Predictive Radio Access Networks for Vehicular Content Delivery

by

Hatem Abou-zeid

A thesis submitted to the
Department of Electrical and Computer Engineering
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
April 2014

Dedication

To Dad, Mom, and Hafsah.

Abstract

An unprecedented era of "connected vehicles" is becoming an imminent reality. This is driven by advances in vehicular communications, and the development of in-vehicle telematics systems supporting a plethora of applications. The diversity and multitude of such developments will, however, introduce excessive congestion across wireless infrastructure, compelling operators to expand their networks. An alternative to network expansions is to develop more efficient content delivery paradigms. In particular, alleviating Radio Access Network (RAN) congestion is important to operators as it postpones costly investments in radio equipment installations and new spectrum. *Efficient* RAN frameworks are therefore paramount to expediting this realm of vehicular connectivity.

Fortunately, the predictability of human mobility patterns, particularly that of vehicles traversing road networks, offers unique opportunities to pursue *proactive* RAN transmission schemes. Knowing the routes vehicles are going to traverse enables the network to forecast spatio-temporal demands and predict service outages that specific users may face. This can be accomplished by coupling the mobility trajectories with network coverage maps to provide estimates of the future rates users will encounter along a trip.

In this thesis, we investigate how this valuable contextual information can enable

RANs to improve both service quality and operational efficiency. We develop a collection of methods that leverage mobility predictions to jointly optimize 1) *long-term* wireless resource allocation, 2) adaptive video streaming delivery, and 3) energy efficiency in RANs. Extensive simulation results indicate that our approaches provide significant user experience gains in addition to large energy savings. We emphasize the applicability of such predictive RAN mechanisms to video streaming delivery, as it is the predominant source of traffic in mobile networks, with projections of further growth. Although we focus on exploiting mobility information at the radio access level, our framework is a direction towards pursuing a predictive *end-to-end* content delivery architecture.

Co-Authorship

Journal Articles

1. **H. Abou-zeid**, H.S. Hassanein and S. Valentin, "Jointly Optimizing Resource Allocation and Quality for Wireless Video Streaming Using Rate Prediction," *IEEE Trans. Veh. Technol.*, (in preparation).

2. **H. Abou-zeid** and H.S. Hassanein, "Towards Green Media Delivery: Location-Aware Opportunities and Approaches," *IEEE Wireless Commun. Special Issue on Green Media: The Future of Wireless Multimedia Networks*, submitted Dec. 2013.

3. **H. Abou-zeid**, H.S. Hassanein, S. Valentin and M.F. Feteiha, "A Lookback Scheduling Framework for Long-term Quality-of-Service over Multiple Cells," *Wireless Commun. and Mobile Comp.*, to appear.

4. **H. Abou-zeid**, H.S. Hassanein and S. Valentin, "Energy Efficient Adaptive Video Transmission: Exploiting Rate Predictions in Wireless Networks," *IEEE Trans. Veh. Technol.*, 2014, to appear.

5. **H. Abou-zeid** and H.S. Hassanein, "Predictive Green Wireless Access: Exploiting Mobility and Application Information," *IEEE Wireless Commun. Mag. Special Issue on Cooperative and Cognitive Paradigms for Green HetNets*, vol. 20, no. 5, pp. 92-99, Oct. 2013.

Conference Publications

1. **H. Abou-zeid** and H.S. Hassanein, "Efficient Lookahead Resource Allocation for Stored Video Delivery in Multi-Cell Networks," *Proc. IEEE Wireless Commun. And Netw. Conf. (WCNC)*, Apr. 2014, to appear.

2. **H. Abou-zeid**, H.S. Hassanein and N. Zorba, "Enhancing Mobile Video Streaming by Lookahead Rate Allocation in Wireless Networks," *Proc. IEEE Consumer Commun. and Netw. Conf. (CCNC)*, Jan. 2014, pp. 768-773.

3. **H. Abou-zeid**, H.S. Hassanein, and S. Valentin, "Optimal Predictive Resource Allocation: Exploiting Mobility Patterns and Radio Maps," *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 4714-4719.

4. **H. Abou-zeid**, H.S. Hassanein and N. Zorba, "Long-term Fairness in Multi-cell Networks Using Rate Predictions," *Proc. IEEE GCC Conference and Exhibition (GCC)*, Nov. 2013, pp. 131-135.

5. **H. Abou-zeid**, S. Valentin, and H.S. Hassanein, "Context-Aware Resource Allocation for Media Streaming: Exploiting Mobility and Application-Layer Predictions," *Capacity Sharing Workshop*, Oct. 2011.

Patent Filing

1. **H. Abou-zeid**, S. Valentin, and H.S. Hassanein, "Apparatus, methods, and computer programs for a mobile transceiver and for a base station transceiver," *EP Patent App. 20110306323, filed by Alcatel-Lucent*, Oct. 2011.

Acknowledgments

In the name of God, the Most Gracious, the Most Merciful. Praises and thanks are due to God who bestowed upon us endless blessings, and the faculties of seeing, thinking, and learning. It is only with His guidance and sustenance that my endeavors and plans were seen through.

My deep, sincere, gratitude and appreciation are due to my advisor and mentor, Prof. Hossam S. Hassanein. No doubt, your constant guidance, enthusiasm, patience, and understanding led to the completion of this work, and my development as a researcher. I want to thank you for inspiring me, and constantly encouraging me to reach out further; and for your feedback and wisdom at times when it was critically needed. It has been a great pleasure working with you, and I hope to stay in touch both academically, and personally. I am indeed indebted to you for making this an enjoyable and fruitful experience.

I have also been fortunate to have a great thesis committee: Prof. Abouelmagd Noureldin, Prof. Karen Rudie, Prof. David Skillicorn, and Prof. Halim Yanikomeroglu. Thank you for your constructive feedback and support throughout the past years. My thanks are also due to Prof. Rudie for encouraging me to undertake several teaching fellowship positions at Queen's University. Teaching has been a very rewarding experience for me, and your guidance is much appreciated.

During 2011, I had the great opportunity of spending 6 months at Bell Labs-Stuttgart for an internship. I would like to express my thanks to Dr. Stefan Valentin for his supervision, and this invaluable work experience in an exceptional industrial environment. My participation in the Context-Aware Resource Allocation (CARA) project exposed me to patenting and practical design aspects, and has been a fulfilling experience.

Since I arrived in Canada, I have been supported by friends and colleagues who made living in Kingston a pleasant experience. In particular, I want to thank Khalid Elgazzar and Sharief Oteafy for their invaluable support and advice with settling in, and friendship thereafter. To Mahmoud Ouda and Mohamed Salah, thank you for being such great friends. The stress of the endless hours spent at work would have been exhausting without you. I would like to thank Muhammad Muhaimin too for his great help in generating realistic mobility traces that I used for evaluation. My thanks are also due to all the members of the Telecommunications Research Lab (TRL) for their friendship, collaboration, valuable discussions, and fun times. Special thanks to Ms. Basia Palmer for all what she has done to make work more productive and efficient. To Ms. Debie Fraser, thank you for your devotion to helping all the graduate students, and patience with our constant requests.

The endless forms of support I received from my parents since my childhood are unimaginable, immeasurable, and incomprehensible. I shall always be forever grateful and thankful for what you have done and sacrificed, not only to make this thesis possible, but for my life as a whole. To Sameh and Dalia, you've been great siblings too, and your adorable children have given me true smiles to the heart!

To my dear wife Hafsah, the past three years would never have been the same without you. Your genuine understanding, support, and encouragement have made my life, and this thesis so much better. I want to also thank you for my new family in London, Ontario: Dr. Monthir, Dr. Salwa, Omar, Mariam, Fatima, Nusaiba, Suhaib, and Lujane. It has been a great pleasure being part of the family, and I will always cherish the times spent together.

Last, but certainly not least, I would like to thank Ahmed Khalil. You have been such a great, supportive friend for over 20 years. May we maintain this friendship, and may I return your constant favors!

Hatem Abou-zeid
Kingston, Ontario
April 2014.

Contents

Dedication i

Abstract ii

Co-Authorship iv

Acknowledgments vi

Contents viii

List of Tables xii

List of Figures xiii

Chapter 1: Introduction 1
- 1.1 Research Statement . . . 3
- 1.2 Thesis Contributions . . . 4
- 1.3 Thesis Organization . . . 5

Chapter 2: Background and Overview 7
- 2.1 Enabling Mobility Predictions . . . 7
- 2.2 Exploiting Mobility Predictions in Cellular Networks . . . 9
  - 2.2.1 Location Management . . . 9
  - 2.2.2 Handoff Resource Reservation . . . 10
- 2.3 Toward Predictive Radio Access Networks . . . 10
  - 2.3.1 Network Performance Maps . . . 11
  - 2.3.2 Illustrative Use Cases . . . 13
  - 2.3.3 Operational Overview . . . 15

Chapter 3: Long-term Radio Resource Planning: Exploiting Mobility Predictions and Radio Maps 17
- 3.1 Introduction . . . 17

3.2 Related Work . . . 19
3.3 System Models . . . 20
3.3.1 Notational Conventions . . . 20
3.3.2 Overview . . . 21
3.3.3 Network and Mobility Models . . . 22
3.3.4 Radio Map and Mobility Information . . . 23
3.3.5 Resource Sharing Model . . . 25
3.3.6 Data Traffic . . . 25
3.4 Multi-cell Predictive Resource Allocation (PRA): Optimal Problem Formulations . . . 26
3.4.1 Preliminaries and Assumptions . . . 26
3.4.2 Predictive Maximum-Rate Allocation . . . 26
3.4.3 Predictive Max-Min Fairness . . . 28
3.4.4 Predictive Throughput-Fairness Trade offs . . . 29
3.5 Numerical Results and Discussion . . . 32
3.5.1 Evaluation Set-up . . . 32
3.5.2 PRA-MaxNetRate and PRA-MaxMin . . . 34
3.5.3 Predictive Proportional Fairness . . . 37
3.6 Summary . . . 44

Chapter 4: Enhancing Wireless Video Streaming Delivery 46
4.1 Introduction . . . 46
4.1.1 Progressive Video Download Enhancements . . . 47
4.1.2 Adaptive Video Streaming Enhancements . . . 48
4.1.3 Related Work . . . 50
4.1.4 Chapter Outline . . . 52
4.2 Predictive Resource Allocation for Video Streaming . . . 53
4.2.1 Preliminaries . . . 53
4.2.2 Optimal Problem Formulation: Video Degradation Minimization 55
4.2.3 Proposed VDMin Algorithm . . . 57
4.2.4 Performance Evaluation . . . 60
4.3 Predictive Adaptive Streaming: Jointly Optimizing RA and Quality Planning . . . 65
4.3.1 System Overview . . . 65
4.3.2 Adaptive Video Streaming Model . . . 65
4.3.3 Optimal Problem Formulation . . . 66
4.3.4 Proposed PAS Algorithms . . . 72
4.3.5 Performance Evaluation . . . 78
4.3.6 Testbed Measurements . . . 84
4.4 Summary . . . 87

Chapter 5: Energy-Efficient Predictive RANs: Application to Stored Video Transmission **89**
5.1 Introduction . . . 89
5.1.1 Chapter Outline . . . 92
5.2 Review on Green Wireless Access Techniques . . . 92
5.2.1 Time Domain Approaches . . . 93
5.2.2 Frequency Domain Approaches . . . 93
5.2.3 Network Reconfiguration . . . 94
5.2.4 Potential of Cooperative Mechanisms . . . 94
5.3 Predictive Green Wireless Access (PGWA) Overview . . . 95
5.3.1 BS Power Consumption Model . . . 96
5.3.2 PGWA: Approach for Stored Video Transmission . . . 97
5.4 Minimizing BS Power Consumption for Video Transmission . . . 98
5.4.1 Optimal Problem Formulation . . . 99
5.4.2 Distributed Heuristic Solution . . . 100
5.4.3 Simulation Results . . . 102
5.5 Joint Power-Video Degradation Optimization for Video Transmission 104
5.5.1 Optimal Problem Formulation . . . 104
5.5.2 Centralized and Distributed Algorithms . . . 106
5.5.3 Performance Evaluation . . . 109
5.6 Joint Power-Quality Planning Optimization for Adaptive Video Transmission . . . 113
5.6.1 System Overview . . . 113
5.6.2 Optimal Problem Formulation . . . 114
5.6.3 Multi-stage Algorithm . . . 122
5.6.4 Performance Evaluation . . . 129
5.7 Implementation Considerations . . . 139
5.8 Summary . . . 141

Chapter 6: Conclusion and Future Directions **143**
6.1 Summary . . . 143
6.2 Future Directions . . . 145
6.2.1 Modeling Uncertainty . . . 145
6.2.2 Robust Predictive Solutions . . . 145
6.2.3 Distributed Approaches and Signaling . . . 146
6.2.4 Practical Implementation Considerations . . . 146
6.2.5 Leveraging Predictions for End-to-End Content Delivery . . . 147
6.3 Concluding Remarks . . . 148

Bibliography **150**

Appendix A: Towards Predictive End-to-End Content Delivery **160**
A.1 Preliminaries: Scalable Video Coding 161
A.2 Predictive In-Network Caching 162
A.3 Predictive Prefetching/Content Pushing 164

List of Tables

3.1 Summary of Frequently Used Symbols 21

4.1 Summary of Frequently Used Symbols in PAS 68

5.1 Summary of Frequently Used Symbols in PGS 116

A.1 Summary of Location-Awareness Potential in End-to-End Media Delivery. 167

List of Figures

2.1 Use cases for Predictive Radio Access Network (P-RAN). 14

2.2 Operations and information exchange in the P-RAN. 16

3.1 Network and mobility models considered. 22

3.2 Example of the predictive RA made to a user. 24

3.3 Example of the resource sharing factor $x_{i,n}$ in Predictive Resource Allocation (PRA). 27

3.4 Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for full buffer traffic on the road network. Note that, Max-Rate and PRA-MaxNetRate overlap. 35

3.5 Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for full buffer traffic with RWP mobility on the 19 cell network. Note that, Max-Rate and PRA-MaxNetRate overlap. 36

3.6 Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for file download on the road network. 38

3.7 Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for file download with RWP mobility on the 19 cell network. 39

3.8 Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for full buffer traffic. 41

3.9 Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for file download traffic. 42
3.10 Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. α for 40 users and file download traffic. 43

4.1 Prebuffering and video degradation as observed from *cumulative* data allocated to a user. 54
4.2 Pareto-optimal trade-off in Video Degradation (VD); RWP scenario with $M = 100$. 62
4.3 Network Video Degradation and Jain's Fairness of Video Degradation vs. number of users with RWP mobility. 63
4.4 Network Video Degradation and Jain's Fairness of Video Degradation vs. number of users with road network mobility. 64
4.5 System models and notation. 67
4.6 Joint rate allocation and segment quality with (a) Predictive Adaptive Streaming, (b) traditional non-predictive approach. 70
4.7 PAS video quality and freezing for the highway scenario. 81
4.8 PAS video quality and freezing for the Random Way Point (RWP) scenario. 82
4.9 Effect of prediction errors on the Predictive Adaptive Streaming (PAS) schemes in the highway scenario, $M = 12$. 83
4.10 Testbed setup with 4 IEEE 802.11g Access Points (APs), 1 Control Computer, and 4 Android Handhelds 85
4.11 Experimental results from real-time testbed measurements. 86

5.1 Sample user allocation with time illustrating power minimization. (a) In traditional allocation, airtime is divided equally among users. (b) In PGWA, allocations are low if the user rate is increasing, and high when the rate is high, to avoid inefficient future allocations. (c) similar to (b) but using only one BS. Although air-time in (c) is larger than (b), more energy is saved as one BS is switched off. 98
5.2 Average air-time with (a) varying streaming rates for 40 users; (b) varying number of users for 1.2 Mbit/s streaming. 103
5.3 Video degradation $\mathrm{VD}_{\mathrm{Net}}$ and BS power consumption for varying number of users. 110
5.4 Trade-off of $\mathrm{VD}_{\mathrm{Net}}$ and BS power consumption for varying PGWA-MinAirVD-LP and PGWA-MinAirVD-Alg air-time weights. 111
5.5 Pareto-optimal trade-off of $\mathrm{VD}_{\mathrm{Net}}$-BS Power Consumption for the different allocation algorithms; $M = 30$. 112
5.6 Predictive Green Streaming Operation. 114
5.7 Sample Base Station (BS) power consumption with time: $P_0 = 200\,\mathrm{W}$ and $P_m = 1300\,\mathrm{W}$. (a) In traditional operation, BS airtime is inefficient utilized. (b) With PGS-MinAir, BS airtime is minimized by opportunistic allocations. In (c) PGS-MinPower groups user allocations to allows deep sleep modes. 121
5.8 The Proposed Multi-Stage Predictive Green Streaming (PGS) Solution. 123
5.9 PGS results for the single-cell scenario. 132
5.10 PGS results for the multi-cell scenario. 134

5.11 Effect of shadowing prediction errors on the PGS schemes in the highway scenario, $M = 20$. 136
5.12 Effect of shadowing prediction errors and fast fading on the PGS schemes in the highway scenario, $M = 20$. 138
5.13 Key elements and functions of PGWA. The subscripts i and n represent users and time slots respectively. 140

A.1 Potential of location awareness and mobility predictions for end-to-end content delivery. 161

Chapter 1

Introduction

The phenomenal growth of mobile traffic in recent years is unprecedented. Global mobile data traffic grew 81% in 2013, and is forecasted to increase nearly 11-fold by 2018 [1]. This is driven by the recent popularity of data-intensive applications and social networking websites. In particular, mobile video accounted for over 50% of the traffic in 2012, with projections of a 14-fold increase by 2018 [1]. Consequently, operators are facing formidable resource management challenges to cope with this demand.

In cellular networks, the data path for downlink traffic typically starts at the Internet or content servers, and then traverses the wireless carrier's Core Network (CN) and Radio Access Network (RAN), before reaching the mobile user. Congestion at any point throughout the network results in Quality of Service (QoS) degradations, thereby impacting the users' experiences [2]. Among the network elements, alleviating RAN congestion is particularly important as it can postpone investment in additional Base Station (BS) installations and backhaul networks, or in new spectrum. This limits the capital expenditure (CapEx) increases to operators since radio equipment installations make up to 70% of CapEx [3]. An 'intelligent', or more

efficient RAN, also reduces overall energy consumption significantly as BSs account for more than 50% of the network energy consumption [4]. Therefore, devising novel RAN frameworks is paramount to maintaining costs while satisfying the increasing Telecom market demands.

In this thesis, we investigate how predictions of user *location* can be used to improve the performance of RANs. Knowing the routes users are going to traverse enables the network to forecast spatio-temporal demands and predict service outages that specific users may face. For instance, if a user watching a YouTube video is moving towards the cell edge or a tunnel, the network can increase the allocated wireless resources in advance to prebuffer some additional video content. Mobility predictions also facilitate higher network efficiency. A user running a delay-tolerant application and moving towards a BS can be delayed transmission until getting closer. This allows the BS to save energy by optimizing transmission times when users are at their highest achievable rates.

We believe that the upcoming era of ubiquitous location information and the capability of anticipating events can be jointly leveraged to design efficient RANs. In essence, this research is motivated by:

1. the availability of accurate user location information in mobile networks with over 770 million GPS-enabled mobile phones in use today [5]. Recent analyses on human travel patterns also reveal that people follow particular routes with a predictability of up to 93% [6, 7, 8].

2. the projected growth of the automotive telematics market. An era of connected cars is anticipated where entertainment, maintenance, safety, and navigation applications are run from within vehicles [9]. This will dramatically increase

the content delivery to vehicular users.

3. the increasing feasibility of generating accurate, real-time, coverage and radio environment maps. Developments in smartphones and 3rd Generation Partnership Project (3GPP) standardization allow users to periodically collect and report radio measurements, and QoS indicators, at different locations, and then report them to network operators [10, 11].

4. the emergence of context-aware network architectures and self-organizing functionalities in cellular networks [12] is enabling efficient context signaling and in-network adaptation in future networks.

1.1 Research Statement

Current research on Resource Allocation (RA) in RANs has primarily focused on distributing the available network resources to meet the immediate application requirements of mobile users. The adoption of predictive techniques that facilitate long-term RA planning and QoS provisioning in wireless networks remains limited. We believe that:

> Coupling user location predictions with network performance maps can enable the development of proactive resource allocation and content delivery schemes in future wireless networks.

The primary goal of this thesis is to develop a collection of methods that leverage mobility predictions to optimize 1) wireless resource allocation, 2) video streaming delivery, and 3) energy efficiency in RANs.

1.2 Thesis Contributions

The major contributions of this thesis are the following:

1. *Predictive Resource Allocation (PRA)*: We propose a *predictive* RA framework that extends the RA planning horizon to tens of seconds, spanning multiple cells. This is accomplished by exploiting predictions of users' future data rates and application demands. Through PRA, we introduce the notion of *long-term* multi-cell cooperation, where allocations made to users in one cell impact the amounts allocated in future cells traversed by the users.

2. *Video Prebuffering Control*: We introduce the concept of video *prebuffering* control based on rate predictions. To this effect, we formulate a mathematical framework that models stored video transmission and develop PRA schemes that exploit rate predictions to enhance video streaming. In essence, the PRA schemes strategically buffer content in advance in the users' devices, to minimize video degradations. Optimal benchmark solutions are derived and polynomial time algorithms are developed.

3. *Joint RA and Quality Planning for Adaptive Video Streaming*: We propose a novel *in-network* Predictive Adaptive Streaming (PAS) approach that leverages wireless rate predictions to *jointly* optimize resource allocation and video segment quality over a time horizon. Predictive adaptive streaming improves the experienced video quality while simultaneously eliminating video stalls. We study implementation feasibility and performance on a testbed with real wireless links and real videos. Our results demonstrate that PAS is a promising function for future cellular networks and content delivery platforms.

4. *Predictive Green Wireless Access (PGWA) Framework*: We investigate how mobility awareness can minimize BS downlink power consumption. This is accomplished through a mathematical framework for both constant quality and adaptive video streaming. We also consider predictive RA that groups user allocations in consecutive blocks, allowing BSs to subsequently turn off in deep-sleep energy saving modes. The PGWA framework jointly determines multi-user resource allocation, video segment quality levels, BS power levels, and BS on/off status for a multi-cell network. We also develop a multi-stage algorithm to solve the problem in polynomial time. Finally, we discuss several implementation issues and future directions towards a predictive end-to-end content delivery framework.

1.3 Thesis Organization

In this chapter we presented and motivated the primary research problem, and discussed our major contributions towards predictive RA and content delivery in RANs. The rest of this thesis is organized in several chapters outlined below.

Chapter 2 surveys prior research efforts in exploiting location awareness to enhance different mobile network functions such as paging, handover and resource reservation. An overview of the proposed predictive RAN framework is then presented.

Chapter 3 first introduces the system model and assumptions that will be used throughout the thesis. Based on these models, the optimal PRA problem is then formulated to maximize network throughput and provide varying degrees of fairness. Simulation results are presented to demonstrate the potential gains of PRA over baseline approaches that do not incorporate rate predictions.

Chapter 4 investigates the application of PRA to video streaming delivery. First a constant quality video stream is assumed and the optimal problem is formulated. Then a prebuffering algorithm that *trades off* overall network streaming quality and *fairness* in individual user quality is developed. The second part of this chapter tackles the joint RA and quality adaptation problem, and accordingly the optimal problem is formulated and an efficient real-time algorithm proposed.

In Chapter 5 we build on the video streaming models developed in Chapter 4 to address the problem of energy efficiency. A mathematical framework is laid out to transmit video streams while consuming minimum BS power, and enabling a trade-off between video quality and energy consumption. Several centralized and distributed algorithms with close to optimal performance are also developed.

Chapter 6 presents a summary of the topics addressed in this thesis, and an outlook on potential future research directions.

Chapter 2

Background and Overview

Investigating the potential of mobility predictions to optimize cellular network operations has been an important research topic in the past decade. This was based on the vision that future mobile devices and networks are likely to be capable of accurately predicting user mobility.

After a brief overview on approaches to predicting human mobility, this chapter discusses the two primary cellular network operations where such predictions have been successfully applied. Then, in the second part of the chapter, we introduce the proposed Predictive Radio Access Network (P-RAN) and overviews its potential uses and operation.

2.1 Enabling Mobility Predictions

LTE networks support a wide range of location positioning methods with varying levels of granularity including the assisted-Global Navigation Satellite System (A-GNSS). A dedicated LTE Positioning Protocol (LPP) is also devised to coordinate signaling between the User Equipment (UE) and the BS [13]. On the other hand, a plethora of navigation hardware and software is also available in today's smart phones

enabling users to report their current location. With this evolution in network and handset hardware, determining a mobile user's location is readily possible.

As discussed earlier, recent analyses on human mobility traces indicate that people tend to follow particular routes regularly, thereby enabling high predictability [6]. Generally, the prediction of user trajectories is possible by mapping position, speed and direction of travel onto street maps, particularly for highways and rural areas. To predict mobility patterns for a longer time interval, user input of the destination may be provided, either directly or through navigation software. Alternatively, the network may also generate databases of user mobility profiles to facilitate such predictions. Moreover, for users on public transportation, the routes of buses and trains are known in advance. Input from real-time traffic information can also be used to account for variations arising from the time of day, weather, or accidents.

There have been several promising research efforts that predict user mobility based on real data from both cellular networks and GPS traces. Some works focus on short-term mobility prediction (next cell) [14], while others attempt to predict the full trajectory [15],[16]. Approaches used include data mining [14], Markov renewal processes [15], and learning routes between common destinations via string matching [17]. User behavior profiling [18] is also a promising complementary approach that can provide more context to the mobility prediction models. Today, profiling is garnering increasing interest from industry as it brings opportunities for innovative Location Based Services (LBSs), which are expected to generate a revenue of $13.5 billion by 2015 [19].

2.2 Exploiting Mobility Predictions in Cellular Networks

Several research efforts have been made to exploit mobility predictions to improve cellular network operations. The most prominent of these are in location management and handoff resource reservation.

2.2.1 Location Management

Locating mobile users and devices in cellular networks is done using a combination of location updates (by the mobile) and paging (by the network). The paging scheme determines how and where to search for a mobile user given the latest location update from that user [20]. Predicting a user's location can therefore increase the efficiency of paging by reducing both the frequency of the updates and the size of the paging zone. For example, Liang and Hass [21] propose a location prediction scheme that uses a Gauss-Markov mobility model to capture the correlation of the user's velocity with time. Based on this prediction scheme, the network pages the user around the predicted location. An analytical framework that includes the cost of making these location updates/predictions is developed by the authors, and the performance advantage of prediction in location management is assessed for various system configurations. A similar effort is made by Zang and Bolot [20], and Taheri and Zomaya [22] where clustering and data mining techniques are used to extract patterns of user movement and thereafter decrease the signaling requirements for location updates. Real network data with 300 million call records was used to validate the proposed prediction-based solution in reference [22].

2.2.2 Handoff Resource Reservation

In traditional handoff prioritization schemes, a portion of the radio capacity is permanently reserved for handoffs. Mobility predictions have been proposed to improve the efficiency of handoff prioritization schemes by dynamically adjusting the reservations based on knowledge of user trajectories [23],[24]. A similar concept is followed by Chlamtac *et al.* [25] where instantaneous traffic load is estimated in terms of handoff call arrival rate and call departure rate. This is based on a mathematical formulation that predicts a mobile user's movement behavior in terms of boundary crossing probability and cell residence probability. The work by Yu and Leung [26] is based on more realistic assumptions where the users' movement *history* is used to predict the cell to which the terminal will handoff to, as well as the handoff time. With this information, bandwidth reservations are made to guarantee some target handoff dropping probability. A similar proposal is also made by Choi and Shin [27]. On the other hand, Bolla and Repetto [28] introduce the idea of coupling macroscopic vehicle traffic forecasting to predict the load in different cells. This model can then also be used for resource reservation.

2.3 Toward Predictive Radio Access Networks

While mobility predictions have provided promising results in location management and handoff prediction, limited work has been conducted towards *long-term* resource allocation. The approaches proposed in handoff management are mostly concerned with optimizing the reserved resources needed for imminent handoffs. In addition, the application focus has been voice calls.

Today, mobile application usage is changing. This is driven by both the higher

connectivity speeds as well as the larger screen sizes of pads, tablets, and more recently phablets. For example, watching a full length movie or a 20 minute sitcom on a phone screen was not previously foreseen as the ideal viewing experience. However, statistics conducted in 2013 indicate that Netflix's downstream traffic share in North America almost doubled from 2.2% to 4.0% in just a year [29]. Such trends of long, data-intensive sessions are anticipated to continue to grow with the emergence of telematics and infotainment systems in vehicles.

In an effort to address these challenges, we investigate how predictive mechanisms can be incorporated into radio access networks to plan better, more efficient transmission. In the proposed P-RAN framework, we introduce the use of:

- *Network performance maps*: which indicate the signal strengths and achievable data rates at different geographical locations.
- *Application context*: which enables forecasting user data demands in the near future; e.g., the quality level and duration of a video clip can be used to estimate future user requests.

Therefore, P-RAN is a cross-layer framework where future estimates of Physical layer (PHY) information are coupled with context information from the application layer to optimize long-term RAN functionality. The rest of this chapter first discusses how network performance maps are typically generated, and then highlights some illustrative use cases of P-RAN that we will address in detail in this thesis.

2.3.1 Network Performance Maps

Network operators commonly conduct road drive tests to geographically measure radio signal strengths and other network performance metrics at different locations.

This information is then processed to generate maps that estimate the correlation between location and channel capacity (or other QoS metrics) [30]. These maps are commonly referred to as coverage, radio, or network performance maps. Openly accessible radio maps are also available online such as the OpenSignal Project [11] where signal strength is crowdsourced from users. While such maps provide a reasonable estimation of the wireless data rates, they do not accurately capture the dynamics of network congestion or environmental changes, and therefore require constant updating, particularly in urban areas where traffic demand fluctuates. To address this, a recent feature known as "Minimization of Drive Tests (MDT)" is defined in Long Term Evolution (LTE) Rel-10 that exploits the ability of UE to include location information as part of the radio measurement reporting. With MDT, UEs log radio measurements during their idle states and send periodic reports to the network [10], facilitating real-time map updates.

In addition to experimental location-rate mapping, several analytical studies have also been made to model and predict the correlation between location and channel capacity. The work in [31] presents a comprehensive survey of methods that model path loss for different terrains/environments. Sampling strategies and interpolation techniques that incorporate practical measurements into the derived models are also presented. The authors conclude that online learning strategies and data mining approaches that use measurement-based models are likely to provide the most robust radio mapping frameworks. An example of such an approach is the work in [32] which characterizes the impact of different environments, and sampling positions on the wireless channel predictability.

Some practical studies investigating the correlation between location and received

data rates have also been conducted [33],[34]. Yao *et al.* analyze bandwidth traces collected from two independent cellular providers for routes running through different radio conditions including terrestrial and underwater tunnels [33]. Their findings confirm the correlation between mobile bandwidth and location, and indicate that bandwidth uncertainty can be reduced drastically when observations from past trips are used to predict bandwidth. The work by Han *et al.* [34] conducts a similar measurement study, and addresses other contextual factors such as user speed, time of day, and humidity to predict the available bandwidth more accurately.

In this thesis, our goal is not to develop techniques that generate the radio maps themselves, but to propose and evaluate predictive RAN mechanisms that exploit mobility predictions as illustrated in the following section.

2.3.2 Illustrative Use Cases

In addition to providing geographical information on the current supportable data rates, radio maps also enable the estimation of *future* data rates users will experience, provided their mobility trajectories are known. We argue that with this rate information, BSs can plan proactive resource allocations that improve service delivery as highlighted below.

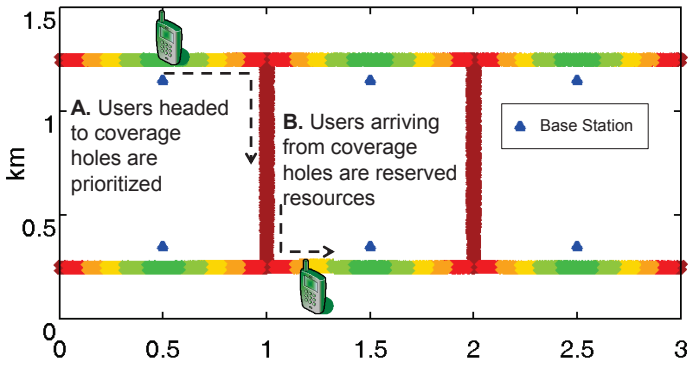
Predictive Resource Allocation (PRA)

Figure 2.1(a) illustrates the relative signal strength (i.e., the colormap of a radio map) along a road network, and shows two use-cases for PRA. In case A, BSs can prioritize users headed to low rate areas (marked in dark colors) and pre-allocate the required data before hand. In case B, BSs can plan resource reservations for users arriving

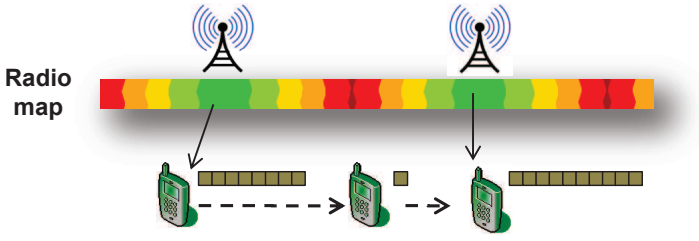
from low rate areas, for instance to accelerate the progress of a file download that was stalled. The goal is to improve user satisfaction by making long-term RA provisions.

Video Delivery

Another practical use of PRA is for stored video delivery such as YouTube, which accounts for a quarter of all mobile network traffic during peak hours [29]. As opposed to live streaming, stored videos can be strategically delivered in advance and pre-buffered at the UE to prevent video stalling as illustrated in Figure 2.1(b). With the use of road maps, the multi-user problem can be solved to optimize the amount of video content delivered to each user, at each time slot, based on their mobility predictions.



(a) User prioritization and resource reservation.



(b) Strategic prebuffering of video streams.

Figure 2.1: Use cases for P-RAN.

Energy Efficiency

Rate predictions can also be used to develop energy efficient RAN transmission. For instance, a user viewing a stored video on a highway traversing two cells, can be pre-allocated the requested video content in the first cell, while the second cell is switched off without causing any video stalling. Furthermore, with some cooperation and signaling, other energy saving schemes can be facilitated. For instance, during handover, users can report the status of the running applications (such as the amount of stored video), and user speed. If the buffer is not empty and speed is considerable, transmission can be momentarily suspended until the users approach the cell center, without degrading the QoS.

2.3.3 Operational Overview

The central idea in P-RAN is to collect and exchange user location and application information and then develop cooperative access and video delivery strategies that improve user QoS and save energy. Figure 2.2 provides a summary of the typical sequence of events and information exchanges required to enable predictive RANs. This includes three main operational stages:

1. collecting and exchanging user location and application information, and radio/QoS reports,

2. predicting long-term user achievable rates and traffic demands, and

3. developing and implementing the P-RAN strategies over multiple cells.

In this thesis, our goal is to formulate and develop predictive RAN solutions with emphasis on video delivery. We address a specific problem in each of the following

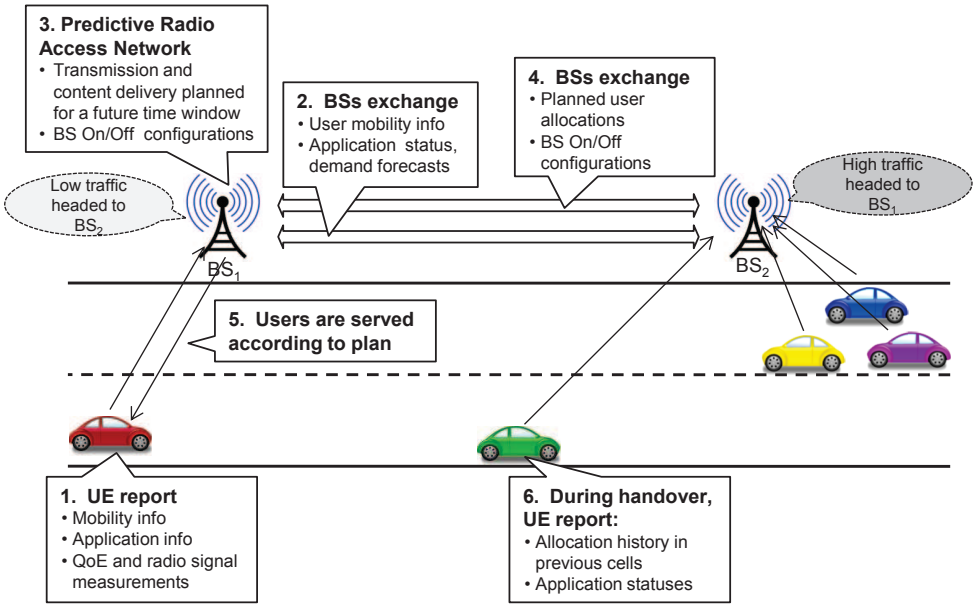


Figure 2.2: Operations and information exchange in the P-RAN.

chapters. Optimal problem formulations are first laid out to determine the benchmark solutions and limits of the QoS gains and energy savings. Then, we develop polynomial-time algorithms for real-time implementation and evaluate their performance through extensive simulations.

