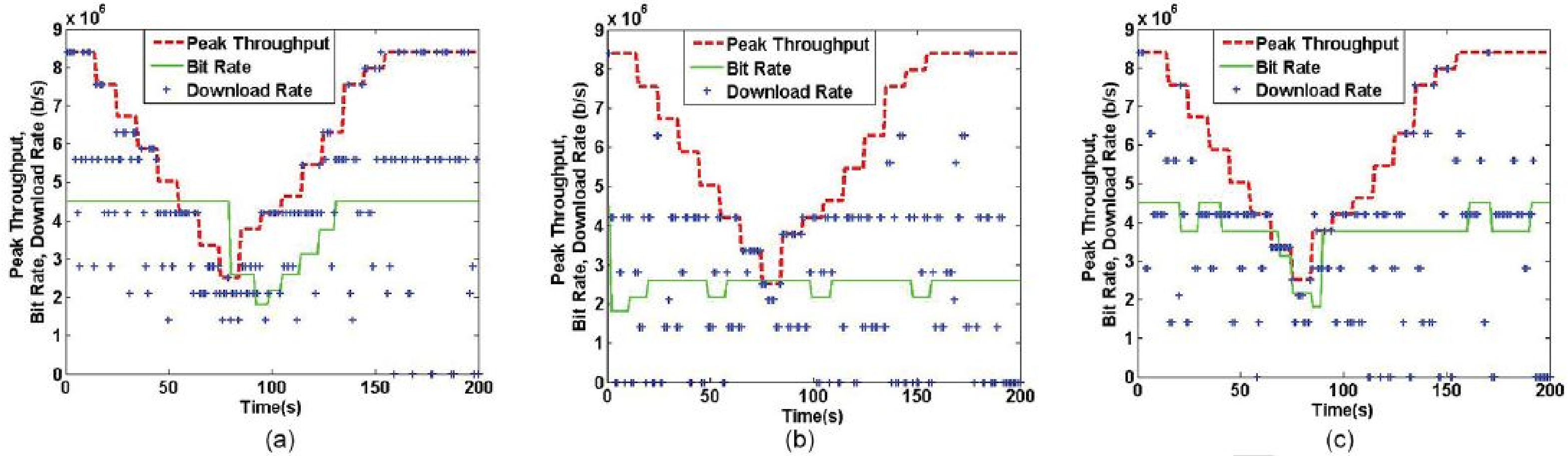


Figure 2.9. Effect of downloading a single video of 200s on bit rate and download rate under variable netowrk load and channel conditions while using (a)RA-DASH, (b)BR-MoDS, and (c)B^{2}R-MoDS.

Table 2.7. Exp_{Time}, VE for RA-DASH and BaSe-AMy, $\%\Delta Exp_{Time}$ $\%\Delta VE$ and for BR-MoDS and B^{2}R-MoDS

	Exp_{Time}(s)	$\%\Delta Exp_{Time}$		VE	$\%\Delta VE$	
	RA-DASH	BR-MoDS	B^{2}R-MoDS	RA-DASH	BR-MoDS	B^{2}R-MoDS
SNR-Variable	1134	46.2	34.8	4.825	-3	-1.9
SNR-High	1286	61.1	41.4	4.821	-3.2	-0.48
SNR-Low	1013	32	20.2	4.78	-2.4	-0.78
	BaSe-AMy	BR-MoDs	B^{2}R-MoDS	BaSe-AMy	BR-MoDS	B^{2}R-MoDS
SNR-Variable	1082	53.2	41.3	4.75	-1.7	0.29
SNR-High	1313	57.7	38.5	4.76	-2	0.733
SNR-Low	977	36.9	24.6	4.54	-2.35	-0.66

Table 2.8. VEL metric values for RA-DASH, BaSe-AMy, BR-MoDS and B^2-MoDS

	VEL		
	RA-DASH	BR-MoDS	B^{2}R-MoDS
SNR-Variable	1	1.42	1.33
SNR-High	1	1.56	1.40
SNR-Low	1	1.29	1.19

	BaSe-AMy	BR-MoDs	B^{2}R-MoDS
SNR-Variable	1	1.5	1.41
SNR-High	1	1.54	1.39
SNR-Low	1	1.4	1.3

when BR-MoDS selects higher bit rates to boost MOS_{Avg}), with the bit rate selected going down as it tracks battery level ratio which decreases as download progresses. However, like BR-MoDS, it also selects the lowest download rate possible. Though we do not illustrate the bit rate selection carried out by BaSe-AMy, it should be noted that BaSe-AMy always selects the highest bit rate possible. BaSe-AMy lowers the bit rate only when battery lifetime remaining is lesser than that required to completely stream the video and the battery level is below a certain threshold.

Next we report on the effect of the DASH based techniques on battery level and quality of video experience. Assuming the battery level is 0.2 at the start of the 200s video download, the battery level reduces by 16.1%, 17.34%, 10.45%, and 12% for RA-DASH, BaSe-AMy, BR-MoDS and B^{2}R-MoDS respectively while achieving a video experience of 4.83, 4.76, 4.66, and 4.793. This shows that the proposed battery aware DASH techniques result in more battery efficient video streaming than the conventional RA-DASH and BaSe-AMy techniques. We also see that BR-MoDS can be more battery efficient than B^{2}R-MoDS as it uses lower bit rates, while B^{2}R-MoDS can achieve higher video experience.

In the next set of experiments, we simulate the video snacking behavior (variable snacking ratio, Table 2.5) by the mobile device downloading video sequence 2 (Table

2.6), starting with battery level 0.2 till the battery gets exhausted, giving the battery lifetime. We report in Tables 2.7 and 2.8 values for Experience Longevity Exp_{Time}, quality of Video Experience VE and VEL metric respectively obtained by RA-DASH and BaSe-AMy when streaming video sequence 2 under variable network load (red dashed line in Fig. 2.9) and variable, high, and low SNR conditions. Also reported in Tables 2.7 and 2.8 are the percentage gains (loss) over RA-DASH and BaSe-AMy in Experience Longevity ΔExp_{Time} and Video Experience $\%\Delta VE$, as well as VEL values, when using BR-MoDS and B^{2}R-MoDS. From Table 2.7 we observe that for variable SNR conditions (row 1), the experience longevity is significantly increased by using BR-MoDS and B^{2}R-MoDS; 46.2% and 34.8% compared to RA-DASH and 53.2% and 41.3% compared to BaSe-AMy. In terms of video experience, BR-MoDS loses 3% and 1.7% compared to RA-DASH and BaSe-AMy while B R-MoDS loses 1.9% and gains 0.29% compared to RA-DASH and BaSe-AMy respectively. As can be expected from the $\%\Delta Exp_{Time}$ and $\%\Delta VE$ results, BR-MoDS and B^{2}R-MoDS show significant gains in VEL compared to both RA-DASH and BaSe-AMy as shown in Table 2.8.

Under high SNR conditions, the longevity of video experience is higher than under variable SNR conditions for all the techniques, including RA-DASH and BaSe-AMy, as less power consuming modes can be used to achieve the required BER. It can be seen from Table 2.7 that by using BR-MoDS and B^{2}R-MoDS, experience longevity increases by 61.1% and 41.4% compared to RA-DASH and by 57.8% and 38.5% compared to BaSe-AMy. In terms of video experience, BR-MoDS loses 3.2% and 2% compared to RA-DASH and BaSe-AMy while B R-MoDS loses 0.4% and gains 0.7% compared to RA-DASH and BaSe-AMy respectively. As expected, BR-MoDS and B^{2}R-MoDS outperform RA-DASH and BaSe-AMy in terms of VEL values (Table 2.8).

Lastly, when channel conditions are bad (low SNR), all the DASH techniques achieve lower battery lifetime compared to high and variable SNR conditions as more power intensive modes have to be used to meet BER requirements resulting in lower battery lifetime. BR-MoDS and B^{2}R-MoDS extend experience longevity by 32% and 20.2% compared to RA-DASH and 36.9% and 24.6% compared to BaSe-AMy. In terms of video experience, BR-MoDS and B^{2}R-MoDS lose 2.4% and 0.78% compared to RA-DASH and 2.35% and 0.66% compared to BaSe-AMy. As before, both BR-MoDS and B^{2}R-MoDS outperform RA-DASH and BaSe-AMy in terms of VEL metric as shown in Table 2.8.

In this chapter, we developed techniques for increasing battery lifetime of mobile devices during video download while ensuring no degradation in user experience. In the forthcoming chapters, we will focus on increasing the power efficiency of base stations in the cellular networks.

2.7 Summary

In this chapter, we presented novel battery aware HTTP video delivery schemes. First, we proposed battery aware video progressive download techniques that dynamically adapt video download rate and transceiver configurations to reduce battery consumption while ensuring user experience. Next, we presented battery aware DASH streaming techniques that aim to maximize both battery lifetime and video quality while ensuring minimum desired video quality by adapting video bit rate in addition to download rate and transceiver configuration. Lastly, we proposed the Video Experience Longevity metric that quantifies the performance of the proposed battery aware DASH techniques in terms of experience longevity and video experience. Our simulation results demonstrated the ability of the proposed techniques to achieve significant increase in battery lifetime, no more than the desired (video quality threshold) loss in video experience and high VEL values as compared to conventional non-battery aware techniques and other battery aware techniques.

While the proposed battery aware video delivery techniques focus on increasing battery lifetime, in future, we aim to investigate techniques that jointly reduce the power consumption at the base station and battery consumption of mobile device while downloading mobile video. We would also like to extend our techniques to explore battery savings when video is streamed and uploaded from mobile devices.

2.8 Acknowledgements

We thank the anonymous WCNC 2013 and IEEE Transactions on Multimedia reviewers for their feedback and comments on the work.

Chapter 2, in part, contains material as it appears in the Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'13). "Rate Adaptation and Base Station Reconfiguration for Battery Efficient Video Download". Ranjini Guruprasad, Sujit Dey. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, contains material as it appears in IEEE Transactions on Multimedia. "Battery Aware Video Delivery Techniques Using Rate Adaptation and Base Station Reconfiguration" ". Ranjini Guruprasad, Sujit Dey. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Dynamic Cell Reconfiguration Framework for Energy Conservation in Cellular Networks

3.1 Introduction

With the explosive growth in wireless communication usage and infrastructure, energy use of cellular wireless networks has lately become a critical issue [46]–[47]. Designers of communication and networking algorithms and protocols have traditionally put less weight on the complexity and power consumption at base stations (BSs) than improving energy efficiency to prolong battery life-time of user equipments (UEs). Today, however, the situation has changed. Pushed by ever increasing energy costs and environmental concerns, all information and communications technology (ICT) industries are seeking ways to reduce energy consumption. In particular, improving the energy efficiency of BSs has become as important as UEs because the BSs have been identified to be the most power consuming equipment, e.g., 60–80% of the total energy consumption in current cellular networks [48].

Energy efficiency with respect to BSs has been considered in many dimensions, spanning from hardware component improvements to network-level solutions. A number of these efforts have focused on novel hardware design and manufacture, e.g., energy-efficient power amplifiers, fanless coolers. Others also considered collocating BSs with renewable energy sources [49]. In the domain of network-level solutions, there are many recent papers, including, for example, the smart deployment at the stage of network planning [50]–[51] by using micro BSs or relays, load-aware dynamic BS switching on/off [52], [48], [53]–[54], and resource management schemes [55]–[56] such as power control and energy-aware user association, etc.

The focus of this chapter is to present network-level solution with emphasis on three energy-saving techniques operating on different control time scales.

- *Active BS selection*: BSs are typically deployed on the basis of peak traffic volume and stay always-on irrespective of traffic load. Recent temporal traffic trace reports that BSs are largely underutilized during low traffic periods such as nighttime [48]. The active BS selection technique operating on a slow time scale (e.g., order of hour or so) allows the system to entirely turn off some underutilized BSs and transfer the imposed loads to neighboring BSs, which leads to huge energy savings.
- *Transmit power budget adaptation*: A typical macro BS spends a small amount of total operational power on the transmit power. However, when the BS reduces its transmit power, a considerable overall energy saving[1] is expected due to its exerting influence on the operational power [58]. The transmit power budget adaptation is a technique in a fast time scale (e.g., order of minutes), which fine tunes the transmit power of BS according to its current cell loading for further energy savings.
- *User association*: The last technique, which determines a proper BS for each user, is necessary to fully exploit the amount of energy savings. The time scale of user association is apparently faster than the above two techniques because some UEs

[1] 1For example, the BS power consumption model in [57] showed that a macro BS can reduce the total power consumption from 766 W to 532 W (i.e., 234 W savings) just by reducing its transmit power from 20W to 10W. Refer to Section 3.2.3 for more details on our power consumption model.

may need to be associated with another BS when a set of active BSs and/or their transmit powers change. Of course, it should be performed whenever a new user arrives.

In this chapter, we consider a problem of minimizing the total power consumption in BSs while satisfying the quality of service (QoS) requirements for all users in the network. To this end, we develop a novel unified framework for energy conservation, called dynamic cell reconfiguration (DCR), linking the above three techniques together into one.

3.1.1 Related Work

Basic concepts of dynamic BS operation, turning BSs on/off based on the temporal and spatial traffic load, have been addressed in [48], [53]–[54]. The authors in [59], [51] also investigated a joint operation and deployment problem to determine where, how many and which type (macro/micro/pico/femto) of BSs need to be deployed in an energy-efficient manner. However, some of the preliminary works [48], [54] did not capture the effect of the signal strength degradation when traffic loads are transferred from the switched-off BS to neighboring BSs. Rather than developing an actual working algorithm, some [48], [51], [53] simply attempted to see how much energy savings can be expected under the deterministic traffic variation over time and moreover, sometimes in a simple network model such as hexagonal or Manhattan model. In order to overcome these weaknesses, we adopt a more sophisticated channel model based on signal to interference plus noise ratio (SINR) reflecting the effect of signal degradation and validate our framework based on a real dataset of BS topology and utilization.

Another piece of technique we investigate in this chapter is the power control, which has been widely studied in literature (see [60] and the references therein). The power control is usually employed to combat the near-far problem in uplink [61] or to maximize user throughput in downlink by exploiting the variation of time and frequency channel (i.e., multi-user diversity) as well as mitigating inter-cell interference. The main purpose of our approach, different from the conventional algorithms, is to reduce the total energy consumption of BSs by adjusting the transmit power budget. Our algorithm is concerned about the entire budget but does not care about how the budget would be actually utilized for multiple users across time/frequency resource in a cell. In this sense, we call it the transmit power budget adaptation. We would like to highlight that it can be superimposed over any power allocation algorithms (e.g., water-filling). For example, once it first adapts the power budget according to the current traffic load, an underlying algorithm distributes power to users within the budget.

There are a few prior works [55]–[62] studying the power control for the purpose of BS energy conservation in slightly different settings. In [55], the authors proposed short- and long-term power controls to exploit the traffic fluctuation, but their analysis was still in a single-cell setting. In [63], the greening effect of interference management with combinations of spatial and temporal power budget sharing is investigated. Niu et al. [62] presented an idea of cell zooming that dynamically adjusts the cell size though BS cooperation, relaying or physical antenna tilt. Our work fills the voids of the previous work in that: We not only propose a practical algorithm in a multi-cell setting, but also address the problem of jointly optimizing the power budget in conjunction with active BS selection and user association. The rest of this chapter is organized as follows. Section 3.2 formally describes our system model and general problem. In Section 3.3, we propose a dynamic cell reconfiguration framework and discuss some of practical implementation issues. In Section 3.4, we present simulation results. In Section 3.5, we conclude the chapter with our notes and observations.

3.2 System Model

3.2.1 Network and Channel Model

Let us consider a cellular wireless network with a set of BSs $\mathcal{B}$. Let x denote a user location lying in the two-dimensional area $\mathcal{L}$ and $i \in \mathcal{B}$ be the index of the i^{th} BS. We assume the same frequency band with bandwidth W in all cells (i.e., reuse factor one) and concentrate on downlink communication that is a primary usage mode for mobile Internet, i.e., from BSs to UEs. However, we would like to mention that some aspects of our work (e.g., user association and active BS selection) can be applied to the uplink scenario as well with a slight modification. Following Shannon's formula, the transmission rate [bits] of a user at location x when associated to BS i, is given by:

$$c_i(x) = W \log_2 \left(1 + \frac{g_i(x) p_i}{\sum_{b \in \mathcal{B}_{on} \setminus \{i\}} g_i(x) p_b + \sigma^2} \right) \tag{3.1}$$

where σ^2 is noise power and p_i [or p_b] is the transmit power of BS i [or BS b], $g_i(x)$ [or $g_b(x)$] denotes the channel gain from BS i [or BS b] to location x, including path loss attenuation, shadowing and other factors if any. Note that the transmission rate $c_i(x)$ depends not only on the set of active BSs B_{on} but also on their transmit power $p = (p_1, p_2, ..., p_{|\mathcal{B}|})$.

3.2.2 Traffic Demand and BS Utilization

We assume that a user at location x has a certain traffic demand, which requires $\gamma(x)$[bits]. To guarantee the QoS of the user, the fraction of radio resource blocks (i.e., time or frequency) need to be allocated by BS i would be $\frac{\gamma(x)}{c_i(x)}$.

We now define an association probability $\pi_i(x)$, which specifies the probability that the user at location x is routed to BS i. As can be seen later in subsection 3.3.1, the optimal $\pi_i(x)$ would be either two extremes 0 or 1. The BS utilization, the average occupied fraction of the BS resource blocks, can be defined as follows:

$$\rho_i \doteq \int_{\mathcal{L}} \frac{\gamma(x)}{c_i(x)} \pi_i(x) dx \tag{3.2}$$

Definition 3.1. *(Feasible Set): When the set of active BSs $\mathcal{B}_{on}$ and their transmit power p are given, the set $\mathcal{F}(B_{on}, p)$ of feasible utilization ρ can be defined as follows:*

$$\mathcal{F}(\mathcal{B}_{on}, \boldsymbol{p}) \doteq \boldsymbol{\rho} = (\rho_1, .., \rho_{\mathcal{B}_{on}}) \mid 0 \leq \boldsymbol{\rho} \leq 1, \tag{3.3}$$

$$\forall x \in \mathcal{L}, 0 \leq \pi(x) \leq 1, \tag{3.4}$$

$$\forall x \in \mathcal{L}, \sum_{i \in \mathcal{B}} \pi_i(x) = 1, \tag{3.5}$$

$$\forall x \in \mathcal{L}, \forall i \in \mathcal{B} \setminus \mathcal{B}_{\wr\backslash}, \tag{3.6}$$

$$\pi_i(x) = 0 \tag{3.7}$$

where we use "$\leq$" to denote element-wise inequality for the vectors. Note that the feasible BS utilization $\boldsymbol{p}$ has the associated probability vector $\pi(x) = (\pi_1(x), \ldots, \pi_{|\mathcal{B}|}(x))$ for all $x \in \mathcal{L}$.

3.2.3 Power Consumption Model

Now let us consider the modeling of the total BS operational power consumption T_i that can capture both dynamic power and static power as follows. The former is proportional to BS's utilization. On the other hand, the latter is the fixed amount of power that a BS dissipates irrespective of its utilization. It is worthwhile mentioning that the static portion of power consumption can be conserved only if the BS is completely shut off.

$$T_i = \underbrace{(1-q_i)\rho_i P_i}_{\text{dynamic}} + \underbrace{q_i P_i}_{\text{static}} \tag{3.8}$$

where $q_i \in [0, 1]$ is the portion of the static power consumption for BS i, and P_i is the maximum power consumption when it is fully utilized with the transmit power p_i. According to [6], P_i is again a function of the transmit power p_i with nonnegative coefficients a_i and b_i:

$$P_i = a_i p_i + b_i \tag{3.9}$$

where the coefficient a_i accounts for the power consumption that scales with the average transmit power and b_i is the offset site power which is consumed independently of the average transmit power. We would like to emphasize that our model given in (3.8) is general enough to grasp a variety of BS power consumption.

- Energy-proportional BS with $q_i = 0$: Assuming ideally equipped with energy-proportional equipment, the BS does not consume any power when idle, and proportionally consumes more power as its utilization increases.

- Non-energy-proportional BS with $q_i > 0$: In practice, several hardware devices inside a BS dissipate standby power even though the BS does not serve any traffic. In an extreme case of $q_i = 1$, the model becomes a constant consumption, which has been widely used in many works in literature [48], [64].

3.2.4 General Problem Statement

We consider a general problem that minimizes the total BS power consumption while all user traffic requirements are guaranteed to be served, in other words, maintaining the BS utilization within the feasible set.

$$[GP] \min_{i \in \mathcal{B}_{on}} T_i \tag{3.10}$$

$$\text{Subject to:} \mathcal{B}_{on} \subseteq \mathcal{B}, \tag{3.11}$$

$$p \leq p^{max}, \tag{3.12}$$

$$\rho \in \mathcal{F}(\mathcal{B}_{on}, p) \tag{3.13}$$

Our ultimate goal is to develop a framework for BS energy conservation that encompasses (i) active BS selection, (ii) user association, and (iii) transmit power budget adaptation. As a first step towards this goal, our own prior work [52] focused on building solutions for the first two sub-problems assuming all BSs are operating at the maximum transmit power, i.e., $p = p^{max}$ without the constraint (3.12).

The active BS selection algorithm presented in this chapter looks similar to GON in [65] because they have been built based on the same system model. However, from the problem formulation standpoint, their objective functions are different (e.g., total power consumption vs. cost minimization with the flow-level performance and the energy consumption), so are their final algorithms. In addition to that by further relaxing the maximum transmit power assumption made in [65], we are able to investigate the interaction between active BS selection and transmit power budget adaptation.

3.3 Dynamic Cell Reconfiguration Framework

In this section, we present details on our framework, called *dynamic cell reconfiguration (DCR)*, that includes the user association, active BS selection, and transmit power budget adaptation algorithms.

3.3.1 User Association

We shall start by considering a given set of active BSs $\mathcal{B}_{on}$ and their fixed transmit power p. In this setting, the static power consumption term can be ignored. So the induced sub-problem of [GP] is to determine which BS each user should be associated to, or equivalently, to find an optimal BS utilization ρ.

$$[UA-P] \sum_{i \in B_{on}} [(1-q_i)P_i\rho_i + L_i(\rho_i)] \tag{3.14}$$

$$\text{Subject to:} \boldsymbol{\rho} \in \mathcal{F}(\mathcal{B}_{on}, p) \tag{3.15}$$

where $L_i(\rho_i)$ is a convex penalty function we intentionally introduce. By adding the penalty into the objective, we can allow the system to balance the traffic load among BSs and avoid a cell getting too congested. Though there may be other methods of penalizing the congested cell for the purpose of load balancing, the work presented in this chapter uses the following penalty function with three configurable parameters.

$$L_i(\rho_i) = \begin{cases} 0, \rho_i < \rho_{th} \\ L_{max}\left(\frac{\rho_i - \rho_{th}}{1-\rho_{th}}\right) \end{cases} \tag{3.16}$$

where $L_{max} \geq 0$ is the maximum penalty value and $\rho_{th} \in [0, 1]$ is the BS utilization threshold from which we start penalizing the BSs; $\beta \geq 1$ controls the sharpness of the penalty function. It is noteworthy that the modified problem given in (3.14) is asymptotically equivalent to the original sub-problem without the penalty function L_i in any of the following conditions: As L_{max} goes to zero, ρ_{th} goes to one, or β goes to infinity.

Lemma 3.2. *The Feasible set $\mathcal{F}(\mathcal{B}_{on}, p)$ in (3.3) is convex.*

Proof. The proof is straightforward by definition of convex set. We refer to [65] for the full proof. □

Theorem 3.3. *When the problem given in (3.14) is feasible, the optimal policy is for user at location x to associate with BS $i^*(x)$, given by*

User association algorithm:

$$i^*(x) = \underset{i \in \mathcal{B}_{on}}{argmax} \quad c_i(x)[(1-q_i)P_i + L_i'(\rho_i^*)]^{-1} \tag{3.17}$$

where ρ^* is the optimal BS utilization

Please refer to Appendix B for the optimality proof.

Remark 1. *But the subtlety is that the optimal policy in (3.17) has a chicken-and-egg dilemma. It requires the optimal utilization ρ^* in advance to calculate the metric for the optimal policy. However, we were able to prove that a distributed algorithm that achieves the global optimum without knowing ρ^* in an iterative manner. A sketch of the proof is given as follows. First we show that the optimal BS utilization ρ^* is the fixed point of a certain mapping. Next we show that the following algorithm (or mapping) produces a descent direction at the current BS utilization $\rho^{[k]}$ (i.e., minimizing the inner-product with the gradient). Thus, it will eventually converge to the global optimal point. The full proof can be obtained via a slight modification of the convergence proof in [66].*

3.3.2 Active BS Selection

In this section, we investigate another piece of subproblem in DCR, namely, active BS selection, where we assume that all active BSs are operating at the maximum transmit power, i.e., $p = p^{max}$. This assumption will be relaxed and the adaptation of transmit power will be covered in the forthcoming section. By solving this problem, we will be able to answer which BSs need to remain active to guarantee the QoS level of users and the others to be turned off for minimizing energy consumption in the network.

$$[BS-P1] \min_{\mathcal{B}_{on} \subseteq \mathcal{B}} UA(\mathcal{B}_{on}) + \sum_{i \in \mathcal{B}_{on}} q_i P_i$$

where $UA(\mathcal{B}_{on}) \doteq \min_{\rho \in \mathcal{F}(\mathcal{B}_{on}, p^{max})} \sum_{i \in \mathcal{B}_{on}} (1 - q_i) \rho_i P_i$ which is the optimal objective value of user association problem.

There is a technical challenge in solving this problem because it can be reduced from a vertex cover problem which is theoretically known as NP-complete [67]. In order to overcome such a high computational complexity, we consider the design of an efficient heuristic algorithm in this section. To that end, we move the static power consumption term in the objective to the constraint with a nonnegative budget Z.

$$[BS-P2] \min_{\mathcal{B}_{on} \subseteq \mathcal{B}} UA(\mathcal{B}_{on}) \tag{3.18}$$

$$\text{subject to} \sum_{i \in \mathcal{B}_{on}} q_i P_i \leq \frac{Z}{\lambda} \tag{3.19}$$

As can be easily noticed, there is a close relationship between [BS-P1] and [BS-P2] as primal/dual problems with a Lagrangian multiplier λ. In order to further convert [BS-P2], let us introduce a diminishing returns property on a set function that is formalized by the concept of submodularity [68].

Definition 3.4. *(Feasible Set): For a real-valued set function H, we define the discrete derivative at $A \subseteq S$ in direction $s \in S$ as $d_s(\mathcal{A}) = H(\mathcal{A} \cup \{s\}) - H(\mathcal{A})$. The H is said to be submodular if*

$$\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \mathcal{S} \implies d_s\mathcal{A}_{\in}\forall s \in S \setminus \mathcal{A}_{\in} \tag{3.20}$$

Similarly, H is supermodular if $-H$ is submodular.