Chapter 3

Long-term Radio Resource Planning: Exploiting Mobility Predictions and Radio Maps

3.1 Introduction[1]

Mobile traffic is not only experiencing unprecedented growth rates, but also becoming more unevenly distributed in space and time [37]. Variable network availability and variable spatio-temporal user demand lead to challenges in providing satisfactory long-term service to mobile users [38]. This is because, while some users may head to favorable network locations, others may move towards congestion zones or locations with poor signal quality. The problem may be even more aggravated when vehicular telematics services become widespread. This spatially varying service is addressed in this chapter through a proposal of Predictive Resource Allocation (PRA) in the P-RAN framework.

In cellular networks, the primary goal of RA is to distribute BS resources (e.g., bandwidth and allocation time slots) efficiently and fairly while satisfying individual user QoS requirements. This is usually accomplished by monitoring average user

[1]Parts of this chapter were previously published in [35],[36].

PHY data rates and queue lengths, and thereafter allocating resources accordingly to trade-off throughput, fairness, and application-specific QoS metrics [13, Ch. 6]. Utility functions are also often used to capture the desired network performance and user satisfaction levels [39],[40]. However, by focusing only on the *current* application demands and user achievable rates, the RA process can only *react* (inefficiently) to sudden changes in user channel gains and application requirements. In an effort to address such limitations, we investigate how user mobility patterns and radio maps can be leveraged to make long-term allocation plans. This provides an additional degree of diversity that translates into higher network efficiency and user QoS. In more detail, this chapter makes the following contributions:

1. Presents PRA which extends the RA planning horizon to tens of seconds, spanning multiple cells, by exploiting predictions of users' future data rates and application demands.

2. Introduces the concept of *long-term* multi-cell cooperation, where allocations made to users in one cell impact the amounts allocated in future cells traversed by the users.

3. Demonstrates the potential of PRA by formulating the allocation problem for three different network objectives. The first maximizes network throughput and the second achieves *max-min* fairness over a specified time horizon. In the third objective, we present a generic long-term fairness RA formulation based on the α-fair utility function [41]. This provides variable degrees of fairness and enables the formulation of a predictive *proportional fair* allocator as a special case.

We conduct extensive performance analyses of the PRA schemes to demonstrate the throughput and fairness gains for different network, mobility, and traffic models.

The rest of this chapter is organized as follows. In the next section we review related work, and follow that with the system models and assumptions in Section 3.3. These models will be used in subsequent chapters as well. In Section 3.4 we formulate the multi-cell PRA problem for several network objectives. We present and discuss the numerical results that demonstrate the potential gains of PRA in Section 3.5. Finally, we conclude in Section 3.6.

3.2 Related Work

Multi-cell resource allocation and BS cooperation is a related area of research, where efforts have focused primarily on instantaneous cooperation to achieve *short-term* objectives. BSs coordinate their transmissions periodically to minimize interference, balance load, or perform joint transmissions to a user [42]. This form of multi-cell RA differs from that proposed in PRA. Our goal is to exploit user rate prediction to make optimal *long-term* allocations over multiple cells.

In related work on long-term QoS, we discuss multi-cell RA that exploits user QoS *history* in previous cells, to make allocations in the current cell [38],[43]. Such a scheme improves long-term QoS by prioritizing users that have had poor service in previously traversed cells . However, it is still considered *reactive* and does not exploit rate predictions to make long-term plans.

The notion of *predictive scheduling* has been proposed in a few works such as references [44] and [45], where predictions of the users' future supported rates are

exploited to improve throughput and fairness. However, these works focus on predictions at a time scale of milliseconds, i.e., they leverage fast fading predictions. On the other hand, the objective of PRA in P-RAN is to exploit mobility-based predictions and radio maps to allocate resources over minutes.

The work by Ali *et al.* [46] relates closely to our work, where predictions of path-loss-related slow variations are used to estimate future data rates. The authors formulate the problem with the objective of increasing system data rate of a single-cell. However, they do not consider the multi-cell problem, or realistic road/mobility scenarios that provide more insight on practical performance measures. The work also focuses on increasing system throughput of infinitely backlogged data traffic, so a discussion on fairness and the implications of different applications is not addressed.

A parallel research effort in the same direction (that is yet to appear) has been conducted by Margolies *et al.* who show that a user's channel state is highly reproducible, and develop a predictive proportional fairness algorithm [47]. As in our work, significant throughput and fairness gains were observed.

3.3 System Models

3.3.1 Notational Conventions

We use the following notational conventions throughout this thesis: $\mathcal{X}$ denotes a set and its cardinality $|\mathcal{X}|$ is denoted by X. We use bold letters to denote matrices, e.g., $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$. $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and ceiling functions respectively and $(x)^+$ denotes $\max\{0, x\}$.

A summary of the commonly used symbols in this chapter is provided in Table 3.1.

Table 3.1: Summary of Frequently Used Symbols

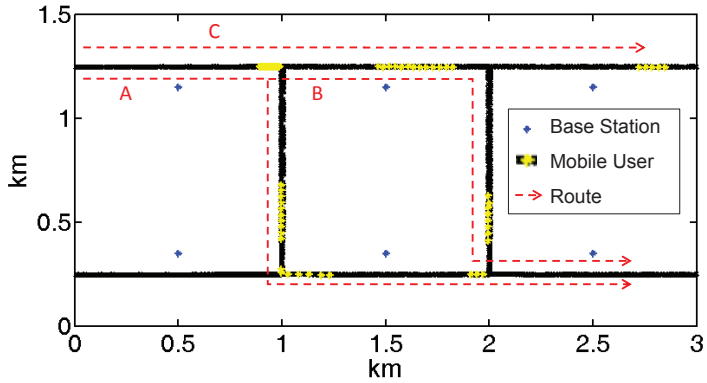
Symbol	Description
α	Fairness control parameter
τ	Time slot duration [s]
i	User index, $i = \{1, 2, \cdots, M\}$
j	BS index, $j = \{1, 2, \cdots, K\}$
n	Time slot index, $n = \{1, 2, \cdots, N\}$
K	Number of BSs in the network
M	Number of users in the network
N	Number of time slots in the prediction window
$\hat{r}_{i,n}$	Predicted link rate of user i at slot n [bits]
$x_{i,n}$	Fraction of BS air-time assigned to user i at slot n
D_i	Total traffic requested by user i during N slots [bits]
R_i	Total traffic received by user i during N slots [bits]
$\mathcal{U}_{j,n}$	Set of the indices of users associated with BS_j at time slot n

3.3.2 Overview

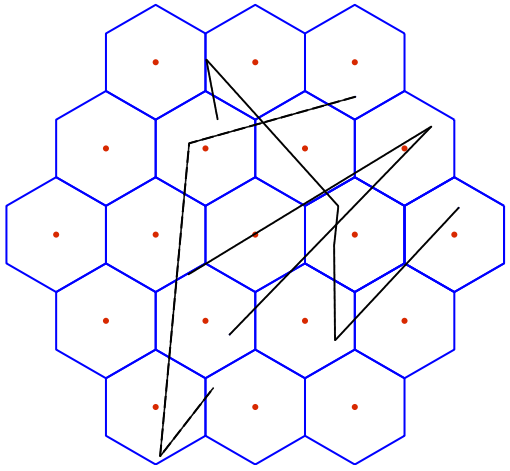
We consider a network with a BS set $\mathcal{K}$ of K BSs and an active user set $\mathcal{M}$ with M users. An arbitrary BS is denoted by $j \in \mathcal{K}$ and a user by $i \in \mathcal{M}$. Time is divided in slots of equal duration τ, during which the wireless channel can be shared arbitrarily among multiple users. We assume that the wireless link is the bottleneck, and therefore the requested content is always available at the BSs for transmission. Users are associated to BSs based on the strongest received signal. The set $\mathcal{U}_{j,n}$ contains the indices of all the users associated with BS j at time slot n.

3.3.3 Network and Mobility Models

Two network and mobility scenarios are considered. The first is a network of six BSs shown in Figure 3.1(a), that covers the illustrated road network. To provide realistic vehicular mobility we use the Simulation of Urban Mobility (SUMO) microscopic road traffic simulation package [48] to generate traces for the three routes denoted by A, B and C. Vehicles enter the network at a flow of 1 vehicle per second, and traverse the



(a) Road network with mobility routes.



(b) The 19 cell network with RWP mobility.

Figure 3.1: Network and mobility models considered.

three routes with equal probability. We refer to this as the *road network* scenario. A special case of the road network is the *highway* scenario where vehicles follow route C only.

To provide performance insights in a larger and more random mobility setup, we also model the 19 cell network illustrated in Figure 3.1(b). Users move according to the RWP mobility model [49] with a constant speed, zero pause time between the waypoints, and no wrap-around. While this model is not practical, it enables the study of PRA when users experience different sequences of data rate fluctuations.

3.3.4 Radio Map and Mobility Information

The radio map assumed to be typically available at the service provider would contain the average data rates at different network locations. In order to abstractly model such a radio map, we use a standard outdoor propagation path loss model $\mathrm{PL}(d) = 128.1 + 37.6 \log_{10} d$, where the user-BS distance d is in km [50]. The feasible link rate is then computed using Shannon's equation with Signal to Noise Ratio (SNR) clipping at 20 dB to account for practical modulation orders. Therefore, a user i at slot n, will have a feasible data rate transmission of

$$r_{i,n} = \tau B \log_2(1 + P_{\mathrm{rx}_{i,n}}/N_o B) \quad [\text{bits}] \tag{3.1}$$

where P_{rx}, N_o and B are the received power, noise power spectral density, and the transmission bandwidth respectively.[2] The slot user rate $r_{i,n}$ gives the number of bits

[2] Although we only consider path loss, a more complex channel model that includes shadowing and interference may be used to determine P_{rx}. The proposed PRA approaches are generic and only require an estimate of the radio map which can be determined either through measurements or radio propagation models.

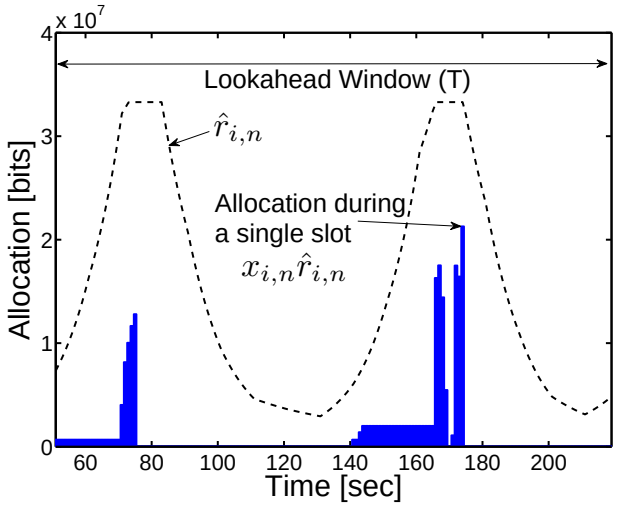


Figure 3.2: Example of the predictive RA made to a user.

that can be transmitted during a time slot, i.e., the transmission rate normalized with slot duration τ.

We assume that user mobility information is known accurately for the upcoming T seconds, which we call the lookahead, or *prediction window*. The future link capacities can therefore be determined by computing the expected received power based on the user-BS distances, and then substituting in Eq. (3.1). Note that there are $N = T/\tau$ time slots within the prediction window which we denote by the set $\mathcal{N} = \{1, 2, \cdots, N\}$. This will generate a matrix of future link rates as defined by $\hat{\mathbf{r}} = (\hat{r}_{i,n} : i \in \mathcal{M}, n \in \mathcal{N})$.

Figure 3.2 illustrates an example of $\hat{r}_{i,n}, \forall n$ for a user traversing two BSs along a highway. We can see that the predicted rate follows two cycles of rate increases and decreases. This is based on the user motion across the two BSs.

3.3.5 Resource Sharing Model

BS airtime can be shared arbitrarily among the active users in any slot n, during which the achievable data rate is assumed to be constant (as determined from the radio map). A typical value of such a slot interval τ is 1 s for vehicle speeds up to 15 m/s, during which average wireless capacity is not significantly affected. We define the resource sharing matrix $\mathbf{x} = (x_{i,n} \in [0,1] : i \in \mathcal{M}, n \in \mathcal{N})$ which gives the fraction of time during each slot n that the BS bandwidth is assigned to user i. The rate received by each user, at each slot, is the element-wise product $\mathbf{x} \odot \hat{\mathbf{r}}$. Therefore, $\mathbf{x}$ controls both the *per-slot* and total *long-term* rates users receive over the N slots. A sample allocation $x_{i,n}, \forall n$ for a user i is illustrated in Figure 3.2. Here the bars indicate the proportion of $\hat{r}_{i,n}$ allocated to that user. Therefore, $x_{i,n}$ is the optimization variable that defines user allocations over the BSs. Note that since a user can traverse multiple cells during N, BS cooperation is needed to make the allocation plan. This is assumed to be possible via an inter-BS interface such as the X2-interface in LTE networks.

3.3.6 Data Traffic

The data traffic requested by user i at time slot n is denoted by $D_{i,n}$, and the total data requested during the N slots is denoted by D_i. We consider two cases for user traffic: (i) full buffer traffic, where $D_{i,n} \to \infty \; \forall \, i, n$, and (ii) file download traffic, where D_i is finite $\forall \, i$. Full buffer traffic is used to illustrate the limits of the fairness performance.

3.4 Multi-cell Predictive Resource Allocation (PRA): Optimal Problem Formulations

3.4.1 Preliminaries and Assumptions

In this chapter, our goal is to formulate the Predictive Resource Allocation problems and evaluate them. We are interested in quantifying the upper bounds of the throughput and fairness gains, and therefore assume ideal predictions of user rates. In upcoming chapters, this assumption is relaxed and the effects of prediction errors are investigated. Thus, the matrix of predicted user rates $\hat{\mathbf{r}}$ is assumed to be accurate. The output of the PRA, i.e., the resource sharing factor $x_{i,n} \in [0, 1]$, is determined for the N upcoming time slots, for all users. Since a user can traverse multiple BSs during the N slots, $\hat{\mathbf{r}}$ is assumed to be available at a central coordinating BS which plans user allocations for all the cooperating BSs. This is depicted in Figure 3.3 where sample values of $x_{i,n}$ are presented for a network with two BSs and four users. In the following formulations, we demonstrate how users are assigned resources depending on the network objectives and their application requirements.

3.4.2 Predictive Maximum-Rate Allocation

In order to maximize the network throughput over the upcoming N time slots, the sum of the rates allocated to all the users during N should be maximized, i.e., $\sum_{n=1}^{N} \sum_{i=1}^{M} \hat{r}_{i,n}\, x_{i,n}$. We refer to this long-term RA scheme as PRA-MaxNetRate, which can be formulated as the following Linear Program (LP):

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{n=1}^{N} \sum_{i=1}^{M} \hat{r}_{i,n}\, x_{i,n} \tag{3.2}$$

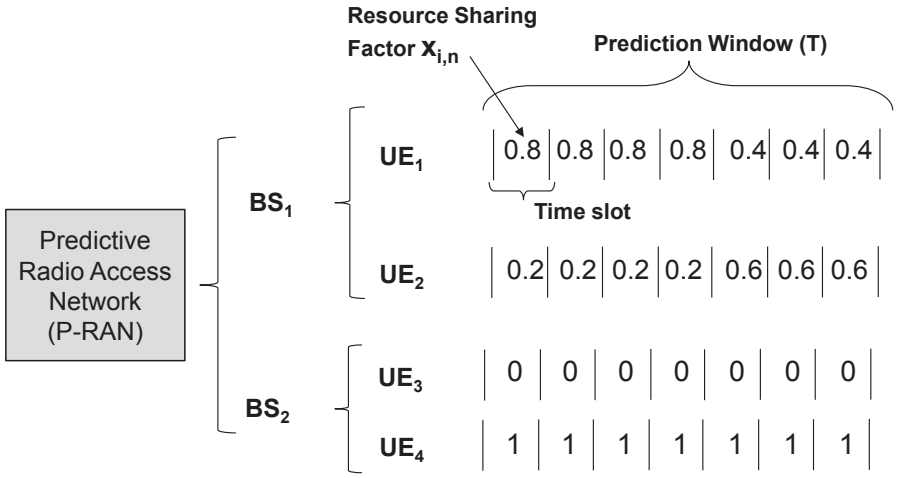


Figure 3.3: Example of the resource sharing factor $x_{i,n}$ in PRA.

$$
\begin{aligned}
\text{subject to:}\quad & \text{C1:} \quad \sum_{i \in \mathcal{U}_{j,n}} x_{i,n} \leq 1, && \forall\, j \in \mathcal{K},\, n \in \mathcal{N} \\
& \text{C2:} \quad \sum_{n=1}^{N} \hat{r}_{i,n}\, x_{i,n} \leq D_i, && \forall\, i \in \mathcal{M} \\
& \text{C3:} \quad x_{i,n}\, \hat{r}_{i,n} \geq \text{MBR}, && \forall\, i \in \mathcal{M},\, n \in \mathcal{N} \\
& \text{C4:} \quad 0 \leq x_{i,n} \leq 1 && \forall\, i \in \mathcal{M},\, n \in \mathcal{N}.
\end{aligned}
$$

Note that the outer summation over time in the objective is included as PRA optimization is made for all time slots in the prediction window. Constraint C1 expresses the resource limitation at each BS by ensuring that the sum of the resource sharing factors of all users associated with each BS does not exceed 1 at every time slot. Since C1 is applied at each BS, and every time slot, there are KN constraints in total. C2 limits the amount of data assigned to each user during the N time slots to the total amount request. This leads to M constraints. In C3, an optional Minimum Bit Rate (MBR) is defined to provide a lower limit of user allocation for each time slot. This is to ensure that while considering the long-term allocation, the per-slot

user application needs are also be met. If this is set to zero, we get an upper bound of the PRA-MaxNetRate objective. Finally, C4 provides the bounds for the resource sharing factor. This gives a total of $KN + M + MN$ constraints and $2MN$ variable bounds.

It is important to note that the objective function couples rates received by users from several BSs during the N slots, and is solved centrally for all cooperating BSs. In PRA-MaxNetRate, a user with a finite traffic request that is moving from a congested region to a low-density, high-rate region, will only be served with the MBR until the high-rate conditions commence. This to ensure maximum network utilization when the user arrives at the high-rate region. PRA-MaxNetRate provides the upper bound on the throughput of any PRA formulation.

3.4.3 Predictive Max-Min Fairness

We now focus on the objective of achieving long-term $\max - \min$ fairness among the users as they traverse multiple BSs To achieve this, resources are allocated such that the minimum user throughput during the prediction window is maximized. This PRA-MaxMin problem can be formulated as follows:

$$\underset{\mathbf{x}}{\text{maximize}} \quad \min_{i} \frac{1}{D_i} \sum_{n=1}^{N} \hat{r}_{i,n}\, x_{i,n} \tag{3.3}$$

$$\text{subject to:} \quad \text{C1, C2, C3, C4,}$$

where the constraints C1-C4 are similar to the PRA-MaxNetRate constraints in Eq. (3.2). Note that the objective in Eq. (3.3) achieves a max-min allocation of the *sum* of the data allocated to each user over the N slots, and not the individual

data at each slot n. This achieves the desired long-term fairness, while C3 provides the MBR at each time slot. The variable D_i in the optimization objective is used to normalize the allocated user data by the total data requested during the N slots. We are interested in maximizing the minimum of this ratio over all users. Since PRA-MaxMin has a piece-wise linear concave objective function, it can be expressed as an equivalent LP by introducing an optimization variable Y and constraint C5, as follows:

$$\underset{\mathbf{x},Y}{\text{maximize}} \quad Y \tag{3.4}$$

$$\text{subject to:} \quad \text{C1, C2, C3, C4}$$

$$\text{C5:} \quad -\frac{1}{D_i}\sum_{n=1}^{N}\hat{r}_{i,n}\,x_{i,n} + Y \leq 0, \qquad \forall\, i \in \mathcal{M}.$$

The PRA-MaxMin allocation can be used for best effort, delay tolerant applications such as FTP downloads or software updates, where MBR may even be zero. The network can then schedule long-term predictive allocations to ensure that users are served equally as they move across multiple cells. Although $\max - \min$ fairness provides a strictly fair allocation, it is generally not a practical fairness objective. Thus, a more general predictive fair allocator is needed, which we discuss next.

3.4.4 Predictive Throughput-Fairness Trade offs

In this section, we formulate predictive resource allocation that provides a trade-off between throughput and long-term fairness.

α-Proportional Fairness

Achieving fairness among users generally comes at the cost of reducing system throughput. For example, $\max-\min$ fairness ensures that the minimum data rate that any user receives is maximized. In other words, an allocation is said to be $\max-\min$ fair if any allocation a_i cannot be increased without decreasing some a_j which is greater than or equal to a_i. This measure gives absolute priority to fairness over system throughput. Proportional fair RA [51] on the other hand provides a trade-off between fairness and throughput and its usefulness in scheduling has been of significant interest in literature and industry. To achieve an arbitrary degree of the fairness-throughput trade-off, a generalization of proportional fair and $\max-\min$ fair allocation was presented in [41]. This is known as the α-proportional fair allocation where the parameter α controls the degree of fairness in the allocation. α-proportional fair allocation is based on defining the following utility function:

$$\phi_\alpha(y) = \begin{cases} \frac{y^{(1-\alpha)}}{(1-\alpha)}, & \text{if } \alpha \geq 0, \alpha \neq 1, \\ \log y, & \text{if } \alpha = 1, \end{cases} \tag{3.5}$$

and thereafter solving the optimization problem:

$$\underset{\mathbf{a}}{\text{maximize}} \quad \sum_{i=1}^{N} \phi_\alpha(a_i) \tag{3.6}$$

$$\text{subject to:} \quad \mathbf{a} \in \mathcal{S},$$

where $\mathbf{a}$ is the user allocation vector and $\mathcal{S}$ is the feasible region of $\mathbf{a}$. Note that when $\alpha \to 0$, α-proportional fairness is reduced to maximum throughput allocation, and

when $\alpha \to 1$, proportional fairness is achieved from the logarithmic utility. It has also been shown that when $\alpha \to 2$, potential delay minimization is obtained, and when $\alpha \to \infty$, $\max - \min$ fairness is achieved [41].

Problem Formulation

In order to plan an allocation that achieves long-term α-proportional fairness over multiple cells, the optimization problem in Eq. (3.6) can be formulated as follows (for $\alpha \neq 1$):

$$\underset{\mathbf{x}}{\text{maximize}} \quad \sum_{i=1}^{M} \frac{\left(\sum_{n=1}^{N} \hat{r}_{i,n}\, x_{i,n}\right)^{(1-\alpha)}}{1-\alpha} \tag{3.7}$$

$$\text{subject to:} \quad \text{C1}, \text{C2}, \text{C3}, \text{C4}.$$

As in Eq. (3.2), the inner summation over time in the objective is introduced to maximize the utility achieved from the total data a user receives during the prediction window. We refer to this formulation as the Predictive α-Proportional Fairness (PPF) which is a convex optimization problem since the constraints are linear, and the objective function is an increasing, concave function. It is important to note that when $\alpha \to 0$ in PPF, the result is a *predictive* Max-Rate scheduler, similar to that formulated earlier in Section 3.4.2. Similarly, the corresponding predictive allocator applies to other values of α.

The problem can also be equivalently represented as follows by introducing additional optimization variables R_i that denote the total data assigned to each user:

$$\underset{\mathbf{x},\mathbf{R}}{\text{maximize}} \quad \sum_{i=1}^{M} \frac{R_i^{(1-\alpha)}}{1-\alpha} \tag{3.8}$$

$$\text{subject to:} \quad \text{C1}, \text{C2}, \text{C3}, \text{C4}$$

$$\text{C5:} \quad \sum_{n=1}^{N} \hat{r}_{i,n}\, x_{i,n} = R_i \qquad \forall\, i \in \mathcal{M}.$$

For the case when $\alpha = 1$, the resulting PPF problem becomes:

$$\underset{\mathbf{x},\mathbf{R}}{\text{maximize}} \quad \sum_{i=1}^{M} \log R_i \tag{3.9}$$

$$\text{subject to:} \quad \text{C1}, \text{C2}, \text{C3}, \text{C4}, \text{C5},$$

which we refer to as predictive proportional fairness. This is also a convex problem due to the concave log utility function.

3.5 Numerical Results and Discussion

3.5.1 Evaluation Set-up

The simulation setup is based on the system models presented earlier in Section 3.3. The evaluation is divided into two stages. We first present and discuss the performance of PRA-MaxNetRate and PRA-MaxMin compared to traditional resource allocators (discussed below) that do not incorporate predictions. This is achieved by solving the problems in Eq. (3.2) and Eq. (3.4.3) using Gurobi [52], a state-of-the-art mathematical programming solver for LPs. We then investigate the performance of the PPF allocator by solving the problems in Eq. (3.4.4) and Eq. (3.9) for various values of α. As PPF is not an LP, we use MOSEK [53] with CVX [54], which is a package for specifying and solving convex programs.

Reference Resource Allocators

To provide a performance reference we consider the following resource allocators.

- **Max-Rate Allocation:** In Maximum Rate (MR) allocation, the user i^* with the highest data rate $\hat{r}_{i^*,n}$ at each slot n is granted full channel access, i.e., $x_{i^*,n} = 1$. This maximizes the network throughput but makes no effort to serve users fairly.

- **Equal Share Allocation:** In Equal Share (ES) allocation, BS air-time is shared equally among the users at each time slot n. If there are $M_{j,n}$ users associated with BS j at time n (i.e., $M_{j,n} = |\mathcal{U}_{j,n}|$), then $x_{i,n} = 1/M_{j,n}$ for each user $i \in \mathcal{U}_{j,n}$. The received rate for each user will therefore be $\hat{r}_{i,n}/M_{j,n}$.

Performance Metrics

- T_{Net}: the average downlink network throughput calculated as the sum of the average data rate of all the users.

- J_{Net}: Jain's fairness index [55] for user throughput, and is computed as:

$$\frac{(\sum_{i=1}^{N} \mathrm{T}_i)^2}{N \sum_{i=1}^{N} \mathrm{T}_i^2}, \tag{3.10}$$

  where T_i is the average user throughput during the N slots. Fairness is computed based on the total data a user receives during N since we are interested in long-term fairness.

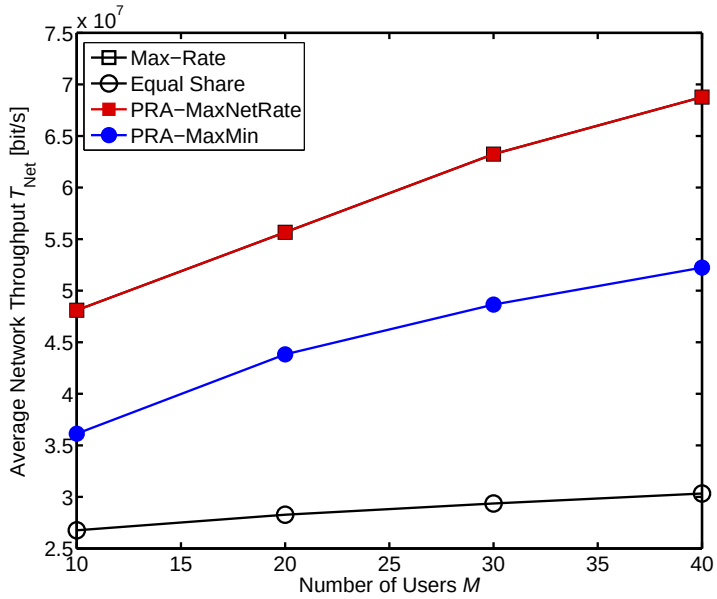
3.5.2 PRA-MaxNetRate and PRA-MaxMin

We evaluate the performance of PRA-MaxNetRate and PRA-MaxMin for both networks of Figure 3.1 where the inter-BS distance is set to 1 km for the 19 cell network. BS transmit power is 40 W and the bandwidth is 10 MHz. In the RWP scenario, user speed is 10 m/s and file download size is 1 Gbit. For the road network, user speed is variable as obtained from the SUMO trace file, and file download size is 750 Mbits. The prediction window T is 250 s for both networks with $\tau = 1$ s. PRA-MaxNetRate and PRA-MaxMin have an MBR of zero to illustrate their performance bound.

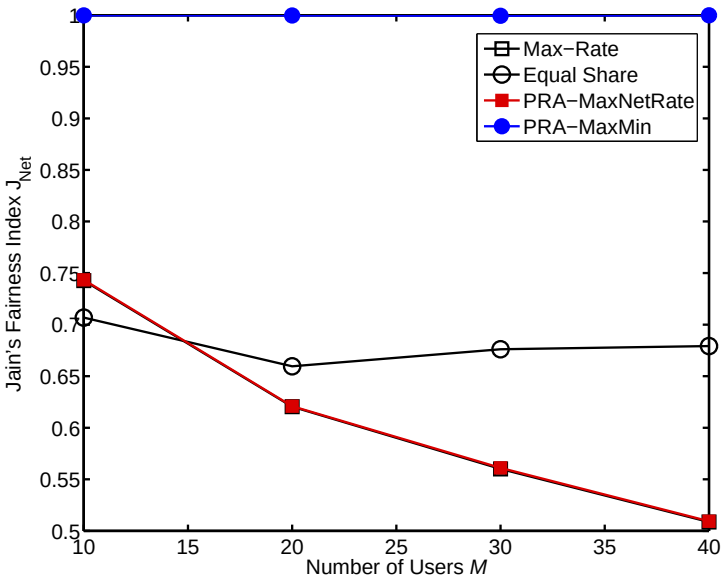
Full Buffer Traffic

Figure 3.4 and Figure 3.5 show the results of PRA-MaxNetRate and PRA-MaxMin with full buffer traffic for the road network and RWP mobility, respectively. In Figure 3.4(a) and Figure 3.5(a) we see that the network throughput T_{Net} of PRA-MaxNetRate converges to the reference Max Rate allocator. This is expected, since with full buffer traffic the throughput cannot be increased further by long-term RA planning. However, significant throughput gains of PRA-MaxMin over the Equal Share allocator are observed, which is due to the opportunistic planning of resources when users are at their peak rates.

Figure 3.4(b) and Figure 3.5(b) compare the fairness levels of the allocators in the road network and RWP mobility respectively. The figures show how PRA-MaxMin ensures a Jain's fairness index of 1 for both mobility scenarios which is a result of the optimization objective in Eq. (3.3). The fairness gain over other allocators is larger in the RWP scenario when the network is saturated since users follow routes that have large variances in the average SNR values, resulting in poor fairness. Also note

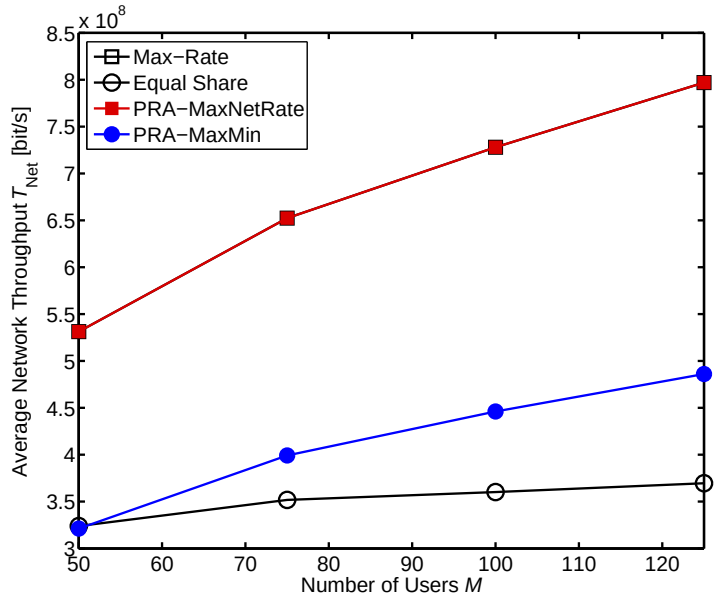


(a) Network throughput.

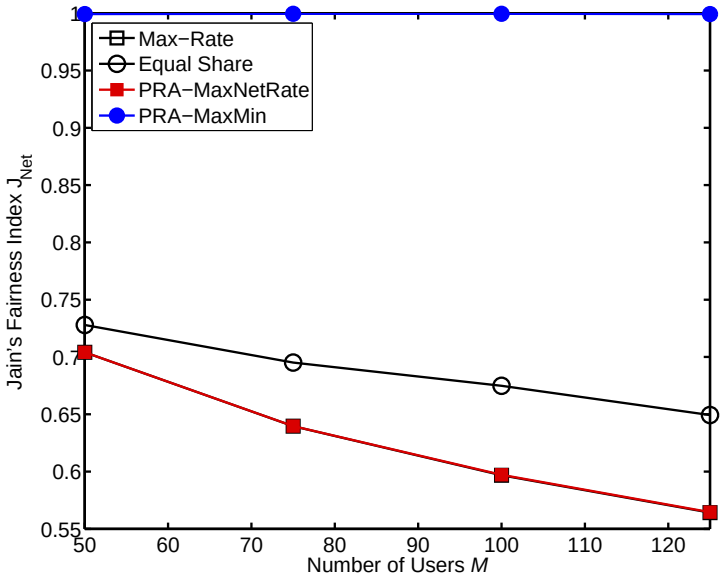


(b) Jain's fairness index.

Figure 3.4: Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for full buffer traffic on the road network. Note that, Max-Rate and PRA-MaxNetRate overlap.



(a) Network throughput.



(b) Jain's fairness index.

Figure 3.5: Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for full buffer traffic with RWP mobility on the 19 cell network. Note that, Max-Rate and PRA-MaxNetRate overlap.

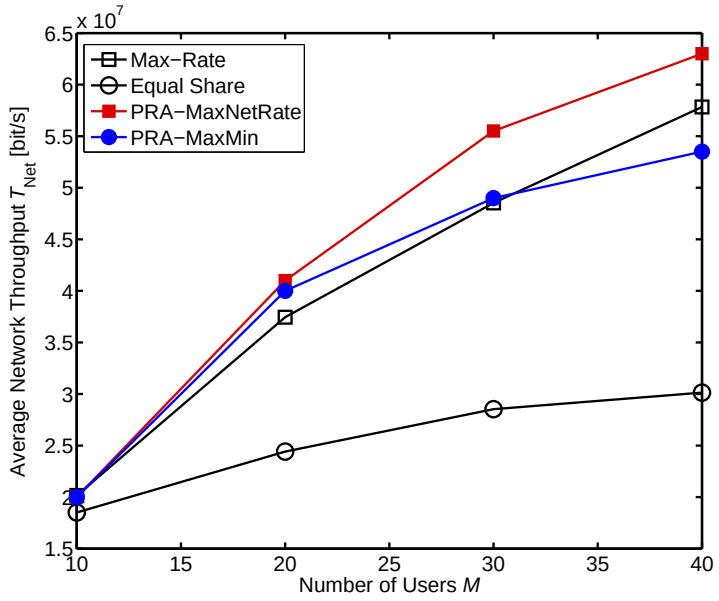
in Figure 3.4(a) that PRA-MaxMin has a high throughput in addition to the high fairness.

File Download Traffic

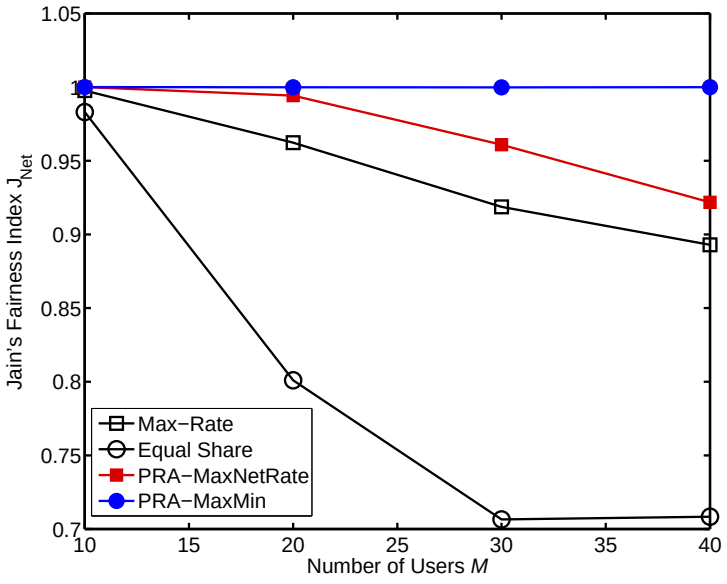
Figure 3.6 and Figure 3.7 illustrate the results of PRA-MaxNetRate and PRA-MaxMin with file download traffic for the road network and RWP mobility respectively. The throughput gains of PRA-MaxNetRate are now apparent in Figure 3.6(a) and Figure 3.7(a), as opposed to the full buffer case where no increase in T_{Net} was possible. As the number of users increases, we see an increase in T_{Net} compared to the MR allocation scheme. This is because PRA-MaxNetRate can delay serving a user in a congested cell, if the user is moving to a lower density, high data-rate region of the network. Such occurrences are more frequent in the 19-cell network with RWP mobility, and therefore the T_{Net} increase is more apparent in this case. This indicates that when users have finite data requests, long-term RA planning can increase overall network throughput. Figure 3.7(b) and Figure 3.6(b) also show that the throughput of PRA-MaxMin can be higher than conventional Max-Rate, while simultaneously achieving a high long-term fairness. This is achieved by opportunistically delaying transmissions to schedule users when they are at their highest rates based on the rate predictions. As the file download size increases, the performance of the PRA schemes will approach the case of full buffer traffic.

3.5.3 Predictive Proportional Fairness

We evaluate PPF for the road network with SUMO generated mobility for the highway scenario (route C). We choose a simpler mobility scenario since we are mainly

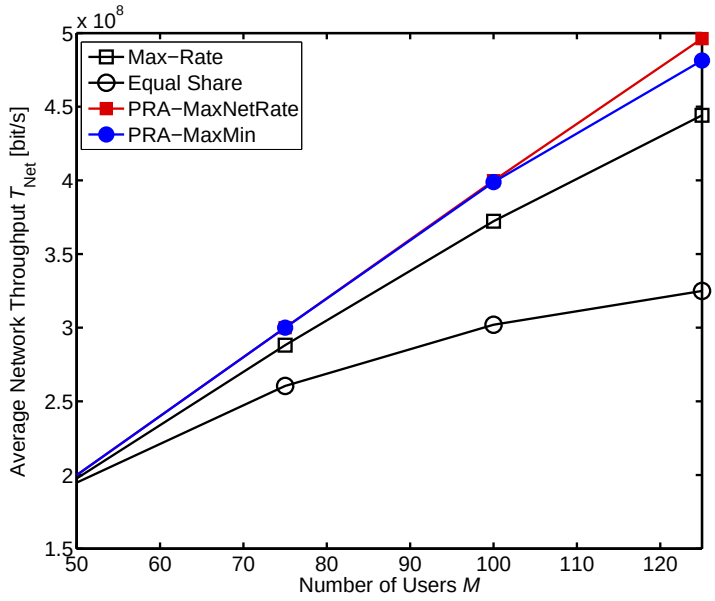


(a) Network throughput.

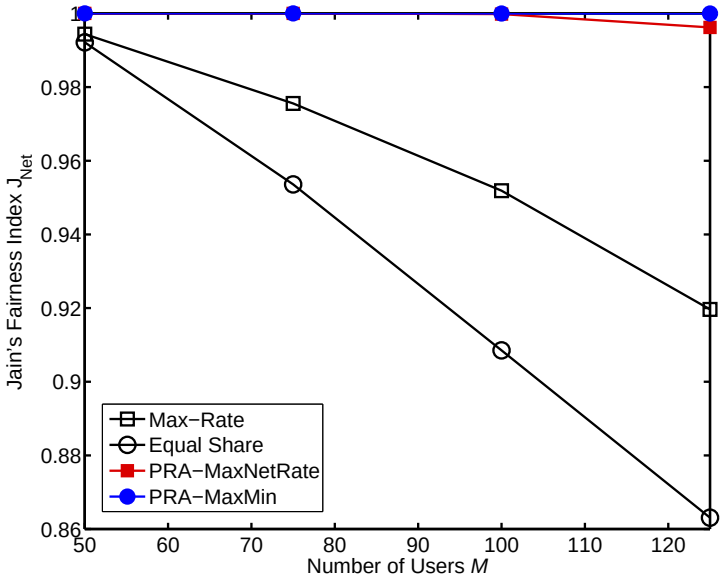


(b) Jain's fairness index.

Figure 3.6: Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for file download on the road network.



(a) Network throughput.



(b) Jain's fairness index.

Figure 3.7: Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for file download with RWP mobility on the 19 cell network.

interested in illustrating the throughput-fairness trade-off that PPF is able to achieve by varying α. BS transmit power is also 40 W and the bandwidth B is 5 MHz. The file download sizes are 1 Gbit and the prediction window is 200 s with $\tau = 1$ s.

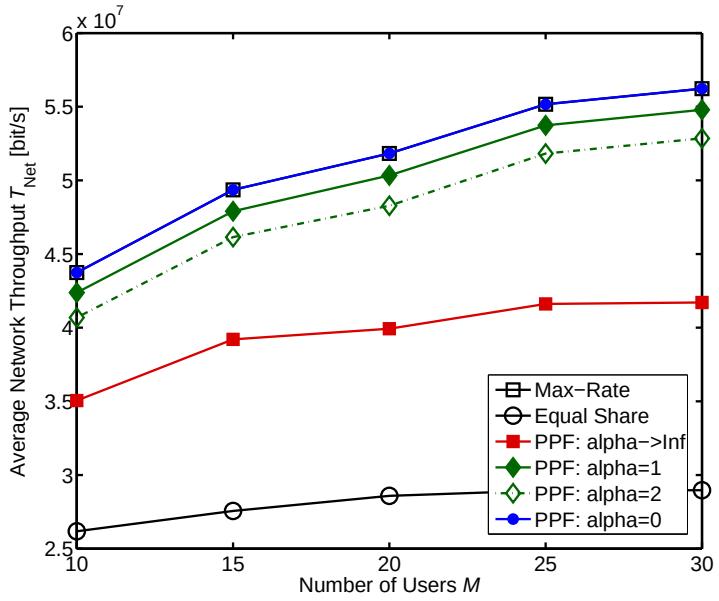
Full Buffer Traffic

Figure 3.8 shows the results of PPF with full buffer user traffic. When $\alpha = 0$, throughput is maximum but fairness is severely affected as illustrated in Figure 3.8(a) and Figure 3.8(b) respectively. On the other hand, when $\alpha \to \infty$, the α-proportional fairness performs as a $\max - \min$ allocator, resulting in the converse behavior of a very high fairness and a low throughput. The results for $\alpha = 1$ (predictive proportional fairness) and $\alpha = 2$ demonstrate the usefulness of the α-proportional fair utility and provide a good trade-off between fairness and throughput.

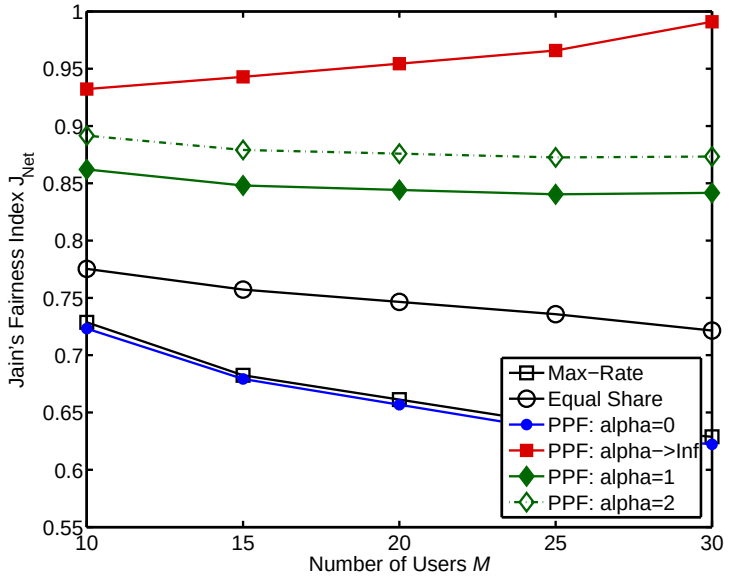
It is worth pointing out that the base-line ES allocator performs poorly in both throughput and fairness. This is due to its lack of knowledge of the rates users will experience, and therefore it does not make opportunistic allocations or planned transmission delays to improve network throughput and user service.

File Download Traffic

In Figure 3.9, we present the results when users have finite traffic requests, which is the more practical case. Here we can see that at low load, network performance is not very different for the α-proportional fairness variants. However, there still is a considerable performance gain over the non-predictive ES allocation scheme. As the load increases, the behavior tends to follow the full buffer traffic scenario. The benefits derived from setting α to 1 or 2 are also apparent from the satisfactory

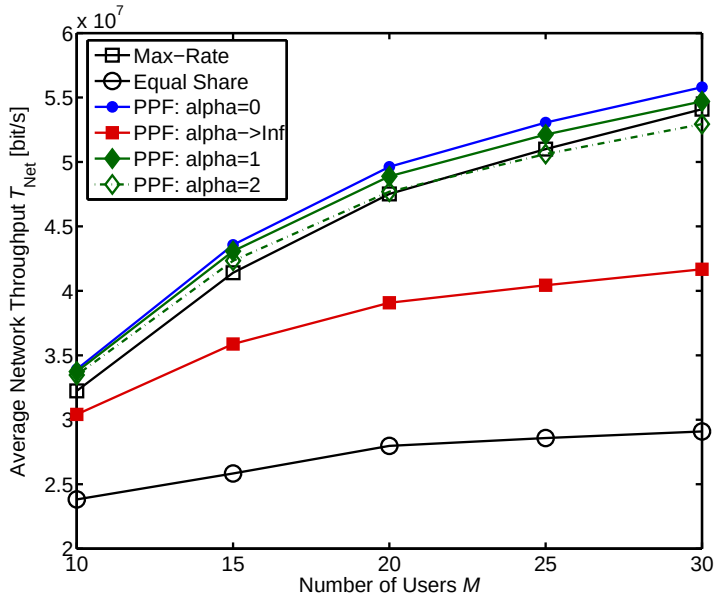


(a) Network throughput.

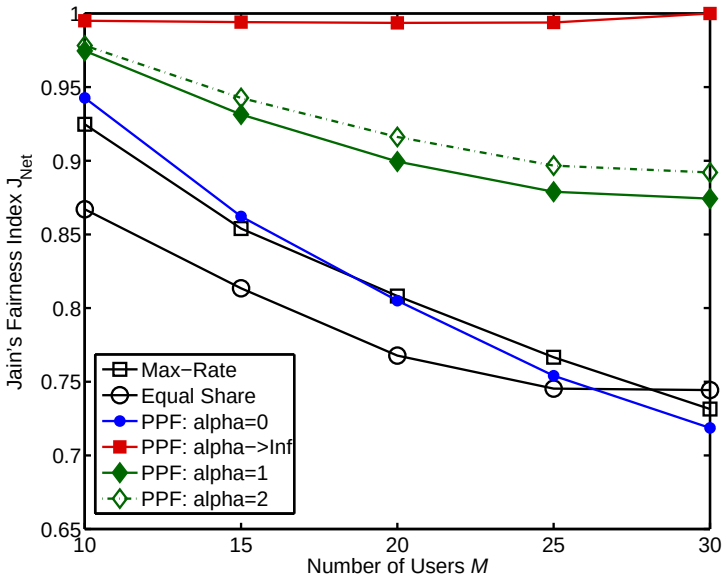


(b) Jain's fairness index.

Figure 3.8: Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for full buffer traffic.

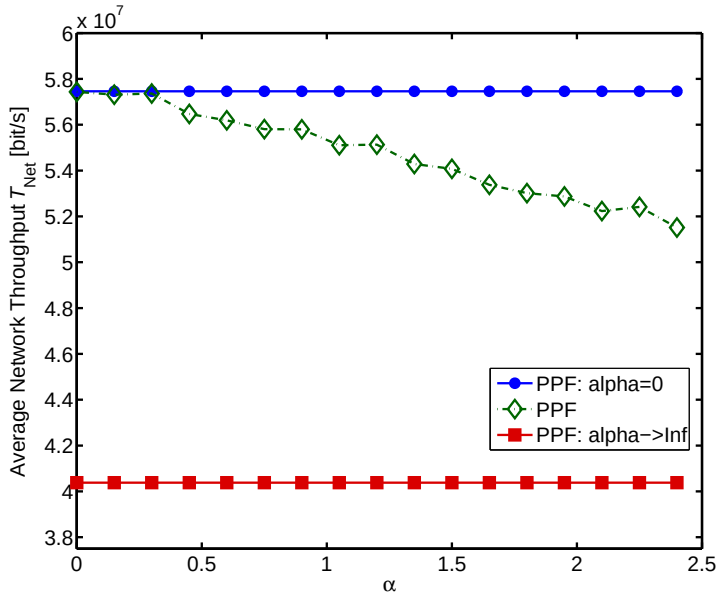


(a) Network throughput.

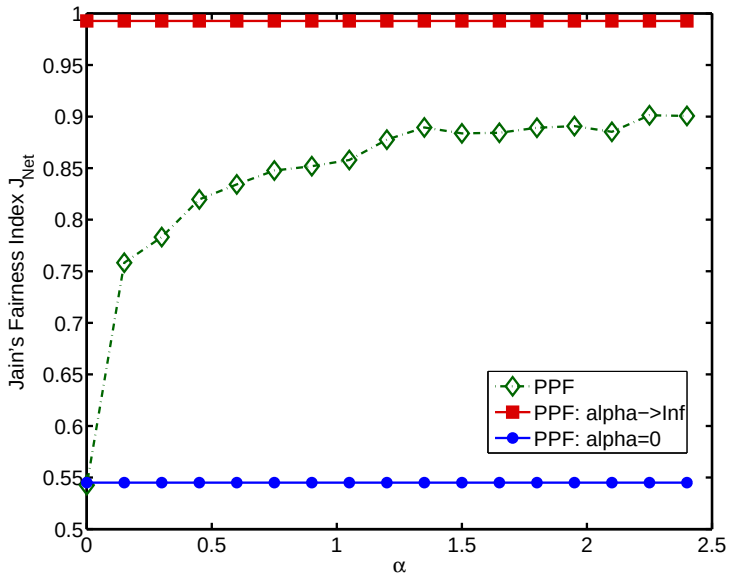


(b) Jain's fairness index.

Figure 3.9: Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. number of users M for file download traffic.



(a) Network throughput.



(b) Jain's fairness index.

Figure 3.10: Network throughput T_{Net} and Jain's Fairness Index J_{Net} vs. α for 40 users and file download traffic.

throughput-fairness trade-off.

Effect of α

Figure 3.10 illustrates the throughput and fairness results for varying values of α with a fixed number of users $M = 40$. We can see that as α increases from 0 to 1, a significant fairness improvement is achieved with a relatively small throughput loss. This indicates that predictive proportional fairness ($\alpha = 1$) can provide a good operating point. As α increases further, the rate of increase in fairness decreases while throughput begins to decrease considerably, thereby indicating that may not be a desirable operating direction. Figure 3.10 also shows the extent by which predictive $\max - \min$ fairness sacrifices throughput to achieve fairness.

3.6 Summary

This chapter introduced *predictive resource allocation*, that leverages user mobility patterns and radio maps to make long-term RA plans spanning multiple cells. The motivation of proposing such an allocation approach is the plethora of navigation and context information available in today's smart phones which can facilitate predictive access schemes. We demonstrated how being aware of a user's upcoming rate allows the network to jointly plan more spectrally efficient rate allocations and improve fairness. In particular, the α-fair utility of the Predictive α-Proportional Fairness allocator provides the flexibility to obtain the desired long-term tradeoff between fairness and throughput. The presented numerical results provide a benchmark of the PRA performance in realistic and random user mobility scenarios. Our findings suggest that significant network and user fairness gains are observed compared to RA

schemes that do not utilize any predictions.

The potential of leveraging mobility patterns and radio maps to develop PRA schemes is not limited to improving throughput and fairness, i.e., PRA formulations are also possible for other QoS objectives. In particular, PRA can be effectively applied to the transmission of stored video content, which is the focus of the following chapter.

Chapter 4

Enhancing Wireless Video Streaming Delivery

4.1 Introduction[1]

The increasing popularity of online media content and video sharing through social networking websites is imposing a myriad of challenges to network operators. Mobile video is now forecasted to account for over 70 percent of the mobile data traffic by 2016 [1]. Much of this is pre-recorded video content such as movies, TV shows, and short clips delivered from popular sites such as YouTube and Netflix. As video content dominates the overall traffic, it shall become the major contributor to network congestion. Furthermore, in vehicular environments, the rapid channel fluctuations and road structures, such as tunnels, result in streaming disruptions as users traverse the network. Consequently, novel paradigms for video delivery are imperative to maintain acceptable quality of experience. In this chapter, we investigate how mobility predictions can be leveraged to enhance both constant-bit-rate and adaptive video streaming.

[1]Parts of this chapter were previously published in [56],[57].

4.1.1 Progressive Video Download Enhancements

Progressive download has become the most popular delivery mechanism for stored videos, accounting for over 50 percent of the Internet traffic in the US [58]. Media files are divided into fragments or chunks and most commonly delivered using HTTP over TCP. Once sufficient content is buffered at the receiver, the media file begins to play. In this chapter, we present video transmission schemes that improve the streaming experience by looking *ahead* at the future rates users are anticipated to receive. Being aware of a user's upcoming rate allows the network to plan rate allocations that prevent or reduce video degradations. For instance, if a user is moving towards the cell edge or a tunnel, the network can increase the allocated wireless resources allowing the user to buffer more video content. *Pre-buffering* this additional data then provides smooth video streaming since the user can consume the buffer while being in poor radio coverage. If, on the other hand, the user is approaching the BS, transmission can be delayed, provided sufficient video content has previously been buffered. Therefore, by coupling knowledge of the users' buffer status and future data rates, the BS can devise *predictive* rate allocation strategies that enhance the streaming experience.

In more detail, our approach makes the following contributions:

1. We develop a Predictive Resource Allocation (PRA) scheme that exploits rate predictions to enhance video streaming, while enabling a *trade-off* between overall network streaming quality and *fairness* in individual user quality. The problem is formulated as a multi-objective LP which provides a benchmark solution.

2. To efficiently solve the aforementioned PRA problem, we develop a polynomial-time algorithm with a performance close to the LP benchmark, at a fraction of the memory and computation requirements. The algorithm is also tunable and can effectively follow the Pareto-optimal trade-off of the multi-objective LP.

4.1.2 Adaptive Video Streaming Enhancements

In progressive streaming, the media is delivered to the client at the same rate or 'quality level' irrespective of the fluctuating capacity of the core network and/or the wireless link between the BS and the UE. This can result in video stalling when a high quality fragment is transmitted during bad channel conditions. To overcome this limitation, multi-quality Adaptive Video Streaming (AVS) has emerged. The essence of AVS is to seamlessly adapt streaming quality to the current wireless data rate. In HTTP-based implementations, such as HTTP Live Streaming (HLS) [59] or Dynamic Adaptive Streaming over HTTP (DASH) [60], the video content is divided into a sequence of small file segments, each containing a short interval of playback time. Each segment is made available at multiple bit rates, and depending on the channel capacity, the suitable segment quality is selected for transmission [61]. This reduces video freezing and is particularly suited for mobile video delivery where users experience channel gain fluctuations.

Typically, in AVS, each client tries to estimate the available bandwidth (e.g., by measuring the average arrival rate of data at the HTTP layer) and then choosing the video rate accordingly. However, making accurate estimations are challenging during congestion, resulting in poor user experience [58],[62]. In a mobile environment, determining the appropriate quality is even more challenging due to the rapid link

fluctuations and uneven spatial traffic distribution [63]. This causes frequent quality variations and video stalls, thereby reducing the long-term Quality of Experience (QoE). In an effort to address these challenges, and enhance AVS, we propose to incorporate user rate predictions based on mobility patterns. First, this information can help plan the required long-term segment qualities that ensure smooth streaming. With such an approach, a user headed to a tunnel will prebuffer several *low* quality segments in advance, even if *current* channel conditions permit high quality video. However, to profit from rate prediction, it is not sufficient to only adapt the video quality of upcoming segments. The wireless channel resources can be controlled as well to ensure that pre-buffering receives the wireless capacity it needs. We therefore propose *in-network* Predictive Adaptive Streaming (PAS), which jointly optimizes RA and segment quality planning using rate predictions. Our work in this direction makes the following contributions:

1. The PAS problem is first modeled and formulated as an Mixed Integer Linear Program (MILP) that enables i) the joint multi-user RA and segment quality optimization, and ii) buffer control of the user device to prevent over buffering.

2. We show that the MILP formulation can be reduced to a two stage LP-based solution. This near-optimal formulation shows a 1% performance gap with respect to the benchmark MILP. However, it holds only for the special case when segment quality levels can be approximated by linear bit rate increases.

3. We present a polynomial-time PAS algorithm for the general case. Results demonstrate that this heuristic exhibits close to optimal performance in eliminating video freezing, with *higher* robustness to rate prediction *errors* compared to the MILP.

4.1.3 Related Work

Rate Predictions in Progressive Video Download

Leveraging rate predictions to optimize RA for wireless video streaming was proposed very recently in a limited number of parallel research efforts. Lu and de Veciana [64] use rate predictions to minimize system utilization and avoid streaming delays of constant bit rate video streams. The authors present a detailed buffer model and formulate the multi-user, single cell case, as a non-convex problem. Then, optimal algorithms for the single user case are developed and significant reductions in BS resource utilization are observed. Kolios *et al.* [65] also propose a similar concept, and discusses the potential energy savings that can be achieved by mobility-aware wireless access. An overview of a single-user algorithm is presented to minimize energy of constant bit rate videos. Both of these works address single-cell, *low-load* scenarios where system utilization can be reduced. However, they do not focus on the multi-user, high-load case where video degradations are to be minimized. In additional, algorithms that provide video quality *fairness* among users are not considered.

Rate Predictions in Adaptive Video Streaming

Exploiting user mobility trajectories coupled with network coverage maps to optimize AVS delivery has shown promising results in several works [66, 67, 68, 69, 70]. For example, Yao *et al.* develop a rate adaptation algorithm that proactively switches to the predicted transmission rates by consulting a stored bandwidth map [66]. However, unlike in our work, the authors do not intend to prebuffer content based on predictions, but to improve TCP rate control and throughput by faster convergence to the available capacity.

Predicting network outages and adapting video rate to provide smooth streaming has been initially proposed by Curcio *et al.* [67]. The results of the work indicate that both the number of rebufferings and the cumulative length of rebuffering delays can be significantly reduced by incorporating network coverage maps. However, the presented mechanisms are based on a distributed *user-centric* solution that calculates the (prediction-based) buffering parameters and reports them to the streaming server. Therefore, they do not address the multi-user resource allocation problem where the BSs can plan resources and segment qualities. In addition, the primary application of the work is real-time media delivery where video rates are adapted. Therefore, methods for long-term prebuffering of stored video content into UE caches is not the primary focus of the work. Thereafter, a more practical approach was followed by Riiser *et al.* [69] for multi-quality, stored video delivery over HTTP. A working prototype is developed with a bandwidth look-up service that is constantly updated by users traversing the network. The results of real-life experiments and extensive simulations demonstrate the effectiveness of *client-based* quality level planning. A similar practical prototype is also evaluated with more emphasis on optimizing the granularity of the bandwidth maps, and averaging algorithms to cluster regions with similar bandwidth by Singh *et al.* [70]. More recently, the effects of bandwidth prediction errors were investigated by Fardous and Kanhere [68]. The authors argue that if the lookahead window is large, the impact of potential errors in bandwidth estimation is compounded, and therefore explore the design space to determine if there is an optimal window size that achieves the best possible performance.

Although the preceding works all demonstrate the use of predictions to optimize AVS quality planning, they focus on distributed single-user solutions. This prevents

1) obtaining network-wide objectives or efficiently trading-off video quality among multiple users, and 2) jointly optimizing BS resource allocation to reserve the required resources for the selected quality levels. To this end, PAS investigates the direction towards *in-network*, or network-assisted, solutions that improve service in multi-user scenarios, and allow prioritization in AVS delivery over cellular networks.

4.1.4 Chapter Outline

After this brief introduction of the proposed prediction-based enhancements to video streaming, we present the details of PRA for progressive video streaming in Section 4.2. The optimal problem formulation is first developed in Section 4.2.2 as a multi-objective LP, after which a near-optimal algorithm is presented in Section 4.2.3. The resulting video streaming enhancements of the proposed schemes are then evaluated in Section 4.2.4 under several network settings.

The second part of this chapter focuses on the PAS framework where both RA and quality are jointly optimized. Section 4.3.3 describes the MILP formulation of PAS, while Section 4.3.4 presents the proposed near-optimal algorithms. In Section 4.3.5, we study the performance and robustness of the PAS schemes by extensive simulation under realistic vehicular mobility. We then present experimental results based on a testbed implementation with real videos and wireless links in Section 4.3.6. Finally, we conclude the chapter in Section 4.4.

4.2 Predictive Resource Allocation for Video Streaming

4.2.1 Preliminaries

We consider a network with a BS set $\mathcal{K}$ and an active user set $\mathcal{M}$. An arbitrary BS is denoted by $k \in \mathcal{K}$ and a user by $i \in \mathcal{M}$. Users request stored video content that is transported using an HTTP-based progressive download mechanism. We assume that the wireless link is the bottleneck, and therefore the requested video content is always available at the BS for transmission. We also follow the system models presented earlier in Section 3.3.

Definitions

If the streaming bit rate is denoted by V [bps], the cumulative number of bits that must be transmitted by time slot n, is $D_{i,n} = V\tau n$, in order to ensure smooth streaming. This is illustrated in Figure 4.1 as the minimum target cumulative data for smooth video playback. The cumulative data received by slot n, given a user resource allocation $x_{i,n}$, is defined as $R_{i,n} = \sum_{n'=1}^{n} x_{i,n'} \hat{r_{i,n'}}$. If this is higher than $D_{i,n} \forall n$, then the video is played smoothly. The difference $R_{i,n} - D_{i,n}$, represents the amount of video content that is *pre-buffered* at the end user at slot n. On the other hand, if $R_{i,n} < D_{i,n}$, the user experiences video stalling, or a lower quality video, and therefore the video experience is degraded as defined in the following metric.

Video Degradation Metric

The Video Degradation (VD) metric quantifies the difference between the cumulative number of bits the user is requesting at the streaming bit rate, and the cumulative rate allocated to the user. It is computed for each user at each time slot, and therefore

a matrix of $\mathrm{VD}_{i,n}$ values can be obtained according to:

$$\mathrm{VD}_{i,n} = [V\tau n - \sum_{n'=1}^{n} x_{i,n'}\,\hat{r}_{i,n'}]^{+}. \tag{4.1}$$

When $\sum_{n'=1}^{n} x_{i,n'}\,\hat{r}_{i,n'} > V\tau n$ it means more is allocated than required, indicating that future video content is pre-buffered, and $\mathrm{VD}_{i,n} = 0$. On the other hand, if the converse holds, then the user experiences video degradation as illustrated in Figure 4.1 between 150 s and 220 s. In other words, VD represents the amount of unfulfilled video demand.

The average network VD over N slots is therefore:

$$\mathrm{VD}_{\mathrm{Net}} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{i=1}^{M}\mathrm{VD}_{i,n}. \tag{4.2}$$

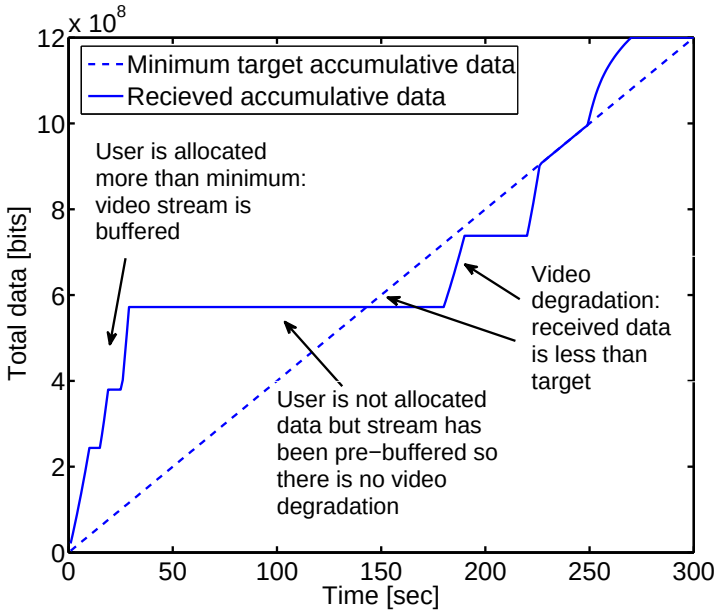


Figure 4.1: Prebuffering and video degradation as observed from *cumulative* data allocated to a user.

We can also compute the *future* VD that a user at slot n will experience (denoted by $\tilde{\text{VD}}_{i,n}$) based on the current cumulative allocation at slot n, and a tentative air-time allocation for the slots $n+1, n+2, \cdots, N$. This is obtained as follows

$$\tilde{\text{VD}}_{i,n} = \sum_{n'=n}^{N} [V\tau n' - \sum_{n''=1}^{n'} x_{i,n''}\, \hat{r}_{i,n''}]^{+}, \tag{4.3}$$

where n' and n'' are dummy variables. A high value of $\tilde{\text{VD}}_{i,n}$ indicates that the user does not have content pre-buffered, and that the tentative future allocation is insufficient (possibly because the user is headed towards poor channel conditions/congested zones). This measure will be used by the algorithm in Section 4.2.2 to prioritize such users and pre-buffer their video content before the poor conditions prevail.

4.2.2 Optimal Problem Formulation: Video Degradation Minimization

The objective of this PRA scheme is to make the optimal pre-buffering allocations to users, in advance, so that they all experience smooth playback. If this is not possible (e.g., at high-load), then the goal is to determine the resource sharing or rate allocation matrix $\mathbf{x}$ that minimizes the VD of the users. This is achieved by exploiting the rate prediction matrix $\hat{\mathbf{r}}$ such that users opportunistically pre-buffer video content during their peak rates, before poor channel conditions prevail.

This optimization problem can be formulated as:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{n=1}^{N} \sum_{i=1}^{M} [V\tau n - \sum_{n'=1}^{n} x_{i,n'}\, \hat{r}_{i,n'}]^{+} \tag{4.4}$$

$$\begin{aligned} \text{subject to:} \quad & \text{C1:} \quad \sum_{i \in \mathcal{U}_{k,n}} x_{i,n} \leq 1, && \forall\, k \in \mathcal{K}, n \in \mathcal{N}, \\ & \text{C2:} \quad \sum_{n=1}^{N} \hat{r}_{i,n}\, x_{i,n} \leq \tau N V, && \forall\, i \in \mathcal{M}, \\ & \text{C3:} \quad 0 \leq x_{i,n} \leq 1 && \forall\, i \in \mathcal{M}, n \in \mathcal{N}. \end{aligned}$$

Constraint C1 expresses the resource limitation at each base station. It ensures that the sum of the air-time of all users associated with BS k is equal to 1 at every time slot. C2 limits the amount of video content delivered to a user during the N slots to the total amount request by that user. Finally, C3 provides the bounds for the resource sharing factor.

While solving the problem in Eq. (4.2.2) will minimize the total VD over the prediction window, fairness is not accounted for. In order to jointly minimize the *total* network VD and the *individual* user video degradations during the N slots, we define the following multi-objective optimization problem:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \frac{\alpha}{MN} \sum_{i=1}^{M} \sum_{n=1}^{N} \frac{\text{VD}_{i,n}}{\tau n V} + \frac{\beta}{N} \max_{i} \sum_{n=1}^{N} \frac{\text{VD}_{i,n}}{\tau n V} \tag{4.5}$$

$$\text{subject to:} \quad \text{C1}, \text{C2}, \text{C3}.$$

The first term in the objective is the normalized network-wide VD, while the second term represents a *min-max* objective for the normalized individual user VD, and $\alpha, \beta \in [0, 1]$. Although the constraints in Eq. (4.5) are linear, the objective is not, due to the $[\cdot]^+$ operator in Eq. (4.1), and the *min-max* component of the objective function. However, by introducing auxiliary variables we can express the problem

defined in Eq. (4.5) as the following equivalent LP:

$$\underset{\mathbf{x},\mathrm{Def},Y}{\text{minimize}} \quad \alpha \sum_{i=1}^{M} \sum_{n=1}^{N} \frac{\mathrm{Def}_{i,n}}{\tau n V} + \beta\, Y \tag{4.6}$$

$$\begin{aligned}
&\text{subject to:} && \mathrm{C1}, \mathrm{C2}, \mathrm{C3} \\
&\forall\, i, n && \mathrm{C4}: \quad \tau n V - \sum_{n'=1}^{n} x_{i,n'}\, \hat{r}_{i,n'} - \mathrm{Def}_{i,n} \leq 0, \\
&\forall\, i && \mathrm{C5}: \quad \sum_{n=1}^{N} \frac{\mathrm{Def}_{i,n}}{\tau n V} - Y \leq 0, \\
&\forall\, i, n && \mathrm{C6}: \quad \mathrm{Def}_{i,n} \geq 0.
\end{aligned}$$

In this reformulation, minimizing the additional variables $\mathrm{Def}_{i,n}$ is now equivalent to minimizing $\mathrm{VD}_{i,n}$ as computed in Eq. (4.1). The *min-max* objective component is replaced with the variable Y and C5. Constraint C5 ensures that Y takes the value of the largest user VD, and therefore minimizing it is equivalent to minimizing the max of user VD. Although the problem is now linear, it has a large number of constraints and optimization variables, requiring large-scale LP solvers, and significant memory and time to solve. Therefore, solving Eq. (4.2.2) can only serve as an offline performance benchmark, which we refer to as VDMin-Opt. For real-time implementation, we present a heuristic algorithm in the following section.

4.2.3 Proposed VDMin Algorithm

The main idea of the proposed VD minimization algorithm is to first keep track of the *cumulative* rates allocated to users up to slot n. Then, the *future* video degradations users are predicted to experience are computed, and resource allocations $x_{i,n}$ are made such that the user VDs are minimized.

Algorithm Steps

As in the VDMin-Opt formulation, this algorithm jointly minimizes $\mathrm{VD}_{\mathrm{Net}}$ and individual user VD. The essence of the algorithm is the definition of a new *rate allocation metric* to make air-time allocations $x_{i,n}$. The metric can also be *tuned* to trade off fairness in user VD with network $\mathrm{VD}_{\mathrm{Net}}$. In more detail, it consists of the following steps:

- *Step 1*: Initialize $x_{i,n} = 0$ for all the users and time slots.
- *Step 2*: Compute the *future* $\tilde{\mathrm{VD}}_i$ each user will experience at slot n, using Eq. (4.3).
- *Step 3*: Each base station k allocates the full air-time at slot n to the user i^* (i.e., $x_{i^*,n} = 1$) that satisfies

$$i^* = \arg\max_i \; r_{i,n} \, \tilde{\mathrm{VD}}_{i,n}^{\gamma} \quad \forall i \in \mathcal{U}_{k,n}. \tag{4.7}$$

  The intuition of this allocation metric is to prioritize users with *both* a high *current* channel quality and a high *future* video degradation. Therefore, users will opportunistically pre-buffer their content when their channel quality is good, and before poor future conditions prevail. The parameter γ controls the influence of the future user VD in the metric. A higher γ will prioritize users with VD and provide more fairness.

- *Step 4*: Repeat steps 2 and 3 for all $n \in \mathcal{N}$.
- *Step 5*: Calculate $\mathrm{VD}_{\mathrm{Net}}$ using Eq. (4.2).

- *Step 6*: Repeat steps 2-5 until there is no more decrease in VD_{Net}.

Note in the first iteration, $\mathbf{x} = 0$ in the computation of Eq. (4.3), and therefore step 3 will not exploit future VD information. However, subsequent iterations of steps 2-5 will allocate $x_{i,n}$ based on the values of $x_{i,n'}$, $\forall n' = n+1, n+2, \ldots N$ of the previous iteration. As $\hat{r}_{i,n}$ does not change over iterations, the selection in step 3 changes in the direction of decreasing VD. It was found that the algorithm converges within four to six iterations, as observed for the various network and mobility settings in Section 4.2.4. The complete procedure is presented in Algorithm 1, which we refer to as VDMin-Alg.

Computational Complexity

The computational complexity of VDMin-Alg is primarily dominated by the computation of $\tilde{\text{VD}}_{i,n}$ in Eq. (4.3) which takes $O(N^2)$ time for each user. Therefore,

Algorithm 1 PRA Video Streaming Algorithm: VDMin-Alg

Require: $\hat{r}_{i,n}, \mathcal{U}_{k,n}, V, \tau, , M, K, N$
1: Initialize $x_{i,n}, \tilde{R}_{i,n} = 0 \quad \forall\, i, n$
2: **repeat** {allocation iterations}
3: Calculate VD_{Net} using Eq. (4.2).
4: **for all** time slots n **do**
5: Reset $x_{i,n} = 0 \quad \forall i$.
6: **for all** base stations k **do**
7: **for all** users $i \in \mathcal{U}_{k,n}$ **do**
8: Calculate the *future* VD using Eq. (4.3)
9: **end for**
10: Set $x_{i^*,n} = 1$ to i^* with the highest $\hat{r}_{i,n}\, \tilde{\text{VD}}_{i,n}^{\gamma}$.
11: **end for**
12: **end for**
13: Calculate VD_{Net} after allocation.
14: **until** {no more decrease in VD_{Net}}
15: **return** $\mathbf{x}$

the overall complexity is of the order $O(MN^3)$ to make the allocation plan for the upcoming N slots in a network with M users.