We rewrite [BS-P2] in the standard form of submodular maximization problem as follows.

$$[BS-P3] \max_{\mathcal{A}\subseteq\mathcal{B}_{on}} H(\mathcal{A}) \tag{3.21}$$

$$\text{subject to} c(\mathcal{A}) - \sum_{i\in C} c_i \leq C \tag{3.22}$$

where $\mathcal{A} = \mathcal{B}_{on} \setminus \mathcal{B}_{init}$, $\mathcal{B}_{init}$ is an any initial BS set, $H(\mathcal{A}) = UA(\mathcal{B}_{init})$ $-UA(\mathcal{B}_{init} \cup \mathcal{A}), c(i) = q_i P_i$, and $C = \frac{Z}{\lambda} - \sum_{i\in\mathcal{B}_{init}} c(i)$

It is worthwhile mentioning that there exists an intuitive yet efficient greedy algorithm for [BS-P3] only if H is a non-decreasing submodular. It works as follows: Starting from the empty set $\mathcal{A} = \emptyset$, it iteratively adds the element with the highest value of metric $(H(\mathcal{A}\cup i) - H(\mathcal{A}))/c(i)$ while the total cost is within the budget C. Mathematically, it has been shown in [68], [69] that this greedy heuristic can give a suboptimal solution with an approximation factor of $(1-1/e)$.

Though it is quite difficult to prove the submodularity of H in general cases, it is indeed possible under some reasonable assumptions. We first assume that all BSs have the same q_i and P_i values for mathematical simplicity and ignore the penalty function $L_i(\rho_i)$ artificially introduced earlier. Then, the user association given in (3.17) becomes

$$i^*(x) = \underset{i\in\mathcal{B}_{on}}{argmax} \quad c_i(x) \tag{3.23}$$

where the decision is purely based on the transmission rate (or SINR).We further make an assumption of marginal interference as follows.

Assumption 1. *Adding (resp. removing) one BS has marginal impact on the total amount of interference. In other words, the increment (resp. decrement) of interference is almost negligible to users.*

Theorem 3.5. *Under Assumption 1, a set function $H(\mathcal{A})$ is nondecreasing and submodular.*

Proof. By the definition of $H(\mathcal{A})$, the other terms not having ρ_i can be ignored since they are either constant or irrelevant to the set $\mathcal{A}$. Hence, the proof of Theorem 3.5 is equivalent to proving the following two statements.

1. $\sum_{i \in \mathcal{B}_{on}} \rho_i$ is montonically decreasing as $\mathcal{B}_{on}$ increases.

2. $\sum_{i \in \mathcal{B}_{on}} \rho_i$ is supermodular as a function of $\mathcal{B}_{on}$.

Please refer to Appendix B for the full proof. □

The implication of Theorem 3 is that the greedy heuristic mentioned earlier would also work well to solve our active BS selection problem. After some tweaks to suit the problem [BS-P1] better, we propose the following active BS selection algorithm that borrows the metric (i.e., the decrement per unit cost when removing BS i) from the greedy heuristic.

Our proposed algorithm starts from the point where all BSs are turned on and finds the best BS candidate which will yield the maximum energy savings when turned off. Note that the denominator *is the amount of static power consumption saving* from turning off BS i. On the other hand, the numerator *is the increment of dynamic power consumption*, which comes from the fact that UEs originally associated with the switched-off BS would see possibly lower transmission rate $c_i(x)$ due to father distance to the

Table 3.1. Active BS selection algorithm

```
1. Initialize $\mathcal{B}_{on} = \mathcal{B}$
2: Repeat:
3:        Find $i^* = \underset{i\in\mathcal{B}_{on}}{argmin}\frac{UA(\mathcal{B}_{on}\backslash\{i\})-UA(\mathcal{B}_{on})}{q_iP_i}$
4:        $UA(\mathcal{B}_{on}\setminus\{i\})-UA(\mathcal{B}_{on}) < q_iP_i$, then $\mathcal{B}_{on}\rightarrow\mathcal{B}_{on}-\{i^*\}$
5:        Else, go to Finish
6: Finish: $\mathcal{B}_{on}$ is the set of active BSs
```

new serving BS. In line 4, the algorithm checks whether there is a net energy saving (in other words, the decrement in static power consumption is larger than the increment in dynamic power consumption). If so, we shut off BS i and repeat the loop. Otherwise, we stop the algorithm.

3.3.3 Transmit Power Budget Adaptation

After the active BS selection finds and turns on the minimum number of BSs (operating at their maximum transmit power $p = p^{max}$), there is still room for further energy reduction. There may be a scenario where some of active BSs has light traffic load (i.e., clearly $\rho_i < \rho_{th}$), but it is not possible[2] to turn off any of those BSs since reducing the set of active BSs will lead to QoS violation. In this section, we will discuss the last DCR technique, i.e., transmit power budget adaptation, which is a finer level tuning than the coarse BS on/off control. Given $\mathcal{B}_{on}$, we decompose the original problem into the intra-cell problem as follows, in which each BS locally controls the transmit power based on its own traffic load.

$$[TX-P]\ \min_{p_t\le p_t^{max}} T_i = (1-q_i)\rho_iP_i + q_iP_i \tag{3.24}$$

$$\text{Subject to:}\rho_i \le \rho_{th} \tag{3.25}$$

[2] If possible, it should have been done in the stage of active BS selection.

Plugging (3.9) into (3.24), the total BS operational power consumption can be rewritten as:

$$T_i = (a_i p_i + b_i)[(1 - q_i)\rho + q_i] \tag{3.26}$$

There are a couple of important observations from (3.26). Looking at the term inside the first parentheses (a linear relationship with p_i), we can notice that reducing p_i will have positive impact towards energy savings. However, on the other hand, it has negative impact in the term inside the second parentheses because the BS utilization ρ_i will increase due to the reduced transmission rate (see (3.1) and (3.2) for the definitions). Thus, it should be mentioned that it is not always beneficial to keep reducing the transmit power. In addition, there will exist the minimum transmit power level to meet the constraint (3.25).

We shall start by deriving how much the transmit power budget each BS can reduce providing that the interference from other BSs are fixed. Later we will relax this fixed interference condition in our final algorithm. In general, since $c_i(x)$ is a concave function of transmit power $p_i(x)$, the following equality holds:

$$\rho_i(p_a) \leq \rho_i(p_b) \cdot \frac{p_b}{p_a} \quad \text{for any} \quad p_a \leq p_b \tag{3.27}$$

Note that equality holds if $c_i(x)$ is a linear function of $p_i(x)$ (i.e., low SINR regime), which has been assumed to derive rate/power control algorithms in some references [70].

After we substitute $p_a \longrightarrow p_i$ satisfying $\rho_i(p_i) = \rho_{th}$ and $p_b \longrightarrow p_i^{max}$ into (3.27), we have the following minimum transmit power level $\underline{p}$ to meet the constraint (3.25).

$$p_i \geq \frac{\rho_i(p_i^{max})}{\rho_{th}} p_i^{max} = \underline{p} \tag{3.28}$$

Now we find the optimal transmit power level to minimize the total BS operational power consumption. As discussed earlier, reducing the transmit power is not always beneficial. We will see shortly that the total power consumption is upper bounded by a convex function of the transmit power, so there exists a minimizer. Applying the inequality (3.27) again to the total power consumption (3.26), we have

$$\begin{aligned} T_i \leq & a_i(1-q_i)\rho_i(p_i^{max})p_i^{max} + b_i q_i + a_i q_i p_i \\ & + b_i(1-q_i)\rho_i(p_i^{max})p_i^{max}/p_i \end{aligned} \tag{3.29}$$

The right-hand side of inequality is convex because it is the weighted summation of an affine function of p_i and another convex function $1/p_i$. This can be also confirmed by its second order with respect to p_i, i.e., $2b_i(1-q_i)\rho_i(p_i^{max})p_i^{max}p_i^{-3} \geq 0$. Thus, a minimizer $\hat{p}$ is given by

$$\hat{p} = \sqrt{\frac{b_i(1-q_i)\rho_i(p_i^{max})p_i^{max}}{a_i q_i}} \tag{3.30}$$

Together with the maximum transmit power p_i^{max} and the minimum transmit power in (3.28), we can obtain a suboptimal solution p^* (optimal when the equality holds in (3.27)).

$$p^* = \min[\max[\hat{p}, \underline{p}], p_i^{max}] \tag{3.31}$$

Remark 2. *When $q = 1$ (constant BS power consumption), the solution becomes $p^* = \underline{p}$. On the other extreme case of $q = 0$ (energy-proportional BS), the solution is $p^* = p_i^{max}$, which implies that no power adaptation is required.*

So far we have considered one-shot power adjustment starting from p_i^{max}. If each

Table 3.2. Transmit power budget adaptation algorithm

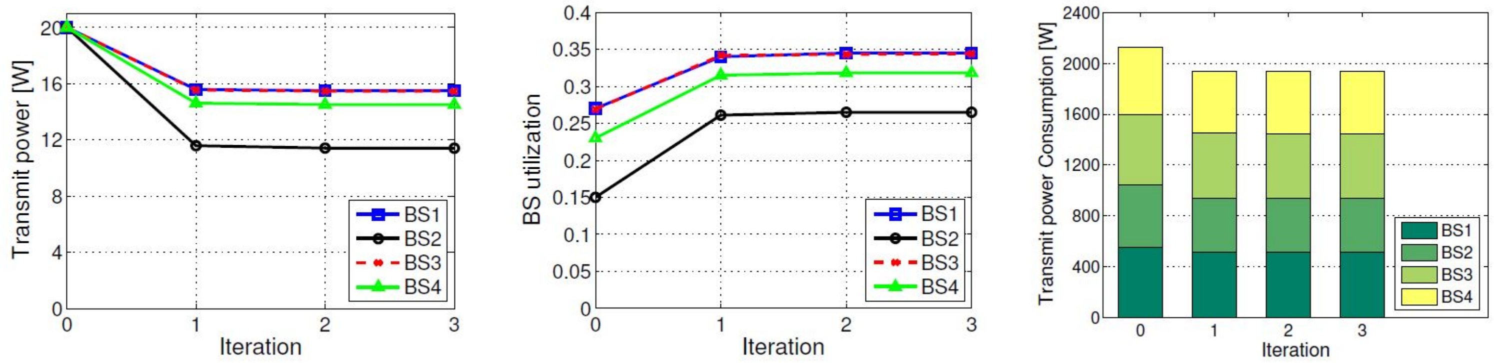
```
1. Initialize $k = 0$ and $p_i[0] = p_i^{max}$
2: Repeat:
3:        Update the interference from neighboring BS $j \neq i$
4:        $\overline{p} \longrightarrow p_i[k], \underline{p} \longrightarrow \frac{\rho_i(p)}{\rho^{th}} \cdot \overline{p}$ and $\hat{p} \longrightarrow \sqrt{\frac{b_i(1-q_i)\rho_i(\overline{p})\overline{p}}{a_i q_i}}$
            $p_i[k+1] \longrightarrow \min[\max[\hat{p}, \underline{p}, \overline{p}]$
5:        If $| T_i[k] - T_i[k+1] | > \epsilon, k \longrightarrow k+1$
6:        Else, go to Finish
7: Finish: $p_i[k+1]$ is a suboptimal transmit power budget
```

BS reduces its transmit power, then the users will experience different SINR. They will usually see higher SINR due to reduced interference from neighboring BSs, but it is also possible to see lower SINR depending on the power reduction ratio between the home BS (the users are associated with) and the other BSs.

This offers an opportunity to further adjust the transmit power. In other words, the power adaptation needs to be iteratively carried out with the updated interference till there is no further savings in terms of the total power consumption. This is the basic principle of our transmit power budget adaptation algorithm.

Our algorithm is shown in Table 3.2 works as follows. Starting from its maximum transmit power (step 1), each active BS i adjusts its transmit power based on $\overline{p}$, $\underline{p}$, and $\hat{p}$ (step 4). If the reduction of total operational power consumption in this iteration is greater than a small constant $\epsilon > 0$ (step 5), then the BS i updates interference from other BSs based on $p_j[k+1]$ for $j \neq i$ and repeat the loop. Otherwise, the transmit power budget adaptation algorithm stops (step 6).

In Fig. 3.1, we provide an example to illustrate how the proposed algorithm adapts the transmit power in a network topology of 4.5x4.5km^2 (see Section 3.4 for detailed parameter settings and Fig. 3.3 for the layout of BSs). For this particular example, we have considered that only four BSs are active, each of which has the maximum transmit



(a) Transmit power

(b) BS utilization

(c) Total power consumption

Figure 3.1. Variation of (a) transmit power, (b) utilization and (c) total BS power when transmit power is adapted in an iterative manner

power $p_i^{max} = 20W$. Figs. 3.1a and b show the transmit power adaptation of the BSs and their utilizations, respectively, and Fig. 3.1c shows the resulting total BS operational power consumption. After the first iteration, the transmit power of the least utilized BS 2 is reduced by about 8W whereas BSs 3 and 4 reduce the power about 4W. The reduction of the transmit power naturally leads to the increase in the BS utilization, however, it is still a way lower than our threshold $\rho_{th} = 0.7$. The changes are nominal in subsequent iterations 2 or 3, and the algorithm exits in the next iteration since there is no further saving in total power consumption.

Based on our empirical data, we would like to highlight that the transmit power budget adaptation algorithm converges quickly. Even in different configurations (with a different number of active BSs), we could observe similar a convergence trend, e.g., typically within a few iterations. As can be seen in Fig. 3.1c, transmit power budget adaptation brings about 10% of the total power savings (from 2,131 W to 1,936 W) by the fine-tuning of transmit power.

The iterative transmit power adaptation algorithm performs well when the traffic load is fixed or decreases over time as active BSs keep reducing the transmit powers until the convergence. However, in general scenarios where there exists a mixture of load-increasing and decreasing BSs with respect to time, some BSs have to increase their transmit powers. As a result, it would bring more interference to users in other cells, which also makes neighboring BSs increase the power to meet the QoS requirement of users. This can lead to oscillatory behavior and pose a technical challenge in terms of convergence. Additionally, since the transmit power adaptation might not suffice to cater for the temporal/spatial-varying load, balancing the traffic load via changing the user association and/or turning on additional BSs (or a different set of active BSs) may be required.

3.3.4 Integrated Approach: Dynamic Cell Reconfiguration (DCR)

So far we have developed three pieces of energy saving techniques (i.e., active BS selection, transmit power budget adaptation and user association) in a static scenario. To tackle the challenge of time-varying traffic with a mixture of load-increasing and decreasing BSs mentioned above, this section presents an integrated DCR framework. This framework jointly optimizes all of our techniques developed so far in a systematic way towards a single goal, i.e., energy savings while ensuring that the QoS requirements of all users are met. In the proposed DCR, three techniques with different control time scales interact with each other as follows. Please see the flowchart in Fig. 3.2 for a pictorial description.

The active BS selection algorithm described in Section 3.3.2 periodically (every T_p time units, e.g., half hour in our simulations[3]) determines a minimal set of BSs to remain active and turns off the other BSs, followed by the user association update. For each active BS $i \in \mathcal{B}_{on}$, if ρ_i does not exceed ρ_{th}, the transmit power budget adaptation described in subsection 3.3.3 can play a role in reducing further power consumption. On a much faster time scale than the active BS selection, the BS adapts its transmit power according to the current BS utilization ρ_i. We would like to mention that the transmit power adaptation is carried out in a manner transparent to the users, in other words, it does not change user association unless the BS utilization reaches ρ_{th}. In this way, unnecessary handover can be avoided.

As time goes on, BS i may experience high cell loading $\rho_i \geq \rho_{th}$ due to the increased traffic. In this case, the transmit power is immediately reset to p_i^{max}, i.e., a fast fallback to continue guaranteeing the QoS. After increasing the transmit power

[3] Many measurement studies (e.g., [48]) reported that the traffic load clearly varies over time (as well as space) but could be assumed almost constant during a certain period of time, e.g., typically one hour. Since the time scale for determining the set of active BSs would be similar to the order of traffic changing, its period is set at half hour in our DCR framework.

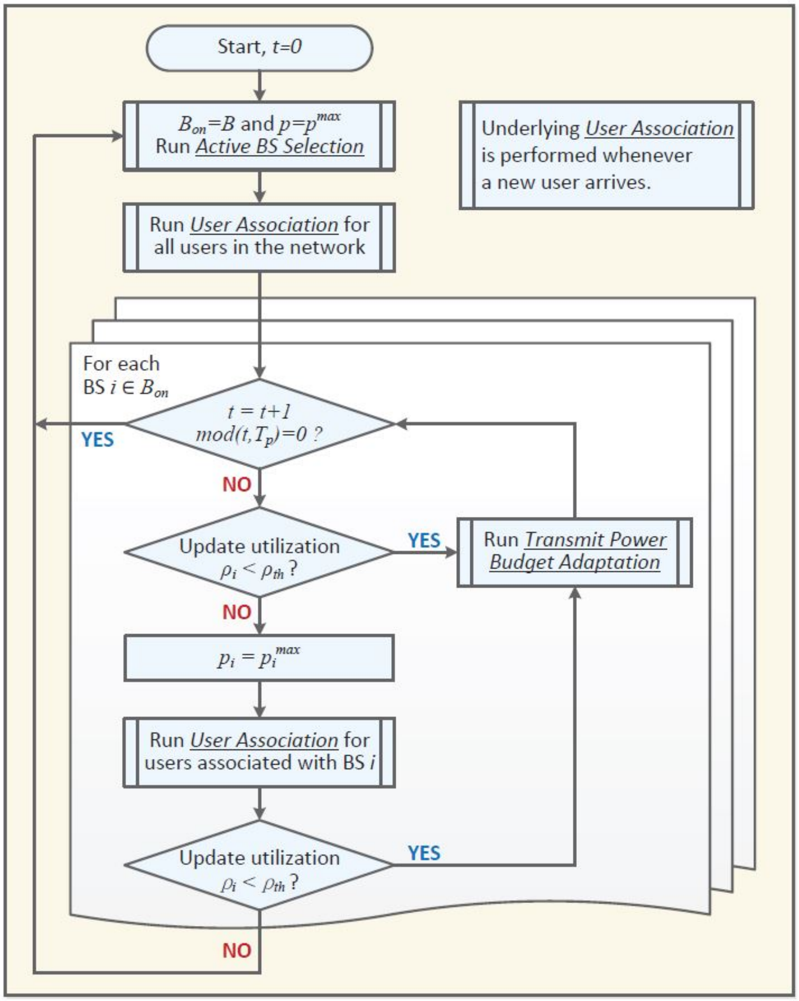

Start, t=0
B_{on}=B and p=p^{max}
Run Active BS Selection
Underlying User Association is performed whenever a new user arrives.
Run User Association for all users in the network
For each BS $i \in B_{on}$
$t = t+1$
$mod(t,T_p)=0$?
YES
NO
Update utilization $\rho_i < \rho_{th}$?
YES
Run Transmit Power Budget Adaptation
NO
$p_i = p_i^{max}$
Run User Association for users associated with BS i
Update utilization $\rho_i < \rho_{th}$?
YES
NO

Figure 3.2. Flowchart of the integrated DCR framework

to its maximum, if the BS utilization ρ_i goes down below the threshold ρ_{th}, then the power adaptation can now be carried out. Otherwise (even the maximum transmit power cannot lower the utilization enough), we recall the active BS selection algorithm to find a different set of active BSs to be switched on. To this end, we reset $\mathcal{B}_{on} = \mathcal{B}$ to consider the entire BS set as candidate active BSs, which allows us to have a wider choice of selection and may lead to a more energy-efficient solution. Lastly, there is an underlying user association algorithm, which is performed whenever a new user arrives to the network.

3.3.5 Discussion on the implementation of DCR

Complexity: It is worth analyzing the applicability of the proposed framework in terms of computational complexity. In particular, we concentrate on the active BS selection since it is relatively more complex than the other two techniques. Given the number of candidate BSs $\mid \mathcal{B} \mid$, there are $\binom{|\mathcal{B}|}{n}$ possible combinations to choose n active BSs. The total complexity of an optimal algorithm that finds the best set of BSs through exhaustive search is $\sum_{n=1}^{|\mathcal{B}|} \binom{|\mathcal{B}|}{n}$, which grows exponentially with the number of BSs, i.e., $O(2^{|\mathcal{B}|})$. On the other hand, however, the proposed active BS selection algorithm only requires $O(\mid \mathcal{B} \mid^2)$ (i.e., see the pseudo code in subsection 3.3.2: linear complexity in the line 4x the number of iterations at most $\mid \mathcal{B} \mid$), which makes it much easier to be implemented in practice. This linear complexity is because the proposed algorithm turns off the BS with a given metric one by one until there is a net energy saving.

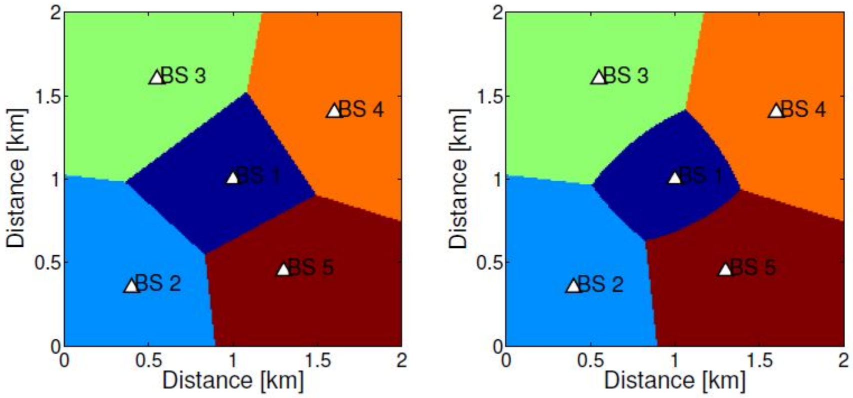
Our framework assumes a centralized network controller for running a centralized piece of DCR framework, i.e., active BS selection. Such a centralized controller can be radio network controller (RNC) in the 3G universal mobile telecommunications system (UMTS) access network or mobility management entity (MME) in the 4G long term evolution (LTE) access network. Each RNC or MME, running one instance of the active BS selection, is responsible for controlling the BSs (nodeB in UMTS or enodeB in LTE) that are connected to it. In practical systems, the typical number of BSs connected to the RNC or MME is a couple of dozen. This would give an idea of how much complexity reduction our algorithm offers. For instance, $O(2^{30} \approx 10^9)$ vs. $O(900)$ when $\mid \mathcal{B} \mid = 30$.

Group handover: When a BS is turned off for energy-saving purpose, UEs served by the BS need to be transferred to one of its neighboring BSs according to the user association algorithm presented in Section 3.3. This procedure is nothing new compared to the conventional handover except the fact that many UEs should be handed over simultaneously which implies a lot of control signalling.

There have been some studies done on the group handover [71], originally targeted to support passengers on mass transportation such as buses or trains. If this type of technique is used together with our framework, then, it would help reducing the possible performance degradation due to excessive control overhead.

3.4 Simulation Results

We evaluate the performance of the proposed DCR framework though simulations. Typical maximum transmit power for macro BSs and their maximum operational power are considered to be $p_i^{max} = 20W$ and $P_i^{max} = 865W$ (with the coefficients $a_i = 22.7$ and $b_i = 411$) according to [58], respectively. The static power portion q_i is assumed to be 0.5, but we will examine the effect of varying this parameter in subsection 3.4.4. In generating the user traffic, all intersection points on a rectangular grid with 30 m in the network are considered as a set of candidate locations for the user arrival. Each user arrives at location x following a Poisson point process with arrival rate $\lambda(x)$ and generates one file request with mean $1/\mu(x) = 100$Kbyte. We vary the traffic demand $\gamma(x) = \lambda(x)/\mu(x)$ [bits] by changing its arrival rate $\lambda(x)$. Other parameters including channel modeling for the simulations follow the urban macro model as presented in the 3GPP technical report [72].



(a) Without penalty ($\rho_{\text{th}} = 1$) (b) With penalty ($\rho_{\text{th}} = 0.5$)

Figure 3.3. Snapshots of coverage: the maximum penalty $L_{max} = \sum_{i \subseteq P_i^{max}}$ and sharpness of the penalty function $\beta = 2$: (a) Without penalty ($\rho_{th} = 1$) and (b) with penalty ($\rho_{th} = 0.5$)

3.4.1 Load Balancing via Penalty-based User Association

We shall start by demonstrating the effectiveness of the proposed user association algorithm. A simple network composed of five active BSs in 2x2km^2 and the spatially heterogeneous traffic load are considered, i.e., the required rate $\gamma(x) \propto (max(r) - r)^5$ where r is the distance from the center. So the area in the center, mostly covered by BS 1, can be interpreted as hotspot. In order to see how the proposed user association algorithm balances traffic loads, we plot Fig. 3.3 illustrating snapshots of BSs' coverage areas for the cases (a) without and (b) with penalty function. We can easily notice the effect of introducing the penalty function L_i into the reconfiguration algorithm by comparing the two figures. With penalty, some users leave the congested BS 1, as indicated by the shrinking of cell 1 in Fig. 3.3, and associate with neighboring BSs 2-5, which are actually under-utilized.

Such a load balancing comes at the cost of slight increase in dynamic power consumption. In order to show this tradeoff, we manually calculate the delay performance

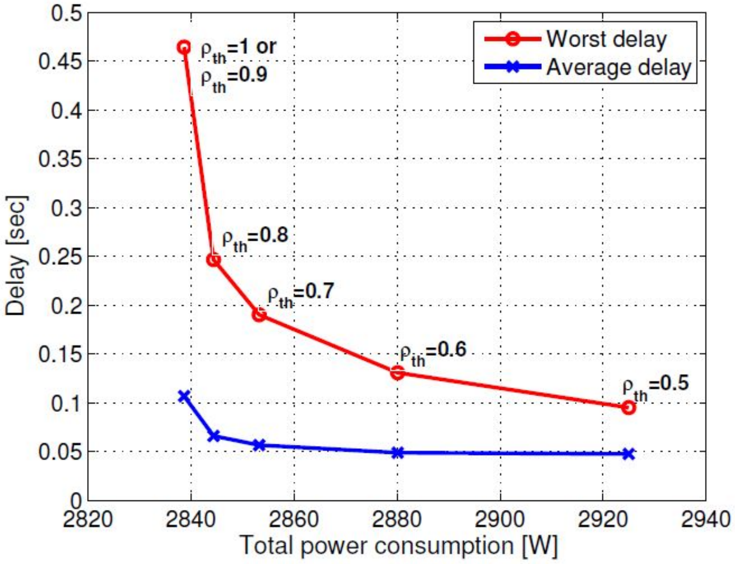


Figure 3.4. Tradeoff between delay and total power consumption by varying the BS utilization threshold ρ_{th} from 0.5 to 1.0

as a yardstick of load balancing by assuming $M/M/1PS$ queue [4] In Fig. 3.4, the average delay is the average performance of these five cells and the worst delay is the highest delay among five cells (usually, happens in the hot spot cell covered by BS 1). The less delay means the less congestion (i.e., the more effective load balancing). As shown in Fig. 3.4, the power cost is marginal compared to the delay benefit we can expect. For example, in the case of $\rho_{th} = 0.7$, there are 39% and 47% reductions in the average and worst delay, with 0.56% (2,838 W to 2,854 W) increase in power consumption. Note that this tradeoff graph may also be used to choose ρ_{th} in practice based on the maximum tolerable delay. In the rest of simulation study, we set $\rho_{th} = 0.7$ as it gives the most of benefits from load balancing with minimal power cost.