4.2.4 Performance Evaluation

Simulation Set-up

We consider two network and mobility scenarios for evaluation. The first is the six BS road network with vehicular mobility shown in Figure 3.1(a), and the second is the 19 cell network illustrated in Figure 3.1(b), where users move according to the RWP mobility. The inter-BS distance is set to 1 km, with a BS transmit power of 40 W, a center carrier frequency of 2 GHz, and a bandwidth of 10 MHz. The user speed S is set to 10 m/s for the RWP mobility model, and the video streaming rate V is set to 3 Mbits. We consider a prediction window N of 200 slots with a slot duration τ of one second. Simulations are repeated 50 times to obtain the average values of the following metrics.

- VD_{Net}: the average network VD as defined in Eq. (4.2).
- $J_{\text{Net}}^{\text{VD}}$: Jain's fairness index for user VD over the N slots, and is computed as

$$\frac{(\sum_{i=1}^{M} \text{VD}_i)^2}{M \sum_{i=1}^{M} \text{VD}_i^2}, \tag{4.8}$$

  where VD_i is the average individual user VD during the prediction window N.

We compare the performance of the PRA-VDMin schemes against two baseline approaches that do not exploit rate predictions: ES and Rate-Proportional (RP). In ES, air-time is shared equally among the users at each time slot. If there are $N_{k,n}$ users associated with BS k at time n, then $x_{i,n} = 1/N_{k,n}\, \forall i \in \mathcal{U}_{k,n}$. The RP allocator

is designed to be more spectrally efficient but not completely fair to users. Here, the air-time assigned to each user i, at slot n, is in proportion to the achievable data-rate $\hat{r}_{i,n}$ of that user. Therefore, $x_{i,n} = \hat{r}_{i,n} / \sum_{i \in \mathcal{U}_{k,n}} \hat{r}_{i,n}$ in RP.

Numerical Results and Discussion

Figure 4.2 illustrates the Pareto-optimal trade-off between VD_{Net} and $J_{\text{Net}}^{\text{VD}}$ that the PRA-VDMin mechanisms achieve. First, we observe the significant improvements in *both* VD_{Net} and VD fairness of the proposed predictive schemes compared to the RP and ES baseline allocators. We can see that video degradation can be reduced by 50% without sacrificing fairness. The figure also demonstrates how the proposed VDMin-Alg closely follows the Pareto-optimal benchmark curve of the VDMin-Opt that is solved offline. By varying γ in VDMin-Alg, the $\text{VD}_{\text{Net}} - J_{\text{Net}}^{\text{VD}}$ trade-off can be effectively controlled. Further, the deviation from the benchmark VDMin-Opt is lower in the region that offers a good trade-off, with $\gamma = 1$ providing a suitable operating point.

We now discuss the impact of network load on the gains, and on the performance of the VDMin-Alg algorithm. Figure 4.3(a) shows the effect of varying the number of users on the experienced video degradation for the RWP scenario. To illustrate the lowest achievable VD, we set $\alpha = 1, \beta = 0$ and refer to this setting as VDMin-Opt-Greedy. We can clearly see the large VD savings which increase with network load compared to the base-line RP and ES allocators. While the baseline RP outperforms the ES, it has a lower fairness in VD as illustrated in Figure 4.3(b). VDMin-Opt-Greedy also performs poorly in terms of VD fairness, particularly at higher loads. In Figure 4.3(a) and Figure 4.3(b) we also depict the performance of the proposed

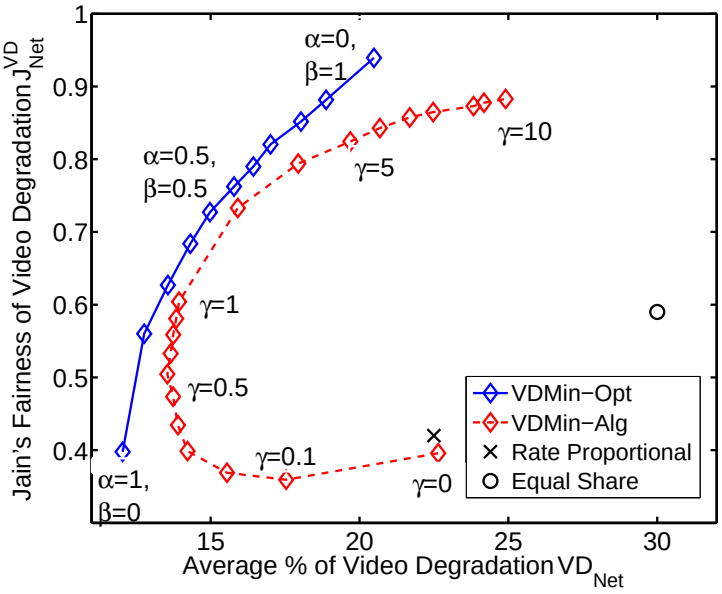
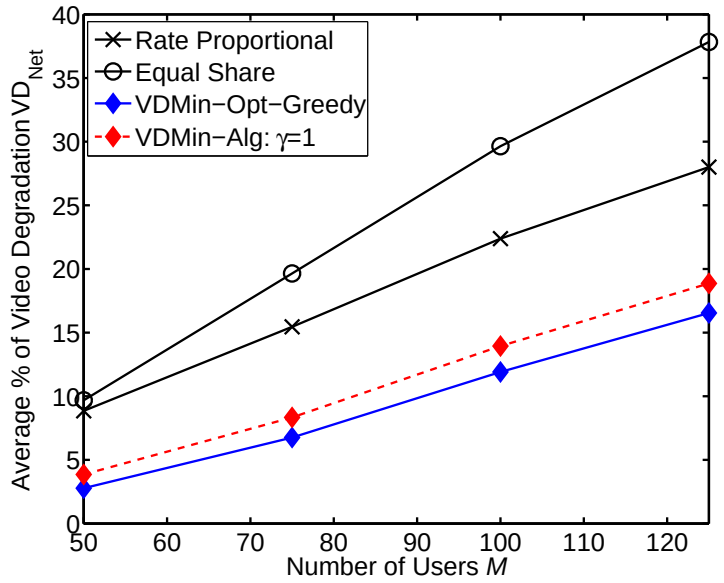


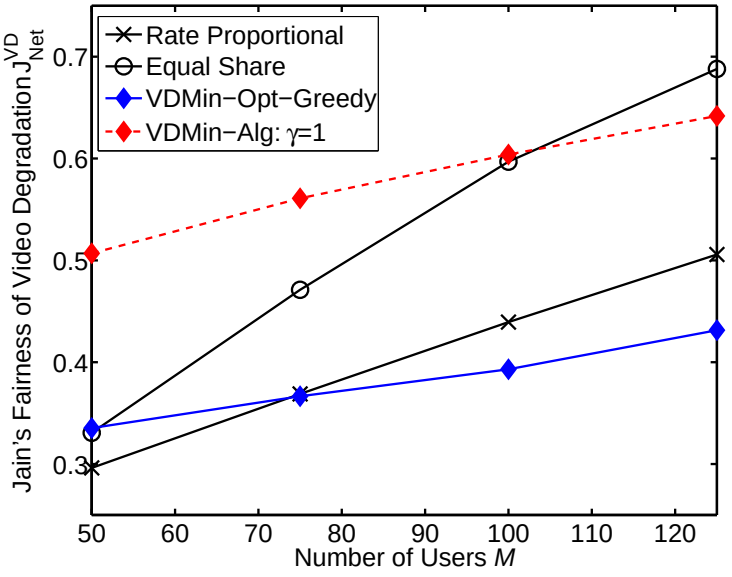
Figure 4.2: Pareto-optimal trade-off in VD; RWP scenario with $M = 100$.

PRA algorithm (VDMin-Alg) at the selected operating point of $\gamma = 1$. We can see that it follows the benchmark VDMin-Opt-Greedy closely in reducing VD, while simultaneously providing significant fairness gains.

In Figure 4.4 a similar study is conducted on the road network of Figure 3.1(a). We observe that VDMin-Alg follows VDMin-Opt even more closely in terms of VD minimization, thereby supporting its generality. The fairness gains are less in this case, which is expected, since all the users follow similar trajectories and therefore, all the allocators provide a reasonable degree of fairness. Nevertheless, VDMin-Alg still provides some fairness gains at close to optimal VD minimization.

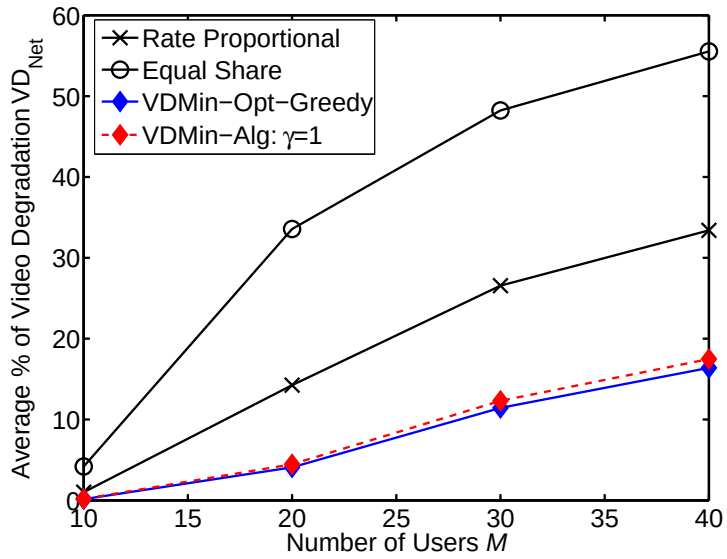


(a) Video Degradation

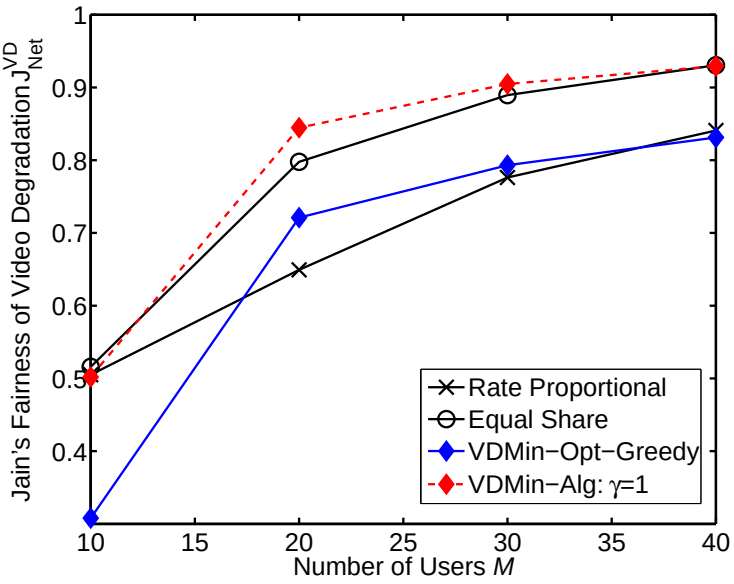


(b) Video Degradation Fairness

Figure 4.3: Network Video Degradation and Jain's Fairness of Video Degradation vs. number of users with RWP mobility.



(a) Video Degradation



(b) Video Degradation Fairness

Figure 4.4: Network Video Degradation and Jain's Fairness of Video Degradation vs. number of users with road network mobility.

4.3 Predictive Adaptive Streaming: Jointly Optimizing RA and Quality Planning

We now investigate how predictions can be used not only in resource allocation, but in video segment quality planning as well. Knowledge of future user rates are used to enhance traditional HTTP-based adaptive video streaming by *jointly* optimizing rate allocation and video quality over a time horizon.

4.3.1 System Overview

Consider a network with a BS set $\mathcal{K}$ and an active user set $\mathcal{M}$. Users request stored video content that is transmitted using adaptive bit rate streaming over HTTP, e.g., HLS [59] or DASH [71]. Time is divided in slots of equal duration τ, during which the wireless channel can be shared among multiple users. The system objective is to jointly optimize rate allocation and video bit rate planning for all the users in the network. The optimal long-term plan is achieved by utilizing knowledge of the future channel gains that each user will experience. We assume that the wireless link is the bottleneck, and therefore the requested video content is always available at the BS for transmission. The link model and resource sharing is implemented as in the system models of Section 3.3, and a summary of the frequently used symbols can be found in Table 4.1

4.3.2 Adaptive Video Streaming Model

In AVS over HTTP, the video content is divided into a sequence of small HTTP-based file segments. Each video segment is then pre-encoded in multiple versions, each with a specific video bit rate and resolution, or 'quality level'. Higher quality segments

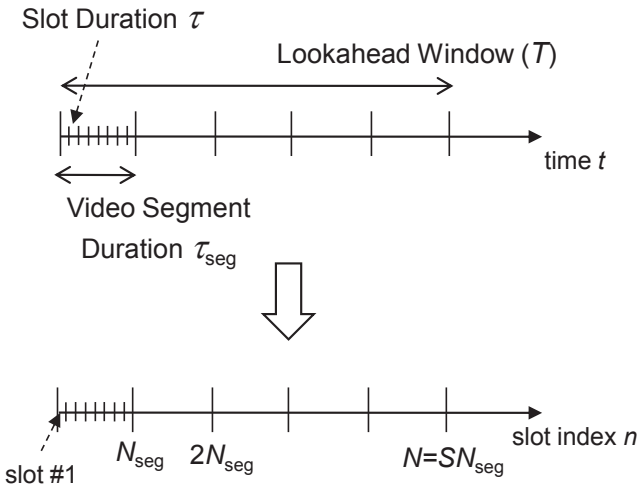
will be larger in size but represent similar playback durations [72]. We denote the segment quality levels by $l \in \mathcal{Q}$, where $\mathcal{Q} = \{1, 2, \ldots, q_{\max}\}$, and $q_{\max}$ is the maximum quality level. The function $f^{Q}_{\text{rate}}(\cdot)$ maps the quality level to the corresponding bit rate. Higher segment qualities will require higher bit rates for successful reception, and therefore $f^{Q}_{\text{rate}}(\cdot)$ is an increasing function of l. If V denotes the maximum quality bit rate, then $f^{Q}_{\text{rate}}(q_{\max}) = V$. An example of $f^{Q}_{\text{rate}}(\cdot)$ with equal bit rate increases of $V/q_{\max}$ is illustrated in Figure 4.5(b).

To assign the quality level of each user segment, we define the binary decision matrix $\mathbf{q} = (q_{i,s,l} \in \{0, 1\} : i \in \mathcal{M}, s = \{1, 2, \ldots, S\}, l \in \mathcal{Q})$. If there are three quality levels, and $q_{i,s,1} = 1$, this means that user i will receive segment s at quality level 1; and the remaining quality level indices should be zero, i.e., $q_{i,s,2} = 0$ and $q_{i,s,3} = 0$. Therefore, to ensure that only one quality level is selected for each user $\sum_{l=1}^{q_{\max}} q_{i,s,l} = 1 \; \forall i, s$.

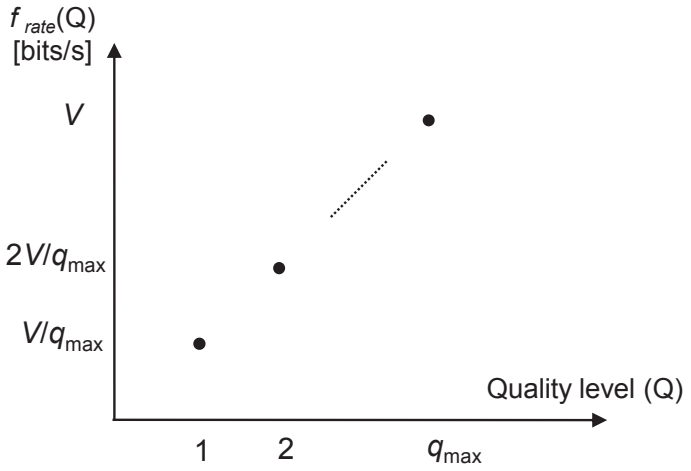
We denote the duration of each video segment by τ_{seg}, which is a multiple of τ. The prediction window T is also selected to be divisible by τ_{seg} as shown in Figure 4.5(a). In terms of time slots, $N_{\text{seg}} = \tau_{\text{seg}}/\tau$ denotes the number of slots that make up one video segment. Also, $S = N/N_{\text{seg}}$ denotes the number of video segments during T, as illustrated in Figure 4.5(a).

4.3.3 Optimal Problem Formulation

In this section, we formulate the PAS problem as an MILP to provide performance benchmarks for evaluation. The objective is to maximize video quality over all users, under the constraint that no video stalling occurs. This is achieved with the knowledge of the future user channel gains over a predefined prediction window. The MILP



(a) Relationships between the defined time durations and slot indices.



(b) Sample quality level to bitrate mapping function.

Figure 4.5: System models and notation.

Table 4.1: Summary of Frequently Used Symbols in PAS

Symbol	Description
i	User index, $i = \{1, 2, \cdots, M\}$
k	BS index, $k = \{1, 2, \cdots, K\}$
n	Time slot index, $n = \{1, 2, \cdots, N\}$
$q_{\max}$	Maximum quality level
s	Segment index, $s = \{1, 2, \cdots, S\}$
K	Number of BSs in the network
M	Number of users in the network
N	Number of slots in the prediction window [slots]
N_{seg}	Number of slots in a video segment [slots]
S	Number of segments in the prediction window
T	Duration of the prediction window [s]
V	Video streaming rate at $q_{\max}$ [bps]
τ	Duration of a time slot [s]
τ_{seg}	Duration of a video segment [s]
L_i	Buffer limit of user i [bits]
$q'_{i,s}$	Quality level of segment s for user i, $q'_{i,s} \in [1, q_{\max}]$
$q_{i,s,l}$	Binary variable for quality level l of segment s for user i
$\hat{r}_{i,n}$	Link rate of user i at slot n [bits]
$R_{i,n}$	Cumulative data allocated to i by slot n [bits]
$\mathcal{U}_{k,n}$	Set containing the indices of users associated with BS k at slot n
$\text{VD}_{i,n}$	Video degradation perceived by user i at slot n
$x_{i,n}$	Fraction of air-time assigned to user i at slot n

formulation jointly determines the optimal air-time $\mathbf{x}$ and segment quality $\mathbf{q}$ matrices, provided the user-rate matrix $\hat{\mathbf{r}}$ is given. To do so, we first derive the relationship between the cumulative allocated resources and long-term segment quality planning.

Joint Rate Allocation and Video Segment Quality Planning

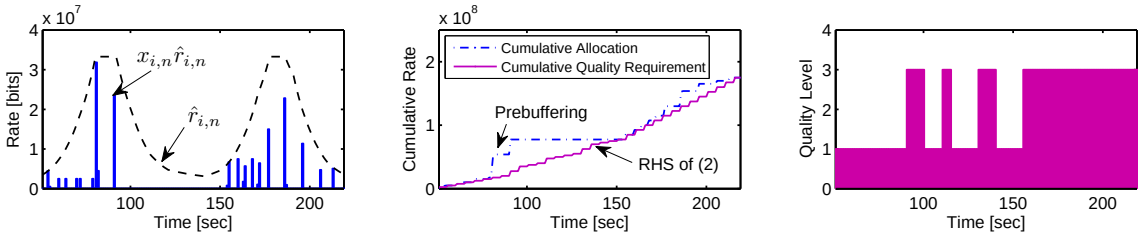
Consider a user steaming a stored video at the *maximum* quality level $q_{\max}$. For the video to play back without interruptions, the user should receive a constant rate of V bits every second. Alternatively, a bulk of video content can be transmitted at once and buffered at the user's device, after which transmission can be suspended momentarily without causing video stalling. Therefore, we are interested in the cumulative video content transmitted (and stored) at the user's device. At time slot n, this is given by $\sum_{n'=1}^{n} x_{i,n'}\,\hat{r}_{i,n'}$, which should be greater than $n\tau V$, for all n, to ensure smooth playback. With the knowledge of $\hat{r}_{i,n'}$, rate allocation can be made to ensure smooth playback, by pre-buffering future content for a user heading towards poor channel conditions.

However, during high network load it may not be possible to provide uninterrupted playback at the highest quality, even with predictive buffering. In this case, the quality level of a select number of video segments should be lowered. This joint relationship between the cumulative allocated rate and segment quality selection that ensures smooth playback is captured in the following constraints:

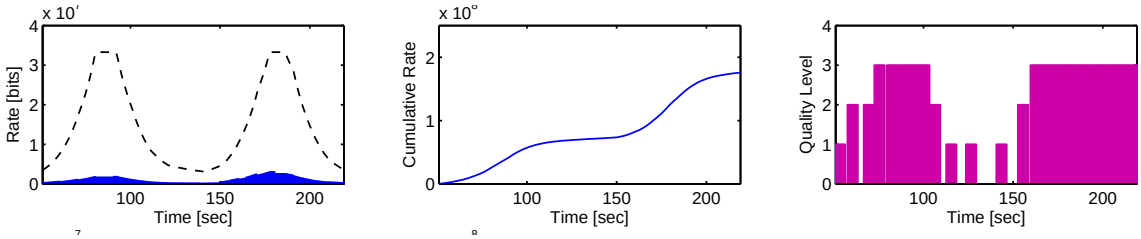
$$\sum_{n'=1}^{sN_{\text{seg}}} x_{i,n'}\,\hat{r}_{i,n'} \geq \tau_{\text{seg}} \sum_{s'=1}^{s} \sum_{l=1}^{q_{\max}} q_{i,s',l}\, f^{Q}_{\text{rate}}(l) \quad \forall\, i, \forall s, \tag{4.9}$$

$$\sum_{l=1}^{q_{\max}} q_{i,s,l} = 1 \qquad \forall\, i \in \mathcal{M}, \forall s \in \{1, 2, \ldots, S\}. \tag{4.10}$$

The left hand side (L.H.S.) of Eq. (4.9) gives the cumulative bits allocated to user i at the slots that correspond to the end of each segment, whereas the right hand side (R.H.S.) expresses the cumulative bits *required* to stream up to s video segments



(a) Predictive Adaptive Streaming (PAS).



(b) Traditional non-predictive RA.

Figure 4.6: Joint rate allocation and segment quality with (a) Predictive Adaptive Streaming, (b) traditional non-predictive approach.

at the quality levels specified by $q_{i,s,l}$. Note that the constraint indicates that for uninterrupted playback, segment s must be fully downloaded by time slot sN_{seg}. The primary objective is to exploit the knowledge of $\hat{r}_{i,n}$ such that segment quality is maximized, without violating Eq. (4.9).

An illustrative example of joint rate-quality optimization for a system of 30 users traversing two cells is shown in Figure 4.6. Figure 4.6 (a) shows that with PAS, as the user approaches the first cell center, more air-time is granted and content is pre-buffered. However, the video segments are pre-buffered at the optimal mix of high and low quality levels that will not cause playback interruptions. This is compared to the case in Figure 4.6 (b) where users share the air-time equally, and segment quality is based on the *current* link capacity with no regard to future planning. We can see that although the user enjoys a higher quality video initially, several buffer underflows

occur at the cell edges.

PAS MILP Problem Definition

In addition to the key constraint in Eq. (4.9), a limit can also be imposed on the number of bits that are pre-buffered at the user's device. This may be due to the video client and network policy, or device limitations. If L_i denotes the limit for user i, then we have the constraint

$$\begin{aligned}\sum_{n'=1}^{n} x_{i,n'}\,\hat{r}_{i,n'} - \tau_{\text{seg}} \sum_{s'=1}^{\lfloor n/N_{\text{seg}} \rfloor} \sum_{l=1}^{q_{\max}} q_{i,s',l}\, f^{Q}_{\text{rate}}(l) \\ - \frac{\tau_{\text{seg}}}{N_{\text{seg}}}(n \bmod (N_{\text{seg}})) \sum_{l=1}^{q_{\max}} q_{i,\lceil n/N_{\text{seg}} \rceil,l}\, f^{Q}_{\text{rate}}(l) \leq L_i \quad \forall\, i, \forall n.\end{aligned} \tag{4.11}$$

The L.H.S. of Eq. (4.11) determines the difference between the cumulative allocated bits and the played back stream at every slot n, and therefore denotes the buffered bits. In particular, the second term accounts for the bits of previously played video segments, and the third term represents the portion of the current segment that has been played.

Additionally, we have the BS resource constraint which limits the sum of the user air-time fractions to unity, i.e.:

$$\sum_{i \in \mathcal{U}_{k,n}} x_{i,n} \leq 1, \qquad \forall\, k \in \mathcal{K},\, \forall n \in \{1, \ldots, N\}. \tag{4.12}$$

This constraint is applied at each BS, where the summation is over all users i associated with BS k at slot n.

Finally, the total number of bits allocated to a user during the N slots should be

equal to that specified by the segment quality plan, which is expressed by

$$\sum_{n=1}^{N} x_{i,n}\,\hat{r}_{i,n} = \tau_{\text{seg}} \sum_{s=1}^{S} \sum_{l=1}^{q_{\max}} q_{i,s,l}\, f^{Q}_{\text{rate}}(l) \qquad \forall\, i \in \mathcal{M}. \tag{4.13}$$

Hence, the PAS-MILP problem can be formulated as

$$\underset{\mathbf{x},\mathbf{q}}{\text{maximize}} \quad \sum_{\forall i \in \mathcal{M}} \sum_{s=1}^{S} \sum_{l=1}^{q_{\max}} q_{i,s,l}\, f^{Q}_{\text{rate}}(l) \tag{4.14}$$

$$\begin{aligned} \text{subject to:} \quad & \text{Constraints}: \; Eq.\ (4.9), Eq.\ (4.10), Eq.\ (4.11), Eq.\ (4.12), Eq.\ (4.13), \\ & 0 \le x_{i,n} \le 1, \;\; \forall\, i \in \mathcal{M}, \forall n \in \{1, \ldots, N\}, \\ & q_{i,s,l} \in \{0, 1\} \;\; \forall i \in \mathcal{M}, \forall s \in \{1, \ldots, S\}, \forall l \in \mathcal{Q}. \end{aligned}$$

Clearly, solving large instances of the MILP defined in Eq. (4.14) is computationally intractable, so we now develop two approximate, but tractable solutions that provide close to optimal results.

4.3.4 Proposed PAS Algorithms

LP-Based Solution for Linear Quality Bit rates

In some cases, the bit rates corresponding to the quality levels can be approximated to follow a linear increase, as illustrated in Figure 4.5(b). We use this property to present a two stage PAS solution where, first, an LP is solved to determine the rate allocation and, then, it is followed by an algorithm to determine the discrete segment quality levels. To do so, we define the segment quality matrix as $\mathbf{q}' = (q_{i,s} \in [1, q_{\max}] : i \in \mathcal{M}, s = \{1, 2, \ldots, S\})$, which now takes real values instead of binary values, which facilitates having two indices instead of three. Since the quality-rate relationship is

assumed to be linear, we can rewrite Eq. (4.9) as the following linear constraint

$$-\sum_{n'=1}^{sN_{\text{seg}}} x_{i,n'}\,\hat{r}_{i,n'} + \frac{V\tau_{\text{seg}}}{q_{\max}}\sum_{s'=1}^{s} q'_{i,s'} \leq 0, \quad \forall\, i, \forall s. \tag{4.15}$$

With this definition of $\mathbf{q}'$, the problem in Eq. (4.14) can be reformulated as the LP

$$\underset{\mathbf{x},\mathbf{q}'}{\text{maximize}} \quad \sum_{\forall i\in\mathcal{M}}\sum_{s=1}^{S} q'_{i,s} \tag{4.16}$$

$$\begin{aligned}
\text{subject to:}\quad & \text{Constraints}: \ Eq.\ (4.12), Eq.\ (4.15),\\
\forall\, i \qquad & -\sum_{n=1}^{N} x_{i,n}\,\hat{r}_{i,n} + \sum_{s=1}^{S} q'_{i,s} = 0,\\
\forall\, i,n \qquad & \sum_{n'=1}^{n} x_{i,n'}\,\hat{r}_{i,n'} - \frac{V\tau_{\text{seg}}}{q_{\max}}\sum_{s'=1}^{\lfloor n/N_{\text{seg}}\rfloor} q'_{i,s'} - \frac{V(n \bmod N_{\text{seg}})}{q_{\max}} q'_{i,\lceil n/N_{\text{seg}}\rceil} \leq L_i,\\
\forall\, i,n \qquad & 0 \leq x_{i,n} \leq 1,\\
\forall\, i,s \qquad & 1 \leq q'_{i,s} \leq q_{\max}.
\end{aligned}$$

The solution to Eq. (4.16) determines the rate allocation matrix $\mathbf{x}$, but $\mathbf{q}'$ has real values. Thus, we present the following algorithm that uses the values of $\mathbf{x}$ to generate an integer solution of $\mathbf{q}'$. We refer to this two stage PAS solution, as PAS-LP-QAlg.

Segment Quality Algorithm (QAlg) The objective of this algorithm is to determine the segment quality plan that maximizes quality while providing smooth playback. The idea is to iterate over the segments in sequence and greedily maximize the current segment quality, while ensuring that the future segments can be streamed, at least, at the lowest quality level. This is achieved as follows: all the segments are first initialized to a value of 1. Then, at the start of each segment, $q'_{i,s}$ is set to $q_{\max}$

Algorithm 2 Segment Quality Algorithm (QAlg)

Require: $x_{i,n}, \hat{r}_{i,n}, V, \tau, N_{\text{seg}}, \tau_{\text{seg}}, q_{\max}, M, N, S$

Initialize all segments to lowest quality level $q'_{i,s} = 1 \; \forall \, i, s$.
for all users i **do**
  for all segments s **do**
    Set current segment quality to highest level $q'_{i,s} = q_{\max}$
    while $q'_{i,s} \geq 0$ **and** Constraint defined in Eq. (4.15) is not met for the current or remaining segments $s, s+1, \ldots, S$ **do**
      Set $q'_{i,s} = q'_{i,s} - 1$ (lower current segment quality)
    end while
  end for
end for
return $\mathbf{q}'$

and a check is made to ensure that the constraint defined in Eq. (4.15) is satisfied for $s, s+1, \ldots S$. If this is not the case, the current segment quality is iteratively decremented, until the constraints are met, or the quality level is zero. The complete procedure is outlined Algorithm 2.

Note that Algorithm 2 is applied to each user independently since the resource allocation has already been determined. A practical property of QAlg is that it ensures users experience the highest quality level as soon as possible, and for the longest possible duration. This is not guaranteed by solving the problem in Eq. (4.14) since, when a mix of low and high quality segments are pre-buffered, they can be ordered arbitrarily while remaining equivalently optimal. Therefore, Algorithm 2 can also be used to 'post-process' the optimal result of $\mathbf{x}$ in the problem of Eq. (4.14), to generate $\mathbf{q}$ solutions that favor 'early' high quality streaming.

In general, an LP with a polynomial number of constraints and variables can be solved in polynomial time. On the average, even the widely-used Simplex algorithm [73] can solve the LP in such time. Therefore, the PAS-LP-QAlg provides a

computational advantage over PAS-MILP, but is restricted to the case of linearly increasing quality bit rates. Next, we introduce an approach to solve the general case of PAS-MILP in polynomial time.

General Heuristic Solution

In this approach, we develop a heuristic to determine the rate allocation $\mathbf{x}$, and then use Algorithm 2 to determine the video segment qualities. This is based on the intuition that developing a good rate allocation scheme will allow users to pre-buffer content and avoid playback interruptions. Thereafter, segment qualities can be explicitly planned for each user based on the allocated bits.

Rate Allocation Heuristic This heuristic is divided into two steps. In the first step, users that do not have enough content pre-buffered to stream the video at the *lowest* quality level are prioritized and their demands fulfilled. This is to ensure that the 'no freezing' constraint in Eq. (4.9) is satisfied. In the second step, the remaining air-time is granted to the user that has both a high channel and a high future video degradation based on the VD metric proposed in Section 4.2.1. Note that the second step follows a procedure similar to Algorithm 1.

In detail, the heuristic performs the following steps:

- *Step 1*: Initialize $x_{i,n}, R_{i,n} = 0 \ \forall i, n$.
- *Step 2*: Determine the set of priority users $\mathcal{P}_{k,n}$ that have insufficient cumulative allocation $R_{i,n}$ to play the video at the lowest quality level. Sort $\mathcal{P}_{k,n}$ in descending order of $\hat{r}_{i,n}$, and allocate the air-time needed in sequence according

to

$$x_{i,n} = \frac{V\tau n/q_{\max} - R_{i,n}}{\hat{r}_{i,n}}, \tag{4.17}$$

where $R_{i,n} = \sum_{n'=1}^{n} x_{i,n'}\,\hat{r}_{i,n'}$. Usually, this stage will only consume part of the BS air-time, since only the deficient users are served, and at the lowest quality level.

- *Step 3*: Compute the future VD each user will experience using Eq. (4.3). The remaining BS air-time is allocated to the user i^* that satisfies

$$i^* = \arg\max_i \hat{r}_{i^*,n}\,\tilde{\mathrm{VD}}_{i^*,n} \quad \forall i \in \mathcal{U}_{k,n}. \tag{4.18}$$

  This follows the same intuition in Algorithm 1.

- *Step 4*: Repeat steps 2 and 3 for all BSs and time slots.

- *Step 5*: Calculate the network VD using Eq. (4.2), and repeat steps 2-5 until there is no further decrease in $\mathrm{VD}_{\mathrm{Net}}$.

The convergence of the iterations also follows a similar approach to that of Algorithm 1.

To complete this general heuristic, segment qualities are determined using Algorithm 2, with the constraint defined in Eq. (4.15) generalized to:

$$-\sum_{n'=1}^{sN_{\mathrm{seg}}} x_{i,n'}\,\hat{r}_{i,n'} + \tau_{\mathrm{seg}} \sum_{s'=1}^{s} f^{Q}_{\mathrm{rate}}(q'_{i,s'}) \leq 0, \tag{4.19}$$

to account for non-linear bit rate increases. The complete procedure is presented in Algorithm 3, which we refer to as PAS-Heuristic-QAlg.

Algorithm 3 PAS-Heuristic-QAlg: General Heuristic Solution

Require: $\hat{r}_{i,n}, \mathcal{U}_{j,n}, V, \tau, N_{\text{seg}}, q_{\max}, M, K, N$

Initialize $x_{i,n}, R_{i,n} = 0 \quad \forall\, i, n = 1, 2, ..N$
repeat {allocation iterations}
  Calculate VD_{Net} using Eq. (4.2) before the allocation iteration.
  for all time slots n **do**
    Reset $x_{i,n} = 0 \quad \forall i$.
    for all base stations k **do**
      Initialize BS air-time to 1, $\text{BS}_{\text{air}} = 1$.
      Find set of priority users $\mathcal{P}_{k,n}$, and sort $\mathcal{P}_{k,n}$ in descending order of $\hat{r}_{i,n}$.
      for all users $i \in \mathcal{P}_{k,n}$ **do**
        if $\text{BS}_{\text{air}} > 0$ **then**
          $x_{i,n} = \frac{V\tau n/q_{\max} - R_{i,n}}{\hat{r}_{i,n}}$.
          $\text{BS}_{\text{air}} = \text{BS}_{\text{air}} - x_{i,n}$
        end if
      end for
      if $\text{BS}_{\text{air}} > 0$ **then**
        for all users $i \in \mathcal{U}_{k,n}$ **do**
          Given the current $x_{i,n}$, calculate the *future* video degradation $\tilde{\text{VD}}_{i,n}$, using Eq. (4.3).
        end for
        Set $x_{i^*,n} = \text{BS}_{\text{air}}$ to user i^* with the highest $\hat{r}_{i,n}\,\tilde{\text{VD}}_{i,n}, \quad \forall i \in \mathcal{U}_{k,n}$
      end if
    end for
  end for
  Calculate VD_{Net} after allocation.
until {no more decrease in VD_{Net}}
Determine the segment qualities using Algorithm 2.
return $\mathbf{x}, \mathbf{q}$

Computational Complexity We first evaluate the time complexity of the rate allocation heuristic. Determining the set of priority users and sorting them in Step 2 takes $O(M+M\log M)$ time, and overall, Step 2 has a $O(M\log M)$ runtime. In Step 3, the core computational function is computing Eq. (4.3) which has a time complexity of $O(N^2)$, and therefore step 3 takes $O(MN^2)$ time. This gives a complexity per time slot of $O(MN^2 + M\log M)$, and for N slots, we have $O(MN(N^2 + \log M))$. As

previously discussed, typically only 4 to 6 iterations are required for convergence.

In QAlg, the core step is to evaluate the constraint in Eq. (4.15), which has as a time complexity $O(N+S)$ for a single user. This step is repeated at most $q_{\max}$ times when the constraint is violated, and repeated for each segment and each user. The resulting complexity order is $O(q_{\max}MS(N+S))$. In the worst case $S=N$, and $q_{\max}$ is typically less than 6, which gives a worst case runtime of $O(MN^2)$. Therefore, the rate allocation runtime dominates the complexity of PAS-Heuristic-QAlg, leading to an overall runtime of $O(MN(N^2+\log M)) \approx O(MN^3)$.

4.3.5 Performance Evaluation

In this section, we use simulations to compare the performance of the PAS solutions and investigate the effects of prediction errors.

Simulation Setup

We first consider the highway scenario with three BSs at an inter-BS distance of 1 km. For system evaluation in a more general, albeit less practical setting, we also consider the 19 cell network in Figure 3.1(b). Users have RWP mobility with a speed of 15 m/s, zero pause time between the waypoints, and no wrap-around. BS transmit power is 40 W, and bandwidth is 5 MHz. The slot duration $\tau = 1$ s, $T = 200$ s, and $L_i = VT$ (i.e., no buffering limit to demonstrate the bounds of the gains). We consider a video format with linear bit rate increases $\{1.2, 2.4, 3.6\}$ Mbit/s, and $\tau_{\text{seg}} = 10$ s. Gurobi 5.1 [52] was used to solve the optimization problems, and Matlab was used as a simulation environment.

We compare the performance of the PAS schemes against baseline approaches

that do not exploit rate predictions. The baseline allocators have two stages: rate allocation, then quality adaptation. Two rate allocation schemes are considered, MR and ES, as described in previously in Section 3.5.1. Segment quality is then adapted based on the allocated rate at the start of the current segment, and the highest supportable level selected. We refer to these approaches as MR-AdaptQ, and ES-AdaptQ.

The network-wide video quality metrics are defined as:

- Q_{Net}: the total quality of all delivered segments, divided by the number of requested segments.
- F_{Net}: the average percentage of playback time where the video is stalled, over all users.

Numerical Results and Discussion

Figure 4.7 and Figure 4.8 present the results for both network scenarios, where we can see that the PAS schemes provide significant gains over the non-predictive approaches in both video quality and stream freezing. These gains are due to 1) exploiting the knowledge of user rate variations to generate multi-user rate allocation plans that increase network video quality, and 2) the long-term segment quality planning that prevents freezing from occurring. We can make the following observations from the results:

1. Over 15% quality improvement is attained at high-load, while completely eliminating 15 s of freezing for every 100 s of playback time.
2. The PAS-LP-QAlg scheme is near-optimal with less than a 1% performance gap

from the benchmark MILP results.

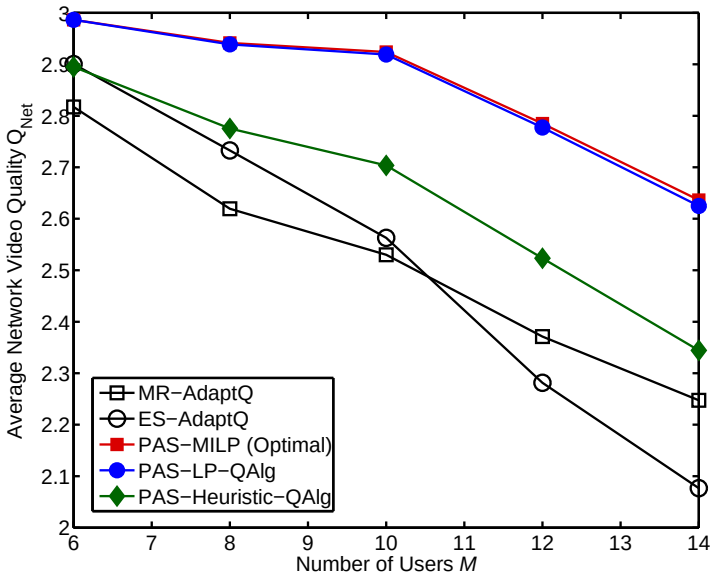
3. The PAS-Heuristic-QAlg successfully eliminates freezes without compromising network quality compared to the baseline schemes. However, the PAS optimization-based approaches achieve higher quality. This indicates that there is room to further improve the rate allocation heuristic developed in Section 4.3.4.

4. The gains in the RWP scenario are higher due to the significant capacity fluctuations both, *between users*, and along *individual* user paths. The baseline MR-AdaptQ performs very poorly with RWP mobility, due to its greediness that results in starvation for users that remain in poor channel conditions.

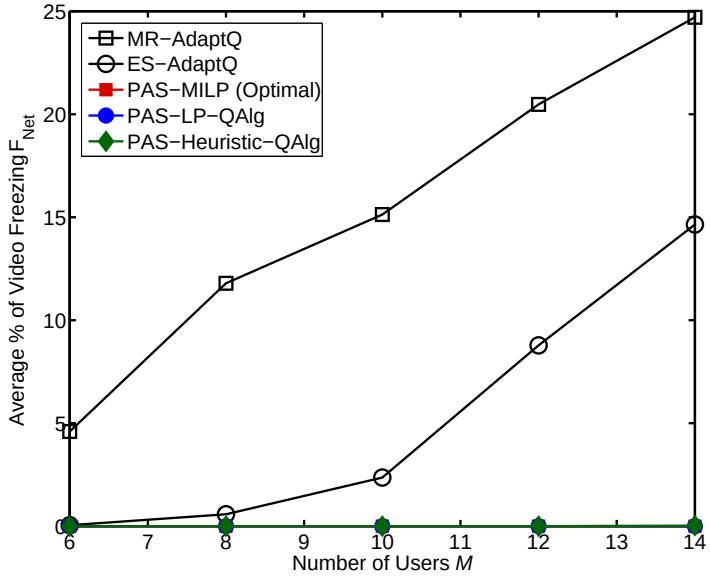
Effect of Prediction Errors

To evaluate the effect of prediction errors, we introduce a Gaussian random variable with a mean of zero and a variance σ^2 to the predicted user SNR, and denote the resulting user rate matrix as $\tilde{\mathbf{r}}$. Therefore, while the PAS schemes use $\hat{\mathbf{r}}$ for allocation, the actual rates received are determined by $\mathbf{x} \odot \tilde{\mathbf{r}}$. Figure 4.9 illustrates the impact of such errors, where we can see that the PAS schemes are quite robust, indicating that even *trends* in the future user rates can provide considerable QoS gains.

It is important to note that the PAS-Heuristic-QAlg exhibits the highest robustness to prediction errors for video freezing. This is mainly because the PAS optimization approaches make discrete allocation bursts as observed in Figure 4.6. While being optimal, these bursts can be spaced out in time, and therefore when the predicted rate is less, the user has to wait until the next allocation to resume playback. In contrast, the PAS rate allocation heuristic performs a more steady flow of allocation across multiple slots, for cases when users do not have any buffered segments.

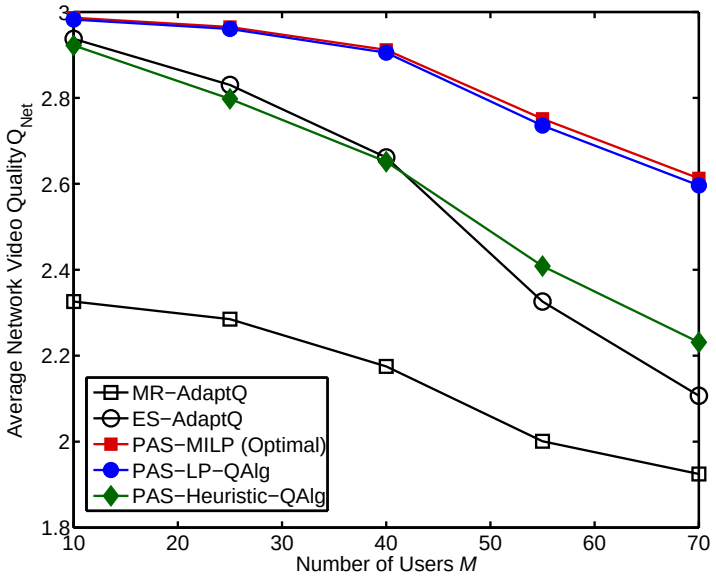


(a) Network Quality.

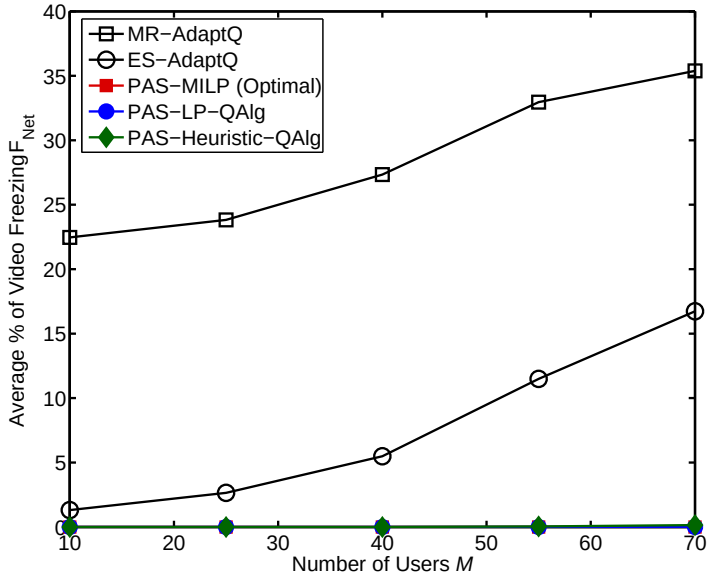


(b) Video Freezing.

Figure 4.7: PAS video quality and freezing for the highway scenario.

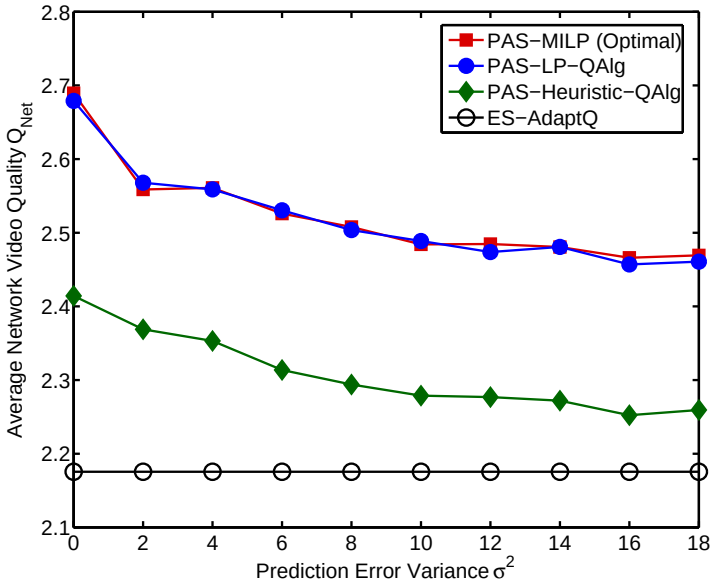


(a) Network Quality.

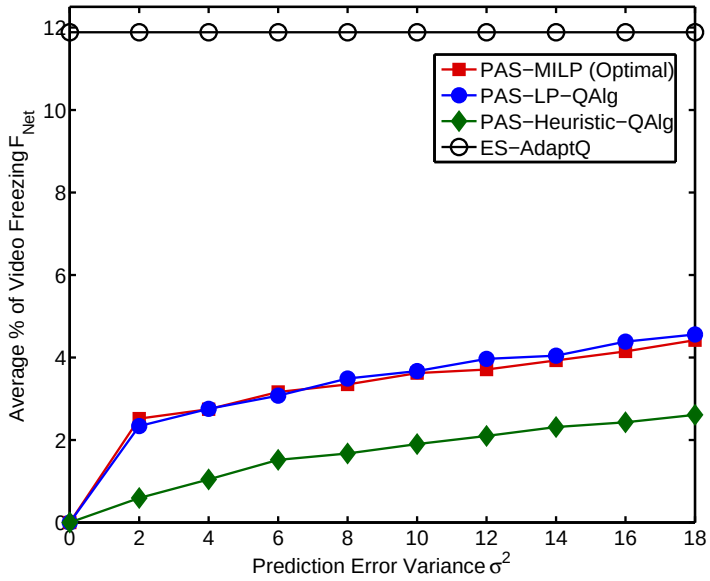


(b) Video Freezing.

Figure 4.8: PAS video quality and freezing for the RWP scenario.



(a) Network Quality.



(b) Video Freezing.

Figure 4.9: Effect of prediction errors on the PAS schemes in the highway scenario, $M = 12$.

4.3.6 Testbed Measurements

To gain additional insight on the performance of PAS under real hardware constraints, and with real videos, we present results of a simple scenario implemented on a testbed at Bell Labs-Stuttgart, Germany. Before discussing video quality results, we describe the experimental setup.

Testbed Setup

The aim of the testbed design is to study the performance of mobile users in a tractable and reproducible manner. To this end, the handhelds never physically move but their mobility is emulated. This is done by limiting the wireless data rate according to their configured distance d using the path-loss and link rate models from Section 3.3. After resource allocation, the testbed allocates the resulting rate to the handset by using traffic shaping. Similarly, handovers among the BSs are triggered depending on the emulated position of the user. This design as "testbed on a table" allows reproducible experiments on real hardware and software, under real time constraints. It is, thus, very helpful to study the performance of optimization algorithms in a complex, yet controlled environment.

Figure 4.10 illustrates the testbed setup which consists of 1 control computer, 4 BSs, and 4 handhelds. Handhelds and BSs are connected via IEEE 802.11g, a wireless LAN [74], while the BSs are connected to the control computer via Ethernet. The handhelds are 2 tablets and 2 Smartphones, all running Android 4.1 as the operating system [75] and VLC [76] as video player. The BSs are typical Linux computers operating as IEEE 802.11g Access Points using built-in traffic shaping functions to allocate wireless rates. The control computer manages testbed operation, computes

Figure 4.10: Testbed setup with 4 IEEE 802.11g Access Points (APs), 1 Control Computer, and 4 Android Handhelds

optimal solutions using the solvers in [77], and executes algorithms implemented in the Python programming language.

Before a BS can deliver the videos to the handhelds over the air, it receives the video stream via Ethernet from the control PC. Here, an unmodified Apache web-server delivers a video stream using the HLS protocol [59] with a segment length of $\tau_{\text{seg}} = 10\,\text{s}$. All users stream the same high-motion video [78], which is encoded using standard MPEG-4 codecs at 3 different quality levels. This leads to the video bit rates $f^{Q}_{\text{rate}} = \{1416, 2952, 3608\}$ kbit/s, which are approximately linear.

Experimental Results

We configure a simple scenario of two users moving along a horizontal line of 4 km across two BSs. The first BS is placed at a coordinate of $(500, 300)\,\text{m}$ with respect to the user starting point, and the second is at $(d, 300)\,\text{m}$, where d is varied to assess

performance at different BS proximities. User speed is 20 m/s during which 200 seconds of video are streamed. We run PAS-MILP on the testbed and compare to the baseline ES-AdaptQ scheme.

Figure 4.11 shows how the non-predictive baseline cannot smoothly stream the video when an increasing d leads to a significant coverage gap between the two BSs. In contrast, PAS pre-buffers sufficient video segments at a low quality while it is still covered by the first BS. Minor streaming glitches were also observed for PAS, as handover delays and TCP's congestion control occasionally decrease the delivered throughput leading to decoding errors at the handheld's video players. Such practical consequences may be accounted for by introducing conservative measures in the PAS approaches, in addition to feedback during run-time. Nevertheless, our observed results verify that significant QoS gains can be achieved by exploiting rate predictions in a real system setting.

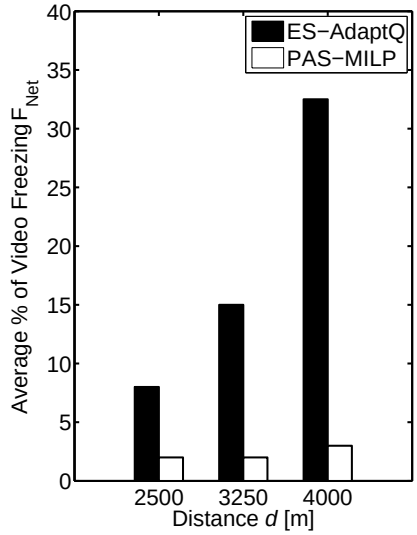


Figure 4.11: Experimental results from real-time testbed measurements.

4.4 Summary

In this chapter, we presented predictive video delivery approaches that can significantly enhance the mobile video streaming experience. Our first focus was to formulate a PRA scheme that exploits rate predictions to minimize video disruptions. To do so, long-term video delivery planning that strategically *pre-buffers* content was introduced. This is opposed to the conventional RA approach that does not look ahead at network conditions, and fulfills instantaneous requirements only. To provide an evaluation benchmark, we first formulated the problem as a multi-objective LP that captures a trade-off between minimizing total video degradation and achieving fairness in video quality among users. Then, we presented an algorithm that closely follows the benchmark solution in polynomial time, and with a significantly lower memory requirement. The numerical evaluations conducted indicate that pre-buffering content improves the video streaming quality for various user mobility patterns and network loads.

We then investigated how knowledge of future wireless data rates can improve adaptive video streaming. We formulated the optimal joint allocation of rate and video quality as an MILP. Then, we presented an LP based solution that achieves less than 1% performance loss for the case of linear bit rate increases, and a general polynomial-time heuristic. We carried out extensive performance evaluation for all three approaches and significant QoS gains were consistently observed when rate prediction is used. Our results clearly demonstrate that strategic buffering and segment quality planning can maintain smooth video streaming for mobile users. These promising results were confirmed by experiments on a IEEE 802.11g-based testbed

with handheld devices. Solving the proposed PAS in real time shows that the developed solutions have practical value. Therefore, we believe that PAS is a promising function for future cellular core networks and content delivery platforms.

Chapter 5

Energy-Efficient Predictive RANs: Application to Stored Video Transmission

5.1 Introduction[1]

The dense deployment of BSs, and the growing mobile traffic, has made mobile network energy efficiency a very active research topic in the last decade. In addition, the variety of tablets, pads, and ultrabooks, is adding more pressure to the energy drain at both the network and end user. The result is increased operational expenditures (OpEx) for network operators and a negative environmental impact [82]. Consequently, research and standardization efforts are focusing on devising green mechanisms to save energy across the network, as well as on user devices. In particular, among the wireless network elements, BSs account for more than 50% of the network energy consumption [83]. Therefore, devising *efficient* BS transmission mechanisms will have a significant impact on overall energy savings, operational costs, and CO_2 emissions. Furthermore, energy-efficient rate allocations can improve the spectral efficiency and provide more resources for other services, and thus, impact

[1]Parts of this chapter were previously published in [79],[80],[81].

overall network performance.

As networks are typically designed to satisfy peak user demands, radio access energy can be reduced in a number of ways at times of lower demand. This includes putting BSs to intermittent short sleep modes during low load, as well as adaptively powering down select BSs completely where demand is low for prolonged time periods. Evidently, if the network is aware of 1) the user temporal and spatial traffic demands, and 2) the spatial variability in network capacity, or supported data rates (e.g., a radio map), it can make better adaptations that reduce energy consumption. To this end, we now investigate the energy saving potential of predictions in RANs, which we refer to as Predictive Green Wireless Access (PGWA). Although we focus on stored video transmission, similar approaches may be applied to other delay tolerant traffic as well.

Being aware of a user's upcoming rate allows the network to plan spectrally efficient rate allocations, without violating user streaming demands. For instance, if it known that a user is approaching the BS, transmission can be delayed, provided sufficient video segments have previously been buffered. This allows the BS to save energy by 'sleeping' as the user approaches, and then 'waking up' for a short period, during which a high data transmission is possible. The main idea is to grant users more air-time access at their highest achievable data rates, and less access when they are at lower achievable rates. This allows the BS to transmit more data in less time, and consume lower transmit energy.

In more detail, we address the following problems in this chapter:

1. We first consider the problem of minimizing BS power consumption for constant

quality videos, during *low load*. The problem is first formulated for the multi-user, *multi-cell* case as an LP, and then we present a distributed heuristic that reduces BS power significantly without requiring long-term predictions or signaling overhead.

2. We then address the problem of medium load, constant quality videos, where video degradations occur. We extend the previous formulation to capture the trade-off between BS power consumption and overall video degradation. Centralized and distributed multi-cell algorithms are also proposed to solve the aforementioned problem.

3. We then consider the general case of adaptive video streaming and the energy-quality trade-off. In addition to saving power by minimizing utilization, we also incorporate deep sleep modes where BSs can be switched off in the problem definition. This is formulated as a detailed multi-user, multi-cell optimization framework for energy-efficient *adaptive streaming*. An approximate multi-stage heuristic algorithm is then developed to solve the problem in polynomial time.

As previously discussed in Section 4.1.3, a few parallel research efforts have explored leveraging rate predictions to optimize RA for wireless video streaming [64], [65]. Our work differs from, and extends, these works in several aspects: 1) we address the multi-cell problem and provide centralized and distributed user-assisted approaches to solve the problems, 2) we also consider the delivery of adaptive video streams and, thus, address the joint quality-energy trade-off, 3) in addition to saving power by minimizing utilization, we consider deep sleep modes, and 4) we formulate a comprehensive optimization framework for energy efficient AVS, and develop an algorithm that is robust to prediction errors.

5.1.1 Chapter Outline

The rest of this chapter is organized as follows. In Section 5.2, we review some of the conventional methods of energy saving in RANs. This provides more context on how PGWA can be applied to specific energy saving RAN functionalities. We then present an overview on the proposed PGWA in Section 5.3. Following this, we address the three energy saving problems discussed in the introduction, in Section 5.4, Section 5.5, and Section 5.6. Within each section, the optimal problem is first formulated, and then approximate solutions are developed. Simulation results are then presented to investigate the derived energy saving gains and compare the solutions. Thereafter, in Section 5.7 we discuss some implementation considerations of PGWA, with emphasis on the key functionalities needed and the required information exchange and signaling. Finally, we conclude the chapter with a summary in Section 5.8.

5.2 Review on Green Wireless Access Techniques

Radio access power consumption is distributed among the BS Power Amplifier (PA), signal processing, air conditioning, and the power supply [83]. Power consumption can be reduced in a number of ways including improving individual hardware component efficiency and developing more efficient baseband signal processing schemes. However, an alternative approach is to power down the hardware components themselves during low traffic load. This can be referred to as traffic-aware energy efficiency, as it requires knowledge of traffic demand such that user service is not compromised [84]. The importance of such techniques is that networks are generally designed to support peak demands, which last for only a small fraction of the day. During the rest of the operation time, when traffic demand is lower, BS power consumption can be reduced

in a number of ways [83].

5.2.1 Time Domain Approaches

BSs transmit data to users by allocating units of bandwidth over time slots. During low to medium traffic, the number of required units decreases. Unused resources can be aggregated in a way that creates complete time slots without any data transmission, enabling transceiver hardware to be deactivated and save energy during micro-sleeps of up to 214 micro seconds [85]. Additionally, more advanced sleep modes that reduce the transmission frequency of reference and control signals have been proposed, such as the extended-cell Discontinuous Transmission (DTX) in LTE. The benefit of such sleep modes is dependent on how fast deactivation and reactivation can be supported by the PA, power supply, and signal processing [83]. Similarly, the UE can also enter a Discontinuous Reception (DRX) mode to save power by monitoring the Down Link (DL) control channel less frequently, and going to sleep when there is no packet scheduled for the UE. A recent comparative study on 3GPP UE sleep mechanisms is available in reference [86].

5.2.2 Frequency Domain Approaches

As BS transmit power is distributed across the bandwidth, scheduling only a limited number of subcarriers at a given time allows the BS to lower the PA supply voltage, resulting in energy savings. However, the PA is not completely shut down, and therefore energy savings are limited. To address this, BSs can be implemented such that groups of carriers are served by individual PAs. In this case, the unused PAs can be turned off completely when the corresponding aggregate carriers are not scheduled

for transmission. This is known as the carrier aggregation approach in LTE [83]. A detailed comparison between the time and frequency domain power savings has been conducted by Desset *et al.* [87].

5.2.3 Network Reconfiguration

Although the aforementioned approaches provide energy savings, BSs still consume a considerable amount of fixed load-independent power to remain functional. Therefore, significant energy savings can be obtained if a select subset of BSs is dynamically switched off *completely*, when traffic is low for prolonged periods [88],[4], [89]. However, services will be affected during this setup, and a transitional period is needed before inactive BSs can re-operate. Nevertheless, this operation is particularly suited for heterogeneous network deployments, where a macro cell overlays smaller cells that serve users during high traffic demand. When load is low, it is possible to selectively switch off some of the smaller cells, while radio coverage is guaranteed by the macro-cell. For example, Ismail and Zhuang [90] propose an optimal on-off switching framework that adapts to fluctuations in traffic to maximize energy savings under user service constraints.

5.2.4 Potential of Cooperative Mechanisms

Network cooperation is envisioned to play a significant role in improving network efficiency and long-term user experience [91],[38]. Furthermore, self-organizing networks (SON) introduced in the 3GPP TS 32.521 [12] enable heterogeneous mobile networks to self-optimize and reconfigure, thereby improving user experience and reducing network operational and management costs. Studies have shown that such

cooperation can further improve energy efficiency by increasing the potential savings of traffic-aware radio access. For example, inter-BS cooperation is proposed to configure network layout by powering down select BSs, partially or fully, depending on network traffic [90],[92],[84], and to dynamically adjust cell sizes according to traffic load [93].

In summary, the goal of the aforementioned energy saving strategies is to make transmission/network adaptations that reduce BS power consumption without compromising user QoS. In this chapter, we investigate how *predictions* of user locations coupled with network performance maps, can facilitate higher energy savings.

5.3 Predictive Green Wireless Access (PGWA) Overview

The general idea of PGWA is to leverage knowledge of the *future* temporal and spatial user achievable rates, and traffic demands, to enable networks to make better *long-term* adaptations that reduce energy consumption. To do so, the resource sharing model is first defined, e.g., a time slotted system where users share air-time in arbitrary fractions in each slot. Then, depending on whether time or frequency domain resource sharing is implemented (or both), the load-power consumption model is derived. To incorporate BS On/Off switching, the value of deep-sleep power consumption is required, in addition to constraints on the minimum time duration needed before a BS can be turned back on. Thereafter, based on the user application requirements and corresponding rate predictions, the PGWA determines 1) how BS resources will be allocated among users over a predefined prediction window T, and 2) the BS On/Off configuration for each BS during T.

5.3.1 BS Power Consumption Model

We consider a BS downlink power consumption model based on the common linear load-power dependency [87],[94]. According to this model, power is proportional to the BS load, with a fixed power required at minimum load. For BS k at slot n, this can be represented as follows:

$$p_{k,n} = \begin{cases} P_0 + (P_m - P_0)\,\mathrm{BS}_{k,n}^{\mathrm{load}}, & 0 < \mathrm{BS}_{k,n}^{\mathrm{load}} \leq 1, \\ P_{\mathrm{sleep}}, & \mathrm{BS}_{k,n}^{\mathrm{load}} = 0, \end{cases} \tag{5.1}$$

where P_m and P_0 are the power consumption at the maximum and minimum non-zero load, and BS load is computed as $\mathrm{BS}_{k,n}^{\mathrm{load}} = \sum_{i \in \mathcal{U}_{k,n}} x_{i,n}$. When there is no load, the BS can enter a sleep mode which consumes $P_{\mathrm{sleep}}\,[W]$. Advanced BS designs and hardware components allow P_{sleep} to be a fraction of P_0, or even zero, i.e., a complete BS switch-off [87]. Therefore, we assume $P_{\mathrm{sleep}} = 0$. BSs entering this deep sleep mode are required to remain off for at least n_{off} time slots to allow sufficient time before a wake-up is possible. We denote the BS power per slot matrix by $\mathbf{p} = (p_{k,n} \in [0, P_m] : k \in \mathcal{K}, n \in \mathcal{N})$, and the BS on/off binary decision matrix by $\mathbf{b} = (b_{k,n} \in \{0, 1\} : k \in \mathcal{K}, n \in \mathcal{N})$.

5.3.2 PGWA: Approach for Stored Video Transmission

As previously discussed, stored videos can be strategically delivered ahead of time and cached at the user equipment, after which transmission can be momentarily suspended while the user consumes the buffer. If we consider a user requesting a stored video at slot $n = 1$, with a streaming rate of V [bit/s], then the minimum cumulative video content for smooth streaming is $D_{i,n} = V\tau n$, which is represented with a dashed line in Figure 5.1. The cumulative allocation made to a user i by slot n is denoted by $R_{i,n} = \sum_{n'=1}^{n} x_{i,n'}\, \hat{r_{i,n'}}$. To experience smooth streaming, $R_{i,n} \geq D_{i,n}$, $\forall\, n$ for user i.

Figure 5.1 illustrates how BS transmission time can be minimized by leveraging user rate predictions. We can see in Figure 5.1(a) that a traditional RA scheme (e.g., that distributes BS air-time equally among users), will continue to serve the user even when in poor channel conditions. Notice how the user cumulative rate allocation denoted by $R_{i,n}$ is significantly higher than the required allocation $D_{i,n}$, implying that the video content is being pre-buffered, even when channel rates are poor. On the other hand, as shown in Figure 5.1(b), a *predictive* scheme that is aware of the user's future rate, will wait to make bulk transmissions at times of high channel conditions, while making the minimal transmissions that ensure $R_{i,n} \geq D_{i,n}$ at other times. This achieves lower airtime usage, resulting in lower power consumption or more resources for other services. Figure 5.1(c), demonstrates that it is also possible to consider the case where BS_2 is switched off, by modifying the predicted rate to account for this. In this case, it is possible to completely pre-buffer the video content before the user leaves BS_1. Although the total air-time is higher than that in Figure 5.1b, this can achieve more energy savings as BS_2 is switched off completely.

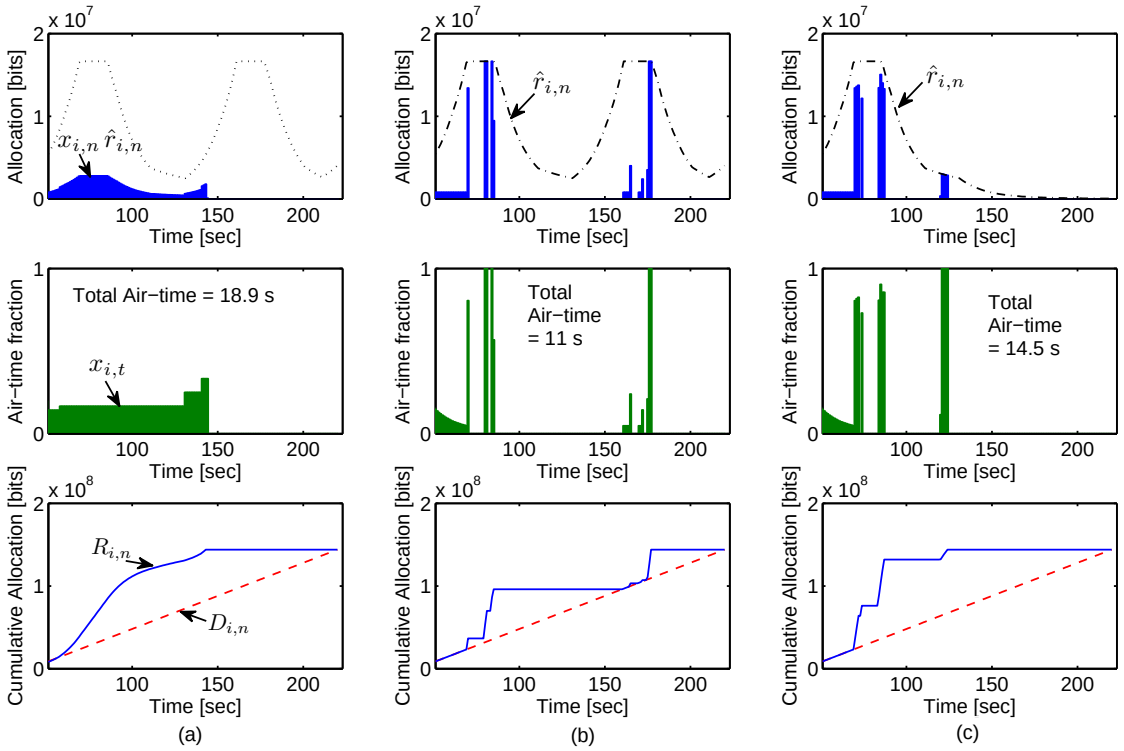


Figure 5.1: Sample user allocation with time illustrating power minimization. (a) In traditional allocation, airtime is divided equally among users. (b) In PGWA, allocations are low if the user rate is increasing, and high when the rate is high, to avoid inefficient future allocations. (c) similar to (b) but using only one BS. Although air-time in (c) is larger than (b), more energy is saved as one BS is switched off.

5.4 Minimizing BS Power Consumption for Video Transmission

In this section, we formulate the PGWA problem for stored video transmission that exploits user rate predictions over multiple cells. The objective is to minimize network-wide BS power consumption, without causing any streaming discontinuities. This is possible when network load is relatively low so a resource allocation solution exists where no buffer underruns occur.

5.4.1 Optimal Problem Formulation

Consider a network with a BS set $\mathcal{K}$ and an active user set $\mathcal{M}$. An arbitrary BS is denoted by $k \in \mathcal{K}$ and a user by $i \in \mathcal{M}$. Users request stored video content that is transported using an HTTP-based progressive download mechanism. We assume that the wireless link is the bottleneck, and therefore the requested video content is always available at the BS for transmission.

The problem of minimizing network wide BS power without violating streaming requirements is equivalent to minimizing the BS air-time due to the linear load dependent power model of Eq. (5.1). For a multi-user, *multi-cell* scenario, and a prediction window of N slots, this can be formulated as the LP:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{n=1}^{N} \sum_{i=1}^{M} x_{i,n} \tag{5.2}$$

$$\begin{aligned} \text{subject to:} \quad & \text{C1:} \quad \sum_{i \in \mathcal{U}_{k,n}} x_{i,n} \leq 1, && \forall\, k \in \mathcal{K}, n \in \mathcal{N}, \\ & \text{C2:} \quad D_{i,n} - R_{i,n} \leq 0, && \forall\, i \in \mathcal{M}, n \in \mathcal{N}, \\ & \text{C3:} \quad 0 \leq x_{i,n} \leq 1 && \forall\, i \in \mathcal{M}, n \in \mathcal{N}. \end{aligned}$$

Constraint C1 expresses the resource limitation at each base station. It ensures that the sum of the air-time of all users associated with a BS k is equal to 1 at every time slot. C2 ensures that the cumulative video content requirement is not violated at each time slot. Finally, C3 provides the bounds for the resource allocation factor. Note that the outer summation over time slots in Eq. (5.2) is to minimize the sum of air-time consumed during the window of N slots.

Although the formulation in Eq. (5.2) is an LP, generating the constraint matrix

can have a very large memory requirement and significant computational power due to the long-term planning horizon and multiple cells involved. Further, as it is centralized, a signaling overhead is incurred. We therefore present the following lightweight, distributed heuristic that achieves close to optimal performance at low load.

5.4.2 Distributed Heuristic Solution

As previously illustrated in Figure 5.1, BS air-time is minimized when users are granted more air-time access at their highest achievable data rates and less access when they are at lower achievable rates. The following distributed heuristic is divided into two stages. In the first stage, minimum air-time is granted to each user, to ensure smooth playback (i.e., $D_{i,n} = R_{i,n}$). If $R_{i,n-1} > D_{i,n}$ then no air-time is granted to this user in this stage. In the second stage, excess BS air-time is allocated to users whose achievable data rate is going to decrease (i.e., they are moving away from the BS). This is to opportunistically pre-buffer as much video content as possible before the user's achievable rate decreases any further. The heuristic, which we refer to as PGWA-LowLoad performs the following steps at each BS in every time slot:

- *Step 1:* Sort the users in descending order of their achievable rates.
- *Step 2:* Grant each user the required air-time that satisfies constraint C2 in Eq. (5.2). This is computed as: $x_{i,n} = \max(0, D_{i,n} - R_{i,n-1})/r_{i,n}$, where $r_{i,n}$ is the current achievable rate.
- *Step 3:* Determine the subset of users, whose rate is going to decrease, and that have remaining video content to be delivered. If the set is empty, end the step; otherwise sort the set in descending order of the achievable user rates.

- *Step 4:* Use the remaining BS air-time to pre-buffer as much future video content as possible, to the first user in the sorted set of step 3. Remove this user from the set of users with decreasing rates.

- *Step 5:* Repeat step 4 if there is additional BS air-time.

Remarks

1. In step 2, generally speaking, it is not necessary to allocate BS air-time sequentially to the ordered users, as there should be sufficient air-time to serve all the users with their minimal video content (as we are considering saving air-time during low network load).