[4] Under $M/M/1$ processor sharing (PS) queue, the expected number of flows in cell i is $\rho_i(1-\rho_i)$. With the help of Little's law, dividing it by the system arrival rate that is the integration over $\lambda(x)$ over its coverage area, we can obtain the expected per flow delay in the cell.

3.4.2 Power Savings Under Static Traffic Load Scenario

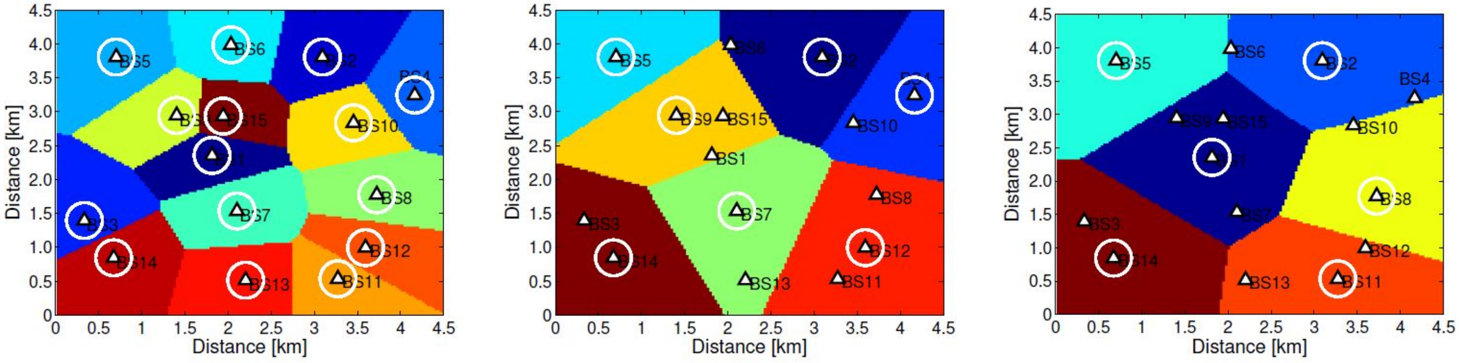
Effectiveness of active BS selection: Let us first investigate the performance of active BS selection together with the user association (i.e., UA-BS). To have more realistic results, a topology with fifteen BSs in 4.5x4.5km^2, a part of 3G network in metropolitan area [73], is adopted (see Fig. 3.5). For comparison, we also consider three other schemes:

- All-on (conventional scheme): always turning on all BSs.
- Util-based: turning off the least utilized BS one by one which is shown to be an effective heuristic in [65].
- Exhaustive: finding an optimal set of BSs through an exhaustive search.

Fig. 3.5 shows snapshots of the active BSs and their coverage areas at the normalized traffic load[5] = 0.3 for different schemes. All-on keeps all BSs turned on at such a low load, which naturally leads to energy inefficiency. However, the proposed active BS selection algorithm (with linear complexity) and exhaustive scheme (with exponential complexity) turn off eight and nine BSs for energy conservation, respectively. As a consequence, the remaining BSs dynamically reconfigure their cells (i.e., cell zooming).

In our simulations under various configurations, the proposed algorithm often finds a near-optimal solution that has the same number of active BSs as exhaustive and just one or two more in the worst case. It is also worthwhile investigating the static and dynamic power consumption breakdown: UA-BS ($4.46kW = 3.03kW + 1.43kW$) vs. exhaustive ($4.25kW = 2.60kW + 1.65kW$). US-BA consumes more static power

[5] In our simulation, the normalized traffic load [no unit] is the traffic load normalized by the traffic load at peak time.

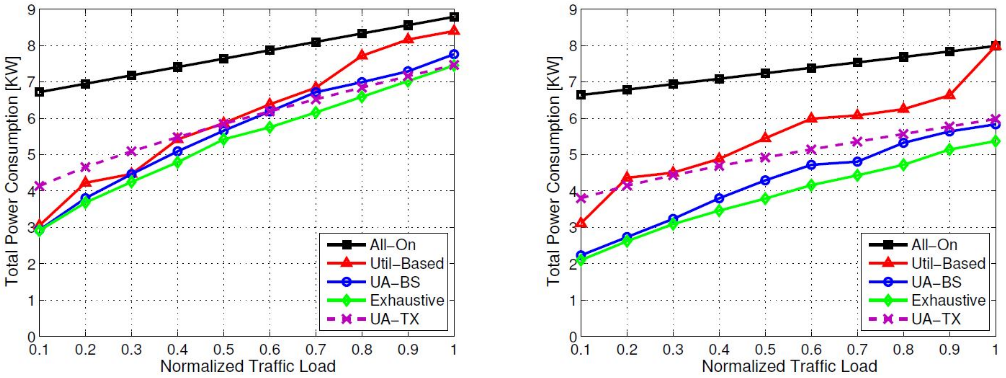


(a) All-On

(b) UA-BS

(c) Exhaustive

Figure 3.5. Snapshots of the active BSs and their coverage areas at the normalized traffic load = 0.3 for different schemes. The BSs with and without white circles are active and inactive BSs, respectively. Different colored regions represent the coverage areas of active BSs. Util-based is not shown here because it finds the same solution as the UA-BS does in this particular scenario: (a) All-on, (b) UA-BS, and (c) exhaustive.



(a) Uniform traffic distribution

(b) Non-uniform traffic distribution

Figure 3.6. Total power consumption with different schemes under static traffic load: (a) Uniform traffic distribution and (b) non-uniform traffic distribution

than exhaustive scheme due to the higher number of active BSs, while it consumes less dynamic power.

The total power consumption of the cellular network as a function of the static normalized traffic load in both (a) uniform and (b) non-uniform[6] traffic distribution is evaluated in Fig. 3.6. Our results show that a brute-force util-based works well in the uniform environment, but not in non-uniform environment. However, UA-BS always outperforms util-based, and moreover its performance is very close to that of the exhaustive search solution. Compared to the static All-on scheme, it yields the potential energy savings of 10–60% depending on the amount of traffic and its spatial distribution.

Further from Fig. 3.6, we can see that UA-BS can clearly reduce more power consumption compared to All-on and util-based in non-uniform environment than uniform environment. This is because the non-uniform environment has more spatial variations (e.g., extremely under-utilized BSs and high-utilized BSs in different areas at the same time), which allows the active BS selection algorithm to turn more BSs off in sparse areas, as opposed to the environment where all BSs have a similar level of utilization.

[6] A linearly decreasing traffic along the diagonal direction from left top to right bottom in Fig. 3.5 is considered to generate non-uniform environment.

Effectiveness of transmit power budget adaptation: We will now validate the performance of another piece of our energy-saving techniques, which is the transmit power budget adaptation coupled with user association (i.e., UA-TX). In order to see the pure benefit of adapting transmit power, we do not consider turning off BSs here, but the other simulation environment remains the same. In Fig. 3.6, the dotted line shows the total power consumption with the transmit power budget adaptation. As can be seen, compared to All-on operating at maximum transmit power, the proposed algorithm can reduce the total power consumption by 1.37 kW (when normalized traffic load = 1) ≈ 2.53 kW (when normalized traffic load = 0.1) under uniform environment, and by 1.95 kW≈ 2.79 kW under non-uniform environment, respectively.

We will next the compare the two proposed algorithms, UA-TX and UA-BS. Under uniform environment, we see that the performance gap between the two schemes is much lower than that under non-uniform load conditions. This is due to the difference of utilization levels under uniform and non-uniform traffic distributions, which allows less or more opportunity for UA-BS to switch off BSs as explained in the previous subsection. Above a certain traffic load condition, it is not easy for UA-TX to get more savings. This is mainly because we cannot turn off a BS unless we can ensure its traffic to be transferred to neighboring BSs. On the other hand, however, UA-TX can get some savings (even if little) as long as there is any unused power budget. This explains why UA-TX outperforms or performs comparably to UA-BS as the load increases.

We will conclude the comparison with a note on the overheads introduced by switching off BSs in UA-BS and transmit power adaptation in UA-TX. The user association adaptation is common to both the algorithms and overhead due to user handover is discussed in subsection 3.3.5. Another overhead is the exchange of message/control information. UA-BS primarily requires utilization information whereas UA-TX mainly relies on interference information from neighboring BSs. The amount of information

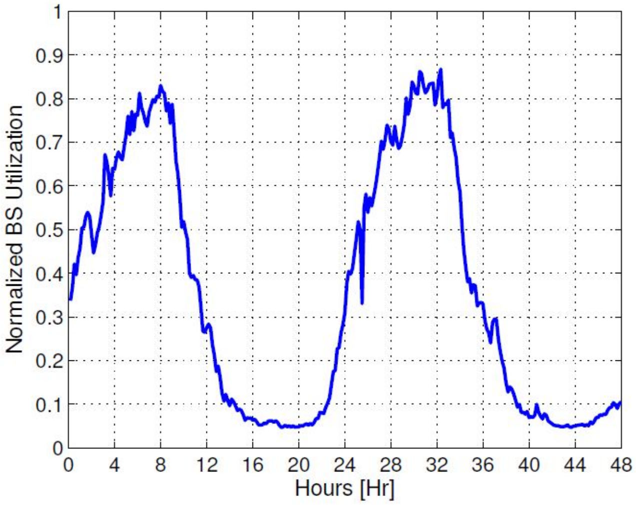


Figure 3.7. A sample real-traffic trace during 48 hours

would not be a big deal and may be considered to be marginal compared to the large volume of data traffic. However, the frequency of information exchange could be costly, especially in UA-TX, as transmit power adaptation is carried out at much faster time scales (e.g., an order of minutes) than switching off BSs (e.g., an order of hour or so). In order to implement the algorithm in practice, more attention needs to be devoted to this kind of overhead problem.

3.4.3 DCR Framework Under Dynamic Traffic Load Scenario

In this section, we will discuss the performance of the whole DCR framework including user association, active BS selection and transmit power budget adaptation. In order to have more realistic results and at the same time to examine the potential savings in response to time-varying load, we adopt a sample traffic trace [48] shown in Fig. 3.7. The trace, originally obtained from an anonymous cellular operator, gives the variation of BS utilization with a temporal granularity of 10 min across 48 h in a metropolitan area. The other simulation settings, such as the network topology, channel and power consumption modeling, are exactly the same as the ones used in previous

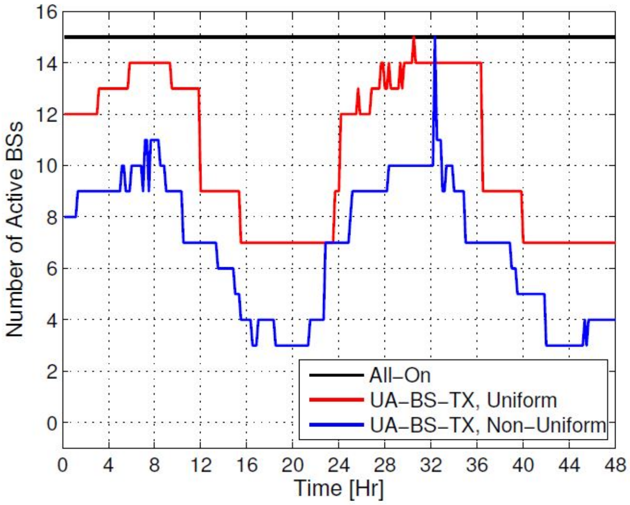
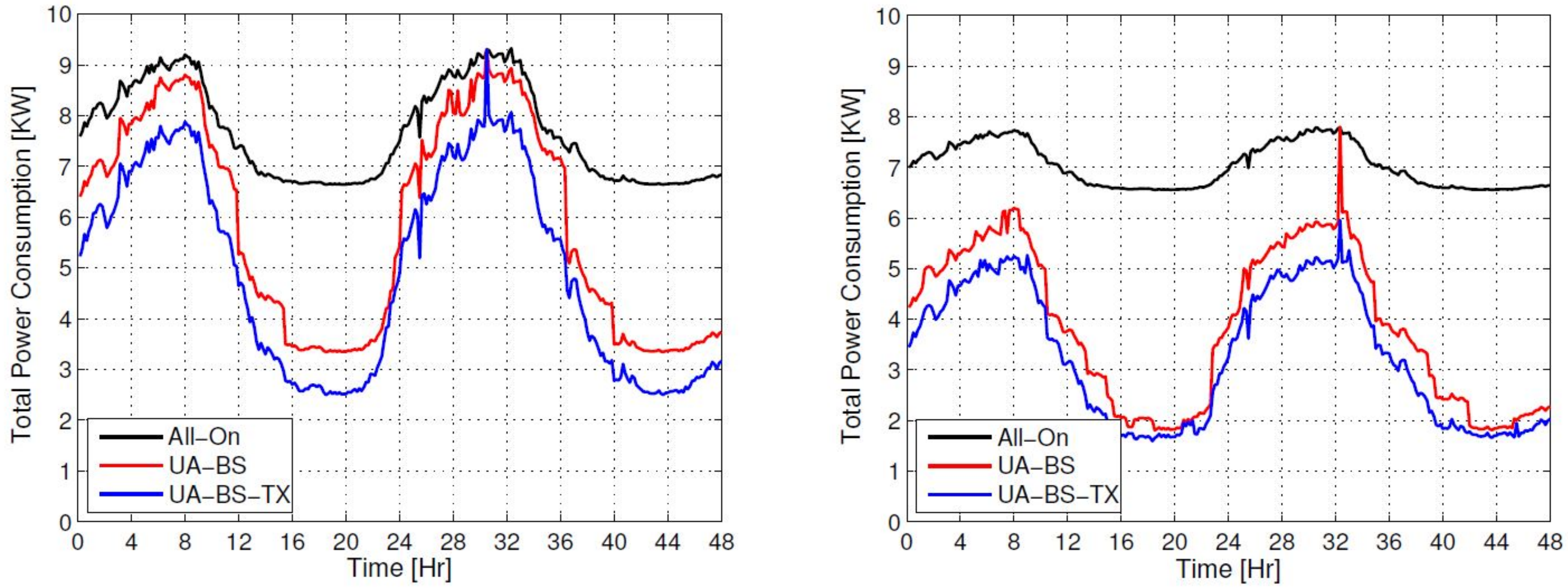


Figure 3.8. Number of active BSs

sections. For performance comparison, we consider All-on as a baseline, and we compare its performance with (i) active BS selection technique used in conjunction with user association (i.e., UA-BS) and (ii) the integrated DCR framework including all algorithms we have proposed so far (i.e., UA-BS-TX).

Fig. 3.8 shows the number of active BSs selected by UA-BS-TX under uniform and non-uniform environment. For reference, we also plot the number of active BSs for All-on, which is always equal to the total number of BSs. As the transmit power adaptation is carried out based on $\mathcal{B}_{on}$ given by active BS selection, UA-BS-TX and UA-BS have the identical number of active BSs. Therefore, we illustrate only the result of UA-BS-TX here. As can be seen in Fig. 3.8, there is room to turn off some BSs most of the time except at peak time. For example, under the uniform (resp. non-uniform) traffic distribution, up to 8 (resp. 12) BSs can be turned off during low traffic periods for energy conservation by the proposed UA-BS-TX. This is in contrast to the energy-inefficient scheme, all-on, which turns on all the 15 BSs at all times irrespective of the distribution and the amount of load.

Fig. 3.9 shows the total power consumption of the cellular network in response



(a) Uniform traffic distribution

(b) Non-uniform traffic distribution

Figure 3.9. Total power consumption with different schemes under dynamic traffic load: (a) Uniform traffic distribution and (b) non-uniform traffic distribution

Table 3.3a. Total energy use of different schemes and energy savings compared to All-on scheme in different hours for the uniform environment

	Overall	Peak	Off-Peak
All-On	367.80 kWh	143.98 kWh	109.59 kWh
UA-BS	281.63 kWh (23.43%)	134.26 kWh (6.75%)	59.29 kWh (45.90%)
UA-BS-TX	238.09 kWh (35.27%)	118.95 kWh (17.38%)	45.59 kWh (58.40%)

Table 3.3b. Total energy use of different schemes and energy savings compared to All-on scheme in different hours for the non-uniform environment

	Overall	Peak	Off-Peak
All-On	337.37 kWh	123.47 kWh	107.63 kWh
UA-BS	185.06 kWh (45.15%)	91.56 kWh (25.84%)	36.04 kWh (66.51%)
UA-BS-TX	157.97 kWh (53.18%)	79.56 kWh (35.56%)	30.16 kWh (71.98%)

to the time varying load. As expected, All-on scheme consumes the highest total power for all load conditions. UA-BS achieves lower total power consumption than All-on as it switches on/off BSs dynamically depending on the traffic load. However, since it operates at $p = p^{max}$, its total power consumption is higher than that of UA-BS-TX, which additionally adapts the transmit power as well.

Tables 3.3a and 3.3b summarizes the total energy use of different schemes in different hours: Overall (during 48 h), peak times (2-10 and 26-34 h) and non-peak times (14-22 and 38-46 h). The numbers in parentheses represent the percentage of energy savings compared to All-on scheme. In overall, UA-BS-TX can provide a significant amount of energy savings, e.g., 35.27%and 53.18% under uniform and non-uniform traffic distribution. The energy savings are mostly obtained from turning some BSs off and the transmit power adaptation contributes to about 10% extra savings. More savings under the non-uniform environment is due to less number of active BSs than the uniform environment as shown in Fig. 3.8.

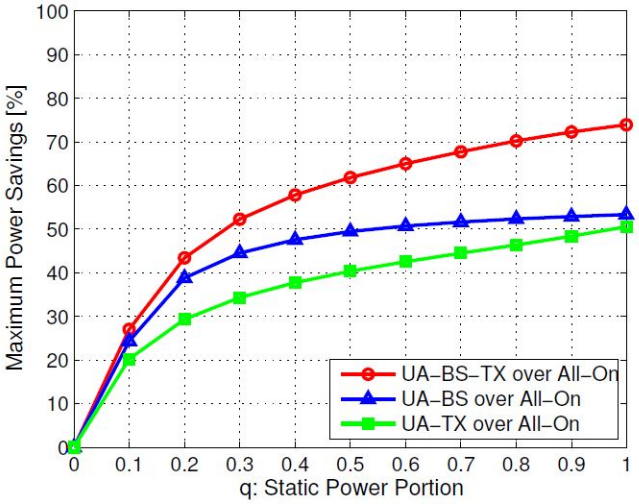


Figure 3.10. Effect of static power portion q_i on maximum energy savings

It is also worthwhile mentioning that the transmit power control relatively becomes dominant at peak times in terms of the percentage of the energy savings. For instance, in the case of uniform traffic distribution, more than half of the energy savings comes from adapting the transmit power.

3.4.4 Effect of the Portion of Static Power Consumption q_i

Fig. 3.10 illustrates the effect of varying the static power consumption weight q_i on maximum power savings possible (at a low load ≈ 0.2) over All-on. As expected, there is no gain at $q_i = 0$ because energy-proportional BSs have no standby power dissipation. The savings achieved by UA-BS is always better than UA-TX but the performance gap decreases as the contribution of static power increases (i.e., q_i value increases). For example, at $q_i = 1$, nearly 50% savings for either UA-BS or UA-TX are possible while the integrated DCR framework including all three energy saving techniques (i.e., UA-BS-TX) can obtain more than 70% savings. Given that current and near future BSs are operating in the high q_i range, the proposed DCR energy saving techniques would bring huge benefit to the cellular networks.

3.5 Summary

In this chapter, we proposed a novel dynamic cell reconfiguration framework for BS energy saving that encompasses active BS selection, transmit power budget adaptation and user association in cellular wireless networks. Through analytical and simulation studies, we demonstrate the effectiveness of our DCR framework. The proposed framework can achieve significant savings during periods of low traffic such as at night and provide considerable savings even at peak time. We also made an interesting observation that high savings are expected, especially, when the portion of static power consumption of BSs is high. The proposed framework brings many interesting research opportunities, for example, we are currently investigating the impacts presented by DCR on the cellular uplink.

Though the DCR techniques developed result in significant savings in cellular network power consumption, the underlying operation of BS on/off requires tens of minutes for completion and will not be able to respond to finer time scale variation in BS load. This could potentially lead to coverage holes and thereby degradation in user experience. We address this in the next chapter by developing dynamic RF chain switching techniques that minimize the power consumption of cellular networks while ensuring that there are no coverage holes in the cellular networks.

3.6 Acknowledgements

We thank the anonymous Journal of Communications and Networks (JCN): Special issue on Green Communication reviewers for their feedback and comments on the work.

Chapter 3, in part, contains material as it appears in Journal of Communications and Networks (JCN): Special issue on Green Communication. “Dynamic Cell Reconfig-

uration Framework for Energy Conservation in Cellular Wireless Networks". Kyuho Son, Ranjini Guruprasad, Santhosh Nagraj, Mahasweta Sarkar, Sujit Dey. The dissertation author was the primary student investigator and author of this paper.

Chapter 4

User QoS-aware Adaptive RF Chain Switching for Power Efficient Cooperative Base Stations

4.1 Introduction

By 2022, the expected number of mobile subscriptions and the resulting mobile traffic is expected to reach 8.9 billion subscriptions and 69 Ebytes respectively [74]. To cater to the explosive growth in mobile data subscriptions and traffic, it is estimated that the total number of base stations (BSs) in cellular networks all over the world will grow to 11.2 million by 2020 [3], a 47% increase compared to the number of BSs deployed in 2014. Further, deployment of massive number of antennas at BSs is seen as a promising paradigm to increase data rates [4]. This is expected to increase the electricity consumption and thereby, decrease the energy efficiency of cellular networks [4]. Specifically, the electricity consumption of BSs which constitutes 80% of electricity consumption of cellular networks is estimated to increase from 84TWh to 109TWh by 2020 (38% increase from 2014) if measures are not taken to reduce the power consumption of BSs. The increasing electricity consumption has two effects - (a) the carbon equivalent emissions is estimated to increase to 235 Mto CO_{2e} by 2020 (a 37% increase from 2014) [3] and (b) the electricity bill which currently contributes to 10-15%

of the operating expenses in developed markets and about 50% [5] in developing markets will further increase. Hence, increasing the power efficiency of base stations becomes a critical requirement to reduce growing operating cost for mobile operators and to comply with the trending global desire to reduce energy consumption and carbon footprint, and increase sustainability.

Amongst many components of the BS, the power amplifier (PA) in RF chain consumes about 65% [75] of the total power consumption in the BS. Further, multi-input multi-output (MIMO) BS providing high data rates and enhanced coverage uses multiple RF chains which increase the contribution of RF chain power consumption. Consequently, to reduce BS power consumption, it is vital to develop techniques that can lower RF chain power consumption.

The total power consumption due to RF chains is determined by the number of active RF chains, transmission power, transmission bandwidth and duration of transmission required to satisfy the Quality of Service (QoS) i.e., throughput and block error rate (BLER) requirements of the users. Given the user association (UA), there may exist multiple combinations of the above-mentioned BS resources that satisfy the users' QoS requirements and which result in varying levels of BS resource utilization and RF chain power consumption [76].

Moving from single BS to cluster of BSs which have overlapping coverage areas, there may be multiple users located in the coverage area of more than one BS. This implies that there may exist multiple combinations of UA across the cluster BSs which will satisfy the QoS requirements of all the users associated with the cluster BSs. Different combinations of UA can result in different BS resource utilizations and hence RF chain power consumption.

In this chapter, we propose a cooperative adaptive RF chain switching technique which explores the BS resource and UA spaces to maximize the number of RF chains

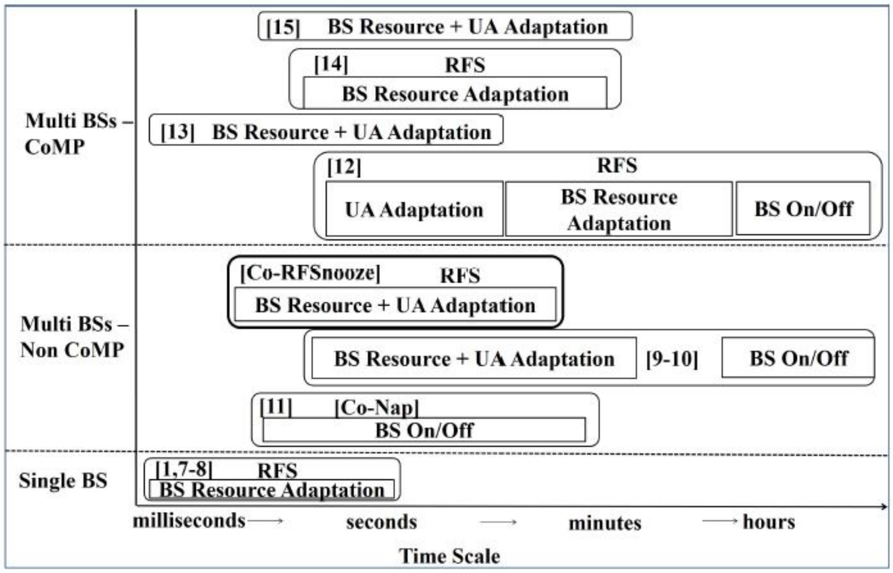


Figure 4.1. Comparison of related work with the proposed Co-RFSnooze technique

that can be switched off to minimize RF chain power consumption and thereby power consumption of the BSs in the cluster. While trying to adapt the BS resources and UA, the proposed technique ensures that individual BS utilization constraints are not violated and QoS requirements of all the users in the cluster are satisfied.

4.1.1 Related Work

In this section, we will briefly describe prior work related to BS resource and UA adaptation to achieve adaptive RF chain switching (RFS) and power efficient operation of cellular networks. The relevant techniques are grouped in to three categories based on (a) the number of BSs considered for applying the BS on/off, BS resource and UA adaptation techniques and (b) the use of coordinated multi-point (CoMP) transmission. Note that, though BS on/off switches RF chains, it is not adaptive as BS on/off either switches on or off all RF chains. Further, in each category, techniques are distinguished based on time scale of operation. We will refer to time scales of milliseconds to minutes as short time scale and tens of minutes to hours as long time scale. The above described grouping is shown in Fig. 4.1.

We will first discuss the techniques applicable to a single BS as shown in the bottom row of Fig. 4.1. The technique (termed Min-Cost in [76] and RFSnooze in this chapter) proposed in the preliminary version of this work [76] adapts the number of RF chains, time slots and frequency blocks while satisfying both the users' throughput and BLER requirements as well as BS utilization constraints. Authors in [77] propose data rate, power, RF chain and subcarrier allocation in a manner that maximizes the energy efficiency of data transmission of a single BS. The technique proposed in [78] jointly maximizes transmitter and receiver energy efficiency of a single BS and associated users. In contrast to the above single BS techniques, the proposed short time scale Co-RFSnooze technique is applicable to cluster of cooperating BSs. It extends [76] to jointly adapt the individual BS resources as well as the UA of all the cluster users (Section 4.3.4) to maximize the number of RF chains that can be switched off in the entire cluster and minimize the cluster power consumption. We will next discuss the techniques which are applicable to a cluster of cooperating BSs that do not use CoMP transmission (middle row, Fig. 4.1).

Dynamic BS on (active)/off (inactive) techniques switch BSs on or off based on number of associated users [79] and the estimated savings in power consumption due to switching off of BSs [80]. The above techniques switch off all the components of a BS which takes tens of minutes and can be classified as a long time scale operation. Though short time scale operations of BS resource and UA adaptation are applied to the subset of active BSs, long time scale switching off of BSs could potentially lead to coverage holes. Coverage holes are a major concern for the operators as a user in the coverage hole will not receive coverage. In contrast, our proposed approach adapts BS resources and UA on a short time scale enabling finer tracking of the BS load and finer control on BS power consumption without degrading coverage capabilities.

The Co-Nap technique proposed in [81] implements short time scale BS on/off by adapting the number of ”nap” (sleep) time slots for the cluster BSs in a coordinated manner. As all the BS RF chains are switched off in the ”nap” time slots, it reduces BS power consumption. Unlike the Co-Nap strategy which adapts only the on/off pattern of BSs, the proposed Co-RFSnooze technique jointly adapts BS resources and UA to achieve adaptive RFS. We will demonstrate in Section 4.4.2 that this joint adaptation achieves higher power efficiency compared to Co-Nap.

Next, we will discuss techniques that are applicable to cluster of cooperating BSs using CoMP transmission (top row, Fig. 4.1). The long time scale technique in [82] determines the BS and RF chain on/off pattern, UA and power allocation and the short time scale technique in [83] exploits the varying delay tolerance of users to enable time slot based BS sleep. The throughput requirements of the users associated with the inactive BS in [82]-[83] are met through CoMP transmission by the active BSs in the cluster. The authors in [84] propose a resource allocation algorithm for full-duplex, distributed antenna, multi-user communication network that minimizes the power consumption of cluster of BSs by dynamically switching off RF chains while satisfying the QoS requirements of downlink and uplink users. The above techniques require sharing of the channel state information (CSI) and data of all the users in the cluster via the backhaul to compute the multi-cell precoding matrix to perform CoMP transmission. The proposed Co-RFSnooze technique does not utilize CoMP transmission and instead proposes novel heuristics and combination of centralized-decentralized framework that requires sharing of only the user QoS and association information to significantly reduce the communication via the backhaul. As shown in Fig. 4.5b (Section 4.4.2), there are 270 users in the cluster during high load and the techniques [82]-[84] will require sharing CSI information and data of all the 270 users whereas the proposed technique requires user QoS and association information of only 35 users (users transferred shown in Fig.

4.6b).

The technique proposed in [85] determines the BS-user association for CoMP transmission and performs joint spectrum and power allocation to minimize the total cluster transmission power. However, [85] does not dynamically switch off RF chains and always maintains them in the on state. In contrast, the proposed Co-RFSnooze technique performs BS resource and UA adaptation to dynamically switch off RF chains in the cluster. This can potentially result in higher power savings compared to [85] which always switches on all the RF chains (demonstrated in Section 4.4.2 by significant savings compared to All-On/Co-Nap which switches on all RF chains)

From the above description of the prior art, to the best of our knowledge, this is the first work

- that dynamically switches RF chains in a cluster of cooperating BSs by jointly adapting BS resources and UA on a short time scale to minimize the average cluster power consumption in a transmission frame.

- that jointly adapts BS resources and cluster UA in a manner that the cluster user's QoS requirements and the BS resource utilization constraints are satisfied.

- that does not require BS switching and expensive CoMP data transfer and matrix computations to adaptively switch RF chains in a cluster of cooperating BSs.

The rest of the chapter is organized as follows. Table 4.1a summarizes the notations used. Section 4.2 describes the system model and the optimization problem. In Section 4.3, we propose a heuristic algorithm to solve the optimization problem. In Section 4.4, we provide simulation results under a practical configuration. Finally, we conclude the chapter in Section 4.5.

Table 4.1a. Summary of notations

$\mathcal{B}, BW$	Set of BSs in the network, Transmission bandwidth of BS $b \in \mathcal{B}$
S, R	Maximum number of RF chains at BS and user
P_b^{Tx}, P^{Max}	Transmit power and maximum transmit power of BS b
t^F	Duration of frame
T, T^A, T^I	Number of time slots in a frame, Number of active and idle time slots in a frame
t^O, t^{Sw}	Duration over which all RF chains are off in a frame, RF chain switching duration in a frame
S_t^A, S_t^O, S^{Sw}	Number of active and off RF chains in time slot t, Number of RF chains switching state in a frame
J, ψ_{st}	Number of frequency blocks in time slot $t \in T$, Frequency utilization of RF chain s in time slot t
m, M	Transmission mode and set of all transmission modes
$s_{ib}(m)$	Number of BS RF chains allocated by BS b to the i^{th} user for mode m
$r_{ib}(m)$	Number of RF chains allocated by i^{th} user associated with BS b for mode m
$d_{ib}(m)$	Number of independent data streams received by i^{th} user associated with BS b for mode m
$\gamma_i, BLER_i^{Th}$	Throughput requirement of i^{th} user, Upper bound on BLER requirement of i^{th} user
$\boldsymbol{H}_{ib}, SINR_{ib}$	Channel matrix between i^{th} user and BS b, Signal to interference noise received by i^{th} user from BS b
$TP_{ib}, BLER_{ib}$	Throughput provided by BS b to i^{th} user, BLER provided by BS b to i^{th} user

Table 4.1b. Summary of notations - continued

I_b	Set of users associated with BS b
I_b^{NT}, I_b^{T}	Set of non-transferable and transferable users associated with BS b
$I_b^{T\sim}$	Subset of I_b^T users associated with BS b that require the same set of RF chains and time slots as users I_b^{NT}
P^I, P^O, P^{Sw}	Idle and off power consumption of BS, PA switching power
Δ_p	Power gradient
P_b, P_C	Average power consumption of BS b in a frame, Average cluster power consumption in a frame
$C, \lvert C \rvert$	Set of cluster BSs and number of cluster BSs in cluster C
I_C, I_C^{NT}, I_C^{T}	Set of users in cluster C, Set of non-transferable and transferable users in cluster C
BSU, k_{bi}	BS-user matrix of size $\lvert C \rvert x \lvert I_C \rvert$, entry in BSU matrix of BS b for i^{th} user
E_i	Set of BSs that satisfy i^{th} user's mode SINR threshold
g, E	Transferor BS, set of transferee BSs
RFU	Number of active RF chains to users ratio

4.2 System Model and Problem Formulation

4.2.1 Network, Channel and User QoS Models

Consider the downlink communication in MIMO-Orthogonal Frequency Division Multiple Access (OFDMA) cellular network with set of BSs $\mathcal{B}$ as shown in Fig. 4.2. The overall bandwidth BW is divided in to J equally sized frequency blocks and the transmission frame of duration t^F is divided in to T equally spaced time slots, each of duration $\frac{t^F}{T}$. The maximum number of RF chains that can be active at BS $b \in \mathcal{B}$ and each user device are S and R respectively. We will define a transmission mode m as $m \triangleq (s(m), r(m), d(m))$ where $s(m) \in [1, S]$ is the number of BS RF chains required for mode m, $r(m) \in [1, R]$ is the number of RF chains required at the user device and $d(m) = min(s(m), r(m))$ is the number of independent data streams transmitted by mode m.

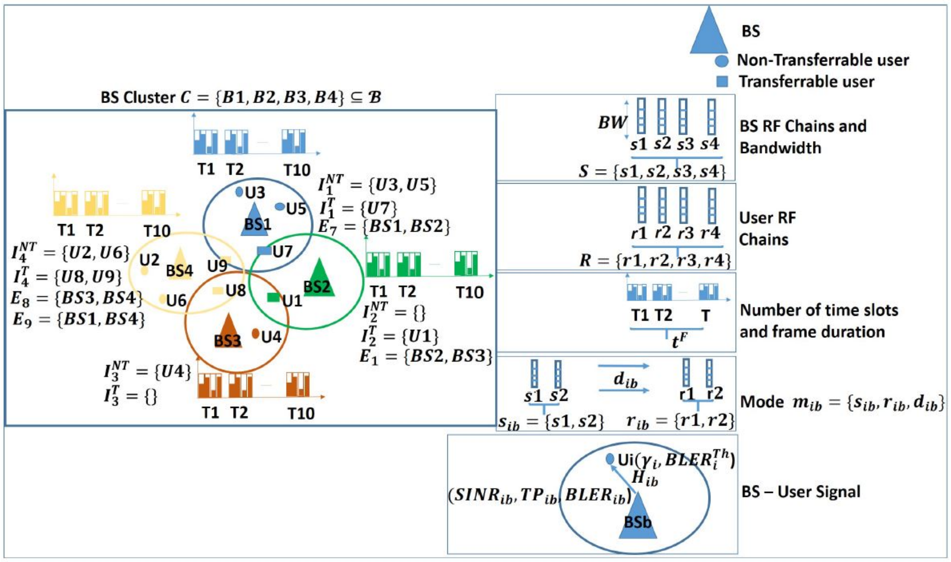


Figure 4.2. System block diagram

We assume single-input single-output (SISO) and Single User-MIMO (SU-MIMO) including spatial multiplexing (SM) and spatial diversity (SD) modes for transmission. We will denote the set of all possible transmission modes as M. In the work presented in this chapter, mode selection is done once every transmission frame and the mode $m_{ib} \in M$ selected for the i^{th} user by BS b does not change within time slots of a frame. Hence, the number of RF chains $s_{ib}(m)$ allocated by BS b to the i^{th} user, number of RF chains $r_{ib}(m)$ allocated by the i^{th} user device and the number of independent data streams $d_{ib}(m)$ received by the i^{th} user remains identical for all the active time slots of the frame.

Let I_b denote the set of users associated with BS b and $I_b^T \subseteq I_b$ denote the subset of 'transferable' users who are in coverage area of BSs $b^{\sim} \in \mathcal{B} \setminus b$ in addition to being in the coverage area of BS b. For cooperative RF chain switching, we propose to adapt the UA of such transferable users which lie in the coverage areas of multiple BSs. This motivates us to consider group or cluster of BSs $C \subseteq \mathcal{B}$ having overlapping areas of coverage enabling cooperation and UA adaptation. In the work presented in this chapter, we adopt the network centric clustering of BSs wherein BSs are grouped together statically based on network planning considerations [86]. Like used extensively in related research [81] and [87], we assume that the set $\mathcal{B}$ can be divided in to disjoint clusters of BSs and the size of each cluster is $|C|$ where $|X|$ denotes the cardinality of set X. We also assume that all the BSs in the cooperative cluster can communicate with each other via the X2 interface.

We assume block fading channel between BS b and the i^{th} user over the entire bandwidth (J frequency blocks) in a frame (T time slots) represented by the complex channel matrix $\boldsymbol{H_{ib}} \in C^{r_{ib}xs_{ib}}$ of rank $A \leq d_{ib}$. The noise at each user's receiver is assumed to be additive white Gaussian with zero mean and variance σ^2. We assume that the user's channel state information (CSI) including channel quality information (CQI) and Rank Indicator (RI) is available at the BS.

Assuming that the transmit power P_b^{Tx} of BS b is equally divided over all frequency blocks and transmit antennas, the signal to interference-noise ratio (SINR) received by the i^{th} user is

$$SINR_{ib} = \frac{P_b^{Tx}}{Js_{ib}} \cdot \frac{\mathbf{H}_{ib}\mathbf{H}_{ib}^H}{\sum_{b^{\sim} \in \mathcal{B} \backslash b} P_{b^{\sim}}^{Tx} \mathbf{H}_{ib^{\sim}} \mathbf{H}_{ib^{\sim}}^H + \sigma^2} \tag{4.1}$$

The throughput TP_{ib} from BS b to i^{th} user is given by

$$TP_{ib} = \frac{BW}{JT} \sum_{t=1}^{T_{ib}} J_{tib} \log_2[\det\{I_{r_{ib}} + SINR_{ib}\}] \tag{4.2}$$

where T_{ib} is the number of time slots and J_{tib} is the number of frequency blocks assigned in time slot $t \in [1, T_{ib}]$ by BS b to the i^{th} user and $I_{r_{ib}}$ is a $r_{ib}xr_{ib}$ identity matrix. The $BLER_{ib}$ achieved for the i^{th} user depends on the BS transmit power P_b^{Tx}, channel $\boldsymbol{H_{ib}}$, and the mode m_{ib}.

$$BLER_{ib} = f(P_b^{Tx}, \boldsymbol{H_{ib}}, m_{ib}) \tag{4.3}$$

In Section 4.3.2, we elaborate how a look up table can be used in lieu of the function in (4.3). Henceforth, user QoS will refer to the user's throughput and BLER requirements.

4.2.2 BS Power Consumption Model

The RF chain consists of PA and RF chain transceiver circuitry. PA is the major contributor to BS power and has four states of operation namely, off, idle, active and switching states [88]. PA is switched off in the off state, and it is on but not transmitting in the idle state. PA transmits in the active state and the power consumption comprises of the idle power and transmission power. The transmission power consumption depends on PA efficiency, transmit power (assumed constant), bandwidth and duration of transmission. The switching power is comparable to idle power, however, the switching duration is much lower than time slot duration. Hence, the contribution of switching power is much lower than that of idle power when power consumption is averaged over the frame duration.

The baseband signal processing, DC-DC conversion, AC-DC conversion and cooling modules of the BS contribute significantly to BS power consumption. As they cannot be switched at the time scale of PA, the power consumption of the above modules has a baseline component independent of the PA state and an additional power component which scales with bandwidth of transmission when PA is transmitting. We adopt the model presented in [89] which captures the characteristics of BS module power consumption described above. The model in [89] is extended to include the off and switching power of PA and is briefly described below.

The frequency utilization ψ_{st} of RF chain $s \in [1, S]$ in time slot $t \in [1, T]$ due to

$\mid I_b \mid$ users is

$$\psi_{st} = \begin{cases} \frac{1}{J}\sum_{i=1}^{|I_b|} J_{sti}, & \text{if PA is in active state} \\ 0, & \text{if PA is in idle or off state} \end{cases} \tag{4.4}$$

where J_{sti} is the number of frequency blocks assigned on RF chain $s \in [1, S]$ in time slot $t \in [1, T]$ to the i^{th} user. As in LTE systems, we consider frequency block allocation on a per time slot basis in a frame [90] to determine ψ_{st}. The number of active RF chains in a time slot t is $S_t^A = \mid \{s : \psi_{st} > 0\} \mid$. The number of active and idle time slots in a frame is given by $T^A = \mid \{t : S_t^A > 0\} \mid$, $T^I = \mid \{t : S_t^A = 0 \wedge \exists s \in [1, S] : s \text{ is on}\} \mid$. Denoting the duration of PA switching as t^{Sw} and the number of RF chains switching in a frame as S^{Sw}, the duration of all the RF chains in the off state in a frame is $t^O = t^F - \frac{t^F}{T}(T^A + T^I) - t^{Sw}S^{Sw}$.

Using the above definitions, the average power consumption of BS b with S RF chains in a frame with T time slots is

$$\begin{gathered} P_b = \frac{1}{t^F}(\sum_{t=1}^{T_b^A}(S_{tb}^A P^I + \Delta_p P^{Max}\sum_{s=1}^{S_{tb}^A}\sum_{i=1}^{|I_b|}\psi_{stib} + \\ (S - S_{tb}^A)P^O) + ST_b^I P^I) + St_b^O P^O + S_b^{Sw} t_b^{Sw} P^{Sw} \end{gathered} \tag{4.5}$$

In the model above, P^O is the BS power consumption when the PA is switched off and includes the idle power consumption of all components excluding the PA and the off state power consumption of PA. The load independent term P^I represents the idle power of PA and the other components. The BS power consumption in the active time slots includes the baseline idle power component given by $S_{tb}^A P^I$ and the active power due to transmission modeled as the load dependent term $\Delta_p \psi_{st} P^{Max}$. The load dependent term $\Delta_p \psi_{st} P^{Max}$ increases linearly with only frequency utilization ψ_{st} as power gradient (slope) Δ_p and maximum transmit power P^{Max} are maintained constant. In the proposed technique, PA is either in the active, off or switching state. Henceforth, $T^I P^I$ is not a contributor to P_b. Defining $S_b^A = \{S_{tb}^A : t \in [1, T_b^A]\}$ and $\psi_b = \{\psi_{sti} : s \in [1, S_{tb}^A], t \in [1, T_b^A], i \in [1, | I_b |]\}$, the average cluster power consumption in a frame is given by

$$P_C = \sum_{b=1}^{|C|} P_b = f(\{(I_b, T_b^A, S_b^A, \psi_b) : b \in C\}) \tag{4.6}$$

4.2.3 Problem Formulation

We can infer from (4.2-4.3, 4.5) that the QoS requirements and channel conditions of I_b users determine the aggregate BS resource utilization and P_b. At the individual BSs, given I_b, the BS resource space formed by number of RF chains S, time slots T and frequency blocks J can be explored during user mode selection to minimize P_b. At the cluster level, adapting the association of users $I_C = \cup_{b \in C} I_b$ will adapt the aggregate BS resource utilization and P_b. However, the association of all the users $I_b \forall b \in C$ cannot be adapted. This is because for every $b \in C$, there may exist a set of non-transferable users $I_b^{NT} \subseteq I_b$ that lie in the coverage area of only BS b and cannot be transferred to any other BS $b^{\sim} \in C \setminus b$ (see Fig. 4.2). The association of set of transferable users $I_b^T = I_b \setminus I_b^{NT}$ can be adapted as they lie in the coverage area of at least one more BS $b^{\sim} \in C \setminus b$ and can be transferred to BSs $\{b^{\sim}\}$. From the above description of I_b^{NT} and I_b^T, we can see that $I_b^{NT} \cap I_b^T = \emptyset \forall b \in C$. Further, assuming that a user is associated with no more than one BS, $I_b^T \cap I_{b^{\sim}} = \emptyset$ even though user $i \in I_b^T$ is located in the coverage area of BS $b^{\sim}$. Using the above, the set of cluster users is given $I_C = I_C^{NT} \cup I_C^T$ where $I_C^{NT} = \cup_{b \in C} I_b^{NT}$ and $I_C^T = \cup_{b \in C} I_b^T$ is the set of non-transferable and transferable cluster users respectively. The sets I_C^T and C together form the UA space that can be explored to adapt the set of users associated with BSs $b \in C$ and affect the individual BS resource utilization.

The objective of the BS and UA resource adaptation is to maximize the number of RF chains that can be switched off in the cluster to minimize P_C while satisfying the QoS requirements in (4.2-4.3) for all the cluster users and not exceeding the BS resource utilization limits. The objective and constraints form the optimization problem stated below. Note, a single cluster C and associated users I_C is considered unless otherwise mentioned.

$$\min \sum_{b=1}^{|C|} \frac{1}{t^F} (\sum_{t=1}^{T_b^A} (S_{tb}^A P^I + \Delta_p P^{Max} \sum_{s=1}^{S_{tb}^A} \sum_{i=1}^{|I_b|} \psi_{stib} + \tag{4.7}$$

$$(S - S_{tb}^A) P^O)) + St_b^O P^O + S_b^{Sw} t_b^{Sw} P^{Sw}$$

Subject to: $TP_{ib} \geq \gamma_i, \forall i \in I_C$ (4.8)

$$BLER_{ib} \leq BLER_i^{Th}, \forall i \in I_C \tag{4.9}$$

$$\frac{t^F}{T} T_b^A + t^{Sw} S_b^{Sw} \leq t^F, \forall b \in C \tag{4.10}$$

$$S_{tb}^A \leq S, \forall t \in [1, T_b^A], \forall b \in C \tag{4.11}$$

$$\psi_{stb} \leq 1, \forall s \in [1, S_{tb}^A], \forall t \in [1, T_b^A], \forall b \in C \tag{4.12}$$

To minimize (4.7), the optimization variables are the sets $I_C^T = \cup_{b \in C} I_b^T$ and $\{T_b^A, \{S_{tb}^A\}, \{\psi_{stb}\} : b \in C, t \in [1, T_b^A], s \in [1, S_{tb}^A]\}$. The idle power and transmission power of the BS due to active RF chains (first and second terms in the summation over T_b^A in (4.7)) are the dominant components of P_b (Section 4.2.2) and thereby, P_C. On the other hand, the off power due to inactive RF chains given by the third term in the summation over T_b^A is much lower than the static and dynamic powers and hence contributes less to the BS power consumption. This implies that the number of active RF chains will have priority in the optimization to minimize P_C. Minimizing the number of RF chains will result in minimizing the first and second terms of the summation over T_b^A while maximizing the third term in the summation over T_b^A. Further, minimizing the number of active RF chains in time slots to zero will maximize the RF chain off duration (t^O) and minimize the number of active time slots T_b^A. This will minimize the first term (entire summation over T_b^A) in (4.7) and maximize the second term (power consumption when all RF chains are off). Therefore, minimizing P_C can be considered equivalent to minimizing (maximizing) the number of active (off) chains. Constraints (4.8-4.9) respectively ensure that the throughput TP_{ib} and the $BLER_{ib}$ provided by BS b satisfies the i^{th} user's required rate γ_i and upper BLER bound $BLER_i^{Th}$. Constraint (4.10) ensures that the sum of duration of transmission and switching is upper bounded by t^F. The number of active RF chains in an active time slot is upper bounded by S in (4.11). The last constraint (4.12) specifies the upper bound on the frequency utilization of every active RF chain. An important point to note here is that satisfying the constraints (4.8-4.9) ensures that every cluster user is associated with a BS and therefore explicit constraints to ensure the same are not required. Henceforth, the optimization will be carried out with the transmission frame as reference.

4.3 Co-RFSnooze Algorithm

4.3.1 Multiple Multidimensional Knapsack Problem

The problem in (4.7-4.12) belongs to the class of Multiple Multidimensional Knapsack Problem (MMKP) as described below. Let the set of cluster users I_C and set of cluster BSs C denote the set of items and knapsacks respectively. UA is equivalent to assigning items to knapsacks and BS resource utilization is equivalent to utilizing the knapsack capacity. The profit of assigning user (item) $i \in I_C$ to BS $b \in C$ (knapsack) is the throughput TP_{ib} and the achievable $BLER_{ib}$ provided by BS b to user i. The number of BS RF chains S denotes the number of dimensions of the knapsack and the capacity of BS b in dimension $s \in [1, S]$ is JT, the total number of frequency blocks in a frame.

The weight of user $i \in I_C$ in dimension $s \in S$ is the total number of frequency blocks assigned to the user in the frame given by $\sum_{t \in T} J_{sti}$. The BS resource and UA adaptation to minimize average cluster power consumption can be seen as MMKP with minimizing the total BS resource utilization, maximizing the users' throughput and minimizing the users' BLER as the criteria for optimization. The problem stated in (4.7-4.12) is a variant of the above multi-criteria MMKP which minimizes BS resource utilization subject to lower bound on throughput provided and upper bound on achieved BLER. As MMKP is a NP-Hard problem [91], we propose a heuristic algorithm that integrates BS resource and UA adaptation heuristics to solve (4.7-4.12).

4.3.2 BS Resource Adaptation - Heuristics and Algorithm

Consider the set of users I_b associated with BS b and let $I =| I_b |$. For brevity of notation, we will drop the subscript b in this subsection. Selection of mode $m_i \in M$ for the user $i \in I_b$ utilizes T_i active time slots, $s_{ti} \forall t \in [1, T_i]$ active RF chains and $J_{sti} \forall s \in [1, s_{ti}], t \in [1, T_i]$ frequency blocks. The mode selection for individual users impacts the overall BS utilization as follows.(i) $T^A = \max_{i=1,..,I} T_i$, (ii) $S_t^A = \max_{i=1,..,I} s_{ti}, \forall t \in [1, T_i]$ and (iii) $\psi_{st} = \sum_{i=1}^{I} \frac{J_{sti}}{J} \forall t \in [1, T_i], s \in [1, S_t^A]$. From the above, it can be inferred that T^A, S_t^A and ψ_{st} can be minimized if each is minimized for every user. However, minimizing each of the BS resource in isolation for every user will lead to an increase in the other BS resources because (a) decreasing T_i increases s_{ti} and J_{sti}, (b) decreasing s_{ti} increases T_i and J_{sti} and (c) decreasing J_{sti} increases T_i and s_{ti} in order to satisfy the QoS of the user. Therefore, joint adaptation of resources allocated to every user is required to minimize BS utilization and P_b.

The RFSnooze (Min-Cost in [76]) algorithm shown in Table 4.2 jointly adapts the BS resources to minimize BS utilization and P_b. The inputs to the algorithm are the required throughput γ_i and BLER threshold $BLER^{Th}$, the rank indicator RI_i and

Table 4.2. RFSnooze algorithm

```
Input: $I_b, \{\gamma_i, BLER_i^{Th}, RI_i, CQI_i, \boldsymbol{H_i} : i \in [1,I]\}, S, J, R, T$
Output: $T^A, \{S_t^A : t \in [1,T^A]\}, \{\psi_{st} : s \in [1,S_t^A], t \in [1,T^A]\}$
1. For all users $i \in [1,I]$
2:     Initialize $M_i^{FS} = \emptyset$, $J_i(m) = 0$, $T_i(m) = 0, \forall m \in M$
3:     For all modes $m \in M$
4:         Scheduler updates $T_i(m) = max_{t\in[1,T]}\{t : J_{ti} > 0\}$,
           $J_i(m) = \sum_{t=1}^{T_i(m)} J_{ti}(m)$ if $TP_i(m, J_i(m), T_i(m))) \geq \gamma_i$
5:         Determine $BLER_i(P^{Tx}, \boldsymbol{H_i}, m)$ using $CQI_i$ entry in LUT
6:         If $BLER_i(P^{Tx}, \boldsymbol{H_i}, m) \leq BLER_i^{Th}, d_i(m) \leq RI_i(m), T_i(m) \leq T, J_{ti} \leq JT$,
           then update $M_i^{FS} = M_i^{FS} \cup m$
7:         Compute $P_i(m)$ using (4.13)
8: Find mode $m^* = argmin_{m\in M_i^{FS}} P_i(m)$
9: Update $T_i = T_i(m_i^*), s_{ti} = s(m_i^*), \psi_{sti} = J^{-1} J_{ti}(m_i^*), \forall s \in [1, s_{ti}], \forall t \in [1, T_i]$
10: Determine $T^A = max_{i\in[1,I]} T_i$
11: For all time slots $t = 1, .., T^A$
12:     Determine $S_t^A = max_{i\in[1,I]} s_{ti}$
13:     Determine off RF chains $S_t^O = S - S_t^A$
14:     Determine $\psi_{st} = J^{-1} \sum_{i=1}^{I} J_{sti}, \forall s \in [1, S_t^A]; \psi_{st} = 0, \forall s \in [1, S_t^O]$
```

the channel quality indicator CQI_i sent as periodic feedback by all the users $i \in [1,I]$ [92], the channel matrix $\boldsymbol{H_i}$, the BS and user device resource upper bounds S, T, J and R. The steps of the algorithm are explained briefly below. The reader can refer to [76] for detailed explanation of the algorithm.