2. Also in step 2, during handover when a user changes BS association, some signaling is required to enable the target BS to compute $x_{i,n} = \max(0, D_{i,n} - R_{i,n-1})/r_{i,n}$. This is because $R_{i,n-1}$ is unknown to the target BS, and should be signaled either from the UE or from the source BS. Additionally, note that $x_{i,n}$ can be computed as $\max(0, V - \text{Buff}_{i,n-1})/r_{i,n}$ where $\text{Buff}_{i,n-1}$ is the amount of video content buffered at the UE. If this is equal to zero, then $x_{i,n} = V/r_{i,n}$, and if it is larger than V, then $x_{i,n} = 0$. This implies that the target BS only requires the video buffer status of the incoming user, which is signaled with minimal overhead without any centralized operation.

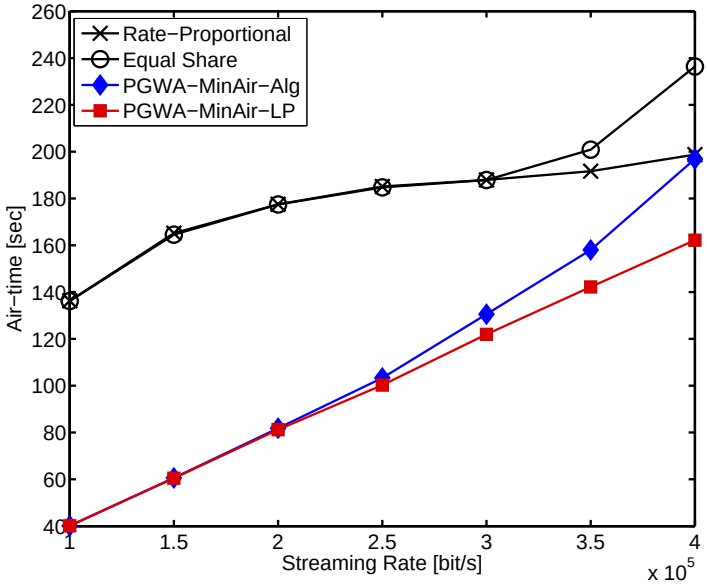
3. No prediction of the entire user rate vector is required. The heuristic only needs to know whether a user rate is increasing or decreasing.

4. The heuristic has a very low computational complexity of $O(M \log M + M)$.

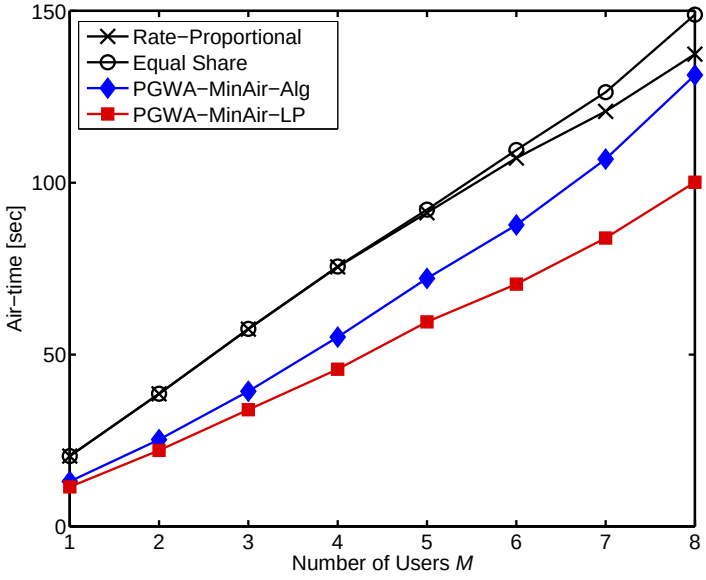
5.4.3 Simulation Results

The potential air-time minimization gains are evaluated for the highway scenario. Transmit power is 40 W, over a 5 MHz bandwidth, and T is 200 seconds with $\tau = 1$ s. To provide a base-line reference we consider the ES and RP allocation schemes that do not incorporate any rate predictions. In ES, airtime is shared equally among the users at each time slot. If there are $N_{k,n}$ users associated with BS k at time n, then $x_{i,n} = 1/N_{k,n}$, $\forall i \in \mathcal{U}_{k,n}$, and the rate allocated to each is $\hat{r}_{i,n}/N_{k,n}$. The RP allocator is designed to be more spectrally efficient but not completely fair to users. Here, the airtime assigned to each user i at slot n is in proportion to the achievable data-rate $\hat{r}_{i,n}$ of that user. Therefore, $x_{i,n} = \hat{r}_{i,n} / \sum_{i \in \mathcal{U}_{k,n}} \hat{r}_{i,n}$, and the rate received is $x_{i,n}\, \hat{r}_{i,n}$.

Figure 5.2(a) considers the scenario of a large number of users ($M = 40$), requesting low rate video streams. We can see a very significant reduction of air-time exceeding 70 percent at low loads. The exact energy savings will depend on the parameters of the BS power model [87]. As the streaming rate increases, the gains decrease since network utilization increases and the degrees of freedom in allocation decrease. Interestingly, the distributed heuristic achieves near optimal performance, with only minimal user-BS cooperation and rate knowledge. In Figure 5.2(b), we investigate the converse case, where a few users are requesting high rate video streams ($V = 1.2$ Mbit/s). In this scenario, although the total network traffic is low, it is not possible to pre-buffer significant portions of video content in advance when the achievable rate is high, since the requested rate is also high. Also, when the achievable rate is low, air-time cannot be reduced significantly due to the high user streaming rate. This implies that air-time reduction shall be less in this case, as confirmed in the results. The performance of the distributed heuristic also deviates from the optimal



(a)



(b)

Figure 5.2: Average air-time with (a) varying streaming rates for 40 users; (b) varying number of users for 1.2 Mbit/s streaming.

allocation as the network load increases.

Nevertheless, it achieves considerable gains at low load, with minimal computation and signaling overhead. However, more advanced heuristics that exploit the rate prediction vector completely would improve the performance.

5.5 Joint Power-Video Degradation Optimization for Video Transmission

At medium to high load, it will not be possible to satisfy constraint C2 in Eq. (5.2) for all users and all time slots. This results in video degradations when $R_{i,n} < D_{i,n}$. In this section, we use the VD definition of Section 4.2.1 to quantify the amount of unfulfilled video demand, and formulate the allocation problem that jointly minimizes BS power and network-wide VD.

5.5.1 Optimal Problem Formulation

The objective now is to exploit rate predictions to determine the optimal pre-buffering allocations to users, that achieve a tradeoff between minimizing the sum user VD and the consumed BS air-time. This is formulated as the following multi-objective optimization problem:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{n=1}^{N}\sum_{i=1}^{M}\left(\frac{w_{\text{VD}}}{D}\text{VD}_{i,n} + \frac{w_{\text{Air}}}{MN}\,x_{i,n}\right) \tag{5.3}$$

$$\text{subject to:} \quad \text{C1, C3,}$$

where w_{VD} and w_{Air} are weights $\in [0,1]$ for VD and air-time minimization respectively. The VD and air-time objective components are normalized by the parameters $D = \sum_{n=1}^{N}\sum_{i=1}^{M} D_{i,n}$, and MN, respectively. When $w_{\text{VD}} = 1$ and $w_{\text{Air}} = 0$, Eq. (5.3)

minimizes user degradations irrespective of the consumed BS air-time. As w_{Air} increases, the problem will trade off the reduction in VD against the additional air-time required. This formulation is particularly useful for cases when users experience prolonged conditions of poor coverage, during which the air-time spent will not result in considerable quality improvements. Note that Eq. (5.3) is non-linear due to the $(\cdot)^+$ operator in the computation of VD. However, since it is piece-wise linear and convex, and the constraints are linear, the problem in Eq. (5.3) can be reformulated as the following LP, which we refer to as PGWA-MinAirVD-LP:

$$\underset{\mathbf{x},\mathbf{Y}}{\text{minimize}} \quad \sum_{n=1}^{N}\sum_{i=1}^{M}\left(\frac{w_{\text{VD}}}{D}Y_{i,n}+\frac{w_{\text{Air}}}{MN}\,x_{i,n}\right) \tag{5.4}$$

$$\begin{aligned}
\text{subject to:}\quad & \text{C1, C3,} \\
& \text{C4: } D_{i,n}-R_{i,n}-Y_{i,n}\leq 0, \quad \forall\, i\in\mathcal{M}, n\in\mathcal{N} \\
& \text{C5: } Y_{i,n}\geq 0, \qquad\qquad\quad \forall\, i\in\mathcal{M}, n\in\mathcal{N}.
\end{aligned}$$

Here we introduce $Y_{i,n}$ as additional optimization variables which we restrict to have positive values in C5. The value of $Y_{i,n}$ therefore captures the degradation only (i.e., when $D_{i,n} > R_{i,n}$), and remains unaffected if content is pre-buffered.

Similar to Eq. (5.2), the multi-objective problem in Eq. (5.4), although linear, has a large number of constraints and optimization variables, which increase dramatically and as N increases. This can be solved with large-scale LP solvers such as Gurobi [52], but requires significant memory and considerable time. Therefore, Eq. (5.4) can serve as an offline performance benchmark, whereas for real-time implementation, we present two heuristic algorithms in the following section.

5.5.2 Centralized and Distributed Algorithms

We first present an iterative algorithm that requires a central BS to make the allocation decisions for all the cooperating BSs. Then, we show how the algorithm can be extended to operate in a distributed fashion.

Centralized PGWA-MinAirVD Algorithm

The objective of this algorithm is to jointly minimize VD_{Net} and BS air-time as in the Pareto-optimal formulation of Eq. (5.3). It consists of the following steps:

- *Step 1*: Initialize $x_{i,n} = 0$ for all the users and time slots.
- *Step 2*: Compute the *future* $\tilde{\text{VD}}_i$ each user will experience at slot n. This is determined based on the current cumulative allocation at slot n, and a tentative air-time allocation for the upcoming slots $n+1, n+2, \cdots, N$, i.e., $\tilde{\text{VD}}_{i,n} = \sum_{n'=n}^{N}[V\tau n' - \sum_{n''=1}^{n'} x_{i,n''}\,\hat{r}_{i,n''}]^{+}$, where n' and n'' are dummy variables.
- *Step 3*: Each BS performs a *greedy* allocation to minimize VD_{Net}. It finds the user that when allocated the full air-time at slot n reduces VD_{Net} the most. To do so, the BSs first compute the *sum* of future VD of all users $\in \mathcal{U}_{k,n}$, that results from allocating the full air-time to user i and nothing to users $i^{'} \in \mathcal{U}_{k,n}\backslash\{i\}$ (the other users in the BS):

$$\tilde{\text{VD}}^{i}_{k,n} = \sum_{i \in \mathcal{U}_{k,n}} \sum_{n'=n}^{N} \text{VD}_{i,n'} \tag{5.5}$$

where $x_{i,n} = 1$ and $x_{i',n} = 0, \quad \forall i^{'} \in \mathcal{U}_{k,n}\backslash\{i\}$. After computing Eq. (5.5) $\forall i \in \mathcal{U}_{k,n}$, the bandwidth is allocated to user i^{*}, that achieves $\tilde{\text{VD}}^{i^{*}}_{k,n} \leq \tilde{\text{VD}}^{i}_{k,n}$.

The idea of this allocation metric is to choose the user that, when allocated the airtime, will result in the lowest overall future BS video degradation. This means that ideally the selected user will have a good *current* channel quality, and poor future conditions, relative to the other users. When selected, the user will achieve the best *reduction* in future BS VD.

- *Step 4*: To introduce the BS airtime trade-off, the user allocation result of step 3 is applied only if the resulting improvement in BS VD, before and after allocation, is larger than a threshold γ, i.e., $\tilde{\text{VD}}_{k,n-1}^{i^*} - \tilde{\text{VD}}_{k,n}^{i^*} > \gamma$. A larger value of γ will introduce more weight to the air-time reduction objective.

- *Step 5*: Repeat steps 2 to 4 for all $n \in \mathcal{N}$.

- *Step 6*: Calculate VD_{Net}.

- *Step 7*: Repeat steps 2-5 until there is no more decrease in VD_{Net}.

The complete procedure is presented in Algorithm 4, which we refer to as PGWA-MinAirVD-Alg.

Distributed PGWA-MinAirVD Algorithm

The goal of the distributed PGWA-MinAirVD algorithm is to allow each BS to perform its own predictive resource allocation. To account for network-wide rate predictions, each BS will have a rate map (or radio map) of the cooperating region of interest (e.g., several BSs along a highway). At the start of the prediction interval, BSs exchange the rate predictions of the users currently under their service. Then, instead of initializing $x_{i,n} = 0$ as in step 1 of the PGWA-MinAirVD-Alg, each BS performs a temporary allocation for the upcoming N slots, where $x_{i^*,n} = 1$ for the

user i^* predicted to have the highest channel rate among all the users at slot n. To do so, the cooperating BS require the initial rate prediction vectors of all the users. Then, with this as the baseline allocation, steps 2 to 5 of PGWA-MinAirVD-Alg are performed independently at each BS, with no further iterations. The intuition of this distributed procedure is that each BS first makes an initial baseline allocation based on a MaxRate scheme, and thereafter uses the VD metrics to adjust the allocations based on the procedure in PGWA-MinAirVD-Alg. An important consideration is that for users handed over, the computation of VD will not be possible in this distributed implementation as the previous values of $x_{i,n}$ are unknown to the target BS, accepting the user. This can be circumvented if the user reports its buffer status during handover for cases of pre-buffered content, or its measure of VD in cases of buffer underflow. Alternatively, the serving BS can also report this information to the target BS during handover. We will refer to this algorithm as PGWA-MinAirVD-Alg-Distr.

Algorithm 4 Centralized PGWA-MinAirVD Algorithm

Require: $\hat{r}_{i,n}, \mathcal{U}_{k,n}, V, \tau, , M, K, N$
1: Initialize $x_{i,n}, R_{i,n} = 0, \quad \forall\, i, n$
2: **repeat** {allocation iterations}
3: Calculate $\mathrm{VD}_{\mathrm{Net}}$ before allocation.
4: **for all** time slots n **do**
5: Reset $x_{i,n} = 0, \quad \forall i$.
6: **for all** base stations k **do**
7: **for all** users $i \in \mathcal{U}_{k,n}$ **do**
8: Calculate $\tilde{\mathrm{VD}}_{k,n}^{i}$ using Eq. (5.5)
9: **end for**
10: Set $x_{i^*,n} = 1$ to i^* that achieves $\tilde{\mathrm{VD}}_{k,n}^{i^*} \leq \tilde{\mathrm{VD}}_{k,n}^{i}$ only **if** $\tilde{\mathrm{VD}}_{k,n-1}^{i^*} - \tilde{\mathrm{VD}}_{k,n}^{i^*} > \gamma$
11: **end for**
12: **end for**
13: Calculate $\mathrm{VD}_{\mathrm{Net}}$ after allocation.
14: **until** {no more decrease in $\mathrm{VD}_{\mathrm{Net}}$}
15: **return** $\mathbf{x}$

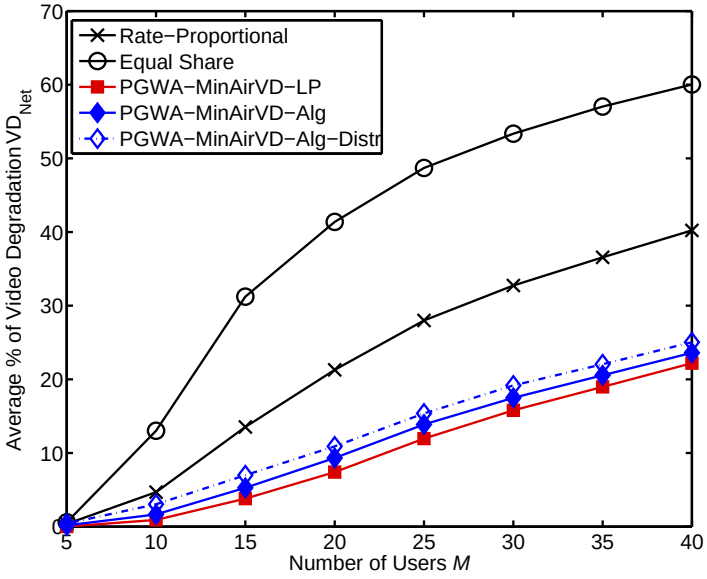
5.5.3 Performance Evaluation

Simulation Setup

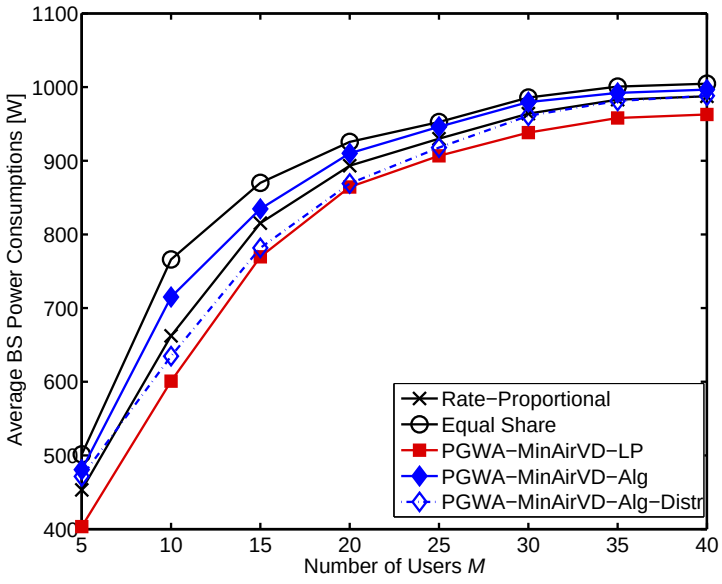
We consider the six BS road network in Figure 3.1(a) for evaluation and use Gurobi 5.1 [52] to solve the optimization problems. We assume a BS transmit power of 40 W, a center carrier frequency of 2 GHz, and a bandwidth of 10 MHz. The video streaming rate V is set to 3 Mbit/s, and the prediction window N to 250 slots, with a slot duration τ of 1 s. BS power consumption at minimum and maximum load is 200 W and 1300 W respectively, as presented for macro BSs employing time-domain duty-cycling in the power model of reference [87]. We compare the performance of the PGWA schemes with the ES and RP baseline allocators presented in Section 5.4.3.

Results and Discussion

Figure 5.3(a) illustrates the video performance of the PGWA schemes with the objective of minimizing degradations without regard to BS power consumption (i.e., $w_{\text{VD}} = 1, w_{\text{Air}} = 0$ and $\gamma = 0$). Significant gains (up to 45% reduction in VD) are observed compared to the traditional schemes that do not look ahead at future user rates. Additionally, both of the proposed centralized and distributed PGWA algorithms achieve close to optimal performance, indicating their effectiveness in minimizing video degradations. Figure 5.3(b) shows the corresponding BS power consumption, where all schemes have somewhat similar performance. This is expected as power reduction was not considered in this setting. In Figure 5.4(a) and Figure 5.4(b) we demonstrate the potential VD-Power trade-off that can be achieved with the proposed PGWA video delivery schemes. As α and γ increase, the LP formulation and the PGWA-MinAirVD algorithms decrease the power consumption at

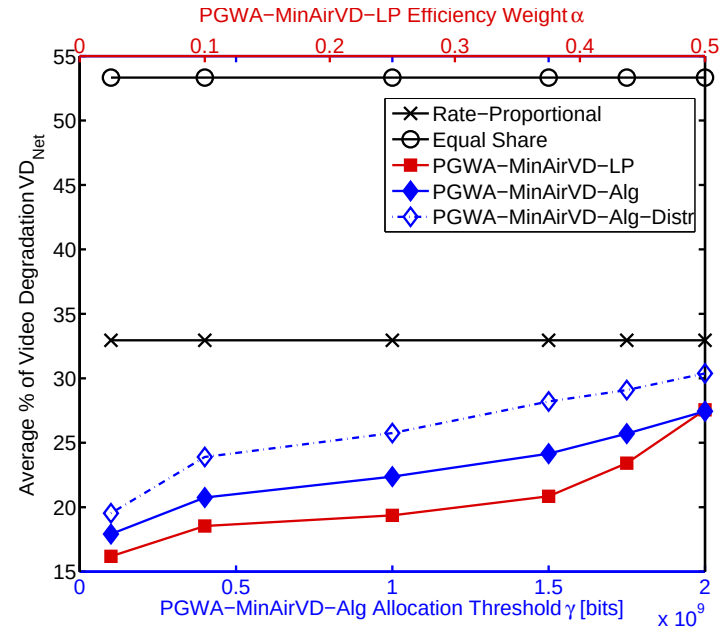


(a) Video Degradation $\mathrm{VD}_{\mathrm{Net}}$: $w_{\mathrm{VD}} = 1, w_{\mathrm{Air}} = 0$ and $\gamma = 0$.

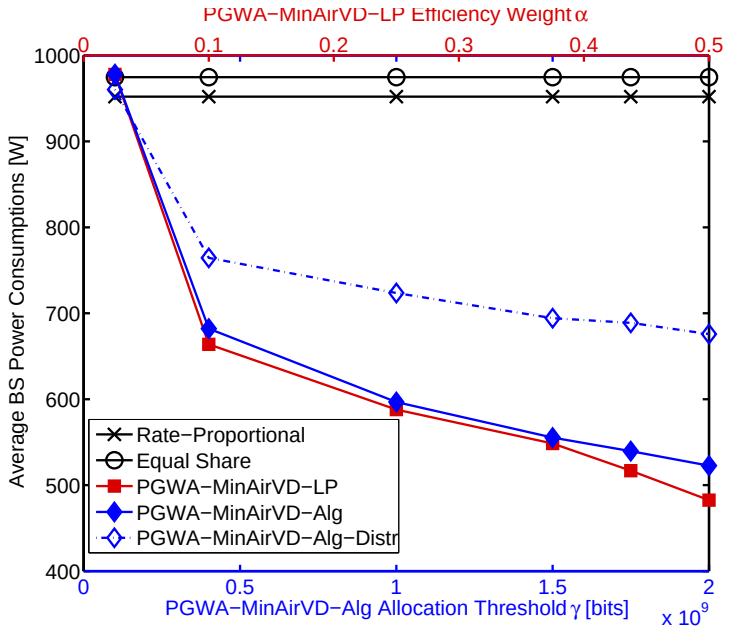


(b) BS Power Consumption: $w_{\mathrm{VD}} = 1, w_{\mathrm{Air}} = 0$ and $\gamma = 0$.

Figure 5.3: Video degradation $\mathrm{VD}_{\mathrm{Net}}$ and BS power consumption for varying number of users.



(a) Video Degradation $\mathrm{VD}_{\mathrm{Net}}$: $w_{\mathrm{VD}} = 1 - \alpha, w_{\mathrm{Air}} = \alpha$.



(b) BS Power Consumption: $w_{\mathrm{VD}} = 1 - \alpha, w_{\mathrm{Air}} = \alpha$.

Figure 5.4: Trade-off of $\mathrm{VD}_{\mathrm{Net}}$ and BS power consumption for varying PGWA-MinAirVD-LP and PGWA-MinAirVD-Alg air-time weights.

the cost of an increase in VD. This trade-off is summarized in Figure 5.5 which illustrates the Pareto-optimal solution between VD_{Net} and BS power. The significant (simultaneous) gains in VD and power are evident over the ES and RP allocation schemes. We observe that with $\alpha = 0.3$, BS power is reduced by almost 50% while VD is simultaneously reduced by 40% compared to RP allocation. The figure also demonstrates how the proposed centralized PGWA-MinAirVD-Alg scheme closely follows the Pareto-optimal benchmark curve of the PGWA-MinAirVD-LP that is solved offline using Gurobi [52]. Finally, the results also indicate that while the distributed scheme (PGWA-MinAirVD-Alg-Distr) offers considerable gains over RP, its performance deviates from the Pareto-optimal benchmark as γ increases.

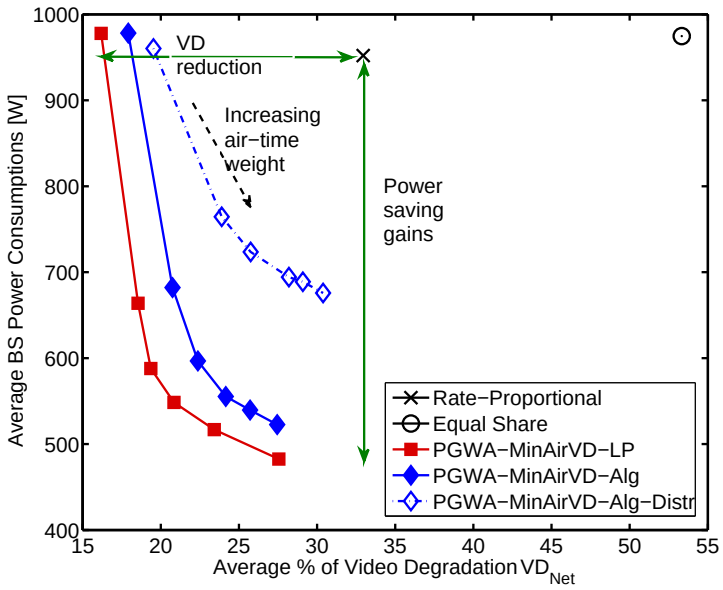


Figure 5.5: Pareto-optimal trade-off of VD_{Net}-BS Power Consumption for the different allocation algorithms; $M = 30$.

5.6 Joint Power-Quality Planning Optimization for Adaptive Video Transmission

In this section, we now consider the general case of PGWA for video streaming where:

1. users request stored video content that is transmitted using *adaptive* bitrate streaming over HTTP, following the model of Section 4.3.2.

2. BSs can enter deep sleep modes for a minimum duration of time to save additional energy.

The objective is to jointly optimize rate allocation and video segment quality planning for all the users in the network, with energy consumption considerations. To this end, we present a Predictive Green Streaming (PGS) framework that leverages rate predictions to accomplish this goal.

5.6.1 System Overview

Figure 5.6 illustrates the considered architecture for HTTP-based AVS in the wireless network, and outlines the required steps in PGS to conceptualize its operation. First, we assume that user location and navigation information is provided to the BS, which determines the future rates $\hat{\mathbf{r}}$ users will experience by consulting a radio map database. As our focus is to develop the predictive AVS transmission schemes, we assume $\hat{\mathbf{r}}$ is provided to the PGS controller, and thereafter investigate the effect of prediction errors in Section 5.6.4. The PGS framework uses $\hat{\mathbf{r}}$ defined over a time horizon, to minimize power consumption while achieving a target video quality level with no video stalls. To do so, it jointly determines the optimal (i) user rate allocations $\mathbf{x}$, (ii) video segment qualities $\mathbf{q}$, (iii) BS transmit powers $\mathbf{p}$, and (iv) the BS on/off statuses

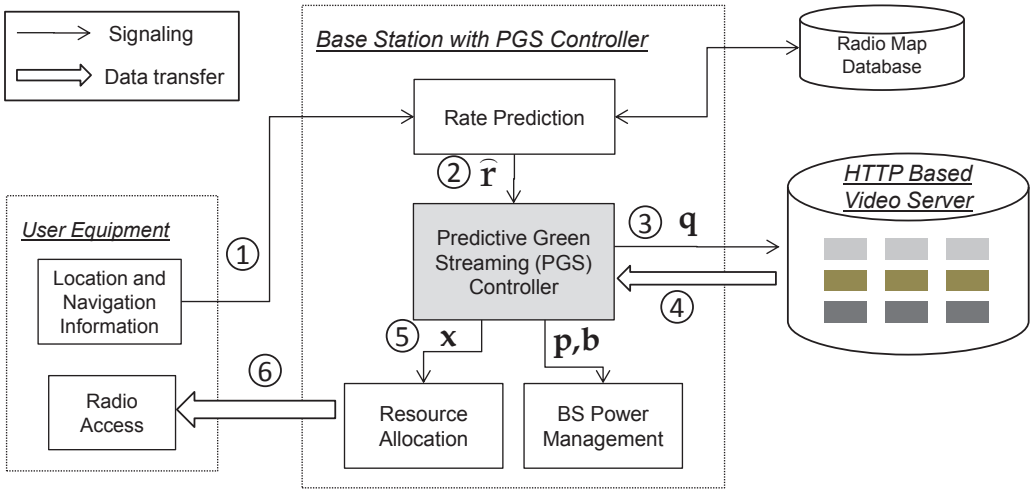


Figure 5.6: Predictive Green Streaming Operation.

b. The required segments as specified in **q** are requested from the HTTP-based video servers. These segments are then delivered to users over time slots in accordance with the determined rate allocation plan in **x**. The PGS controller also determines the deep sleep schedule of the BS that minimizes power consumption without violating user requirements through the optimization variable **b**, which is passed onto the BS power management unit.

5.6.2 Optimal Problem Formulation

We formulate two objectives for PGS as MILPs to provide benchmark solutions for performance evaluation. In the first objective, PGS-MinPower minimizes the total BS power consumption over a network, where BSs can enter deep sleep modes, under the constraint that no users experience video stalling. The MILP formulation is non-trivial due to the tight coupling between the large number of optimization variables. We then present PGS-MinAir with the objective of minimizing transmission airtime, and therefore BS load. However, in PGS-MinAir, BS turn-off is not enabled, and can

be considered a special case of PGS-MinPower.

To formulate the PGS problems several constraints have to be considered, which can be classified into 1) user requirement constraints, and 2) BS operation constraints. A summary of the symbols that are frequently used in the development of these constraints can be found in Table 5.1.

User Constraints

Joint Rate Allocation and Segment Quality for Smooth Streaming: As discussed in Section 4.3.3, the joint relationship between the cumulative allocated rate and segment quality selection that ensures smooth playback is captured in the following constraints, which we re-write for the reader's convenience:

$$\tau_{\text{seg}} \sum_{s'=1}^{s} \sum_{l=1}^{q_{\max}} q_{i,s',l}\, f^{Q}_{\text{rate}}(l) \leq \sum_{n'=1}^{sN_{\text{seg}}} x_{i,n'}\, \hat{r}_{i,n'}, \quad \forall\, i, \forall s, \tag{5.6}$$

$$\sum_{l=1}^{q_{\max}} q_{i,s,l} = 1, \qquad \forall\, i \in \mathcal{M}, \forall s \in \{1, 2, \cdots, S\}. \tag{5.7}$$

Target Quality: If $l_{\text{req}} \in \{1, \cdots, q_{\max}\}$ denotes the desired average quality level for each user, then

$$\sum_{s=1}^{S} \sum_{l=1}^{q_{\max}} q_{i,s,l} \geq l_{\text{req}} S, \qquad \forall\, i \in \mathcal{M}, \tag{5.8}$$

represents the average user quality constraint over all segments.

User Buffer Limit: The buffer limit can play a significant role in energy efficiency by limiting the amount of prebuffered data. This is useful if certain users are known to not complete watching the request video streams. As derived in Section 4.3.3, this

Table 5.1: Summary of Frequently Used Symbols in PGS

Symbol	Description
i	User index, $i = \{1, 2, \cdots, M\}$
k	BS index, $k = \{1, 2, \cdots, K\}$
n	Time slot index, $n = \{1, 2, \cdots, N\}$
$q_{\max}$	Maximum quality level
s	Segment index, $s = \{1, 2, \cdots, S\}$
N	Number of slots in the prediction window
N_{seg}	Number of slots in a video segment
S	Number of segments in the lookahead window
T	Duration of the lookahead window [s]
τ	Duration of a time slot [s]
τ_{seg}	Duration of a video segment [s]
$\mathcal{K}$	Set of BSs in the network
$\mathcal{M}$	Set of users in the network
$\mathcal{N}$	Set of time slots in the prediction window
$\mathcal{N}^s$	Set of time slots belonging to segment s
$b_{k,n}$	Binary decision variable for on/off status of BS k at slot n
$p_{k,n}$	Transmit power of BS k at slot n
$q_{i,s,l}$	Binary variable for quality level l of segment s for user i
$\hat{r}_{i,n}$	Link rate of user i at slot n [bits]
$\text{BS}_{k,n}^{\text{air}}$	Available airtime of BS k at slot n
$D_{i,s}$	Cumulative number of bits required by user i to stream the first s segments [bits]
$R_{i,n}$	Cumulative rate allocated to user i by slot n [bits]
$\mathcal{U}_{k,n}$	Set containing the indices of users associated with BS k at slot n
$x_{i,n}$	Fraction of airtime assigned to user i at slot n

can be expressed by the following constraint.

$$\begin{aligned}&\sum_{n'=1}^{n} x_{i,n'}\,\hat{r}_{i,n'} - \tau_{\text{seg}} \sum_{s'=1}^{\lfloor n/N_{\text{seg}} \rfloor} \sum_{l=1}^{q_{\max}} q_{i,s',l}\, f_{\text{rate}}^{Q}(l) \\ &- \frac{\tau_{\text{seg}}}{N_{\text{seg}}} (n \bmod (N_{\text{seg}})) \sum_{l=1}^{q_{\max}} q_{i,\lceil n/N_{\text{seg}} \rceil,l}\, f_{\text{rate}}^{Q}(l) \leq L_i, \quad \forall\, i, \forall n.\end{aligned} \tag{5.9}$$

Total User Allocation: Finally, the total number of bits allocated to a user during the N slots should be equal to that specified by the segment quality plan, which is expressed by

$$\sum_{n=1}^{N} x_{i,n}\,\hat{r}_{i,n} = \tau_{\text{seg}} \sum_{s=1}^{S} \sum_{l=1}^{q_{\max}} q_{i,s,l}\, f^{Q}_{\text{rate}}(l), \qquad \forall\, i \in \mathcal{M}. \tag{5.10}$$

BS Constraints

BS Resource Limit: The BS resource constraint limits the sum of the user airtime fractions to unity, i.e.,:

$$\sum_{i \in \mathcal{U}_{k,n}} x_{i,n} \leq 1, \qquad \forall\, k \in \mathcal{K},\, \forall n \in \mathcal{N}. \tag{5.11}$$

This constraint is applied at each BS, where the summation is over all users i associated with BS k at slot n.

BS Slot Power Consumption: According to the BS power model of Eq. (5.1), the power consumed by each BS is dependent on (i) the total user airtime, and (ii) whether the BS is kept on or switched off. This is expressed by the following constraint

$$(P_{\text{m}} - P_0) \sum_{i \in \mathcal{U}_{k,n}} x_{i,n} - p_{k,n} + P_0\, b_{k,n} = 0, \;\; \forall\, k \in \mathcal{K},\, \forall n \in \mathcal{N}, \tag{5.12}$$

where the binary decision variable $b_{k,n}$ is multiplied by P_0 to produce zero sleep power when the BS is off.

BS On Constraint: To enforce a BS to be on if there is any load, we apply the following constraint

$$\sum_{i \in \mathcal{U}_{k,n}} x_{i,n} - b_{k,n} \leq 0, \ \ \forall\, k \in \mathcal{K}, \forall n \in \mathcal{N}. \tag{5.13}$$

BS Off Indicator: In order to monitor when a BS is turned off, we introduce an indicator variable $I_{k,n}$ which is equal to 1 only when a BS is turned off. This is achieved by

$$-\, b_{k,n-1} + b_{k,n} + I_{k,n} = 0, \ \ \forall\, k \in \mathcal{K}, \forall n \in \mathcal{N}, \tag{5.14}$$

where $b_{k,0} = 0, \ \forall\, k$. On the other hand, when a BS is switched on, $I_{k,n} = -1$, and if there is no change, $I_{k,n} = 0$. The value of this indicator is used by the following constraint to ensure that a BS remains off for a minimum number of n_{off} slots.

Minimum Off Time: To model the minimum off duration we restrain the BS from turning on for n_{off} slots, once it has been turned off. This can be achieved by

$$b_{k,n} + I_{k,n+c} \leq 1, \ \ \forall\, k \in \mathcal{K}, \forall n \in \mathcal{N}, \forall c, \tag{5.15}$$

where $c = 1, \cdots, n_{\text{off}}$, and $n + c <= N$. The above equation ensures that if the indicator of the previous time slot is 1, then $b_{k,n}$ will have to remain equal to zero for n_{off} slots. This is controlled through the variable c that generates n_{off} constraints to define the status of the upcoming n_{off} slots, for every n. If, on the other hand, the indicator is not 1, then $b_{k,n}$ can take on any value.