In step 4, the output of iterative frequency domain scheduler [93] is extended to allocate $T_i(m)$ time slots, $s_i(m)$ RF chains, $J_i(m)$ frequency blocks for all modes $m \in M$ in a frame for all users $i \in [1,I]$. The $BLER$ in step 5 is determined using the CQI and RI measurements and the Look Up Table (LUT) in [94] (used in lieu of BLER function in (4.3)) that specifies for different CQI values, the SINR threshold $SINR^{Th}(m)$ required for every mode $m \in M$ to result in $BLER \leq 0.1$. For all permissible modes $\{m : d_i(m) \leq RI_i\}$, if $SINR_i \geq SINR^{Th}(m)$ ($SINR_i$ is given by (4.1)), then $BLER_i(m) = BLER_i^{Th}$, else $BLER_i(m)$ is set to value greater than $BLER^{Th}$.

In step 6, the set of feasible modes $M_i^{FS} \subseteq M$ is updated with modes m that satisfies the throughput, BLER, and upper bounds on frequency and time utilization. From (4.5), the power consumption due to feasible mode $m \in M_i^{FS}$ is given by

$$P_i(m) = \frac{1}{t^F}(T_i(m)s_i(m)P^I + \frac{s_i(m)\Delta_p P^{Max}}{J}\sum_{t=1}^{T_i(m)} J_{ti}(m)) \tag{4.13}$$

The power consumption is calculated for every mode $m \in M_i^{FS}$ in step 7 and the mode m_i^* that results in minimum power consumption is chosen in step 8. The number of active time slots T^A, active RF chains $\{S_t^A : t \in [1, T^A]\}$, the frequency utilization $\{\psi_{st} : s \in [1, S_t^A], t \in [1, T^A]\}$ are the algorithm outputs determined in steps 10-14.

From Table 4.2, the complexity of RFSnooze to determine the combination of modes is given by $\mid M \mid O(I)$ and is linear in I. In comparison, complexity of exhaustive search given by $O(\mid M \mid^I)$ is exponential in I.

4.3.3 UA Adaptation - Heuristics

SINR threshold for a mode m is defined as the threshold below which the BLER due to mode m, $BLER(m) > BLER^{Th}$ and can be determined as outlined in [94]. BS b that can provide SINR greater than the minimum of the SINR thresholds of all modes $m \in M$ can service the i^{th} user as there exists at least one mode m for which $SINR_{ib} > SINR^{Th}(m)$. Let E_i denote the set of BSs that can service the i^{th} user. We assume that the cluster users send the CQI and RI information for every BS $b \in C$ to the entire cluster [95]. Using this information, the BS-user assignment matrix $BSU = [k_{bi}]_{|C|x|I_C|}$ with elements $k_{bi} \in [0, \mid C \mid]$ is maintained at all BSs $b \in C$. The value $k_{bi} = 0$ indicates that BS $b \notin E_i$ as it does not satisfy the minimum of mode SINR thresholds for the i^{th} user. Sorting the BSs $b \in E_i$ in the decreasing order of SINR, the values $k_{bi} = 1$ indicates that BS b provides the highest SINR, $k_{bi} = 2$ indicates that BS b provides the second highest SINR to the i^{th} user and so on. Using the BSU matrix, the I_b^{NT} and I_b^T users associated

Table 4.3. Illustration of BSU matrix with $\mid C \mid= 4$ and $\mid I_C \mid= 10$

BS-User	1	2	3	4	5	6	7	8	9	10
1	0	4	1	0	1	3	1	0	2	0
2	1	3	0	0	0	4	2	0	3	2
3	2	2	0	1	0	2	3	0	0	0
4	0	1	0	0	0	1	4	1	1	1
Modified BSU matrix after restricting $E_i = \{b : k_{bi} \in [1,2]\}$										
1	0	0	1	0	1	0	1	0	2	0
2	1	0	0	0	0	0	2	0	0	2
3	2	2	0	1	0	2	0	0	0	0
4	0	1	0	0	0	1	0	1	1	1

with BS b can be defined as

$$I_b^{NT} = \{i : k_{bi} = 1 \wedge E_i = \{b\} \wedge \{v : k_{vi} \geq 2\} = \emptyset\} \tag{4.14}$$

$$I_b^T = \{i : k_{bi} = 1 \wedge E_i = b \cup \{v : k_{vi} \geq 2\}\} \tag{4.15}$$

Table 4.3 shows the BSU matrix for a cluster of size $\mid C \mid= 4$ and $\mid I_C \mid= 10$. Using (4.14-4.15), the sets I_b^{NT} and I_b^T for BSs $b = 1,2,3,4$ can be written as: $I_1^{NT} = \{U3, U5\}$,$I_1^T = \{U7\}$;$I_2^{NT} = \emptyset$,$I_2^T = \{U1\}$;$I_3^{NT} = \{U4\}$, $I_3^T = \emptyset$;$I_4^{NT} = \{U8\}$,$I_4^T = \{U2, U6, U9, U10\}$. Note, for $BS2$, as $I_2^{NT} = \emptyset$ all the RF chains can be switched off by transferring $U1$. We will next discuss heuristics for allocating BS resources to I_b^{NT} and I_b^T users. Without loss of generality, we will consider BS $b \in C$ for the discussion and drop the subscript b for brevity.

From (4.5), the utilization of BS resources is the aggregate utilization due to $I^{NT} \cup I^T$. By allocating resources first to I^{NT} and subsequently to I^T, we can rewrite (4.5) as

$$P = \frac{1}{t^F}\Bigg(\sum_{t=1}^{T^{NT}}(S_t^{NT}P^I + \frac{\Delta_p P^{Max}}{J}\sum_{s=1}^{S_t^{NT}}\sum_{i=1}^{|I^{NT}\cup I^{T^{\sim}}|} J_{sti})$$

$$+\sum_{t=T^A-T^{NT}+1}^{T^A}((S_t^A - S_t^{NT})P^I + \frac{\Delta_p P^{Max}}{J}\sum_{s=S_t^A-S_t^{NT}+1}^{S_t^A}$$

$$\sum_{i=1}^{|I^T\setminus I^{T^{\sim}}|} J_{sti}) + \sum_{t=1}^{T^A}(S-S_t^A)P^O\Big) + t^O S P^O + t^{Sw} S^{Sw} P^{Sw} \tag{4.16}$$

where T^{NT} and S_t^{NT} are the number of active time slots and RF chains in time slot $t \in [1, T^{NT}]$ required to satisfy the QoS requirements of I^{NT} and $I^{T^{\sim}} \subseteq I^T$ users. This implies that $S_t^A - S_t^{NT}$ RF chains can be switched off in time slots $\{t \in [1, T^A] : S_t^A - S_t^{NT} > 0\}$ if $\mid I^T \setminus I^{T^{\sim}} \mid$ users are transferred to feasible cluster BSs. The subset of transferable users $I^{T^{\sim}}$ are updated as non-transferable users as their QoS requirements are satisfied by allocating no more than S_t^{NT} RF chains in time slots T^{NT} allocated to I^{NT} users. The possibility of reducing $\mid I^T \mid$ and complexity of UA is the motivation to allocate BS resources first to I^{NT} users and subsequently to I^T users. Next, we will select the ”transferor” BS g which will transfer users and the ”transferee” BSs E to transfer users to.

Higher the number of RF chains $S_t^A - S_t^{NT}$ that can be switched off, higher the savings in transferor BS power consumption. However, as the number of users $\mid I^T \setminus I^{T^{\sim}} \mid$ that are transferred increases, the number of users that receive less than maximum SINR and the transferee BS power consumption also increases. To maximize $S_t^A - S_t^{NT}$ while minimizing $\mid I^T \setminus I^{T^{\sim}} \mid$ and the increase in transferee BS power consumption, the RF chain-user ratio RFU is defined as

$$RFU = \frac{\sum_{t=T^A-T^{NT}+1}^{T^A} S_t^A - S_t^{NT}}{\mid I^T \setminus I^{T^{\sim}} \mid} \tag{4.17}$$

Larger the RFU ratio, higher will be the savings in transferor BS power consumption and lower will be the number of users receiving less than maximum SINR. Also, large RFU ratio will result in lower increase in transferee BS power consumption. Hence, the BS with the largest RFU ratio is nominated as the transferor BS g.

Amongst the multiple BSs which cover user $i \in I_g^T \setminus I_g^{T^\sim}$, the selection of transferee BS is restricted to that subset of BSs $b \in E_i$ with $k_{bi} = 2$ in the BSU matrix. This has a two-fold effect of reducing (a) the impact on QoS of the user $i \in I_g^T \setminus I_g^{T^\sim}$ and (b) the complexity of UA. The set of transferee BSs corresponding to $I_g^T \setminus I_g^{T^\sim}$ is denoted as E.

The above selection criterion is applied to Table 4.3 resulting in replacing all the entries with $k_{bi} > 2$ with $k_{bi} = 0$ to indicate that BS b is not a transferee BS for the i^{th} user. The bottom portion of Table 4.3 shows the modified BSU matrix. This reduces $\mid E_i \mid$ for i^{th} user and also minimizes the impact on the user QoS. For instance the set of transferee BSs for $U7$ is reduced from $E_7 = \{BS1, BS2, BS3, BS4\}$ to $E_7 = \{BS1, BS2\}$.

We will now discuss the three feasibility conditions that have to be satisfied for transferring users. The first condition is that the QoS requirements of transferrable users of transferor BS and the users of transferee BS have to be satisfied by the transferee BS after the transfer.

$$C1 : \text{satisfy constraints (4.8-4.9)} \forall e \in E, i \in I_e \cup I_g^T \setminus I_g^{T^\sim} \tag{4.18}$$

Let us denote the number of active time slots, active RF chains and frequency utilization of BS b before user transfer as T_b^A, S_b^A, ψ_b and after user transfer as $T_b^{A*}, S_b^{A*}, \psi_b^*$. The second condition is that BS resource utilization of transferee BS e after transfer $T_e^{A*}, S_e^{A*}, \psi_e^*$ should satisfy (4.10-4.12).

$$C2 : \text{satisfy constraints (4.10-4.12)} \forall e \in E \tag{4.19}$$

Denoting the power consumption of BSs after user transfer as P^*, the third condition is that the difference in cluster power consumption before and after transfer should be positive.

$$\begin{aligned} C3: \Bigg(& P_g(I_g, T_g^A, S_g^A, \psi_g) + \sum_{e=1}^{|E|} P_e(I_e, T_e^A, S_e^A, \psi_e) - P_g^*(I_g^{NT} \cup I_g^{T\sim}, \\ & T_g^{A*}, S_g^{A*}, \psi_g^*) - \sum_{e=1}^{|E|} P_e^*(I_e^{NT} \cup (I_g^T \setminus I_g^{T\sim}), I_e^T, T_e^{A*}, S_e^{A*}, \psi_e^*) \Bigg) > 0 \end{aligned} \tag{4.20}$$

4.3.4 Co-RFSnooze Algorithm

The Co-RFSnooze algorithm adopts a bottom-up iterative approach which adapts BS resources at individual cluster BSs and adapts UA at cluster level in an iterative manner. An iteration consists of two key interlinked steps explained below. The first key step is that the Co-RFSnooze algorithm applies the RFSnooze algorithm at each cluster BS to I^{NT} and subsequently to I^T users and determines the RFU ratio. This step (a) minimizes the number of RF chains required to satisfy the QoS requirements of I^{NT} users at the individual BS level, (b) reduces the cardinality of the I^T (Section 4.3.3) to prune the UA space at the cluster level and (c) determines the BS resources required to satisfy the QoS requirements of the $I^T \setminus I^{T\sim}$ users using which the RFU ratio is calculated. The RFU ratio guides the choice of transferor BS and is the crucial link between individual BS resource adaptation and cluster level UA adaptation.

The second key step is the selection of transferor and transferee BSs. The BS with highest RFU ratio is selected as the transferor BS to maximize the savings in power consumption due to switching off RF chains and minimize the impact on users' received SINR. The set of transferee BSs is restricted to BSs that provide the second highest SINR to $I^T \setminus I^{T\sim}$ of transferor BS to reduce UA space. The above two key steps are carried out iteratively by Co-RFSnooze algorithm as described below.

The Co-RFSnooze algorithm is shown in Table 4.4. The algorithm inputs are the set of cluster users, their QoS requirements and the channel state information, the BS

Table 4.4. Co-RFSnooze algorithm

```
Input: $\{I_b^{NT}, I_b^T : b \in [1, | C |]\}, \{\gamma_i, BLER_i^{Th} : i \in [1, | I_C |]\}, \{RI_{ib}, CQI_{ib},$
        $\boldsymbol{H_{ib}} : i \in [1, | I_C |], b \in [i, | C |]\}, S, J, R, T$
Output: $\{I_b, T_b^A, \{S_{tb}^A\}, \{\psi_{stb}\} : s \in [1, S_{tb}^A], t \in [1, T_b^A], b \in [1, | C |]\}$
1. Initialize set of possible transferor BSs $G = C$, set of transferee BSs $E = \{\}$,
        transferor BS $g = \{\}$
2. For all BSs $b \in C$
3:          Initialize $I_b^{NT}$ and $I_b^T$ using (4.14) and (4.15)
4:          Apply RFSnooze to $I_b^{NT}$ to determine BS resource allocation for $I_b^{NT}$
5:          Apply RFSnooze to $I_b^T$ to determine BS resource allocation for $I_b^T$
6:          Determine $I_b^{T\sim} \subseteq I_b^T$ that require no additional time slots
            and RF chains as compared to $I_b^{NT}$
7:          Update $I_b^{NT} = I_b^{NT} \cup I_b^{T\sim}, I_b^T = I_b^T \setminus I_b^{T\sim}$, update $BSU_b$ with $k_{ei} = 0, \forall e \in C \setminus b$
8:          Calculate $P_b$ using (4.5) and $RFU_b$ using (4.17)
9: If $G = \{\}$, then go to step 27, Else
10:         Select transferor BS with highest $RFU$ ratio $g = max_{b \in G} RFU_b$
11:         Update $G = G \setminus g$
12:         Determine subset of BSs $E = \{e : \exists i \in I_g^T \setminus I_g^{T\sim} \wedge k_{ei} = 2\}$
            to which BS $g$ can transfer users $I_g^T \setminus I_g^{T\sim}$
13:For all BSs $e \in E$
14:         Update $I_e^{NT} = I_e^{NT} \cup \{i : i \in I_g^T \setminus I_g^{T\sim} \wedge k_{ei} = 2\}$
15:         Apply RFSnooze to $I_e^{NT}$ to determine BS resource allocation for $I_e^{NT}$
16:         Apply RFSnooze to $I_e^T$ to determine BS resource allocation for $I_e^T$
17:         Determine $P_e^*$ using (4.5) and $\Delta P_e = P_e - P_e^*$
18:         If transfer feasibility condition $C1$ or $C2$ is violated
19:                 Then set $P_e^* = \infty, \Delta P_e = \infty$
20: Apply RFSnooze to $I_g^{NT}$ users of transferor BS $g$ to determine
    BS resource allocation
21: Determine $P_g^*$ using (4.5) and $\Delta P_g = P_g - P_g^*$
22: If transfer feasibility condition $C3$ is true, then for all users $i \in I_g^T \setminus I_g^{T\sim}$,
    for all BSs $e \in E$
23:         Update the $BSU$ matrix $k_{gi} = 0, k_{ei} = 1$
24: Else for all users $i \in I_g^T \setminus I_g^{T\sim}$, for all BSs $e \in E$
25:         Update the $BSU$ matrix $k_{ei} = 0$
26: Go to step 2
27: For all BSs $b \in C$
28:         $I_b = \{i : k_{bi} = 1\}, \{T_b^A, \{S_{tb}^A\}, \{\psi_{stb}\} : s \in [1, S_{tb}^A], t \in [1, T_b^A]\}$ -
            Output of step 4
```

resource upper bounds for the cluster BSs. The algorithm outputs are the set of users associated with each of the cluster BSs and corresponding resource utilization of the BS.

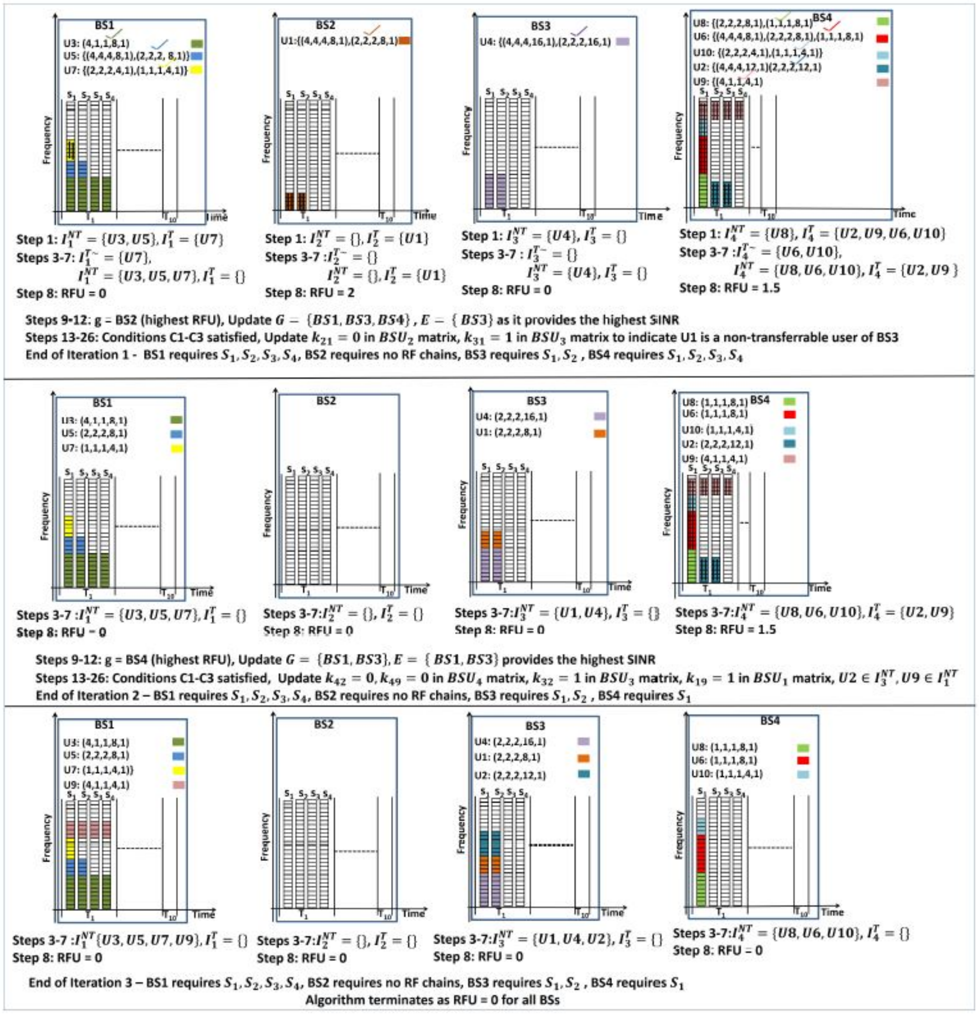
Starting with the set of transferor BSs $G = C$ and set of transferee BSs $E = \emptyset$, the algorithm iterates till the set of transferor BSs $G = \emptyset$. Each iteration starts by allocating individual BS resources first to I_b^{NT} users in step 4 and subsequently to I_b^T users in step 5. The set of users $I_b^{T\sim}$ that can be serviced in T_b^{NT} time slots with $S_{tb}^{NT}, t \in [1, T_b^{NT}]$ RF chains is obtained from step 6. The sets I_b^{NT} and I_b^T are updated in step 7 and the power consumption P_b and the RFU ratio are calculated in step 8.

Using the RFU ratio, steps 10-11 selects the transferor BS g and updates the set of transferor BSs G to exclude the selected BS g. The set of transferee BSs E is selected in step 12 and the corresponding sets of $I_e^{NT}, \forall e \in E$ are updated in step 14 to include the transferable users $I_g^T \setminus I_g^{T\sim}$ of BS g. The update of G and of $I_e^{NT} \forall e \in E$ is of particular importance. By updating the set $G = G \setminus g$ in the current iteration eliminates the selection of BS g as transferor BS in any subsequent iterations. This reduces the cardinality of set of possible transferor BSs G for subsequent iterations and ensures convergence of the algorithm in at most $| C |$ iterations. The update $I_e^{NT} = I_e^{NT} \cup I_g^T \setminus I_g^{T\sim}$ categorizes $I_g^T \setminus I_g^{T\sim}$ of BS g as non-transferable users of BS e. This will not allow oscillatory behavior wherein the users $I_g^T \setminus I_g^{T\sim}$ are assigned back to the transferor BS g in subsequent iterations in which transferee BS e may be selected as transferor BS and BS g as transferee BS.

The BS resource allocation taking in to account the transferred users is determined in steps 15-16 following which the transfer feasibility conditions C1, C2 and C3 (Section 4.3.3) are tested in steps 18-22. Note that condition C1 is implicitly satisfied by the RFSnooze algorithm as it selects feasible modes which satisfies the constraints (4.8-4.9) for each user. Iterative allocation of resources to users as explained in Section 4.3.2, [76] ensures that the BS resource utilization constraints (4.10 -4.12) are satisfied. Given the resource utilization of BSs g and E, C3 is evaluated using (4.20). If conditions C1, C2 and C3 hold, then the BSU matrix entries for users $I_g^T \setminus I_g^{T\sim}$ are updated in step 23 to reflect the disassociation from transferor BS g ($k_{gi} = 1$ to $k_{gi} = 0$) and association with the transferee BS e ($k_{ei} = 2$ to $k_{ei} = 1$). If the conditions do not hold, then the BSU matrix is updated in step 25 to reflect that the users $I_g^T \setminus I_g^{T\sim}$ are non-transferable users of BS g ($k_{ei} = 2$ to $k_{ei} = 0$). In addition the power consumption of all transferee BSs is set to an arbitrarily large number to indicate that the transfer is not feasible. This is carried out for implementation purposes as elaborated in the next subsection. With the updated UA and set of possible transferor BSs G, the next iteration is initiated in step 26.

The iterations terminate when there are no more candidates for transferring users, i.e., $G = \emptyset$. In the final iteration, steps 2-8 are executed, however, since there are no more transferable users, the BS resource allocation obtained in step 4 is the final BS resource allocation. The check in step 9 is true for the final iteration and the algorithm terminates by executing steps 27-28. The outputs of the algorithm are the UA obtained from the BSU matrix and the corresponding BS resource utilization of the cluster BSs.

We will use the example in Table 4.3 (bottom portion) with cluster of size $\mid C \mid= 4$ and $\mid I_C \mid= 10$ users to run through the algorithm steps with the aid of Fig. 4.3. The rows of Fig. 4.3 illustrate the BS resource utilization for each BS at the beginning of an iteration and lists the subsequent steps. The BS resource utilization is shown for one time slot of a transmission frame with $J = 24$ frequency blocks available on each of $S = 4$ RF chains ($S_1, .., S_4$). The maximum number of user RF chains is $R = 4$. The frequency blocks allocated to users are indicated by the color used for the user. Due to lack of space, we have omitted showing multiple time slots in the transmission frame. For each user, the modes $m \in M_i^{FS}$ and the corresponding allocation of time slots and frequency blocks are listed in the legend using a 5-tuple - $(s_i, r_i, d_i, J_i, T_i)$. The I^T of each BS are differentiated by two vertical black colored lines placed on the BS resources allocated. For instance,


BS1
U3: (4,1,1,8,1)
U5: {(4,4,4,8,1),(2,2,2, 8,1)}
U7: {(2,2,2,4,1),(1,1,1,4,1)}
BS2
U1:{(4,4,4,8,1),(2,2,2,8,1)}
BS3
U4: {(4,4,4,16,1),(2,2,2,16,1)
BS4
U8: {(2,2,2,8,1),(1,1,1,8,1)
U6: {(4,4,4,8,1),(2,2,2,8,1),(1,1,1,8,1)
U10: {(2,2,2,4,1),(1,1,1,4,1)}
U2: {(4,4,4,12,1)(2,2,2,12,1)
U9: {(4,1,1,4,1)
S1 S2 S3 S4
Frequency
T1
T10
Time
U3: (4,1,1,8,1)
U5: (2,2,2,8,1)
U7: (1,1,1,4,1)
U4: (2,2,2,16,1)
U1: (2,2,2,8,1)
U8: (1,1,1,8,1)
U6: (1,1,1,8,1)
U10: (1,1,1,4,1)
U2: (2,2,2,12,1)
U9: (4,1,1,4,1)
U7: (1,1,1,4,1))

Step 1: $I_1^{NT} = \{U3, U5\}, I_1^T = \{U7\}$
Steps 3-7: $I_1^{T\sim} = \{U7\}$,
$I_1^{NT} = \{U3, U5, U7\}, I_1^T = \{\}$
Step 8: RFU = 0

Step 1: $I_2^{NT} = \{\}, I_2^T = \{U1\}$
Steps 3-7 :$I_2^{T-} = \{\}$
$I_2^{NT} = \{\}, I_2^T = \{U1\}$
Step 8: RFU = 2

Step 1: $I_3^{NT} = \{U4\}, I_3^T = \{\}$
Steps 3-7 : $I_3^{T-} = \{\}$
$I_3^{NT} = \{U4\}, I_3^T = \{\}$
Step 8: RFU = 0

Step 1: $I_4^{NT} = \{U8\}, I_4^T = \{U2, U9, U6, U10\}$
Step 3-7 :$I_4^{T-} = \{U6, U10\}$,
$I_4^{NT} = \{U8, U6, U10\}, I_4^T = \{U2, U9\}$
Step 8: RFU = 1.5

Steps 9-12: g = BS2 (highest RFU), Update $G = \{BS1, BS3, BS4\}$, $E = \{BS3\}$ as it provides the highest SINR
Steps 13-26: Conditions C1-C3 satisfied, Update $k_{21} = 0$ in BSU_2 matrix, $k_{31} = 1$ in BSU_3 matrix to indicate U1 is a non-transferrable user of BS3
End of Iteration 1 - BS1 requires S_1, S_2, S_3, S_4, BS2 requires no RF chains, BS3 requires S_1, S_2 , BS4 requires S_1, S_2, S_3, S_4

Steps 3-7 :$I_1^{NT} = \{U3, U5, U7\}, I_1^T = \{\}$
Step 8: RFU - 0

Steps 3-7:$I_2^{NT} = \{\}, I_2^T = \{\}$
Step 8: RFU = 0

Steps 3-7:$I_3^{NT} = \{U1, U4\}, I_3^T = \{\}$
Step 8: RFU = 0

Steps 3-7:$I_4^{NT} = \{U8, U6, U10\}, I_4^T = \{U2, U9\}$
Step 8: RFU = 1.5

Steps 9-12: g = BS4 (highest RFU), Update $G = \{BS1, BS3\}, E = \{BS1, BS3\}$ provides the highest SINR
Steps 13-26: Conditions C1-C3 satisfied, Update $k_{42} = 0, k_{49} = 0$ in BSU_4 matrix, $k_{32} = 1$ in BSU_3 matrix, $k_{19} = 1$ in BSU_1 matrix, $U2 \in I_3^{NT}, U9 \in I_1^{NT}$
End of Iteration 2 – BS1 requires S_1, S_2, S_3, S_4, BS2 requires no RF chains, BS3 requires S_1, S_2 , BS4 requires S_1

Steps 3-7 :$I_1^{NT}\{U3, U5, U7, U9\}, I_1^T = \{\}$
Step 8: RFU = 0

Steps 3-7:$I_2^{NT} = \{\}, I_2^T = \{\}$
Step 8: RFU = 0

Steps 3-7:$I_3^{NT} = \{U1, U4, U2\}, I_3^T = \{\}$
Step 8: RFU = 0

Steps 3-7:$I_4^{NT} = \{U8, U6, U10\}, I_4^T = \{\}$
Step 8: RFU – 0

End of Iteration 3 – BS1 requires S_1, S_2, S_3, S_4, BS2 requires no RF chains, BS3 requires S_1, S_2 , BS4 requires S_1
Algorithm terminates as RFU = 0 for all BSs

Figure 4.3. Application of Co-RFSnooze algorithm to example in Table 4.3

$I^T = \{U7\}$ for $BS1$ and two black lines are placed on the yellow blocks on S_1 RF chain.

Initially $G = \{BS1, BS2, BS3, BS4\}, E = \emptyset$. The top row of Fig. 4.3 shows the set of feasible modes M^{FS} (Section IIIB) and the minimum power mode $m*$ (indicated by the tick mark) selected for I^{NT} and I^T of BSs $BS1, BS2, BS3, BS4$ in steps 4 and 5 of iteration 1. The outputs of steps 1-28 for iteration 1 are listed below the BS resource utilization illustration. At the end of iteration 1, the RF chain requirements at $BS1 = \{S_1, S_2, S_3, S_4\}$, $BS2 = \emptyset$, $BS3 = \{S_1, S_2\}$ and $BS4 = \{S_1, S_2, S_3, S_4\}$. Due to transfer of $U1$ from $BS2$ to $BS3$, 2 RF chains are switched off at $BS2$ in iteration 1. This is the initial BS resource utilization of iteration 2 shown in second row of Fig. 4.3. The steps 4-26 of iteration 2 result in transfer of $U2, U9$ from $BS4$ to BSs $BS1, BS3$ and switching off RF chains S_2, S_3, S_4 of $BS4$. This is shown in the third row of Fig. 4.3. The algorithm terminates with the third iteration as RFU ratios $RFU_1 = 0, RFU_2 = 0, RFU_3 = 0, RFU_4 = 0$. We can see that Co-RFSnooze reduces the number of active RF chains from 12 to 7 in the cluster by iteratively applying the RFSnooze algorithm and UA adaptation heuristics.

4.3.5 Complexity Analysis

As exhaustive search of UA space evaluates $| C |^{|I_C^T|}$ combinations, the complexity of UA adaptation is $O(| C |^{|I_C^T|})$. For each UA combination, the exhaustive search of the BS resource space has to evaluate $| M |^{|I_1|} +..+ | M |^{|I_{|C|}|}$ combinations. Therefore, the complexity of joint search of BS resource spaces and UA spaces is given by $O(| C |^{|I_C^T|} (| M |^{|I_1|} +..+ | M |^{|I_{|C|}|}))$. The Co-RFSnooze algorithm evaluates a single combination of UA in an iteration and the maximum number of iterations for convergence of Co-RFSnooze is $| C |$. The complexity of UA space search is $O(| C |)$. In each iteration, the RFSnooze algorithm is executed at most twice for the entire cluster (steps 4-5, 15-16 and 20 in Table 4.4). The number of operations when RFSnooze algorithm (Section 4.3.2) applied to the every BS of entire cluster is $\sum_{b=1}^{|C|} | M || I_b | = | M || I_C |$. The complexity of

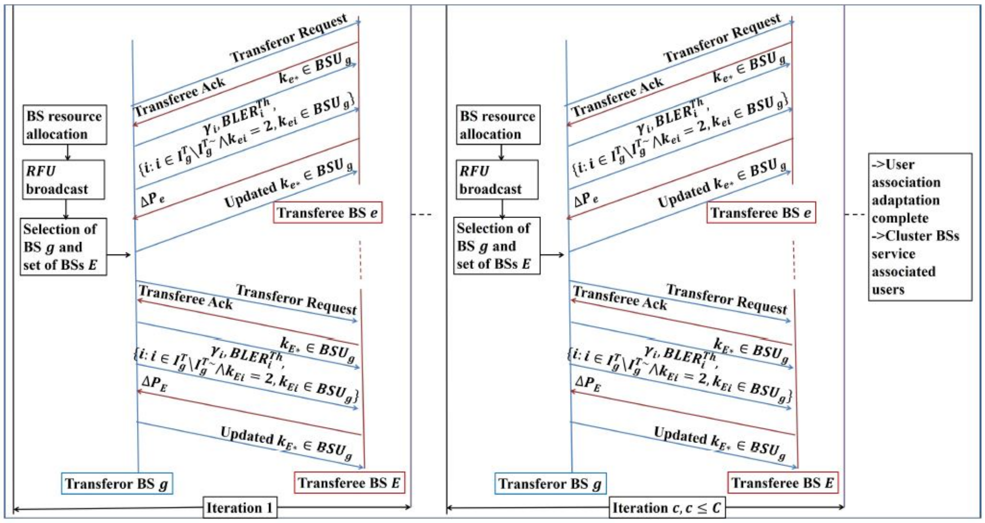

BS resource allocation
RFU broadcast
Selection of BS g and set of BSs E
Transferor Request
Transferee Ack
$k_{e*} \in BSU_g$
$\gamma_i, BLER_i^{Th}$,
$\{i: i \in I_g^T \backslash I_g^{T\sim} \wedge k_{ei} = 2, k_{ei} \in BSU_g\}$
ΔP_e
Updated $k_{e*} \in BSU_g$
Transferee BS e
Transferee Ack
Transferor Request
$k_{E*} \in BSU_g$
$\gamma_i, BLER_i^{Th}$,
$\{i: i \in I_g^T \backslash I_g^{T\sim} \wedge k_{Ei} = 2, k_{Ei} \in BSU_g\}$
ΔP_E
Updated $k_{E*} \in BSU_g$
Transferor BS g
Transferee BS E
Iteration 1
BS resource allocation
RFU broadcast
Selection of BS g and set of BSs E
Transferor Request
Transferee Ack
$k_{e*} \in BSU_g$
$\gamma_i, BLER_i^{Th}$,
$\{i: i \in I_g^T \backslash I_g^{T\sim} \wedge k_{ei} = 2, k_{ei} \in BSU_g\}$
ΔP_e
Updated $k_{e*} \in BSU_g$
Transferee BS e
Transferee Ack
Transferor Request
$k_{E*} \in BSU_g$
$\gamma_i, BLER_i^{Th}$,
$\{i: i \in I_g^T \backslash I_g^{T\sim} \wedge k_{Ei} = 2, k_{Ei} \in BSU_g\}$
ΔP_E
Updated $k_{E*} \in BSU_g$
Transferor BS g
Transferee BS E
Iteration $c, c \leq C$
->User association adaptation complete
->Cluster BSs service associated users

Figure 4.4. Implementation of Co-RFSnooze algorithm

the Co-RFSnooze algorithm for determining the BS resource allocation and UA in $|C|$ iterations is given by $2|C||M|O(|I_C|)$ where $|C|$ and $|M|$ are constants for a given cluster and BS resource configurations. Hence, Co-RFSnooze algorithm achieves linear complexity compared to the exponential complexity of exhaustive search.

4.3.6 Co-RFSnooze Framework

We propose a combination of the centralized approach [96] and the decentralized approach in [95] for the Co-RFSnooze framework to minimize the exchange of user QoS, channel state information (CSI) and control information between the cluster BSs to adapt UA.

The cluster BSs send training sequences to all the cluster users periodically [92]. In response, as implemented in decentralized approach in [95], the users estimate the CSI for each of the BS in the cluster and then send $|C|$ CSI estimates as feedback to every BS in the cluster. In this manner, the cluster BSs have the information about the SINR received by i^{th} user from every cluster BS $b \in C$. This enables the BSs to build and maintain a copy of the BSU matrix locally denoted as BSU_b. With the aid of Table 4.4

and Fig. 4.4, we will next discuss information exchange required for the Co-RFSnooze iterations.

With the inputs required and BSU matrix available at the BSs, steps 2-7 (Table 4.4) are run at every BS $b \in C$ for updating I^T. Subsequently, the BSs broadcast their RFU values to all the other cluster BSs. The BS with highest RFU ratio selects itself as the transferor BS with the other BSs implicitly getting this information from the broadcasted RFU values. Using the updated local copy of BSU matrix, the transferor BS g determines the set of transferee BSs E as in step 12. The above operations are listed in boxes in Fig. 4.4.

We adopt the cooperation protocol in [96] to set up the communication interface between BS g and BSs $e \in E$ shown in Fig. 4.4. The BS g sends the ”Transferor Request” to BSs $e \in E$ which in turn sends the ”Transferee Ack” response to complete the cooperation setup. The BS g transmits to each BS $e \in E$, the row $k_{e*} \in BSU_g$ corresponding to BS e. Note that the row $k_{e*} \in BSU_g$ transmitted by BS g is identical to the row $k_{e*} \in BSU_e$ (local copy of BSU matrix at BS e) except for the entries corresponding to $i \in I_g^{T\sim}$ for which $k_{ei} = 0, k_{ei} \in BSU_g$ (as updated in step 7, Table 4.4) and $k_{ei} = 2, k_{ei} \in BSU_e$. This difference indicates to BS e the reduced set of users $I_g^T \setminus I_g^{T\sim}$ required for steps 13-19. The QoS requirements $(\gamma_i, BLER_i)$ of the users $\{i : i \in I_g^T \setminus I_g^{T\sim}\}$ required as input to RFSnooze algorithm in steps 15-16 are transmitted to the transferee BS. Execution of RFSnooze algorithm in steps 15-16 will implicitly evaluate conditions C1 and C2, which if violated will set the difference power consumption ΔP_e to an arbitrarily large value. The ΔP_e is conveyed to BS g by all BSs $e \in E$ which evaluates condition C3. The BSU_g matrix is updated as per step 23 or step 25 depending on evaluation of condition C3. The updated rows $k_{e*} \in BSU_g$ are transmitted to BSs $e \in E$ and the current iteration ends. The c^{th} iteration consists of the operations indicated by the boxes and information exchange shown in Fig. 4.4. After a cluster BS has been selected as transferor BS, in subsequent iterations, it broadcasts $RFU = 0$ value. In terms of implementation, when all the BSs broadcast $RFU = 0$, the algorithm terminates. Subsequently, the cluster BSs use the updated local BSU matrices to service the associated users.

The overhead due to information exchange among the cluster BSs is as follows. A byte each for mantissa and exponent is sufficient to represent RFU values. The size of BSU row given by $\lceil(\log_2 \mid C \mid)\rceil \mid I_C \mid$ depends on the cluster size and number of cluster users. Two bytes are sufficient to convey the QoS requirements of each of the users $i \in I_g^T \setminus I_g^{T\sim}$. The ΔP_e values can be expressed using a byte each for mantissa and exponent. Analysis in [87] shows that the gains due to adding a BS to the cluster significantly decreases when $\mid C \mid > 4$. Assuming $\mid C \mid = 4$ and $\mid I_C \mid = 300$, the BSU row, RFU byte, ΔP_e value and QoS information will account for $600 + 8 + 16 + 16* \mid I_g^T \setminus I_g^{T\sim} \mid$ bits. Assuming 0.5uW [83] is consumed for every bit transmitted over the backhaul, number of iterations is $\mid C \mid = 4$ and total number of users transferred $\mid I_g^T \setminus I_g^{T\sim} \mid = 35$ (Fig. 4.6b, high load), then the overhead due to information exchange for Co-RFSnooze is 2.368mW. Note that the overhead due to information exchange in iterations has been accounted in the calculation of P_C for the Co-RFSnooze algorithm in Section 4.4.2.

The time scale of BS resource allocation is of the order of milliseconds as current LTE standards allows BS resource allocation every time slot (1ms duration) in a transmission frame. UA adaptation requires user transfer/handover from the transferor BS to the transferee BS. In the work presented in this chapter, it is assumed that the cluster BSs are connected via X2 interface and X2 handovers can be used to achieve the user transfer. Experiments in [97] show that the X2 handovers can take up to 100ms. Therefore, the time required for BS resource adaptation is about f times ($f = 10$ with the values considered) lesser than that required for UA adaptation and results in a two time scale system. The Co-RFSnooze algorithm accommodates the two time scale requirement as follows. Steps 4-5 in Table 4.4 are carried out at periodicity of p_{BR} at individual BSs

Table 4.5. Simulation parameters

Power gradient Δ_p	4.2
Off power P^O, Idle Power P^I	82.75W, 186W
PA switching power P^{Sw}, switching time t^{Sw}	100W, 35us
Maximum transmit power P^{Max}	40W
Bandwidth BW, Number of frequency blocks J	20MHz, 100
Duration of frame t^F, Number of time slots T	10ms, 10
Number of RF chains at BS S and user device R	4, 4
Set of modes M, $\mid M \mid$	{(1,1,1) (SISO), (2,2,2) (SM), (2,2,1) (SD), (4,1,1) (SD), (4,4,4) (SM), (4,2,2) (SM-SD)}, 6
Size of cluster $\mid C \mid$	4
Maximum number of cluster users	300
$BLER^{Th}$ for all cluster users	0.1
Simulation time	24 hours

to adapt BS resource utilization. At periodicity $f \cdot p_{BR} > p_{BR}$, all the iterations of the algorithm executing all the steps in Table 4.4 are carried out to determine the BS resource allocation and UA of cluster BSs. In Section 4.4.2, we evaluate the performance of Co-RFSnooze algorithm at a single time scale using the sample load trace from anonymous operator with granularity of 1 minute. We have chosen a single time scale of 1 minute ($f \cdot p_{BR}$) as it satisfies the time scale requirements of both the adaptations as well reduces the overhead due to user transfer and allows evaluation of the Co-RFSnooze performance in its entirety, i.e, execute all the iterations at every point of the trace. Note, however, the evaluation can be easily extended to show the two time scale operation of Co-RFSnooze.

4.4 Simulation Framework and Results

4.4.1 Simulation Framework

In this section, we describe the simulation framework developed and the simulation parameters listed in Table 4.5. We adopt the topology with 15 BSs in $4.5x4.5km^2$ [98], a part of 3G network in urban environment. The inter-cell distance is 0.5km. The cluster size $\mid C \mid$ is set to 4 and a 16^{th} BS is randomly placed in the considered 15 BS topology to obtain 4 clusters. Without loss of generality, we consider one of the four clusters to evaluate the proposed Co-RFSnooze algorithm. The BS power model presented in Section 4.3.2 is used to estimate the average BS and cluster power consumption in a frame. The BS power consumption parameters are specified in [89] and [88] and listed in Table 4.5. The users (maximum 300) are uniformly and randomly distributed in the cluster. The traffic load is assumed to be spatially heterogeneous with user's required rate $\gamma \propto (\max(d) - d^2)$ where d is the distance between the user and BS. The $BLER$ LUT table in [94] is extended to include the modes (4,4,1) and (4,4,4) and used to determine the $BLER$ of users as explained in Section 4.3.2. Other parameters for the simulations follow the suggestions in the LTE specifications [90]. We consider the COST-231 HATA model for the path loss between the BS and user [99].

For comparing the performance of Co-RFSnooze algorithm, we consider the following algorithm/schemes (Section 4.1.1):

- All-On (conventional scheme): turns on all BS RF chains in active time slots and turns off in off slots.
- RFSnooze [76]: adapts number of active RF chains, time slots and frequency blocks at individual BSs in an uncoordinated manner. RFSnooze [76] has been extended to Co-RFSnooze algorithm in this work.

- Co-Nap [81]: adapts the on/off pattern of the cluster BSs and turns off all BS RF chains to switch off BSs. The short time scale operation of BS switching effected by switching on/off all RF chains in a cooperative manner without using CoMP transmission makes Co-Nap the most relevant prior art technique for comparison.
- Exhaustive Search: yields the combination that switches off the optimal number of RF chains

We will now discuss the implementation details of All-On and Co-Nap. The UA rule for All-On and Co-Nap schemes is that the user is associated with that BS which provides the highest SINR. The scheduling algorithm [93] (Section 4.3.2, [76]) is used to determine the feasible set of modes M^{FS}. As all the RF chains are switched on during the active time slots for All-On and Co-Nap, the mode that utilizes all the RF chains and satisfies the minimum throughput and BLER constraints is selected from the feasible mode set. If the QoS constraints are not satisfied by modes utilizing all the RF chains, then the mode with next highest number of RF chains that satisfies the QoS constraints is selected. The dominant operation in mode selection is determination of M^{FS} and is carried out as explained in Section 4.3.2, [1] for All-On, Co-Nap and RFSnooze. Hence, the the complexity of mode selection for All-On and Co-Nap is given by $\mid M \mid O(\mid I_C \mid)$ (Section 4.3.2).

In case of All-On and Co-Nap, RF chains that are not transmitting in active time slots (in a frame) are in the idle state and by the UA rule, the set $I_b^T = \emptyset, I_b = I_b^{NT} \forall b \in C$. Incorporating the above in to (4.5), the BS average power consumption in a frame is

$$P = \frac{1}{t^F}(\sum_{t=1}^{T^A} SP^I + \frac{\Delta_p P^{Max}}{J} \sum_{s=1}^{S_t^A} \sum_{i=1}^{|I^{NT}|} J_{sti}) + t^O SP^O \qquad (4.21)$$

All-On does not adapt switching of BSs and RF chains. In contrast, Co-Nap adaptively switches on/off BSs and impacts the average power consumption of the cluster as briefly explained below. Co-Nap divides the transmission time into discrete transmission cycles comprising of $| C |$ number of blocks. The BS on/off (flickering) pattern determines the active and inactive (napping) blocks for all the BSs in every transmission cycle. The BS resource allocation is carried out for all the active blocks in a manner that the user QoS requirements are satisfied. Assuming that a block spans over multiple frames, P_b in a frame in an active time block is given by (4.21). For a frame in an inactive block (BS off), (4.21) reduces to SP^O (as $t^O = t^F$). For Co-Nap, the complexity of determining the on/off (1/0) pattern for $| C |$ BSs in $| C |$ blocks and BS resource allocation for $| I_C |$ cluster users is given by $| C | O(2^{|C|}) + | M | O(| I_C |)$.

4.4.2 Simulation Results

We will now present the experimental results obtained using the simulation framework described above. In order to evaluate the performance of the comparison schemes and the proposed algorithm in a practical setting, we adopt the sample traffic trace shown in Fig. 4.5a. The sample traffic trace is the normalized BS utilization measured by an anonymous operator in [100] for 24 hours with granularity of 1 minute. The simulation step is fixed as 1 minute, however, our simulation framework supports simulation step lesser than or greater than 1 minute. Fig. 4.5b shows the number of users in a simulation step. It is given by the product of value of the sample trace and maximum number of cluster users (Table 4.5). Assuming that the number of users and their requirements do not change over the simulation step, the comparison schemes/algorithms and Co-RFSnooze algorithm is run once in every simulation step to determine the BS resource allocation for all the frames and in case of Co-RFSnooze, additionally, the updated UA. The P_C in a simulation step is the power consumption averaged over all the frames in a simulation step and is estimated using (4.6) for the proposed algorithms and using (4.21) in (4.6) for All-On.

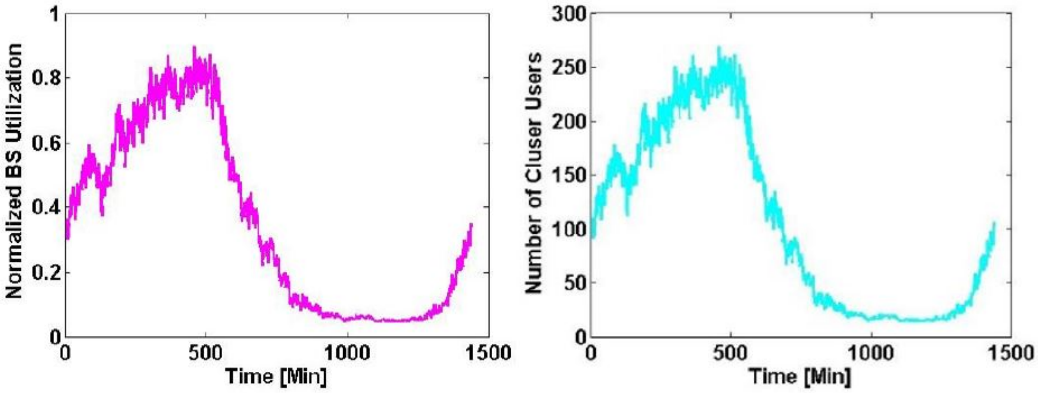


Figure 4.5. (a) Sample traffic trace, (b) number of cluster users

For Co-Nap, the simulation step is equivalent to the transmission cycle and consists of $| C |= 4$ blocks of equal duration. Co-Nap is run once every simulation step to determine the number of active blocks and resource allocation for all the frames in the active blocks. The P_C in a simulation step is equal to the power consumption averaged over the four blocks.

Fig. 4.6a shows the average power consumption of the cluster in a frame P_C for All-On (shown in red), RFSnooze (shown in blue) and Co-RFSnooze (shown in green). All-On consumes higher power than proposed algorithms because, regardless of the load, all the RF chains are on in the active time slots. This increases total RF chain power consumption due to (a) frequency utilization of each active RF chain and (b) idle power of the RF chain transceiver circuitry as all RF chains are either in active or idle state. Joint adaptation of number of active RF chains, frequency and time utilization reduces the cluster power consumption for RFSnooze. The green plot in Fig. 4.6a shows that the savings due to RFSnooze is further extended by Co-RFSnooze. This increase in power savings validates our extension of RFSnooze to Co-RFSnooze which, as elaborated in Section 4.3.4, integrates BS resource adaptation and UA to maximize the number of cluster RF chains that can be switched off. Under high load conditions, RFSnooze

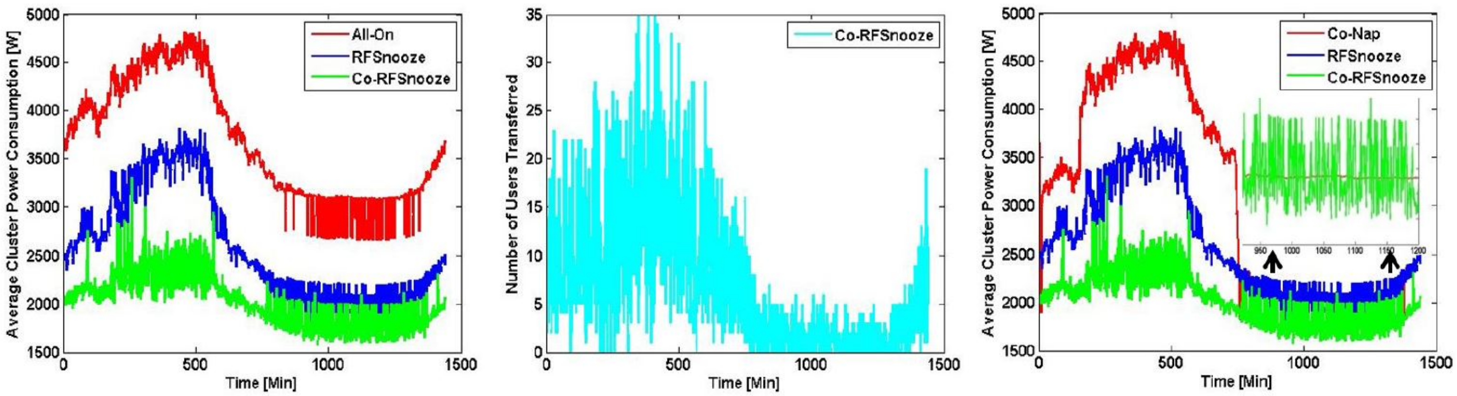


Figure 4.6. (a) comparison of average cluster power consumption of RFSnooze and Co-RFSnooze with that of All-On, (b) number of users transferred by Co-RFSnooze, and (c) comparison of average cluster power consumption of RFSnooze and Co-RFSnooze with that of Co-Nap

achieves up to 35% gains (635^{th} minute) and Co-RFSnooze achieves up to 56% gains (382^{nd} minute) compared to All-On. RFSnooze achieves up to 42% gains (1151^{th} minute) and Co-RFSnooze achieves 49% gains (960^{th} minute) compared to All-On under low load conditions. Note that we refer to the savings in average cluster power consumption as the gains achieved.

We will now compare the performance of RFSnooze and Co-RFSnooze using Figs. 4.6a and 4.6b. Fig. 4.6b shows the number of users transferred by Co-RFSnooze during UA adaptation. Under high load conditions, Fig. 4.6b shows that higher number of users is transferred (up to 35) and Fig. 4.6a shows that Co-RFSnooze achieves up to 43% savings (382^{nd} minute) compared to RFSnooze because higher number of user transfers allows switching off of additional RF chains (Section IIIB,C). Under low load conditions, Co-RFSnooze achieves lower savings of up to 29% (960^{th} minute) because (a) higher number of RF chains are switched off at individual BSs by RFSnooze (b) the number of cluster users (Fig. 4.5b) and transferred users is lower as shown in Fig. 4.6b and (c) higher incidence of instances when no users are transferred resulting in identical performance of RFSnooze and Co-RFSnooze as indicated by corresponding instances in Fig. 4.6a.