PGS-MinPower Optimal Problem Definition

Considering the previously discussed constraints, the PGS-MinPower problem can be formulated as the following MILP:

$$\underset{\mathbf{x},\mathbf{q},\mathbf{p},\mathbf{b}}{\text{minimize}} \quad \sum_{\forall k \in \mathcal{K}} \sum_{n=1}^{N} p_{k,n} \tag{5.16}$$

$$\begin{aligned}
\text{subject to:} \quad & \text{Constraints : } \textit{Eq. (5.6)} \text{ to } \textit{Eq. (5.15)} \\
& 0 \leq x_{i,n} \leq 1, \ \ \forall\, i \in \mathcal{M}, \forall n \in \mathcal{N}, \\
& q_{i,s,l} \in \{0,1\}, \ \ \forall i \in \mathcal{M}, \forall s \in \mathcal{S}, \forall l \in \mathcal{Q}, \\
& 0 \leq p_{k,n} \leq P_m, \ \ \forall\, k \in \mathcal{K}, \forall n \in \mathcal{N}, \\
& b_{k,n} \in \{0,1\}, \ \ \forall\, k \in \mathcal{K}, \forall n \in \mathcal{N}.
\end{aligned}$$

Solving the problem in Eq. (5.16) provides the optimal joint allocations of all the decision variables. However, it is computationally intractable to solve large instances of PGS-MinPower due to the large number of decision variables (M(N+S)+2KN), and the coupling between them. Further, memory requirements are extremely significant as the resulting constraint matrix has a size of (2M+MN+2MS+5KN), which is very large since the duration of the prediction window impacts both N and S.

It is worth noting that overhead may be introduced when turning BSs off/on. This may be accounted for by increasing the value of n_{off} to prevent frequent, short sleeps. Another way to directly incorporate the overhead of turning BSs off/on is through the BS Off Indicator variable $I_{k,n}$ defined in Eq. (5.14). This can be achieved by adding another power consumption term to the objective in Eq. (5.16). The added term would sum over $I_{k,n}$ and multiply the result by a constant that denotes the power

consumption of a single on/off switch. The PGS solution would then minimize the total power consumed while accounting for the overhead of the deep sleep switches.

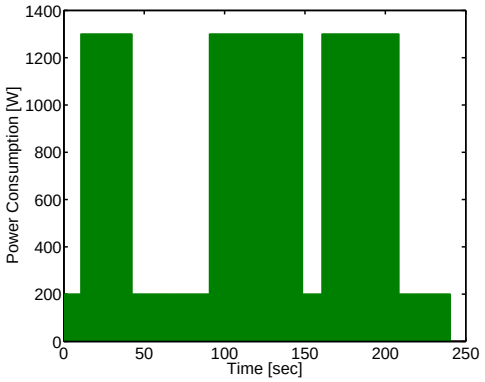
PGS-MinAir Optimal Problem Definition

The PGS-MinAir problem considers the case where BSs cannot be switched off into deep sleep modes, for example due to other types of traffic in the network. PGS-MinAir can therefore be formulated by setting the BS on/off decision variable to 1, and excluding Eq. (5.13) to Eq. (5.15) as required constraints. However, a more compact formulation can also exclude the BS power $p_{k,n}$ variables, and airtime can be minimized directly through user allocations $x_{i,n}$. This is possible due to the linear power model of Eq. (5.1) where BS power is proportional to user airtime. Therefore, the PGS-Airtime MILP problem can be formulated as

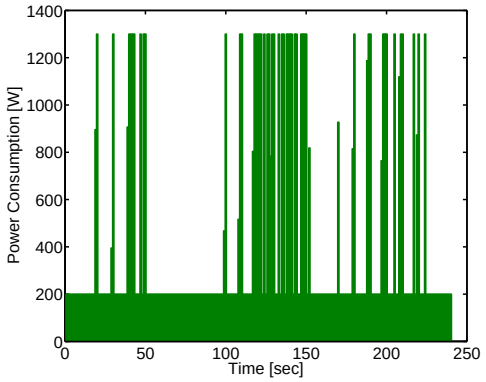
$$\underset{\mathbf{x},\mathbf{q}}{\text{minimize}} \quad \sum_{\forall i \in \mathcal{M}} \sum_{n=1}^{N} x_{i,n} \tag{5.17}$$

$$\begin{aligned} \text{subject to:} \quad & \text{Constraints : } \; Eq.\ (5.6) \text{ to } Eq.\ (5.11), \\ & 0 \leq x_{i,n} \leq 1, \;\; \forall\, i \in \mathcal{M}, \forall n \in \mathcal{N}, \\ & q_{i,s,l} \in \{0,1\}, \;\; \forall i \in \mathcal{M}, \forall s \in \mathcal{S}, \forall l \in \mathcal{Q}. \end{aligned}$$

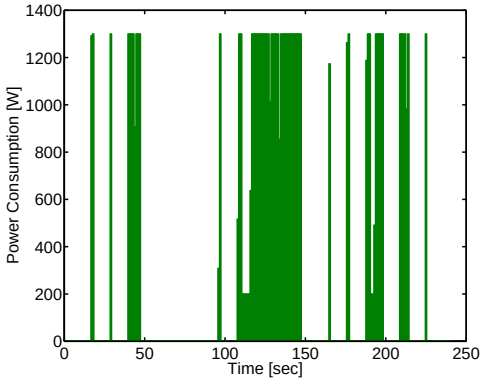
Figure 5.7 depicts the resulting BS power consumption plan for PGS versus a traditional scheme where BS airtime is shared equally among video streaming users. In the scenario considered, vehicular users arrive at the BS in three consecutive groups. In Figure 5.7(a), as long as users are present, BS airtime is completely utilized. However, in Figure 5.7(b) and Figure 5.7(c), PGS allows the BS to transmit in a spectrally efficient way without violating user streaming requirements. Note that while



(a) Traditional non-predictive operation.



(b) PGS-Airtime Minimization.



(c) PGS-Power Minimization.

Figure 5.7: Sample BS power consumption with time: $P_0 = 200\,\mathrm{W}$ and $P_m = 1300\,\mathrm{W}$. (a) In traditional operation, BS airtime is inefficient utilized. (b) With PGS-MinAir, BS airtime is minimized by opportunistic allocations. In (c) PGS-MinPower groups user allocations to allows deep sleep modes.

PGS-MinAir minimizes total transmit time, PGS-MinPower in Figure 5.7(c) is able to strike the optimal tradeoff between serving users when their individual rates are high, and grouping user transmissions together (even if not at their respective best rates) to generate blocks of sleep time. This comes at the cost of increased complexity as observed in the PGS-MinPower formulation, where the optimization variables are tightly coupled. However, at high load, the power saving gains of PGS-MinPower

over PGS-MinAir will decrease and eventually converge to PGS-MinAir. This is due to the decreased ability to generate silent space long enough for a BS switch off. We discuss more details on the tradeoffs involved in the numerical results of Section 5.6.4, but before that, we first present a polynomial-time solution of the PGS the problem that achieves close to optimal results.

5.6.3 Multi-stage Algorithm

In this section, we develop a multi-stage approach to solve the PGS MILPs presented in Section 5.6.2. Figure 5.8 outlines the steps involved. The core stage is a user rate allocation algorithm that assigns BS airtime to users over the prediction window, thereby solving $\mathbf{x}$ and $\mathbf{p}$. Thereafter, segment qualities are explicitly planned for each user based on the allocated bits, and BS on/off statuses are determined from the resulting idle time in $\mathbf{p}$. This approach of decoupling is based on the intuition that an efficient rate allocation scheme (that exploits rate predictions) will provide power savings, while satisfying user quality needs. Before discussing each stage, we introduce the following definitions:

- Cumulative per segment demand $D^{\text{seg}}_{i,s}$: the total number of bits required by user i to stream the first s segments. For a given target quality level l_{req}, $D^{\text{seg}}_{i,s} = s\, f^{Q}_{\text{rate}}(l_{\text{req}})$ [bits], $\forall\, i$.

- User rate percentile $\hat{r}^{y\%}_{i,n}$: the y^{th} percentile of the future user rate, i.e., computed over $\hat{r}_{i,n\leq n'\leq N}$, for each user.

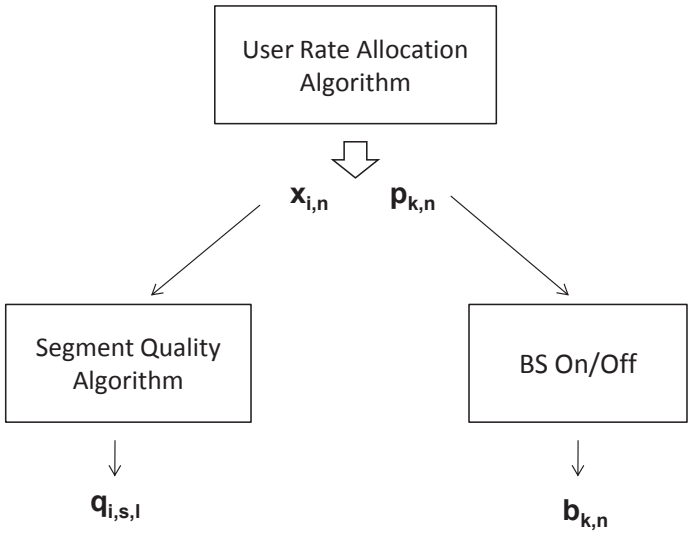


Figure 5.8: The Proposed Multi-Stage PGS Solution.

Rate Allocation

The RA strategy is to divide the problem into a series of allocation subproblems performed at the start of each segment. The idea of this decomposition is to minimize airtime while focusing on satisfying the streaming constraint in Eq. (5.6), that is performed for each segment. If $\mathcal{N}^s$ denotes the set of slots comprising segment s, then $\mathcal{N}^s = \{(s-1)N_{\text{seg}}+1, (s-1)N_{\text{seg}}+2, \cdots, sN_{\text{seg}}\}$, and allocation is made incrementally for each $\mathcal{N}^s$. Each allocation is further divided into two steps 1) airtime minimization, and 2) opportunistic pre-buffering. In the first step, users that do not have enough content pre-buffered to stream the upcoming segment at the target quality level are prioritized and their demands fulfilled with the minimum possible BS airtime. In the second step, users that have exceptionally good channel conditions are granted excess airtime to prebuffer future video content. This will reduce the airtime required later to download upcoming segments. Next, we discuss the details of each step.

Airtime Minimization: At the start of segment s, each BS k determines the set of priority users $\mathcal{P}_{k,s}$ that have insufficient cumulative allocation to play the upcoming video segment at the target quality level. The set $\mathcal{P}_{k,s}$ will therefore not include users that have pre-buffered segments. The required rate allocation $r_i^{\mathcal{P}}$ for user i can then be computed as

$$r_i^{\mathcal{P}} = D_{i,s}^{\text{seg}} - R_{i,sN_{\text{seg}}}, \quad \forall i \in \mathcal{P}_{k,s}, \tag{5.18}$$

where $R_{i,0} = 0 \ \forall i$. To serve the users with these rate requirements, using the minimum BS airtime, we need to solve the following optimization problem

$$\underset{\mathbf{x},\mathbf{Y}}{\text{minimize}} \quad \sum_{n\in\mathcal{N}^s} \sum_{i\in\mathcal{P}_{k,s}} x_{i,n} + \beta \sum_{i\in\mathcal{P}_{k,s}} Y_i \tag{5.19}$$

$$\begin{aligned}
\text{subject to:} \quad & -\sum_{i\in\mathcal{P}_{k,s}} x_{i,n}\hat{r}_{i,n} - Y_i \leq -r_i^{\mathcal{P}}, && \forall n \in \mathcal{N}^s, \\
& \sum_{i\in\mathcal{P}_{k,s}} x_{i,n} \leq 1, && \forall n \in \mathcal{N}^s, \\
& 0 \leq x_{i,n} \leq 1, && \forall\, i \in \mathcal{P}_{k,s}, n \in \mathcal{N}^s, \\
& 0 \leq Y_i, && \forall\, i \in \mathcal{P}_{k,s}.
\end{aligned}$$

The variables Y_i are introduced to capture any unfulfilled rate allocation, for the instances when it is not possible to satisfy all user requirements. As satisfying users with the target quality level has a higher precedence over saving airtime, the weight parameter $\beta > 1$. Generally, at low to moderate loads (where there are potential power savings), solving Eq. (5.19) will yield $Y_i = 0$, and user quality requirements will be met with the minimum BS airtime.

Note that the problem in Eq. (5.19) has a linear objective function with linear

constraints and is therefore an LP, which in general can be solved in polynomial time, with even the widely-used Simplex algorithm [73]. Further, the problem dimension is much smaller than the optimal PGS MILP formulations, and therefore provides significant computational and memory requirement gains.

Alternatively, to avoid the requirement of BSs being equipped with LP solvers, we present the following simple algorithm to solve the problem in Eq. (5.19). First, the set $\mathcal{P}_{k,s}$ is sorted in descending order of user requirements $r_i^{\mathcal{P}}$. Then, each user $i \in \mathcal{P}_{k,n}$ is selected in sequence to transmit at the time slot $n^* \in \mathcal{N}^s$, that has the largest predicted rate for that user, i.e.,

$$n^* = \arg\max_n \hat{r}_{i,n}, \quad \forall n \in \mathcal{N}^s. \tag{5.20}$$

Note that if $\mathrm{BS}_{k,n}^{\mathrm{air}}$ denotes the airtime available in BS k at slot n, then the search in Eq. (5.20) will exclude slots where $\mathrm{BS}_{k,n}^{\mathrm{air}} = 0$. The airtime allocated to the user is then $x_{i,n^*} = \hat{r}_{i,n^*}/r_i^{\mathcal{P}}$, and the remaining BS airtime is updated to $\mathrm{BS}_{k,n^*}^{\mathrm{air}} = \mathrm{BS}_{k,n^*}^{\mathrm{air}} - x_{i,n^*}$. After iterating over all $i \in \mathcal{P}_{k,n}$, $R_{i,n}, \mathcal{P}_{k,s}$ and $r_i^{\mathcal{P}}$ are updated, and the process is repeated until either $\mathcal{P}_{k,s} = \phi$ or there is no remaining BS airtime for $n \in \mathcal{N}^s$. This procedure is outlined in lines 6-14 in Algorithm 5, and numerical results in Section 5.6.4 indicate that it provides almost identical results to the LP of Eq. (5.19).

Opportunistic Pre-buffering: While the airtime minimization stage provides users with their *immediate* needs efficiently, it does not capitalize on granting users their *future* content in advance when their rates are high. Implementing such pre-buffering results in reduced overall airtime as bulk transmissions are made opportunistically in short time durations, and thereafter users are not served. However,

the question is, when is a good time to pre-buffer content to a user? A simple rate threshold will not work well for cases where users have unequal rate distributions over $\mathcal{N}$. We therefore use the previously defined rate percentile $\hat{r}_{i,n}^{y\%}$ metric, as it provides each user with an independent threshold, derived from its own rate statistics. This is applied as follows: for each slot $n \in \mathcal{N}^s$, we first find the user i^* with the largest rate, i.e.,

$$i^* = \arg\max_i \hat{r}_{i,n} \quad \forall i \in \mathcal{U}_{k,n}. \tag{5.21}$$

This rate is then compared to the user's y^{th} rate percentile at n, and if $\hat{r}_{i^*,n} > r_{i^*,n}^{y\%}$, the user is allocated the remaining BS airtime at that slot, and the user airtime is updated to $x_{i^*,n} = x_{i^*,n} + \text{BS}_{k,n}^{\text{air}}$. This completes the two steps of rate allocation performed $\forall n \in \mathcal{N}^s$. The procedure is then repeated by each BS, for each segment in sequence, as outlined in Algorithm 5. The BS power consumption matrix $\mathbf{p}$ is then calculated using Eq. (5.1), where $\text{BS}_{k,n}^{\text{load}} = 1 - \text{BS}_{k,n}^{\text{air}}$.

Setting y: The value of y can affect the resulting power savings and is dependent on the current network load. At low load, a higher y will cause users to only pre-buffer at close to peak rates. This is more efficient provided users do not thereafter fall short of their needs and request airtime before encountering another 'peak'. On the other hand, when load is high, a lower value of y is preferred to allow users to prebuffer more frequently, even if at moderate rates. Although intermediate values $y \in [70, 80]$ provide a good tradeoff, the best value can be determined by iterating the procedure for different values and selecting the rate allocation $\mathbf{x}$ that gives the minimum power consumption.

Algorithm 5 User Rate Allocation Algorithm

Require: $\hat{r}_{i,n}, \mathcal{U}_{k,n}, D_{i,s}, K, M, N, N_{\text{seg}}$

Initialize $x_{i,n}, R_{i,n} = 0 \quad \forall\, i, n = 1, 2, ..N$
for all $y \in \{65, 70, \cdots, 95\}$ **do**
  Reset $x_{i,n} = 0$, $\text{BS}^{\text{air}}_{k,n} = 1 \quad \forall i, k, n.$
  for all segments s **do**
    for all base stations k **do**
      Find set of priority users $\mathcal{P}_{k,s}$, and compute $r_i^{\mathcal{P}}$ as in Eq. (5.18). Sort $\mathcal{P}_{k,s}$ in descending order of $r_i^{\mathcal{P}}$.
      while $\mathcal{P}_{k,s} \neq \phi$ **and** $\sum_{n \in \mathcal{N}^s} \text{BS}^{\text{air}}_{k,n} > 0$ **do**
        for all users $i \in \mathcal{P}_{k,s}$ **do**
          Find slot n^* with the largest rate as in Eq. (5.20).
          Set $x_{i,n^*} = \hat{r}_{i,n^*} / r_i^{\mathcal{P}}$.
          Set $\text{BS}^{\text{air}}_{k,n^*} = \text{BS}^{\text{air}}_{k,n^*} - x_{i,n^*}$
        end for
        Recompute $R_{i,n}, \mathcal{P}_{k,s}$ and$r_i^{\mathcal{P}}$.
      end while
      for all slots $n \in \mathcal{N}^s$ **do**
        Find user i^* with the largest $\hat{r}_{i,n} \quad \forall i \in \mathcal{U}_{k,n}$.
        If $\hat{r}_{i^*,n} > r^y_{i^*}$, then $x_{i^*,n} = x_{i^*,n} + \text{BS}^{\text{air}}_{k,n}$.
      end for
    end for
  end for
  Calculate $p_{k,n}$ using Eq. (5.1), where $\text{BS}^{\text{load}}_{k,n} = 1 - \text{BS}^{\text{air}}_{k,n}$.
  Calculate $P^y_{\text{Net}} = \sum_{k=1}^{K} \sum_{n=1}^{N} p_{k,n}$ for this iteration.
end for
Determine y^* that produces the minimum P^y_{Net}.
return $\mathbf{x}, \mathbf{p}$

Segment Quality Algorithm

After determining the rate allocation matrix $\mathbf{x}$ as specified in Algorithm 5, the user segment quality levels are planned. Since the objective is to determine the segment quality plan that maximizes quality while providing smooth playback, it is similar to Algorithm 2 presented in Section 4.3.4. The idea is to iterate over the segments in sequence and greedily maximize the current segment quality, while ensuring that the

future segments can be streamed, at least, at the lowest quality level.

BS On/Off Switching

To determine the BS on/off status we search each BS for long 'silent' transmission durations over the prediction window, where there is zero load. This is accomplished by the following simple procedure: 1) determine the set of time slots $\mathcal{N}_{\text{On}}$ where $p_{k,n} > P_0$, implying that the BS is on; 2) then determine the difference between the successive time slots in $\mathcal{N}_{\text{On}}$. If this is larger than n_{off}, it means that no transmission occurred for a duration long enough to turn the BS off for that period. A value of zero is subtracted from the first element of $\mathcal{N}_{\text{On}}$ to account for the possibility of switching the BS off before the first start, and the last element of $\mathcal{N}_{\text{On}}$ is subtracted from N to check for a turn off possibility at the end.

This completes the multi-stage PGS solution, which we refer to as PGS-MinPower-Alg. For the case where BSs cannot switch to deep sleep we do not apply the BS On/Off stage, and only airtime is minimized. This solution is denoted by PGS-MinAir-Alg. Finally, when implementing the LP of Eq. (5.19), the algorithm will be denoted by PGS-MinAir-AlgLP.

Computational Complexity

To evaluate the complexity of the PGS multi-stage solution, we first determine the complexity of each stage. The airtime minimization step of the rate allocation involves computing Eq. (5.18) and sorting the set $\mathcal{P}_{k,n}$, which has a time complexity of $O(MN + M \log M)$. Then, rate allocation over the N_{seg} slots takes $O(MN_{\text{seg}})$ time, leading to an overall complexity of $O(MN + M \log M + MN_{\text{seg}})$ for this step.

The subsequent pre-buffering includes computing the future rate percentile and takes $O(N_{\text{seg}}(M + N \log N))$ time. After accounting for S segments for each user, we arrive at an overall complexity of $O(MN^2)$ for rate allocation in Algorithm 5. In the segment quality algorithm, the core step is to evaluate the constraint in Eq. (5.6), which has as a time complexity $O(N + S)$ for a single user. This step is repeated at most $q_{\max}$ times when the constraint is violated, and repeated for each segment and each user. The resulting complexity order is $O(q_{\max}MS(N+S))$. In the worst case $S = N$, and $q_{\max}$ is typically less than 6, which gives a worst case runtime of $O(MN^2)$. As the BS on/off procedure presented earlier has a complexity of $O(KN^2)$, this leads to overall runtime of $O((M + K)N^2)$.

5.6.4 Performance Evaluation

In this section we present numerical results that demonstrate the potential energy savings achieved by exploiting rate predictions in the PGS framework. We also investigate the effects of prediction errors on the performance of the PGS schemes.

Simulation Setup

We consider two network scenarios. The first is a single cell with vehicles moving along a highway that crosses through the cell. For realistic vehicular mobility we use the SUMO traffic simulator [48] to generate mobility traces with a flow of 1 vehicle per second. Second, is a three BS network, also along a highway, with an inter-BS distance of 1 km. Vehicles arrive in groups of ten vehicles each, separated by 60 seconds. This creates the effect of vehicle grouping observed on highways.

BS transmit power is 40 W, and bandwidth is 5 MHz. BS power consumption

at minimum and maximum load is 200 W and 1300 W respectively as presented for macro BSs employing time-domain duty-cycling in the power model of [87]. The minimum off time for a BS is set to 10 s. The slot duration $\tau = 1$ s, and $T = 240$ s. We consider a video format with four quality levels of $\{0.25, 0.5, 0.75, 1\}$ Mbit/s, and a segment length $\tau_{\text{seg}} = 10$ s. Gurobi 5.1 [52] is used to solve the MILP optimization problems, and Matlab is used as a simulation environment.

We compare the performance of the PGS schemes against two baseline approaches that do not exploit rate predictions. These reference schemes have two stages: rate allocation, followed by quality adaptation. Two rate allocation schemes are considered: ES and RP which are implemented as discussed in Section 5.4.3. Segment quality is then adapted based on the allocated rate at the start of the current segment, and the highest supportable level is selected. These approaches are referred to as ES-AdaptQ, and RP-AdaptQ. We also consider a benchmark allocator that exploits rate predictions as in PGS. However, it is energy independent, and its objective is to maximize user quality. This is achieved by solving Eq. (5.17) with the objective of maximizing $q_{i,s}$. This allocator serves as reference to what can be achieved with rate predictions, but without considering energy savings, and is referred to as MaxQuality-MILP.

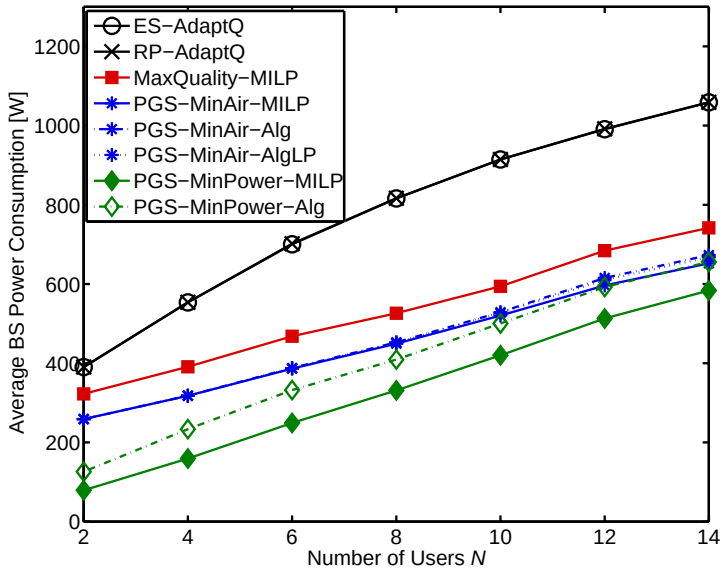
The network-wide video quality and power consumption metrics are defined as:

- Q_{Net}: the total quality of all delivered segments, divided by the number of requested segments.
- F_{Net}: the average percentage of playback time where the video is stalled, over all users.
- P_{Net}: the average downlink power consumption of all BSs over the time window T.

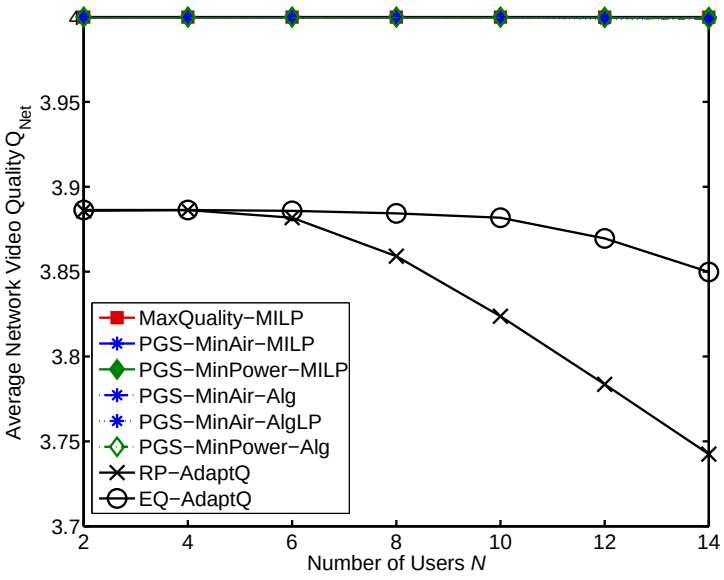
Single-cell Scenario

Figure 5.9(a) shows the average BS power consumption versus the number of users, where as expected, the allocators consume more power with increasing users. The MinAir schemes achieve significant power savings by exploiting rate predictions, without having to power down the BSs. The energy gains can also be viewed as spectral efficiency gains, where the saved airtime can be used for other users or applications. The MinAir-MILP and MinAir-AlgLP exhibit very close performance. This demonstrates the effectiveness of the developed multi-stage PGS framework that is able to achieve close to optimal performance but without significant computational complexity. Also note that the MinAir-AlgLP and the MinAir-Alg achieve very close performance, and therefore the LP formulation of Eq. (5.19) can be replaced with the simple segment airtime minimization algorithm presented in Section 5.6.3 without observable performance loss.

The MinPower-MILP scheme achieves further power savings by switching the BSs to sleep intermittently and making bulk transmissions to users when awake. The sleep times are coordinated such that the users' QoS is not violated. When few users are present, the BS can sleep for prolonged periods of time and therefore the power savings can be very large (approximately one eighth of the baseline allocator power is needed in the case shown in Figure 5.9(a)). However, as expected, with more users, MinPower-MILP gradually converges to MinAir-MILP since the BS cannot find time long enough for a 'sleep session'. The MinPower-Alg performs close to the MinPower-MILP (exact solution) with fewer users, but then deviates and converges to MinAir-Alg. The reason is that MinPower-MILP jointly optimizes BS on/off states with BS airtime minimization, and is therefore able to strike the optimal tradeoff between



(a) BS downlink power consumption for varying number of users in the single-cell scenario.



(b) Average quality level for varying number of users in the single-cell scenario.

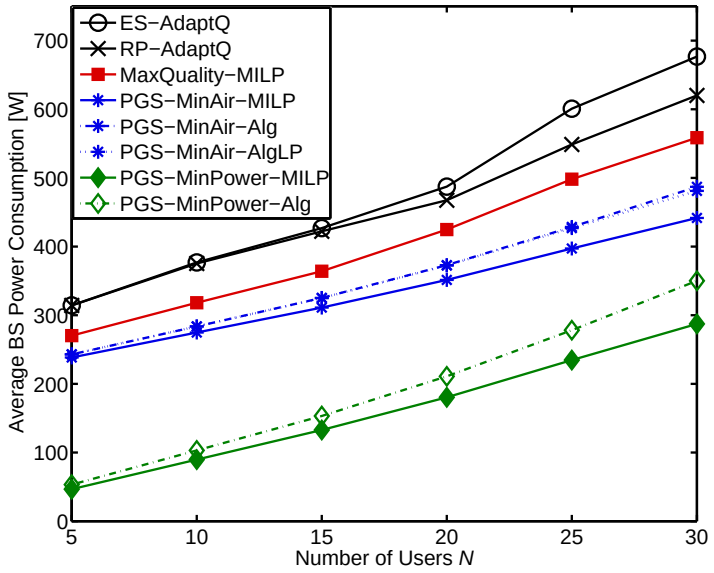
Figure 5.9: PGS results for the single-cell scenario.

serving users when their individual rates are high, and grouping user transmissions together (when not at their respective best rates) to generate blocks of sleep time. This, however, is at the cost of a tightly coupled MILP that can take several minutes to solve.

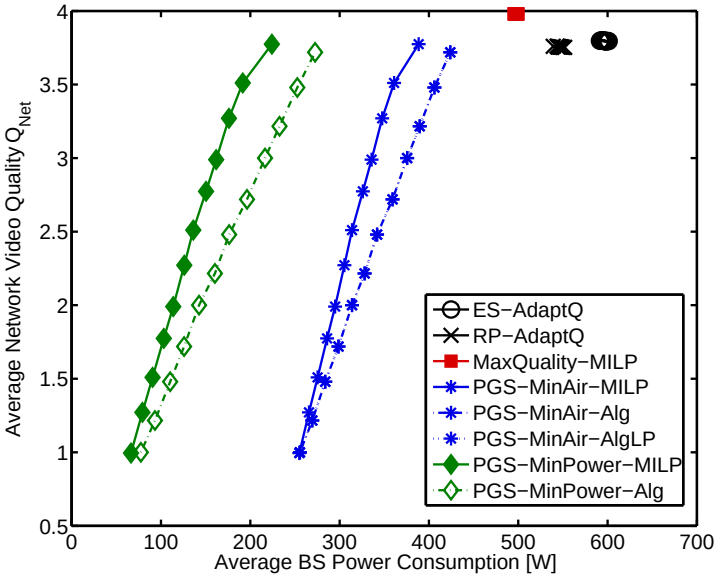
In Figure 5.9(b) we show the corresponding average segment quality level in this scenario. While the rate predictive schemes all achieve the highest quality of 4, the baseline schemes experience a slight quality degradation as the load increases, with the RP-AdaptQ scheme suffering more. The video freezing, which is not depicted, was less than 1% for all allocators.

Multi-cell Highway Scenario

Figure 5.10(a) shows average BS power consumption versus the number of users for the three BS highway scenario. In this multi-cell scenario, the power saving potential of the MinPower-MILP scheme is observed, while all the allocators achieve an average quality level of 3.75. User mobility information allows the BSs to sleep before users arrive in the cells. Further, as the allocation plans are made over the three cells, a user may be granted all the video content in one or two of the BSs and nothing in the third (allowing it to sleep). From Figure 5.10(a) we also note that in this scenario, the MinAir-Alg approaches deviate slightly from the MinAir-MILP solutions with increasing load. This is because with many users in a multi-cell network, the problem becomes more complex and it is more difficult to achieve optimality with the two-step rate allocation algorithm. A similar observation can be made for MinPower-Alg. Note that the scenario also demonstrates the spectral efficiency of the baseline RP-AdaptQ scheme over the EQ-AdaptQ.



(a) BS downlink power consumption for varying number of users in the multi-cell scenario.



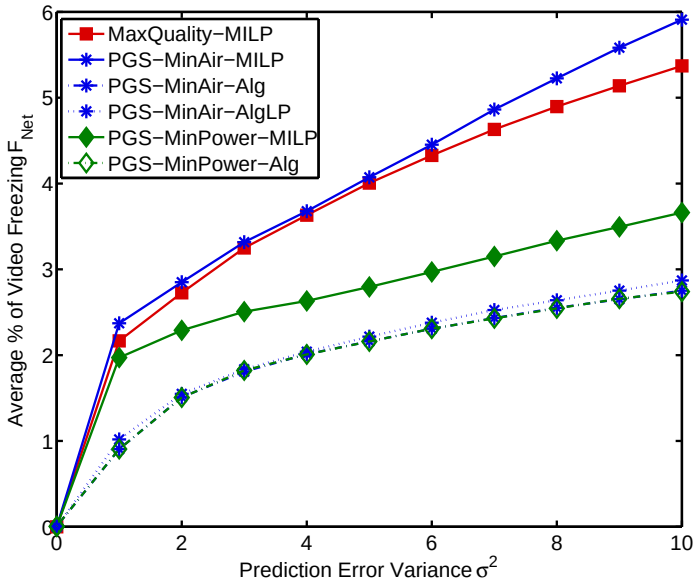
(b) The tradeoff between the average video quality and the BS power consumption in the multi-cell highway scenario.

Figure 5.10: PGS results for the multi-cell scenario.

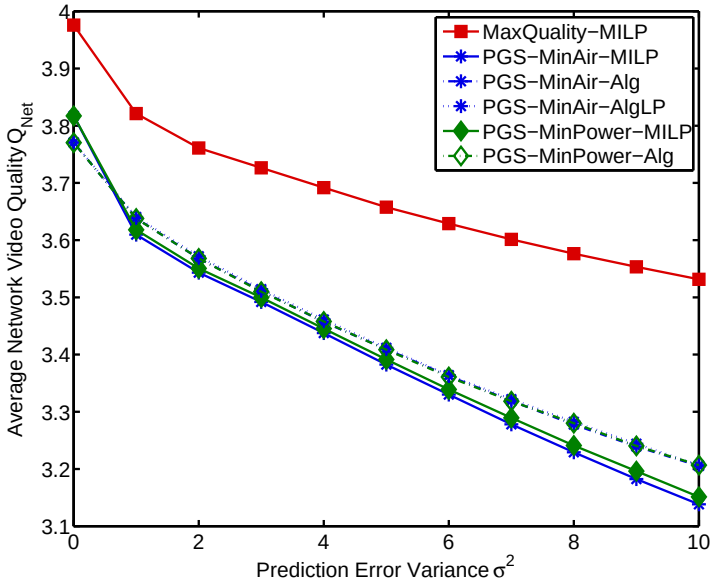
Figure 5.10(b) shows the tradeoff that the PGS framework offers for video quality versus average BS power consumption. As illustrated, the consumed power of MinPower-MILP can be reduced by over 50% as the quality is decreased. The MinAir-MILP scheme also offers significant power reduction, albeit at a lower ratio. We also note that the deviation of the multi-stage algorithm based solutions from the MILP solutions increases as the quality level increases. The reason is that higher quality leads to higher load, where it is more difficult to achieve optimality with the decoupled multi-stage algorithm approach.

Effect of Prediction Errors

To evaluate the effect of prediction errors on the PGS schemes, we add a Gaussian random variable with a mean of zero and a variance σ^2 to the predicted user SNR. This can be considered a shadowing error, and the resulting user rate matrix is denoted by $\tilde{\mathbf{r}}$. Therefore, while the PGS schemes use $\hat{\mathbf{r}}$ to minimize power and maintain user QoS constraints, the actual rates received are determined by $\mathbf{x} \odot \tilde{\mathbf{r}}$. This can degrade video quality and cause video freezing if the resulting allocation does not completely download the segments in their due time. Figure 5.11(a) illustrates the impact of such errors on the video freezing for an increasing error variance σ^2. As expected, a higher error variance increases the video stalls. However, the algorithm based PGS schemes are more robust to prediction errors, and achieve under 3% freezing for a high error variance. This indicates that even trends in the future user rates can provide significant power gains with minimal QoS losses. There are two main reasons behind the larger MILP solution sensitivity to prediction errors. First, since the PGS-MILPs provide lower total airtime, when the observed rates are less than



(a) Video freezing.

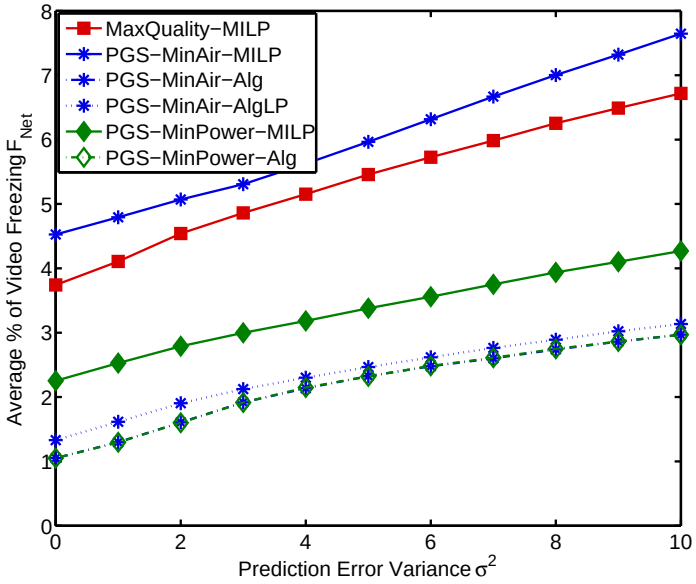


(b) Video quality.