Fig. 4.6c shows the P_C due to Co-Nap (shown in red), RFSnooze (shown in blue) and Co-RFSnooze (shown in green). Under high load, Co-Nap performance is comparable to All-On as it is unable to allow BSs to nap and satisfy the QoS constraints. RFSnooze achieves up to 35% gains (635^{th} minute) and Co-RFSnooze achieves up to 56% gains (382^{nd} minute) compared to Co-Nap under high load conditions. During transition from high load to low load and vice versa, Fig. 4.6c shows the dips in power consumption for Co-Nap (for instance between 50^{th} and 150^{th} minute) as lower load allows napping of BSs. RFSnooze and Co-RFSnooze outperform Co-Nap even in the transition regions by adapting BS resources and jointly adapting BS resources and UA

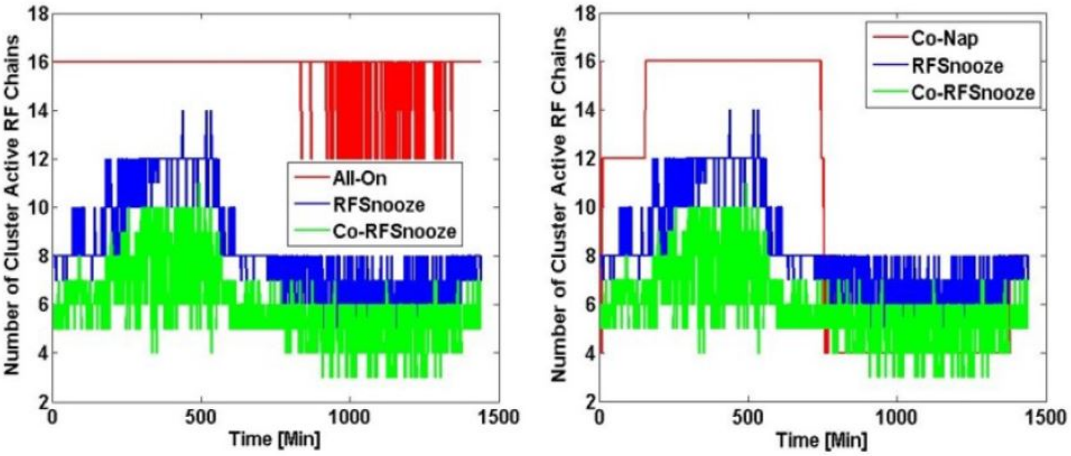


Figure 4.7. Comparison of number of cluster active RF chains of RFSnooze and Co-RFSnooze with (a) All-On, and (b) Co-Nap

respectively. The percentage of gains is lower compared to that under high load conditions at 22% (140^{th} minute) for RFSnooze and 38% (72^{nd} minute) for Co-RFSnooze. Under low load, Co-Nap outperforms RFSnooze as it is able to aggressively nap BSs and satisfy the QoS constraints. Co-RFSnooze outperforms Co-Nap whenever user transfers are possible which allows it to switch off additional RF chains. However, as explained earlier, whenever user transfers are not possible, Co-Nap outperforms Co-RFSnooze. The above behavior of Co-RFSnooze compared to Co-Nap is shown in the inset (zoomed-in section between 900^{th} and 1200^{th} minute) of Fig. 4.6c wherein the green curve repeatedly goes above and below the red curve. Also, due to the bulk of the savings coming from RFSnooze under low load, which underperforms Co-Nap, Co-RFSnooze achieves up to 11% (960^{th} minute) compared to Co-Nap.

Next, we will compare the number of cluster active RF chains used by the proposed algorithms with that used by All-On and Co-Nap in Figs. 4.7a and 4.7b respectively. The number of cluster active RF chains in (a) a frame is the sum of the active RF chains used at individual BSs and (b) a simulation step is the number of cluster active RF chains averaged over all the frames in the simulation step.

Table 4.6. Average percentage savings in P_C of RFSnooze and Co-RFSnooze

	Low Load	High Load	Total
RFSnooze vs All-On	32.74%	26.21%	30%
Co-RFSnooze vs All-On	41.5%	47.38%	44.67%
RFSnooze vs Co-Nap	-16.1%	26%	7.68%
Co-RFSnooze vs Co-Nap	-0.86%	47.25%	25.52%

In Fig. 4.7a, all the cluster BS RF chains are active for All-On under high load whereas RFSnooze uses lesser number of RF chains and the least number are used by Co-RFSnooze. Under low load conditions, there are dips in the number of BS RF chains for All-On because there are no users associated with certain BSs in that instance and we see corresponding dips for RFSnooze and Co-RFSnooze as well. Fig. 4.7b shows that all the cluster RF chains are active for Co-Nap when the load is high as napping of BSs is not possible. Under low load, Co-Nap aggressively reduces the number of RF chains and thereby the power consumption as observed in Fig. 4.6c. RFSnooze consumes higher power than Co-Nap under low load conditions because it uses higher number of RF chains, as is evident from Fig. 4.7b. Further, we can see that the number of active RF chains used by Co-RFSnooze repeatedly goes above and below the number of RF chains used by Co-Nap. This results in similar pattern of P_C of Co-RFSnooze in Fig. 4.6c. During the transition from low load to high load and vice versa, the number of RF chains for RFSnooze and Co-RFSnooze is lower than that of Co-Nap. This is the cause for the trend of P_C of Co-Nap, RFSnooze and Co-RFSnooze during transition periods as seen in Fig. 4.6c.

Table 4.6 presents the percentage of savings in P_C, averaged over 24 hours, for the proposed algorithms with respect to All-On and Co-Nap. Co-RFSnooze outperforms both All-On and Co-Nap when the savings are averaged over 24 hours which includes periods of low, medium and high loads.

We conclude the results by presenting the comparison of Co-RFSnooze and

Table 4.7. Average percentage savings in P_C of Co-RFSnooze compared to Exhaustive Search

	Low Load	Medium Load	High Load
Co-RFSnooze vs Exhaustive Search	0%	-13%	-18%

exhaustive search in Table 4.7. The simulation framework and parameters used is identical to that used for the remaining experiments except the following two changes. As the computational complexity of exhaustive search is exponential in $\mid I_C \mid$ (Section 4.3.5), to keep the simulation time tractable, we have chosen (a) the number of cluster users $\mid I_C \mid= 100$ and (b) low, medium and high load points of $0.1, 0.5, 0.8$ of the sample trace in Fig. 4.5a and the resulting number of users are $10, 50, 80$. We have conducted three runs of Co-RFSnooze and Exhaustive search for each of the load points and report the average percentage savings in P_C of Co-RFSnooze compared to exhaustive search in Table 4.7. The deviation of the Co-RFSnooze P_C from the optimal value achieved by exhaustive search is at most 18% at high load.

4.5 Summary

In this chapter, we presented novel RF switching technique to minimize the average power consumption of a cluster of BSs in a transmission frame while satisfying the cluster users' QoS requirements and BS utilization constraints. Simulation results indicate that the proposed algorithms significantly outperform the conventional All-On scheme while Co-RFSnooze significantly gains over time slot based adaptive BS switching scheme Co-Nap under high and medium loads while being comparable under low load conditions.

In the next chapter, we conclude the thesis with future directions for the work carried out.

4.6 Acknowledgements

We thank the anonymous ICC 2015 and IEEE Transactions on Green Communications and Networking reviewers for their feedback and comments on the work.

Chapter 4, in part, contains material as it appears in the Proceedings of the Internation Conference on Communications (ICC'15). "Power-efficient Base Station Operation through User QoS-aware Adaptive RF Chain Switching Technique". Ranjini Guruprasad, Kyuho Son, Sujit Dey. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, contains material as it appears in IEEE Transactions on Green Communications and Networking. "User QoS-aware RF Chain Switching for Power Efficient Co-operative Base Stations".Ranjini Guruprasad, Sujit Dey. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Conclusion

Cellular networks and mobile devices continue to evolve to offer high bit rates, extensive coverage and processing advanced multimedia applications. Anytime-anywhere connectivity with high data rates and capability to process advanced multimedia applications have revolutionized important sectors of the society and become an integral part of lifestyle of urban and rural populations across the world. The result of the advances in cellular networks and mobile devices is a continued explosive growth in number of mobile subscriptions and volume of mobile data. In this dissertation, we address the challenges in effective and efficient utilization of power/energy resources required to cater to the explosive growth in number of subscriptions and volume of mobile data in an economically and environmentally sustainable manner.

Mobile video is the leading multimedia application and contributes to about two thirds of mobile data traffic. Also, it is a data and compute intensive application which results in significant demands on the components involved in video download and processing the video data. It is shown that the components involved in download consume higher energy than that required for processing. While there is a strong body of research which address the energy consumption due to processing of video data, there is a little research which aims to minimize the battery drain due to video download. In Chapter 2, we developed battery aware techniques for video download and streaming.

We provided discussion on the power models of base station components involved in video transmission and mobile device components involved in video download as well as playback. Also, included in the discussion are bit error rate, channel and user experience and consumption models and battery models to complete the modeling of the ecosystem of video transmission and video download and playback. We proposed two techniques (BR-MoDS and B^{2}R-MoDS) that are applicable for mobile video download and video streaming. Further, we also proposed novel Video Experience Longevity metric which quantifies the gain in battery lifetime and user experience compared to non-battery aware video download and streaming techniques. Experiments showed that the proposed battery aware video download and streaming techniques offer significant savings in battery lifetime compared to non-battery aware video download and streaming techniques with comparable user experience. Further, higher VEL metrics for the proposed BR-MoDS and B^{2}R-MoDS techniques demonstrate that there exists gain in battery lifetime as well as video experience compared to the non-battery aware techniques.

It would be interesting to extend the mobile battery aware techniques to jointly optimize the BS power consumption during video transmission and the battery consumption of mobile devices during video download. Further, such techniques can enable the video consumer to set the priority levels for prioritizing battery drain versus the user experience. On the same lines, a guarantee on battery drain extent can be an extra dimension in pricing of data plans by the video content providers.

Moving from the mobile devices to cellular networks, we identified that reducing power consumption of BSs at the system level is critical for energy efficient operation of cellular networks. In Chapter 3, we developed the an integrated framework for dynamic cell reconfiguration for minimizing the power consumption of cellular networks while satisfying the user Quality of Service (QoS). The dynamic cell reconfiguration framework integrates three techniques namely, BS switch off/on, user association and transmit power budget adaptation. We discussed user QoS, network, channel and BS power consumption models followed by algorithm presentation. We evaluated the proposed dynamic cell reconfiguration techniques and framework for static and dynamic traffic load conditions using actual measurements of BS power consumption and real world traces of BS load. Experiments show that the proposed framework significantly reduces the power consumption of cellular networks while satisfying the QoS requirements of associated users.

An important extension to the work would be the joint optimization of set of active BSs, user association and transmit power budget allocation and energy consumption of the mobile devices. This would result in an end-end framework for energy efficiency of cellular networks and mobile devices. Further, by exploiting the heterogeneity of macro, micro and pico base stations, coverage holes created by switching off macro BSs can be alleviated by offloading users to micro and pico base stations. Such user association presents interesting challenges in tradeoff between increase in power consumption of micro and pico base stations, decrease in macro BS power consumption and satisfying user QoS requirements.

Lastly, we identified that reduction of power consumption of BSs at the component level enables adaptation to load variations on time scale of seconds and minutes. In chapter 4, we discuss the component level BS power consumption model which is centered around the various states of power amplifies in the radio frequency (RF) chain and its impact on the power consumption of other components of the BS. The channel, network and user QoS models were also presented to model the cellular network and associated users interaction and resulting BS power consumption. We first proposed the RFSnooze technique which adapts the number of RF chains, time slots and frequency blocks to minimize the number of active RF chains and thereby, the BS power consumption. The RFSnooze technique achieves the objective while ensuring that the BS resource utilization bounds and user QoS requirements are satisfied. We extended the RFSnooze technique to Co-RFSnooze technique which adapts the number of Rf chains in a cluster of cooperating base stations to minimize the power consumption of cluster of BSs. Co-RFSnooze technique achieves the above by jointly adapting the individual BS resources and user association of users in the cluster of BSs while ensuring that the BS resource utilization bounds and cluster user QoS requirements are satisfied. Experiments using measurements from actual BS component power consumption and real world traces of BS load demonstrate that the proposed Co-RFSnooze technique achieves significant savings in cluster power consumption with no degradation in user QoS levels.

Adaptive RF chain switching to minimize the number of active RF chains can be extended to minimize the number of RF chains in massive MIMO systems which have been identified as one of the key enablers of the 5G cellular networks. Further, the optimization of BS resource utilization and cluster user association presents interesting research problems when the ecosystem includes renewable energy and energy storage systems.

Appendix A

Battery Aware Video Download Techniques

A.1 Battery Efficient Video Download - Framework

The overall framework for information and control data exchange between base station and mobile device, mode selection and reconfiguration during battery efficient video download is shown in Fig. A.1. Each download epoch, T_i consists of the following events, represented by the time duration; (a) T_{Video} - video data transmission by the base station, (b) T_{Status} - channel condition and buffer level status update sent by mobile device, (c) $T_{Mode-Sel}$ - mode selection performed by base station (executing MoDS algorithm) based on mobile device status update, (d) T_{Mode} - mode selected communicated by base station and (e) $T_{Mode-Config}$ - reconfiguration of RF and base band components of base station and mobile device according to the mode selected. The mobile device status update, mode selection and communication and mode reconfiguration are carried out in advance in the current download epoch T_i for the next download epoch T_{i+1}. This ensures that video is transmitted continuously except during T_{Mode}, $T_{Mode-Config}$ and when MoDS selects download idle. We will next discuss the mobile device status update and base station mode update in detail. The Buffer Status Report (BSR) [90] used to report the uplink buffer level reports the mobile device buffer level during download

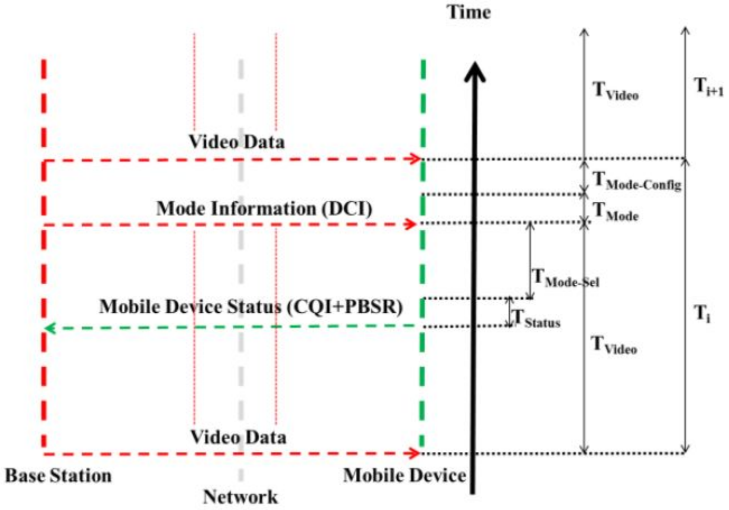


Figure A.1. Overall Framework for battery efficient video download

(downlink) as described below. When the video download session is initiated, Long Buffer Status Report (L-BSR, 3 bytes) conveys video bit rate (V_{BR}) and Short BSR (S-BSR, 1 byte) conveys the mobile device buffer size. As S-BSR is insufficient to report the buffer size in bytes, the buffer size is reported in terms of maximum playback time PBT possible which is the ratio of buffer size, Buf_{Size} to V_{BR}. As V_{BR} and Buf_{Size} are constant for a given video, this information is sent one time at the beginning of the download session. Subsequently, for each download epoch, the buffer level is reported in terms of available PBT calculated using (17) using periodic S-BSR. The MoDS algorithm uses the V_{BR}, Buf_{Size} and PBT information obtained from the BSR to calculate DR^{Min} and DR^{Max}. The periodic Channel Quality Indicator (CQI) [90] reports the channel condition required to obtain the BER values for the modes. Finally, the mode selected is communicated by the base station to the mobile device using the Downlink Control Information (DCI, format 1A, 2A, about 8 bytes) [90]. It should be noted that CQI and DCI information exchange is currently carried out as part of the LTE specifications [90] and the additional data transmitted for buffer levels is nominal – a byte resulting in 1.14mW of power consumption [101] and imposes no more than 0.4% of power overhead even when the mode (1x1, BPSK, CR=1, ZF) is selected. On the other hand, receiving 8 bytes of DCI results in about 2.22μW of power consumption when the mode (1x1, BPSK, CR=1, ZF) is used.

A.2 Scalability Analysis for MoDS under Multi - Client Scenario

In our proposed approach each base station has an instance of MoDS. At the network level, as the number of clients in a network using MoDS grows, they will be geographically distributed across multiple cells of the network, hence will use multiple instances of MoDS associated with the corresponding base station. At the base station level, empirical evidence suggests that the number of clients streaming video concurrently will be rather limited. For example, the study conducted by Motorola in [101] indicates that even with LTE networks, no more than 8 video clients can be supported while downloading video with bit rate of 3.5Mb/s (the high bit rate per video stream reflecting the growing trend of watching higher resolution videos).

As elaborated in Appendix A.1 and shown in Fig. A.1, mode selection time, $T_{Mode-Sel}$ is the time taken by MoDS to select mode for each user in every download epoch of duration, T_i and by studying how $T_{Mode-Sel}$ varies with number of concurrent video clients, we can analyze the performance of a single instance of MoDS running on a single base station serving multiple concurrent video clients. We next present experiments devised to measure $T_{Mode-Sel}$, for each user in the presence of increasing number of concurrent video clients that the base station has to service. We simulate each concurrent user streaming the same video of 183s duration encoded using bit rate of 4.1Mb/s and a snacking ratio of 0.5, under variable SNR conditions (Table 2.5, Section 2.4.4). Note we assume all the users downloading the same video to remove any variability that could arise because of video characteristics, but the experiments can be easily conducted with

Table A.1. $T_{Mode-Sel}$ for each video client in the multi-client scenario

Number of concurrent video clients	Maximum - $T_{Mode-Sel}$(ms)	Mean - $T_{Mode-Sel}$(ms)
1	186.6	148.3
2	371.1	302.7
3	540.7	416.7
4	752.6	590.5
5	857.7	685.3
6	1039.3	819.5
7	1228.9	1074.5
8	1403.7	1289.7
9	1620.9	1447.1
10	1731.3	1636.5
11	1918.3	1784.1
12	2114	1959.6

concurrent users streaming arbitrarily different videos as well. Table T1 below shows the maximum and mean values of $T_{Mode-Sel}$ when number of concurrent video clients is varied from 1 to 6, with the MoDS algorithm running on an Intel Core i7-3632QM CPU operating at 2.2GHz. As expected the time taken by MoDS for each client increases with increasing number of clients. The number of concurrent clients that can be served by a single CPU will be limited by what can be the allowable value for $T_{Mode-Sel}$, which as shown by Fig. A.1, Appendix A.1 is upper-bounded by the duration of each download epoch T_i minus time needed to send status update from the mobile device, T_{Status}, time needed to send the selected mode to the mobile device, T_{Mode} and time needed for mode reconfiguration, $T_{Mode-Config}$. For example, since the download epoch duration value used for the experiment results reported Section 2.4.5 is 2s, it is reasonable to say $T_{Mode-Sel}$ can be up to 1400ms, which means up to 8 concurrent video clients ([102] shows that LTE networks can support up to 8 concurrent video clients) can be served by each Intel Core i7-3632QM CPU, thus requiring 1 such CPU to support 8 concurrent users (at a list price of under $400 per base station).

A.3 Computational Complexity Analysis and Comparison – Video Download Techniques

In order to compare the computational complexity of the three video download techniques, namely, HTTP-PD, EERA and MoDS (considered in Section 2.4.5), we will first analyze computational complexity of each technique in terms of the mode parameters and dimension of the configuration space. As listed in Table 2.4, the mode parameters are channel coding rate, CR, modulation schemes Mod, MIMO encoding rate $MIMO_{Enc-Rate}$, number of transmit antennas N_T, number of receive antennas N_R, MIMO decoding algorithms $MIMO_{Dec}$ and channel decoding algorithms Ch_{Dec}. We denote the number of choices available for coding rate, modulation, MIMO encoding rate, MIMO decoding algorithms and channel decoding algorithms as N_{CR}, N_{Mod}, $N_{MIMO-Enc}$, $N_{MIMO-Dec}$ and N_{Ch-Dec} respectively. For HTTP-PD, the entire search space is traversed for the mode that satisfies the maximum download rate, DR^{Max} and application BER, BER_{App} with computational complexity $O(N_{CR} \cdot N_{Mod} \cdot N_{MIMO-Enc} \cdot N_T \cdot N_R \cdot N_{MIMO-Dec} \cdot N_{Ch-Dec})$. As elaborated in Section 2.3.3, the computational complexity of MoDS is $O(N_{MIMO-Enc} \cdot N_{MIMO-Dec} \cdot N_{Ch-Dec} \cdot I_M \cdot IS_{Dim}^3)$ where I_M denotes the number of iterations required by the optimization tool ”nlopt” to determine the mode that imposes minimum load on battery while satisfying the download rate and application BER constraints and IS_{Dim} denotes the dimension of inner space (consisting of CR, Mod, N_T and N_R). It should be noted that the dimension of inner space is constant and is equal to 4 and from our experiments; we have observed that the typical value of I_M is 5. The computational complexity of EERA is $O(log(N_{CR} \cdot N_{Mod}) \cdot N_{MIMO-Enc} \cdot N_R \cdot N_{MIMO-Dec} \cdot N_{Ch-Dec})$ [20]. From the above discussion, we can see that the computational complexities for the three techniques have certain similar and also different factors; hence it will be difficult to completely compare the complexities. However, when we consider typical values for the parameters as follows, $N_{CR} = 16$ [103], $N_{Mod} = 4$, $N_{MIMO-Enc} = 14$ [103], $N_T = 8$, $N_R = 8$, $N_{MIMO_{Dec}} = 2$, $N_{Ch_{Dec}} = 2$, $IS_{Dim} = 4$ and $I_M = 5$ and compare the complexities for MoDS and EERA, we can see that number of steps required for MoDS is 17920 and that for EERA is 1863. We can therefore conclude that MoDS has higher computational complexity than EERA.

A.4 Computational Complexity Analysis and Comparison – ABR Streaming Techniques

In order to compare the computational complexity of the ABR streaming techniques, namely, ABR-DASH, BaSe-AMy, BR-MoDS and B^{2}R-MoDS, we will first analyze computational complexity of bit rate selection followed by that of download rate and mode selection. Bit rate adaptation by all the techniques considered involves selecting a bit rate version based on certain conditions from a list of 'n' bit rates (n=7 and is the same as the cardinality of the set $V_{BR-ValidSet}$ in our experiments). The reference ABR-DASH technique proposed in [21] uses a multiplicative factor (determined using certain heuristics) to scale down/up the bit rate of the current segment and determine the approximate bit rate for the next segment. BR-MoDS and B^{2}R-MoDS select the approximate bit rate for the next segment depending on the constraints specified in (2.23) and (2.32) and require I_{BR} (number of iterations required by the solver to determine the approximate bit rate, Section 2.5.3) steps. Note that from our experiments, we have observed that I_{BR} has a maximum value of 2. Given the approximate bit rate determined by ABR-DASH, BR-MoDS and B^{2}R-MoDS, selection of valid bit rate from $V_{BR-ValidSet}$ (Table 2.6) requires $logn$ steps resulting in computational complexity of $O(\log n)$ for ABR-DASH and $O(I_{BR} + \log n)$ for BR-MoDS and B^{2}R-MoDS. BaSe-AMy on the other hand has constant complexity ($O(1)$) as it uses certain logic to select either the bit rate positioned above or below the bit rate of the current segment in $V_{BR-ValidSet}$ and does not require to traverse the entire $V_{BR-ValidSet}$. We can therefore conclude that for bit rate selection, BaSe-AMy has lowest computational complexity followed by ABR-DASH and then BR-MoDS and B^{2}R-MoDS. Subsequent to bit rate selection, download rate and mode is selected for every download epoch constituting the segment and we will next discuss the computational complexity of mode selection for all the techniques. For ABR-DASH and BaSe-AMy, the entire search space is traversed to find the mode that satisfies the maximum download rate, DR^{Max} and application BER, BER_{App} with computational complexity $O(N_{CR} \cdot N_{Mod} \cdot N_{MIMO-Enc} \cdot N_T \cdot N_R \cdot N_{MIMO-Dec} \cdot N_{Ch-Dec})$. As BR-MoDS and B^{2}R-MoDS uses MoDS for mode selection, the computational complexity is $O(N_{MIMO-Enc} \cdot N_{MIMO-Dec} \cdot N_{Ch-Dec} \cdot I_M \cdot IS^3_{Dim})$ (Section 2.3.3). From the above discussion, we can infer that the BR-MoDS and B^{2}R-MoDS have lower computational complexity for mode selection than ABR-DASH and BaSe-AMy as BR-MoDS and B^{2}R-MoDS do not have to traverse the entire configuration space. Also, given that n typically has values in the range 7-10 [104] - [105] and the mode selection parameters have values in the range 2-16 (Appendix A.2), we can infer that the computational complexity of mode selection (for instance, the number of steps computed for BR-MoDS using the parameters values in Appendix A.2 is 17920) is significantly higher than that of bit rate selection (for instance BR-MoDS requires $2 \cdot \log 7 = 2.98 steps$). Therefore the ascending order of the techniques in terms of their computational complexities is BR-MoDS and B^{2}R-MoDS, BaSe-AMy and ABR-DASH.

Appendix B

Dynamic Cell Reconfiguration Framework

B.1 Proof of Theorem 3.3

Proof. The problem given in (3.14) is a convex optimization because its feasible set $\mathcal{F}(\mathcal{B}_{on}, p)$ is convex (from Lemma 3.2) and the objective function is also convex (due to the summation of the linear function of ρ_i and convex function $L_i(\rho_i)$). Hence, it is sufficient to show that,

$$\langle \nabla((1-q_i)P_i\rho_i + L_i(\rho_i)), \boldsymbol{\rho} - \boldsymbol{\rho^*} \rangle \geq 0 \tag{B.1}$$

for all $\rho \in \mathcal{F}(\mathcal{B}_{on}, p)$. Let $p_i(x)$ and $p_i^*(x)$ be the associated probability vectors for ρ and ρ^*, respectively. Then, (3.17) generates the deterministic cell coverage, and thus the association rule is given by

$$\pi^*(x) = \mathbf{1}\left\{ i = \underset{j \in \mathcal{B}_{on}}{argmax} \frac{c_j(x)}{(1-q_j)P_j + L_j^{`}(\rho_j^*} \right\} \tag{B.2}$$

and then the inner product (B.1) can be computed as

$$
\begin{aligned}
&\sum_{i\in\mathcal{B}_{on}}((1-q_i)P_i+L_i^{\sim}(\rho_i^*))(\rho_i-\rho_i^*) = \\
&\sum_{i\in\mathcal{B}_{on}}(1-q_i)P_i+L_i^{\sim}(\rho_i^*)\int_{\mathcal{L}}\frac{\gamma(x)}{c_i(x)}(\pi_i(x)-\pi^*(x))dx \\
&=\int_{\mathcal{L}}\gamma(x)\sum_{i\in\mathcal{B}_{on}}\frac{(1-q_i)P_i+L_i^{\sim}(\rho_i^*)}{c_i(x)}(\pi_i(x)-\pi_i^*(x))dx
\end{aligned}
\tag{B.3}
$$

It is clear that the inequality

$$
\sum_{i\in\mathcal{B}_{on}}\frac{(1-q_i)P_i+L_i^{\sim}(\rho_i^*)}{c_i(x)}\pi_i(x)\geq\sum_{i\in\mathcal{B}_{on}}\frac{(1-q_i)P_i+L_i^{\sim}(\rho_i^*)}{c_i(x)}\pi^*(x)
$$

holds from B.2. Substituting this inequality into (B.3) yields the condition in (B.1), which completes the proof. □

B.2 Proof of Theorem 3.5

Lemma B.1. $\sum_{i\in\mathcal{B}_{on}}\rho_i$ *is monotonically decreasing as* $\mathcal{B}_{on}$ *increases.* $\sum_{i\in\mathcal{B}_{on}}\rho_i \geq \sum_{i\in\mathcal{B}_{on}\cup\{b\}}\rho_i$ *holds*

Proof. When we additionally turn on a BS $b\in\mathcal{B}\setminus\mathcal{B}_{on}$, some of users will change their association to the new BS b. Let $\mathcal{L}_b$ denote the coverage area of BS b. For those users $x\in\mathcal{L}_b$, according to (3.23), each users will have higher transmission rate $c_i(x)$ (or better SINR) than before turning on BS b. If not, it should have not switched to the BS b. On the other hand, for the other users $x\in\mathcal{L}\setminus\mathcal{L}_b$, the association will remain unchanged. Thus, each user will have the same signal strength $g_i(x)\cdot p_i$ and at the same time will see the same amount of interference based on Assumption III.1. Consequently, there is no change in their transmission rate $c_i(x)$

Recall the definition of BS utilization P in subsection 3.2.2 that $\sum_{i\in\mathcal{B}_{on}}\rho_i$ is equal to the summation of $\lambda(x)/c_i(x)$ for all users $x\in\mathcal{L}$. As discussed above, $c_i(x)$ is higher than (for users $x\in\mathcal{L}_b$) or equal to (for users $x\in\mathcal{L}\setminus\mathcal{L}_b$) before turning on BS b. Note that $\lambda(x)$ is given and fixed. Hence, $\sum_{i\in\mathcal{B}_{on}}\rho_i\geq\sum_{i\in\mathcal{B}\cup\{b\}}\rho_i$ holds, which completes the proof. □

Lemma B.2. *$\sum_{i\in\mathcal{B}_{on}}\rho_i$ is supermodular as a function of $\mathcal{B}_{on}$*

Proof. According to the equivalent definitions of sub/ supermodular set function (see Proposition 2.1 in [68]), it is sufficient to show that the following inequality holds for all $b\in\mathcal{B}\setminus(\mathcal{B}_{on}\cup\{k\})$.

$$d_b(\mathcal{B}_{on})\geq d_b(\mathcal{B}_{on}\cup\{k\})\tag{B.4}$$

where $d_b(\mathcal{A})=\sum_{i\in\mathcal{A}}\rho_i-\sum_{i\in\mathcal{A}\cup\{b\}}\rho_i$ that is a reduction in the summation of BS utilization by adding BS b.

Let us consider two different sets of active BSs $\mathcal{B}_{on}$ and $\mathcal{B}_{on}\cup\{k\}$ and then investigate how the user association will change in each case when an additional BS b is turned on. For the two different sets, the coverage area of BS b is denoted by $\mathcal{L}_b(\mathcal{B}_{on})$ and $\mathcal{L}_b(\mathcal{B}_{on}\cup\{k\})$, respectively. Since the association is based on the best the transmission rate, the former area is a superset of the latter, i.e., $\mathcal{L}_b(\mathcal{B}_{on})\supseteq\mathcal{L}_b(\mathcal{B}_{on}\cup\{k\})$.

Note that there are two types of area: (i) Common area $\mathcal{L}_{comm}=\mathcal{L}_b(\mathcal{B}_{on})\cap\mathcal{L}_b(\mathcal{B}_{on}\cap\{k\})$, where the users switched to BS b for both starting sets, and (ii) difference area $\mathcal{L}_{diff}=\mathcal{L}_b(\mathcal{B}_{on})\setminus\mathcal{L}_b(\mathcal{B}_{on}\cup\{k\})$, where the users could not be switched to BS b because BS k provides better SINR than BS b.

We rewrite the summation from the perspective of users.

$$d_b(\mathcal{B}_{on}) - d_b(\mathcal{B}_{on} \cup \{k\}) \tag{B.5}$$

$$\begin{aligned}
&= \int_{x\in\mathcal{L}} [\frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}} c_i(x)} - \frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}\cup\{k\}} c_i(x)}]dx \\
&- \int_{x\in\mathcal{L}} [\frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}\cup\{k\}} c_i(x)} - \frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}\cup\{k,b\}} c_i(x)}]dx
\end{aligned} \tag{B.6}$$

It is enough to consider the two areas $\mathcal{L}_{comm}$ and $\mathcal{L}_{diff}$ since $\gamma(x)/c_i(x)$ is unchanged in the other area. Thus, the difference can be computed as follows.

$$\begin{aligned}
&= \int_{x\in\mathcal{L}_{comm}} \left[\frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}} c_i(x)} - \frac{\gamma(x)}{c_b(x)}\right]dx \\
&+ \int_{x\in\mathcal{L}_{diff}} \left[\frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}} c_i(x)} - \frac{\gamma(x)}{c_b(x)}\right]dx \\
&- \int_{x\in\mathcal{L}_{comm}} \left[\frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}\cup\{k\}} c_i(x)} - \frac{\gamma(x)}{c_b(x)}\right]dx
\end{aligned} \tag{B.7}$$

$$\begin{aligned}
&\int_{x\in\mathcal{L}_{comm}} \left[\frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}} c_i(x)} - \frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}\cup\{k\}} c_i(x)}\right]dx \\
&+ \int_{x\in\mathcal{L}_{diff}} \frac{\gamma(x)}{\max\limits_{i\in\mathcal{B}_{on}} c_i(x)} - \frac{\gamma(x)}{c_b(x)}\Bigg]dx
\end{aligned} \tag{B.8}$$

The first integral is non-negative because $\max\limits_{i\in\mathcal{B}_{on}} c_i(x) \le \max\limits_{i\in\mathcal{B}_{on}\cup\{k\}} c_i(x)$ holds. The second integral is also non-negative because $\max\limits_{i\in\mathcal{B}_{on}} c_i(x) \le c_b(x)$ holds for all users $x \in \mathcal{L}_{diff}$. This completes the proof. □

Bibliography

[1] Ericsson mobility report. Technical report, Ericsson, June 2017.

[2] UN Global Population. http://www.un.org/en/development/desa/news/population/2015-report.html. Accessed: 2017-09-09.

[3] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok. The global footprint of mobile communications: The ecological and economic perspective. *IEEE Commun. Mag.*, 49(8):55–62, August 2011.

[4] Qingqing Wu, Geoffrey Ye Li, Wen Chen, Derrick Wing Kwan Ng, and Robert Schober. An overview of sustainable green 5g networks. *CoRR*, abs/1609.09773, 2016.

[5] Nokia Solutions and Networks. Flatten network energy consumption. Technical report, 2013.

[6] Battery Tech. http://www.pocket-lint.com/news/130380-future-batteries-coming-soon-charge-in-seconds-last-months-and-power-over-the-air.html. Accessed: 2017-09-09.

[7] R. Mizouni, M. A. Serhani, A. Benharref, and O. Al-Abassi. Towards battery-aware self-adaptive mobile applications. In *2012 IEEE Ninth International Conference on Services Computing*, pages 439–445, June 2012.

[8] Ericsson consumer lab report. https://www.ericsson.com/en/networked-society/trends-and-insights/consumerlab/consumer-insights/reports/tv-and-media-2016. Accessed: 2017-09-09.

[9] Luca Ardito, Giuseppe Procaccianti, Marco Torchiano, and Giuseppe Migliore. Profiling power consumption on mobile devices. 2013.

[10] D. Feng, C. Jiang, G. Lim, L. J. Cimini, G. Feng, and G. Y. Li. A survey of energy-efficient wireless communications. *IEEE Communications Surveys Tutorials*,

15(1):167–178, First 2013.

[11] J. Joung, C. K. Ho, and S. Sun. Power amplifier switching (pas) for energy efficient systems. *IEEE Wireless Communications Letters*, 2(1):14–17, February 2013.

[12] A. S. Y. Poon. An energy-efficient reconfigurable baseband processor for wireless communications. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 15(3):319–327, March 2007.

[13] Y. Xiao, R. S. Kalyanaraman, and A. Yla-Jaaski. Energy consumption of mobile youtube: Quantitative measurement and analysis. In *2008 The Second International Conference on Next Generation Mobile Applications, Services, and Technologies*, pages 61–69, Sept 2008.

[14] K. J. Ma, R. Bartos, S. Bhatia, and R. Nair. Mobile video delivery with http. *IEEE Communications Magazine*, 49(4):166–175, April 2011.

[15] A. Begen, T. Akgul, and M. Baugher. Watching video over the web: Part 1: Streaming protocols. *IEEE Internet Computing*, 15(2):54–63, March 2011.

[16] S. Haykin. Cognitive radio: brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23(2):201–220, Feb 2005.

[17] H. S. Kim and B. Daneshrad. Energy-constrained link adaptation for MIMO OFDM wireless communication systems. *IEEE Transactions on Wireless Communications*, 9(9):2820–2832, September 2010.

[18] R. Hormis, E. Linzer, and X. Wang. Adaptive mode- and diversity-control for video transmission on mimo wireless channels. *IEEE Transactions on Signal Processing*, 57(9):3624–3637, Sept 2009.

[19] H. Kim, C. B. Chae, G. de Veciana, and R. W. Heath. A cross-layer approach to energy efficiency for adaptive mimo systems exploiting spare capacity. *IEEE Transactions on Wireless Communications*, 8(8):4264–4275, August 2009.

[20] Chi-Yu Li, Chunyi Peng, Songwu Lu, and Xinbing Wang. Energy-based rate adaptation for 802.11n. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, Mobicom '12, pages 341–352, 2012.

[21] G. Tian and Y. Liu. Towards agile and smooth video adaptation in http adaptive streaming. *IEEE/ACM Transactions on Networking*, 24(4):2386–2399, Aug 2016.

[22] Chenghao Liu, Imed Bouazizi, and Moncef Gabbouj. Rate adaptation for adaptive http streaming. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, MMSys '11, pages 169–174, 2011.

[23] Apple http live streaming. https://datatracker.ietf.org/doc/draft-pantos-http-live-streaming. Accessed: 2017-09-09.

[24] Microsoft smooth streaming. http://www.iis.net/downloads/microsoft/smooth-streaming. Accessed: 2017-09-09.

[25] F. Molazem Tabrizi, J. Peters, and M. Hefeeda. Dynamic control of receiver buffers in mobile video streaming systems. *IEEE Transactions on Mobile Computing*, 12(5):995–1008, May 2013.

[26] M. Tamai, N. Shibata, K. Yasumoto, and M. Ito. An energy-aware video streaming system for portable computing devices. In *7th International Conference on Mobile Data Management (MDM'06)*, pages 58–58, May 2006.

[27] M. Kennedy, H. Venkataraman, and G. M. Muntean. Battery and stream-aware adaptive multimedia delivery for wireless devices. In *IEEE Local Computer Network Conference*, pages 843–846, Oct 2010.

[28] D. N. Rakhmatov and S. B. K. Vrudhula. An analytical high-level battery model for use in energy management of portable electronic systems. In *IEEE/ACM International Conference on Computer Aided Design. ICCAD 2001. IEEE/ACM Digest of Technical Papers (Cat. No.01CH37281)*, pages 488–493, Nov 2001.

[29] Min Chen and G. A. Rincon-Mora. Accurate electrical battery model capable of predicting runtime and i-v performance. *IEEE Transactions on Energy Conversion*, 21(2):504–511, June 2006.

[30] Oleg I. Atayero, Aderemi A.and Sheluhin and Yury A. Ivanov. *Modeling, Simulation and Analysis of Video Streaming Errors in Wireless Wideband Access Networks*, pages 15–28. Springer Netherlands, Dordrecht, 2013.

[31] O Sheluhin, A. A. Atayero, Y. A. Ivanov, and J. O Iruemi. Effect of video streaming space–time characteristics on quality of transmission over wireless telecommunication networks. In *IEEE/ACM International Conference on Computer Aided Design. ICCAD 2001. IEEE/ACM Digest of Technical Papers (Cat. No.01CH37281)*, Oct 2011.

[32] S. Shah and V. Sinha. Iterative decoding vs. viterbi decoding: A comparison. In

National Conference on Communications, Feb 2002.

[33] He Wu, Sidharth Nabar, and Radha Poovendran. An energy framework for the network simulator 3 (ns-3). In *Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques*, SIMUTools '11, pages 222–230, 2011.

[34] Shuguang Cui, A. J. Goldsmith, and A. Bahai. Energy-constrained modulation optimization. *IEEE Transactions on Wireless Communications*, 4(5):2349–2360, Sept 2005.

[35] D. Garrett, L. Davis, S. ten Brink, B. Hochwald, and G. Knagge. Silicon complexity for maximum likelihood mimo detection using spherical decoding. *IEEE Journal of Solid-State Circuits*, 39(9):1544–1552, Sept 2004.

[36] Chien-Ching Lin, Yen-Hsu Shih, Hsie-Chia Chang, and Chen-Yi Lee. Design of a power-reduction viterbi decoder for wlan applications. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 52(6):1148–1156, June 2005.

[37] ETSI. Physical channels and modulation. Technical Report v.13.3.0, TS 136 211, 2016.

[38] L. Wang, A. Ukhanova, and E. Belyaev. Power consumption analysis of constant bit rate data transmission over 3g mobile wireless networks. In *2011 11th International Conference on ITS Telecommunications*, pages 217–223, Aug 2011.

[39] Aaron Carroll and Gernot Heiser. An analysis of power consumption in a smartphone. In *Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference*, USENIXATC'10, pages 21–21, 2010.

[40] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. Quantification of youtube qoe via crowdsourcing. In *2011 IEEE International Symposium on Multimedia*, pages 494–499, Dec 2011.

[41] Pablo Ameigeiras, Juan J. Ramos-Munoz, Jorge Navarro-Ortiz, and J.M. Lopez-Soler. Analysis and modelling of youtube traffic. *Transactions on Emerging Telecommunications Technologies*, 23(4), 2012.

[42] The U.S. digital video benchmark—2012 review. Technical report, 2012.

[43] Xin Li, Mian Dong, Zhan Ma, and Felix C.A. Fernandes. Greentube: Power optimization for mobile videostreaming via dynamic cache management. In

Proceedings of the 20th ACM International Conference on Multimedia, MM '12, pages 279–288, 2012.

[44] 3GPP. Transparent end-to-end packet switched streaming service (pss);progressive download and dynamic adaptive streaming over http. Technical Report 26, 3GPP TS 26.247, 2011, 2011.

[45] G. Cermak, M. Pinson, and S. Wolf. The relationship among video quality, screen resolution, and bit rate. *IEEE Transactions on Broadcasting*, 57(2):258–262, June 2011.

[46] R. Q. Hu and Y. Qian. An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems. *IEEE Communications Magazine*, 52(5):94–101, May 2014.

[47] Y. Yi K. Son, H. Kim and B. Krishnamachari. Toward energy-efficient operation of base stations in cellular wireless networks. In *Green Communications: Theoretical Fundamentals, Algorithms, and Applications*. CRC Press, Taylor Francis, Oxford, 2012.

[48] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu. Toward dynamic energy-efficient operation of cellular network infrastructure. *IEEE Communications Magazine*, 49(6):56–61, June 2011.

[49] J. Gong, J. S. Thompson, S. Zhou, and Z. Niu. Base station sleeping and resource allocation in renewable energy powered cellular networks. *IEEE Transactions on Communications*, 62(11):3801–3813, Nov 2014.

[50] F. Richter, A. J. Fehske, and G. P. Fettweis. Energy efficiency aspects of base station deployment strategies for cellular networks. In *2009 IEEE 70th Vehicular Technology Conference Fall*, pages 1–5, Sept 2009.

[51] Q. Zhang H. Chen and F. Zhao. Energy-efficient base station sleep scheduling in relay-assisted cellular networks. In *KSII Transactions on Internet Information Systems*, pages 1074–1086, Mar 2015.

[52] K. Son, S. Nagaraj, M. Sarkar, and S. Dey. Qos-aware dynamic cell reconfiguration for energy conservation in cellular networks. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2022–2027, April 2013.

[53] Marco Ajmone Marsan and Michela Meo. Energy efficient management of two

cellular access networks. *SIGMETRICS Perform. Eval. Rev.*, 37(4):69–73, March 2010.

[54] E. Oh, K. Son, and B. Krishnamachari. Dynamic base station switching-on/off strategies for green cellular networks. *IEEE Transactions on Wireless Communications*, 12(5):2126–2136, May 2013.

[55] S. Luo, R. Zhang, and T. J. Lim. Optimal power and range adaptation for green broadcasting. *IEEE Transactions on Wireless Communications*, 12(9):4592–4603, September 2013.

[56] K. Son and B. Krishnamachari. Speedbalance: Speed-scaling-aware optimal load balancing for green cellular networks. In *2012 Proceedings IEEE INFOCOM*, pages 2816–2820, March 2012.

[57] A. J. Fehske, F. Richter, and G. P. Fettweis. Energy efficiency improvements through micro sites in cellular mobile radio networks. In *2009 IEEE Globecom Workshops*, pages 1–5, Nov 2009.

[58] O. Arnold, F. Richter, G. Fettweis, and O. Blume. Power consumption modeling of different base station types in heterogeneous cellular networks. In *2010 Future Network Mobile Summit*, pages 1–8, June 2010.

[59] Kyuho Son, Eunsung Oh, and Bhaskar Krishnamachari. Energy-efficient design of heterogeneous cellular networks from deployment to operation. *Computer Networks*, 78:95 – 106, 2015. Special Issue: Green Communications.

[60] Mung Chiang, Prashanth Hande, Tian Lan, and Chee Wei Tan. Power control in wireless cellular networks. *Found. Trends Netw.*, 2(4), April 2008.

[61] A. Sampath, P. Sarath Kumar, and J. M. Holtzman. Power control and resource management for a multimedia cdma wireless system. In *Proceedings of 6th International Symposium on Personal, Indoor and Mobile Radio Communications*, volume 1, pages 21–25 vol.1, Sep 1995.

[62] Zhisheng Niu, Yiqun Wu, Jie Gong, and Zexi Yang. Cell zooming for cost-efficient green cellular networks. *Comm. Mag.*, 48(11):74–79, November 2010.

[63] J. Kwak, K. Son, Y. Yi, and S. Chong. Greening effect of spatio-temporal power sharing policies in cellular networks with energy constraints. *IEEE Transactions on Wireless Communications*, 11(12):4405–4415, December 2012.

[64] M. Ajmone Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo. Optimal energy savings in cellular access networks. In *2009 IEEE International Conference on Communications Workshops*, pages 1–5, June 2009.

[65] K. Son, H. Kim, Y. Yi, and B. Krishnamachari. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. *IEEE Journal on Selected Areas in Communications*, 29(8):1525–1536, September 2011.

[66] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam. Distributed $alpha$ -optimal user association and cell load balancing in wireless networks. *IEEE/ACM Transactions on Networking*, 20(1):177–190, Feb 2012.

[67] Richard M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA, 1972.

[68] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, Dec 1978.

[69] A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1):41 – 43, 2004.

[70] B. Radunovic and J. Y. Le Boudec. Optimal power control, scheduling, and routing in uwb networks. *IEEE Journal on Selected Areas in Communications*, 22(7):1252–1270, Sept 2004.

[71] Lei Sun, Hui Tian, and Ping Zhang. Decision-making models for group vertical handover in vehicular communications. *Telecommun. Syst.*, 50(4):257–266, August 2012.

[72] 3GPP. Further advancements for e-utra physical layer aspects. Technical Report 36, 3GPP TR 36.814, 2010, 2010.

[73] K. Son, S. Lee, Y. Yi, and S. Chong. Practical dynamic interference management in multi-carrier multi-cell wireless networks: A reference user based approach. In *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, pages 186–195, May 2010.

[74] Ericsson mobility report on the pulse of networked society. Technical report, Ericsson, November 2016.

[75] Josip Lorincz, Tonko Garma, and Goran Petrovic. Measurements and modelling of base station power consumption under real traffic loads. *Sensors*, 12(4):4281, 2012.

[76] R. Guruprasad, K. Son, and S. Dey. Power-efficient base station operation through user QoS-aware adaptive RF chain switching technique. In *IEEE ICC*, pages 244–250, June 2015.

[77] D. W. K. Ng, E. S. Lo, and R. Schober. Energy-efficient resource allocation in ofdma systems with large numbers of base station antennas. *IEEE Transactions on Wireless Communications*, 11(9):3292–3304, September 2012.

[78] Qingqing Wu, Meixia Tao, and Wen Chen. Joint tx/rx energy-efficient scheduling in multi-radio networks: A divide-and-conque approach. *CoRR*, abs/1502.00052, 2015.

[79] J. Wu, S. Zhou, and Z. Niu. Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks. *IEEE Trans. Wireless Commun.*, 12(8):4196–4209, August 2013.

[80] K. Son, S. Nagaraj, M. Sarkar, and S. Dey. Qos-aware dynamic cell reconfiguration for energy conservation in cellular networks. In *IEEE WCNC*, pages 2022–2027, April 2013.

[81] K. Adachi, J. Joung, S. Sun, and P. H. Tan. Adaptive coordinated napping (CoNap) for energy saving in wireless networks. *IEEE Trans. Wireless Commun*, 12(11):5656–5667, November 2013.

[82] Q. Zhang, C. Yang, H. Haas, and J. S. Thompson. Energy efficient downlink cooperative transmission with bs and antenna switching off. *IEEE Trans. Wireless Commun.*, 13(9):5183–5195, September 2014.

[83] S. Han, C. Yang, and A. F. Molisch. Spectrum and energy efficient cooperative base station doze. *IEEE J. Selected Areas Commun.*, 32(2):285–296, February 2014.

[84] D. W. K. Ng, Y. Wu, and R. Schober. Power efficient resource allocation for full-duplex radio distributed antenna networks. *IEEE Transactions on Wireless Communications*, 15(4):2896–2911, April 2016.

[85] Xueqing Huang and Nirwan Ansari. Joint spectrum and power allocation for multi-node cooperative wireless systems. *IEEE Transactions on Mobile Computing*,

14(10):2034–2044, October 2015.

[86] Nokia. 3gpp setup of comp cooperation areas, r1-090725. Technical report, February 2009.

[87] Thorsten Biermann. *Dealing with Backhaul Network Limitations in Coordinated Multi-Point Deployments*. PhD thesis, Dept. Electrical Engineering, Paderborn Univ., Paderborn, Germany, 2012.

[88] A. Chatzipapas, S. Alouf, and V. Mancuso. On the minimization of power consumption in base stations using on/off power amplifiers. In *IEEE Online Conf. on Green Commun.*, pages 18–23, September 2011.

[89] H. Holtkamp, G. Auer, S. Bazzi, and H. Haas. Minimizing base station power consumption. *IEEE J. on Selected Areas Commun.*, 32(2):297–306, February 2014.

[90] ETSI. Physical channels and modulation. Technical Report v.13.3.0, TS 136 211, 2016.

[91] Bing Han, Jimmy Leblet, and Gwendal Simon. Hard multidimensional multiple choice knapsack problems, an empirical study. *Computers & operations research*, 37(1):172–181, 2010.

[92] ETSI. LTE E-UTRA physical layer procedures. Technical Report v.13.0.0, TS 36 213, 2016.

[93] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel. Multi-QoS-aware fair scheduling for LTE. In *IEEE VTC Spring*, pages 1–5, May 2011.

[94] Mohammad T Kawser, Nafiz Imtiaz Bin Hamid, Md Nayeemul Hasan, M Shah Alam, and M Musfiqur Rahman. Downlink snr to cqi mapping for different multiple antenna techniques in LTE. *IJIEE*, 2(5):757, 2012.

[95] A. Papadogiannis, E. Hardouin, and D. Gesbert. A framework for decentralising multi-cell cooperative processing on the downlink. In *IEEE Globecom Workshops*, pages 1–5, November 2008.

[96] Y. Gao, Q. Wang, and G. Liu. The access network and protocol design for CoMP technique in LTE-Advanced System. In *WiCOM*, pages 1–4, September 2010.

[97] K. Alexandris, N. Nikaein, R. Knopp, and C. Bonnet. Analyzing x2 handover in

LTE/LTE-A. In *WiOpt*, pages 1–7, May 2016.

[98] K. Son, S. Lee, Y. Yi, and S. Chong. REFIM: A practical interference management in heterogeneous wireless access networks. *IEEE J. Selected Areas Commun.*, 29(6):1260–1272, June 2011.

[99] 3GPP. Spatial channel model for multi input multi output (MIMO) simulations. Technical Report 25, 3GPP TR 25.996 2011-03, 2011.

[100] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu. Toward dynamic energy-efficient operation of cellular network infrastructure. *IEEE Commun. Mag.*, 49(6):56–61, June 2011.

[101] L. Wang, A. Ukhanova, and E. Belyaev. Power consumption analysis of constant bit rate data transmission over 3g mobile wireless networks. In *2011 11th International Conference on ITS Telecommunications*, pages 217–223, Aug 2011.

[102] Opportunity and impact of video on lte networks. Technical report, Motorola, June 2009.

[103] I. D. Erotokritov. Space-time block coding for multiple transmit antennas over time selective fading channels, 2006.

[104] Apple http live streaming. http://goo.gl/fJIwC. Accessed: 2017-09-09.

[105] M. Gra, C. Timmerer, H. Hellwagner, W. Cherif, D. Negru, and S. Battista,combined bitrate suggestions for multi-rate streaming of industry solutions. http://alicante.itec.aau.at/am1.html. Accessed: 2017-09-09.