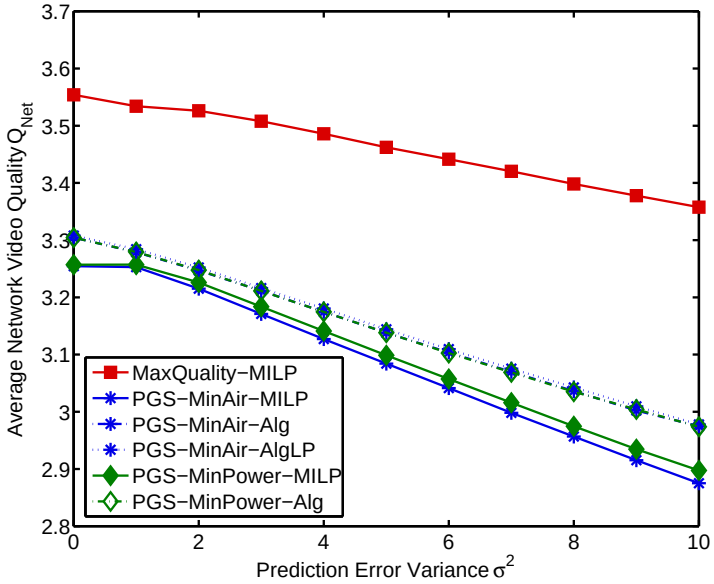
Figure 5.11: Effect of shadowing prediction errors on the PGS schemes in the highway scenario, $M = 20$.

predicted, the user will be allocated an even lower rate, resulting in more freezing. This also explains why MinAir-MILP suffers more than MinPower-MILP (which has a larger airtime, but lower power due to the sleep modes). Secondly, the optimization based approaches make discrete allocation bursts as observed in Figure 5.7(b). While being optimal, these bursts can be spaced out in time (to wait until a user reaches its next peak), and therefore when the predicted rate is less, the user has to wait until the next allocation to resume playback. In contrast, the PGS rate allocation algorithm performs allocation every N_{seg} slots, when a user does not have any buffered segments.

In order to investigate the effect of fast fading, we model the channel with i.i.d. Rayleigh-fading as well. The resultant $\tilde{\mathbf{r}}$ is now computed from an SNR that has an error component and a fast-fading component. The results are shown in Figure 5.12(a) and Figure 5.12(b) where we can see that even with an error variance of zero, the fast fading results in performance losses. Note that the relative effects of errors on the different PGS solutions follow similar trends to the previous results, where the optimal solutions are more sensitive to prediction errors. To improve the performance under effects of fast fading we suggest that a more conservative measure of $\hat{\mathbf{r}}$ can be used while solving PGS. In other words, the values of $\hat{\mathbf{r}}$ can be decreased by a small factor to reduce the error effects on freezing, when the actual rate is less than the predicted rate. Furthermore, in future work we plan to use stochastic channel models along with robust optimization techniques to improve the performance of PGS under uncertainty.



(a) Video freezing.



(b) Video quality.

Figure 5.12: Effect of shadowing prediction errors and fast fading on the PGS schemes in the highway scenario, $M = 20$.

5.7 Implementation Considerations

In the previous sections, we developed PGWA approaches that improve video streaming QoS and reduce BS power consumption. Furthermore, PGWA can also be applied to other non-real time applications as well. However, while we have seen that location predictions and application information can provide valuable wireless access energy savings, several implementation issues need to be addressed to assess the full potential and practical use of PGWA. The gains derived from PGWA are generally dependent on the knowledge of the application's requirements, the user's mobility trajectory, in addition to cooperation between BSs and users. In Section 2.1 and Section 3.3.4 we discussed how location predictions and accurate geographical radio maps are being enabled in current mobile networks. We now outline the remaining functional entities required to facilitate PGWA and illustrate their interaction in Figure 5.13.

Application Demand Information In this module, future rate requirements of users are projected and the status of running applications is classified. For example, the long-term user demand of a stored video stream is predictable based on the streaming rate and duration of the requested video. While the network can infer QoS needs based on the type of traffic, the UE can also provide additional information such as application background/foreground running status, as well as user preferences such as quality versus delay for adaptive streaming videos. In this context, user application profiling can aid the network by registering user preferences and habits to provide additional input to the demand predictor. User application needs are also exchanged between BSs to make long-term allocations and network configuration plans. During hand-over, UEs can update the target cell with the application statuses and allocation

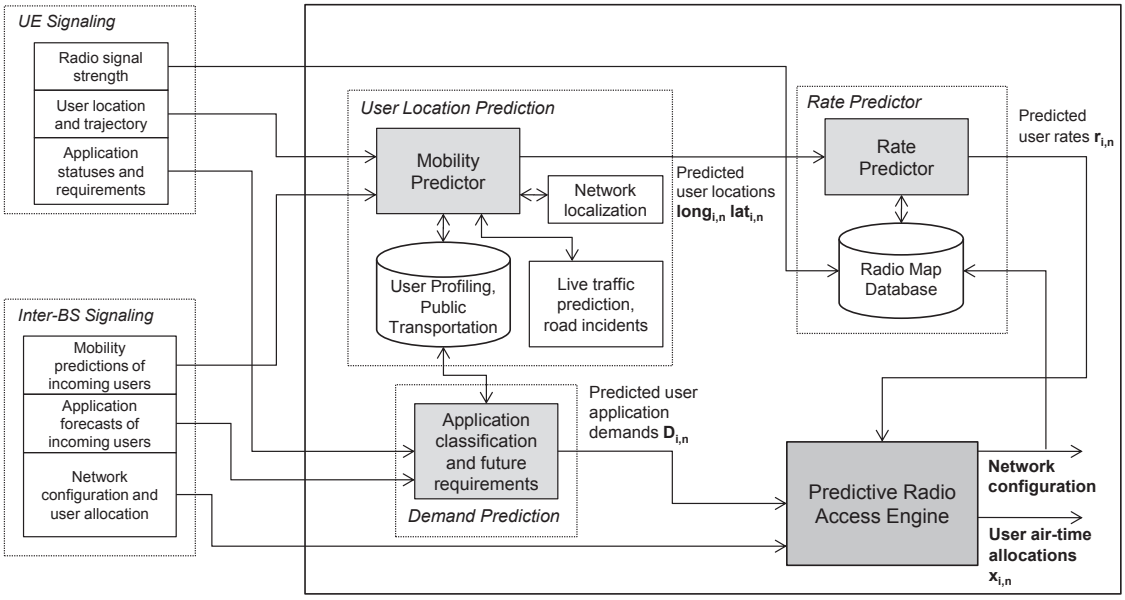


Figure 5.13: Key elements and functions of PGWA. The subscripts i and n represent users and time slots respectively.

history in previous cells, which is particularly useful for distributed operation that minimizes BSs communication.

User Signaling User signaling plays a central role in the PGWA since UEs can be used to aid the prediction processes, i.e., trajectory prediction, rate prediction and application demand forecasting. Provisions for UE involvement in energy efficiency are already being adopted in 3 GPP Release 11. For example, in DRX, the UE assists in determining the favorable connection states. This is because it has the relevant information on the applications running and remaining battery power of the device [86]. However, effective protocols are needed to efficiently communicate the multitude of UE information and context to BSs over existing interfaces.

Inter-BS Cooperation BS cooperation is required at two levels. The first is for the exchange of the UE gathered information at each site, and the second is to collaborate in making the green access allocation decisions and network configuration. The amount of communication depends on whether access decisions are made centrally or are distributed, and on the presence of any iterative procedures to converge to a final decision. Such cooperation can be facilitated in LTE over the special inter-BS X2 interface that allows some form of communication and coordination between BSs.

Predictive Radio Access Engine As depicted in Figure 5.13, the generated predictions are forwarded to the predictive access engine to devise energy efficient transmission and network configurations. Also note that the network configuration output of the green network access engine is fed back to the operator radio map to account for the changes in the network layout. The main focus of this chapter was on developing the energy saving algorithms of the access engine.

5.8 Summary

In this chapter, we discussed how user location predictions and application information can be incorporated into the radio access framework to provide energy savings. We first applied PGWA to stored video transmission during low load when videos can be delivered without streaming stalls. The minimum air-time access problem was formulated as a linear program and a low complexity distributed heuristic was then developed. We then addressed the higher network load case where video degradations are unavoidable, and presented the joint VD-BS power consumption minimization problem. The problem was formulated as a multi-objective LP that captures

the trade-off between minimizing total video degradation and minimizing network-wide BS power consumption. Then, we presented a centralized heuristic algorithm that closely follows the LP solution. A distributed extension of the algorithm was also developed to illustrate the potential gains in a more practical setting with lower signaling requirements.

The general problem of leveraging rate predictions to deliver adaptive video streams and allow BSs to enter deep sleep modes was then investigated in this chapter. The proposed Predictive Green Streaming framework jointly optimizes multi-user rate allocation, video segment quality, and BS on/off status. This was accomplished in an MILP formulation that captures the user video streaming requirements, the BS power consumption, and deep sleep mode operation. As the resulting MILP can be computationally intractable for large problem sizes, a polynomial-time algorithm that decouples the problem into multiple stages was then developed. Performance evaluation results indicated that high network-wide energy efficiency gains are achievable from PGS, and an investigation on the effects of rate predictions and channel fluctuations for the different algorithms was conducted.

Chapter 6

Conclusion and Future Directions

6.1 Summary

In this thesis, we demonstrated the pivotal role of mobility awareness in enabling proactive RAN transmission paradigms. This is primarily motivated by the plethora of location and navigation capabilities of smartphones, and the emergence of self-organizing functionalities in future networks. Instead of focusing on the immediate application requirements, mobility predictions facilitate long-term resource allocation planning and content delivery schemes. Through this research, we demonstrated how both user Quality of Service (QoS) and network operational efficiency can be significantly improved with predictive RANs.

Chapter 1 provided an overview of the research problem tackled in this thesis, and a summary of the thesis contributions. In Chapter 2, we first reviewed previous research efforts that leveraged mobility predictions to improve cellular network performance. Then, we introduced and motivated the development of a Predictive Radio Access Network (P-RAN), and provided an overview of its operational requirements. Extensive simulation results confirmed the expected gains, in both throughput and

long-term fairness guarantees.

The first part of Chapter 3 presented the system models used throughout the thesis. This included the link model, resource sharing model, and mobility and network models. The notion of predictive resource allocation (PRA) was then introduced, where multiple BSs cooperate to plan user air-time usage over a specified time horizon. Several PRA problems were formulated as optimization programs to 1) maximize throughput, 2) provide max-min fairness, and 3) enable a throughput-fairness trade-off.

Chapter 4 extended the premises laid in Chapter 3 to proactive video delivery. We applied PRA to stored video streams and demonstrated how video degradations can be minimized by strategically buffering content in the UEs. We then addressed the problem of adaptive video streaming. Here, an in-network solution that jointly optimizes both long-term resource allocations, and segment quality plans, was proposed. The results of the proposed solutions were significant: 1) video degradations were reduced by up to 50% without sacrificing fairness for constant quality videos, and 2) over 15% quality gains were obtained while eliminating 15 s of freezing for every 100 s of playback, with adaptive video streaming.

Chapter 5 investigated the use of P-RANs to plan energy-efficient BS downlink transmissions for stored video streaming. As in Chapter 4, approximate solutions and optimal problem formulations were developed to assess the potential of predictions in reducing energy consumption. Our findings indicated that even with rough estimates of future channel gains, significant energy savings are possible using polynomial-time algorithms.

6.2 Future Directions

We believe that our thesis made a positive contribution towards developing predictive radio access networks and highlighting their potential to improve QoS and reduce energy consumption. However, there are still many open challenges and implementation issues that need to be tackled in order to fully leverage the gains of predictive RANs. We now highlight some of these avenues for future work.

6.2.1 Modeling Uncertainty

In this thesis, we first assumed that the user rate predictions are accurate, and then assessed the robustness of our solutions to prediction errors. There is a need however to model prediction uncertainty itself, and thereafter develop solutions that incorporate such models. With reference to the PGWA architecture presented in Figure 5.13, there are three main dimensions of uncertainty: 1) location predictions, 2) radio map accuracy, and 3) user demand forecasts (e.g., duration of video a user will continue to view). To this effect, models that capture such variabilities are needed, where stochastic or fuzzy representations of the uncertainties can be introduced. In turn, the degree of uncertainty itself may also be tuned as time progresses depending on the observed discrepancies from the experienced values.

6.2.2 Robust Predictive Solutions

A direct consequence of incorporating models with uncertainty, is to develop solutions that can utilize the additional probabilistic/possibilistic information. For example, if channel variability is high for a particular user, the prediction window may be shortened. Additionally, a more conservative value of the rate predictions may be

selected for users anticipated to suffer from video degradations. This will allocate more airtime to such users at the cost of reduced energy efficiency. Similarly, in adaptive video streaming, the segment quality levels may be lowered when channel variability is high. With respect to application demand, limits on the amount of preallocated content (e.g., seconds of video) may be introduced depending on the degree of certainty that the user will consume this data. It would be interesting to investigate the use of methods from model predictive control, fuzzy systems theory, and stochastic optimization to develop such robust predictive solutions.

6.2.3 Distributed Approaches and Signaling

The scope of the predictive approaches presented in this thesis covers multiple cells, where allocations are jointly made for all the cooperating BSs. In Chapter 5, we saw how distributed solutions can be developed by introducing user-BS signaling during handover. However, the developed approaches deviate from the optimal benchmark solutions, indicating that there is much room for improvement in this regard. In particular, both formal optimization decomposition methods for large-scale systems and real-time algorithms deserve further attention. Herein, it is important to identify what information is signaled among BSs, and between users and BSs, and at what frequency.

6.2.4 Practical Implementation Considerations

In Section 4.3.6, we presented results from a testbed used to demonstrate the practical feasibility and benefits of Predictive Adaptive Streaming. However, several practical

considerations remain to be investigated. First, implementation over a standard compliant LTE system is needed to specifically define where and how the various modules presented in Figure 5.13 can be integrated within LTE. Additionally, AVS typically relies on the client to signal the requested quality levels to the content server [72]. Therefore, the proposed in-network PAS and PGS solutions where the BSs jointly determine both quality levels and RA will require some modifications to traditional AVS. Finally, a large scale simulation study with real road maps, vehicle trajectories, and traffic demand patterns will provide insight on more realistic performance measures and large-scale deployment issues.

6.2.5 Leveraging Predictions for End-to-End Content Delivery

While in this thesis our primary focus has been incorporating rate predictions to devise long-term RA planning, it is also paramount to investigate how predictions can be leveraged across all layers in the delivery process. For example, in Chapter 4 and 5, we demonstrated the effectiveness of joint RA and quality planning to enhance AVS. Similarly, further optimizations of cross layer delivery may lead to significant QoS and efficiency gains. Accordingly, there is a need for an end-to-end predictive delivery framework. We believe this will distribute the incurred signaling and processing overheads across multiple layers, and lead to more effective predictive delivery strategies. To this end, Appendix A provides an overview of such an architecture and outlines further research directions in this regard.

6.3 Concluding Remarks

In this work, we took a step towards investigating the utility of mobility predictions in multi-cell mobile content delivery. In addition to the directions for future work, we summarize some of the key recommendations and lessons learned from this research:

1. Developing approximate polynomial-time P-RAN algorithms is very important for practical solutions. The large time horizon over which optimizations are made in P-RANs makes solving the MILPs (or LPs) in a reasonable time problematic, even with powerful solvers and parallel processing. Furthermore, just constructing the constraint matrices mandates significant memory requirements that may not be available in a BS.

2. When designing approximate algorithms, considerations should be made to account for prediction errors. Making discrete allocation bursts followed by long durations of starvation should be avoided, even if the resulting gains/savings are very high. This is because if the predicted rate is less than anticipated, the video stream will remain stalled until the next planned allocation. In addition, solutions that are re-run periodically based on real-time feedback of the channel and user buffer state are needed.

3. Standardization developments in DASH should be followed closely. There are current efforts towards enabling network and operator involvement in the segment adaptation process (as opposed to the current client-based model). This is being carried out under the MPEG Server and Network assisted DASH Operation (SAND) core experiment [95]. For practical implementations, the functionality of predictive mechanisms, including signaling and intelligence, should

be integrated into the future solutions under SAND.

4. Cross-layer predictive optimizations can lead to high gains. However, it is important to identify the associated signaling requirements and processing overhead. This is particularly important if the cross layer functions are performed across multiple network entities/components.

Bibliography

[1] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018." http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html, 2014. Accessed Apr. 29th, 2014.

[2] N. A. Ali, A.-E. Taha, and H. S. Hassanein, "Quality of service in 3GPP R12 LTE-Advanced," *IEEE Commun. Magazine*, vol. 51, no. 8, pp. 103–109.

[3] iGR, "U.S. regional and small operator network infrastructure Capex and Opex forecast, 2012-2017." https://igr-inc.com/, 2013. Accessed Feb. 20th, 2014.

[4] L. Correia, D. Zeller, O. Blume, D. Ferling, A. Kangas, I. Godor, G. Auer, and L. Van der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Magazine*, vol. 48, no. 11, pp. 66–72, 2010.

[5] Nokia Developer, "Total location: how location services are creating new businesses and user experiences." http://developer.nokia.com/images/uploads/pdfs/insights-03-total-location-report.pdf, 2013. Accessed Feb. 20th, 2014.

[6] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 2008.

[7] C. Song, Z. Qu, N. Blumm, and A. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, 2010.

[8] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, no. 2923, pp. 1–9, 2013.

[9] Accenture, "Perspectives on in-vehicle infotainment systems and telematics, 2011." http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Perspectives-on-In-Vehicle-Infotainment-Systems-and-Telematics.pdf, 2011. Accessed Feb. 20th, 2014.

[10] J. Johansson, W. Hapsari, S. Kelley, and G. Bodog, "Minimization of drive tests in 3GPP," *IEEE Commun. Magazine*, vol. 50, no. 11, pp. 36–43, 2012.

[11] OpenSignal, "The OpenSignal project homepage." http://opensignal.com/, 2013. Accessed Feb. 15th, 2013.

[12] 3GPP, "Telecommunication management; self-organizing networks SON policy network resource model NRM integration reference point IRP; requirements," Technical Specification TS 32.521 v11.1.0, Dec. 2012.

[13] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[14] G. Yavaş, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos, "A data mining approach for location prediction in mobile environments," *Data & Knowledge Engineering*, vol. 54, no. 2, pp. 121–146, 2005.

[15] H. Abu-Ghazaleh and A. S. Alfa, "Application of mobility prediction in wireless networks using markov renewal theory," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 788–802, 2010.

[16] J. Froehlich and J. Krumm, "Route prediction from trip observations," *Society of Automotive Engineers (SAE) World Congress*, pp. 53–58, 2008.

[17] K. Laasonen, "Route prediction from cellular data," in *Proc. Workshop on Context Awareness for Proactive Systems*, pp. 147–158, 2005.

[18] M. A. Bayir, M. Demirbas, and N. Eagle, "Mobility profiler: A framework for discovering mobility profiles of cell phone users," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 435–454, 2010.

[19] Gartner, "Gartner highlights top consumer mobile applications and services for digital marketing leaders." http://www.gartner.com/newsroom/id/2194115, 2012. Accessed Apr. 29th, 2014.

[20] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proc. ACM Int. Conf. on Mobile Comput. and Netw. (MOBICOM)*, pp. 123–134, 2007.

[21] B. Liang and Z. J. Haas, "Predictive distance-based mobility management for multidimensional PCS networks," *IEEE/ACM Trans. on Netw.*, vol. 11, no. 5, pp. 718–732, 2003.

[22] J. Taheri and A. Y. Zomaya, "Clustering techniques for dynamic location management in mobile computing," *Journal of Parallel and Distrib. Comput.*, vol. 67, no. 4, pp. 430–447, 2007.

[23] W.-S. Soh and H. S. Kim, "QoS provisioning in cellular networks based on mobility prediction techniques," *IEEE Commun. Magazine*, vol. 41, no. 1, pp. 86–92, 2003.

[24] M.-H. Chiu and M. A. Bassiouni, "Predictive schemes for handoff prioritization in cellular networks based on mobile positioning," *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, pp. 510–522, 2000.

[25] I. Chlamtac, T. Liu, and J. Carruthers, "Location management for efficient bandwidth allocation and call admission control," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, pp. 1023–1027, 1999.

[26] F. Yu and V. Leung, "Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks," *Computer Networks*, vol. 38, no. 5, pp. 577–589, 2002.

[27] S. Choi and K. G. Shin, "Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks," *IEEE Trans. on Parallel and Distrib. Syst.*, vol. 13, no. 9, pp. 882–897, 2002.

[28] R. Bolla and M. Repetto, "A new model for network traffic forecast based on user's mobility in cellular networks with highway stretches," *Int. Journal of Commun. Syst.*, vol. 17, no. 10, pp. 911–934, 2004.

[29] Sandvine, "Global internet phenomena report, 1h 2013." https://www.sandvine.com/downloads/general/global-internet-phenomena/2013/sandvine-global-internet-phenomena-report-1h-2013.pdf, 2013. Accessed Feb. 20th, 2014.

[30] A. Galindo-Serrano, B. Sayrac, S. B. Jemaa, J. Riihijärvi, and P. Mähönen, "Cellular coverage optimization: A radio environment map for minimization of drive tests," in *Cognitive Communication and Cooperative HetNet Coexistence*, pp. 211–236, 2014.

[31] C. Phillips, D. Sicker, and D. Grunwald, "A survey of wireless path loss prediction and coverage mapping methods," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 255–270, 2013.

[32] M. Malmirchegini and Y. Mostofi, "On the spatial predictability of communication channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 964–978, 2012.

[33] J. Yao, S. S. Kanhere, and M. Hassan, "An empirical study of bandwidth predictability in mobile computing," in *Proc. ACM Int. Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization (WiNTECH)*, pp. 11–18, 2008.

[34] D. Han, J. Han, Y. Im, M. Kwak, T. T. Kwon, and Y. Choi, "MASERATI: Mobile adaptive streaming based on environmental and contextual information," in *Proc. ACM Int. Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization (WiNTECH)*, pp. 33–40, 2013.

[35] H. Abou-zeid, H. Hassanein, and S. Valentin, "Optimal predictive resource allocation: Exploiting mobility patterns and radio maps," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 4714–4719, 2013.

[36] H. Abou-zeid, H. S. Hassanein, and N. Zorba, "Long-term fairness in multicell networks using rate predictions," in *Proc. IEEE GCC Conf. and Exhibition (GCC)*, pp. 131–135, 2013.

[37] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, pp. 882–890, May 2011.

[38] H. Abou-zeid, H. S. Hassanein, S. Valentin, and M. Feteiha, "Lookback scheduling for long-term quality-of-service over multiple cells," in *Proc. IEEE Int. Wireless Commun. and Mobile Comput. Conf.*, pp. 515–520.

[39] G. Song, Y. Li, G. Song, and Y. Li, "Utility-based resource allocation and scheduling in ofdm-based wireless broadband networks," *IEEE Communications Magazine*, vol. 43, pp. 127 – 134, Dec. 2005.

[40] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 1, pp. 66–71, 2009.

[41] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, 2000.

[42] "Multicell cooperation," *IEEE Wireless Commun.*, vol. 20, Feb. 2013.

[43] H. Abou-zeid, H. S. Hassanein, S. Valentin, and M. Feteiha, "A lookback scheduling framework for long-term quality-of-service over multiple cells," *Wireless Commun. and Mobile Comp.*, 2014, to appear.

[44] H. J. Bang, T. Ekman, and D. Gesbert, "Channel predictive proportional fair scheduling," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 482–487, 2008.

[45] J. Hajipour and V. C. Leung, "Proportional fair scheduling in multi-carrier networks using channel predictions," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, pp. 1–5, 2010.

[46] S. H. Ali, V. Krishnamurthy, and V. C. M. Leung, "Optimal and approximate mobility-assisted opportunistic scheduling in cellular networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 6, pp. 633–648, 2007.

[47] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. K. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, 2014, to appear.

[48] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "SUMO-Simulation of Urban MObility: An Overview," in *Proc. Third Int. Conf. on Advances in System Simulation (SIMUL)*, pp. 63–68, 2011.

[49] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Proc. ACM Int. Conf. on Mobile Comput. and Netw. (MOBICOM)*, pp. 85–97, 1998.

[50] 3GPP, "LTE;E-UTRA; radio frequency system scenarios," Technical Report TR 36.942 V11.0.0, 3GPP, Sept. 2012.

[51] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, 1997.

[52] Gurobi, "Gurobi Optimization." http://www.gurobi.com/. Accessed Feb. 11th, 2014.

[53] MOSEK ApS, "The MOSEK Optimization Software." http://www.mosek.com/. Accessed Feb. 11th, 2014.

[54] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta." http://cvxr.com/cvx, Sept. 2013.

[55] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems." Tech. Rep. TR-301, DEC, Sept. 1984.

[56] H. Abou-zeid, S. Valentin, and H. S. Hassanein, "Apparatus, methods, and computer programs for a mobile transceiver and for a base station transceiver." European patent application 20110306929 filed by Alcatel-Lucent, Oct. 2011.

[57] H. Abou-zeid, H. S. Hassanein, and N. Zorba, "Enhancing mobile video streaming by lookahead rate allocation in wireless networks," in *Proc. IEEE Consumer Commun. and Netw. Conf. (CCNC)*, pp. 768–773, 2014.

[58] T.-Y. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari, "Confused, timid, and unstable: picking a video streaming rate is hard," in *Proc. ACM Conf. on Internet Measurement*, pp. 225–238, 2012.

[59] R. Pantos, W. May, and Apple Inc., "HTTP Live Streaming." http://tools.ietf.org/html/draft-pantos-http-live-streaming-11, April 2013. Accessed Jul. 27th, 2013.

[60] 3GPP, "Transparent end-to-end packet-switched streaming service PSS; progressive download and dynamic adaptive streaming over http 3GP-DASH," Technical Specification TS 26.247 V11.2.0, 3GPP, Mar. 2013.

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

[62] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with festive," in *Proc. ACM Int. Conf. Emerging Netw. Experiments and Technologies*, pp. 97–108, 2012.

[63] D. De Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, G. Li, and R. Miller, "Optimization of http adaptive streaming over mobile cellular networks," in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, pp. 898–997, 2013.

[64] Z. Lu and G. de Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, pp. 2806–2814, 2013.

[65] P. Kolios, V. Friderikos, and K. Papadaki, "Energy-aware mobile video transmission utilizing mobility," *IEEE Network*, vol. 27, no. 2, pp. 34–40, 2013.

[66] J. Yao, S. Kanhere, and M. Hassan, "Improving QoS in high-speed mobility using bandwidth maps," *IEEE Trans. Mobile Comput.*, vol. 11, no. 4, pp. 603–617, 2012.

[67] I. D. Curcio, V. K. M. Vadakital, and M. M. Hannuksela, "Geo-predictive real-time media delivery in mobile environment," in *Proc. ACM Workshop on Mobile Video (MoVid)*, MoViD '10, pp. 3–8, 2010.

[68] J. Fardous and S. S. Kanhere, "On the use of location window in geo-intelligent HTTP adaptive video streaming," in *Proc. IEEE Int. Conf. Netw. (ICON)*, pp. 46–51, 2012.

[69] H. Riiser, T. Endestad, P. Vigmostad, C. Griwodz, and P. Halvorsen, "Video streaming using a location-based bandwidth-lookup service for bitrate planning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 3, pp. 24:1–24:19, 2012.

[70] V. Singh, J. Ott, and I. D. Curcio, "Predictive buffering for streaming video in 3G networks," in *Proc. IEEE World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–10, 2012.

[71] ISO/IEC, "Dynamic adaptive streaming over HTTP (DASH) – part 1: Media presentation description and segment formats," International Standard TS 26.247 V11.2.0, Apr. 2012.

[72] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Commun. Magazine*, vol. 50, no. 4, pp. 20–27, 2012.

[73] M. Bazaraa, J. Jarvis, and H. Sherali, *Linear Programming and Network Flows (3rd Edition)*. John Wiley and Sons, 2005.

[74] IEEE, "Standard for information technology – telecommunications and information exchange between systems – LAN/MAN specific requirements – part 11: Wireless LAN MAC and PHY layer specifications – amendment 4: Further higher-speed physical layer extension in the 2.4 GHz band," *IEEE Std 802.11g-2003*, June 2003.

[75] Google Inc., "Android 4.1 (API level 16) Developer Site." http://developer.android.com/about/versions/jelly-bean.html, July 2012. Accessed Jul. 27th, 2013.

[76] VideoLAN, "VideoLAN – Official page for VLC media player." http://www.videolan.org/vlc/index.html, July 2013. Accessed Jul. 27th, 2013.

[77] GNU Project, "GLPK (GNU Linear Programming Kit)." http://www.gnu.org/software/glpk/, June 2012. Accessed Jul. 27th, 2013.

[78] Blender Foundation, "Tears of Steel, Mango Open Movie Project." www.tearsofsteel.org, Sept. 2012. Accessed Jul. 27th, 2013.

[79] H. Abou-zeid and H. S. Hassanein, "Efficient lookahead resource allocation for stored video delivery in multi-cell networks," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2014, to appear.

[80] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, 2014, to appear.

[81] H. Abou-zeid and H. S. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 92–99, 2013.

[82] Z. Hasan, H. Boostanimehr, and V. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524–540, 2011.

[83] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 30–38, 2011.

[84] T. Han and N. Ansari, "On greening cellular networks via multicell cooperation," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 82–89, 2013.

[85] S. Deng and H. Balakrishnan, "Traffic-aware techniques to reduce 3G/LTE wireless energy consumption," in *Proc. 8th Int. Conf. on Emerging Netw. Experiments and Technologies (CoNEXT)*, pp. 181–192, 2012.

[86] M. Gupta, S. Jha, A. Koc, and R. Vannithamby, "Energy impact of emerging mobile internet applications on LTE networks: issues and solutions," *IEEE Commun. Magazine*, vol. 51, no. 2, pp. 90–97, 2013.

[87] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, and H. Holktamp, "Flexible power modeling of lte base stations," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, pp. 2858–2862, Apr. 2012.

[88] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, 2013.

[89] I. Ashraf, F. Boccardi, and L. Ho, "Sleep mode techniques for small cell deployments," *IEEE Commun. Magazine*, vol. 49, no. 8, pp. 72–79, 2011.

[90] M. Ismail and W. Zhuang, "Network cooperation for energy saving in green radio communications," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 76–81, 2011.

[91] B. Soret, H. Wang, K. Pedersen, and C. Rosa, "Multicell cooperation for LTE-Advanced heterogeneous network scenarios," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 27–34, 2013.

[92] F. Han, Z. Safar, W. Lin, Y. Chen, and K. Liu, "Energy-efficient cellular network operation via base station cooperation," in *Proc. IEEE Int. Conf. on Commum. (ICC)*, pp. 4374–4378, 2012.

[93] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Magazine*, vol. 48, no. 11, pp. 74–79, 2010.

[94] EARTH, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," Technical Report INFSO-ICT-247733, Deliverable D2.3, EARTH, Jan. 2012.

[95] ISO/IEC JTC1/SC29/WG11, "Coding of moving pictures and audio, 2013." https://datatracker.ietf.org/documents/LIAISON/liaison-2013-11-04-isoiec-jtc-1sc-29wg-11-cdni-liaison-template-on-mpeg-dash-attachment-3.pdf, 2013. Accessed Feb. 20th, 2014.

[96] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *Proc. ACM Int. Conf. on Mobile Systems, Applications, and Services*, pp. 319–332, 2013.

[97] H. Ahlehagh and S. Dey, "Video caching in radio access network: impact on delay and capacity," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, pp. 2276–2281, 2012.

[98] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5g systems," *IEEE Commun. Magazine*, vol. 52, no. 2, pp. 131–139, 2014.

[99] N. Gautam, H. Petander, and J. Noel, "A comparison of the cost and energy efficiency of prefetching and streaming of mobile video," in *Proc. ACM Workshop on Mobile Video (MoVid)*, pp. 7–12, 2013.

[100] M. Grafl, C. Timmerer, H. Hellwagner, W. Cherif, and A. Ksentini, "Evaluation of hybrid scalable video coding for HTTP-based adaptive media streaming with high-definition content," in *Proc. IEEE World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–7, 2013.

[101] IEEE, "Standard protocol for stream management in media client devices," *IEEE P2200*, June 2012.

[102] X. Wang, T. T. Kwon, Y. Choi, H. Wang, and J. Liu, "Cloud-assisted adaptive video streaming and social-aware video prefetching for mobile users," *IEEE Wireless Commun.*, vol. 20, no. 3, 2013.

Appendix A

Towards Predictive End-to-End Content Delivery

Content delivery over 4G technologies such as LTE typically follows the architecture presented in Figure A.1. For instance, when a video is accessed by a mobile device, it is first requested from content servers. The video stream then traverses the wireless carrier's CN and RAN before reaching the mobile user. Congestion at any point throughout the network results in video quality degradations, thereby reducing the perceived QoE. In addition to efficient RAN schemes, mechanisms for *in-network content caching* [96],[97],[98] and *content prefetching* [99] are also being developed. The objective is to strategically store select content closer to the clients and, if deemed beneficial, in the UE local storage for immediate availability when accessed. The effectiveness of such approaches can also benefit significantly from location-awareness and mobility predictions. While in this thesis our focus has been on predictive RAN functionalities, it is also important to investigate how predictions can be leveraged across all layers in the delivery process. Therefore, there is a need for an end-to-end cross-layer predictive delivery framework. We believe this will distribute the incurred signaling and processing overhead across multiple layers, and lead to more effective predictive delivery strategies. To this end, the following discussion highlights possible

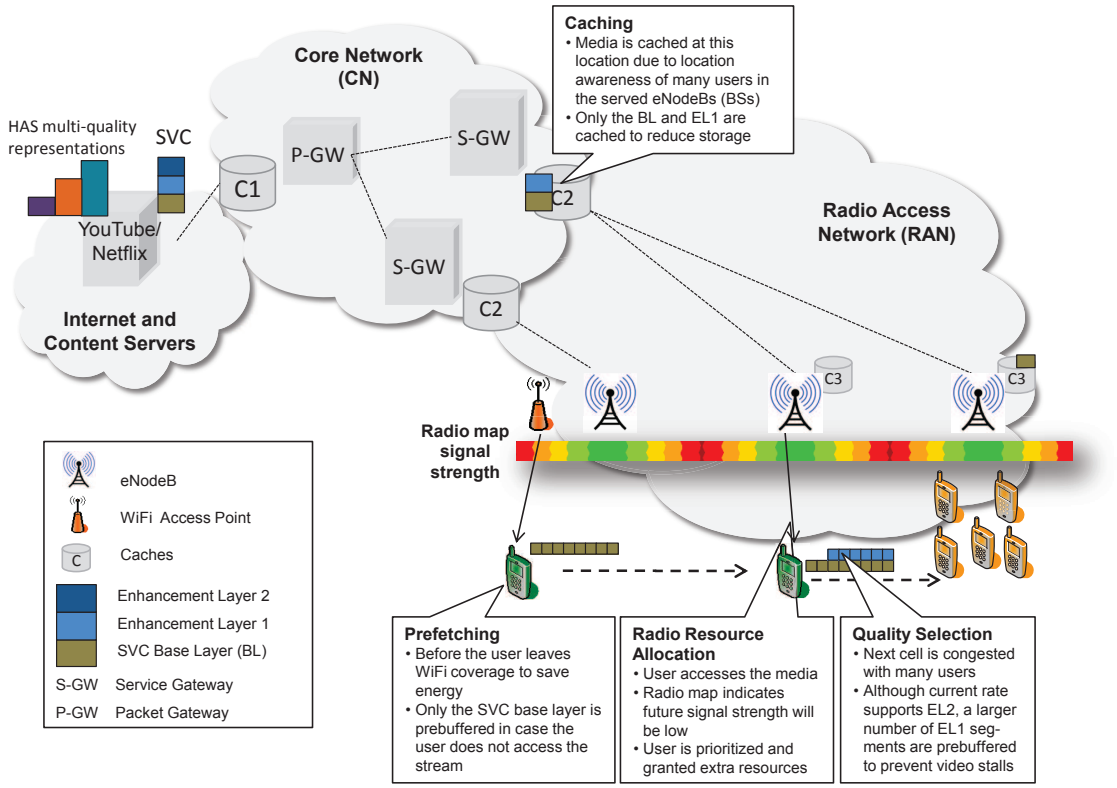


Figure A.1: Potential of location awareness and mobility predictions for end-to-end content delivery.

research directions for predictive in-network caching and content pushing. Issues of cross-layer interaction are also touched upon with particular emphasis on adaptive video streaming delivery.

A.1 Preliminaries: Scalable Video Coding

Scalable Video Coding (SVC) offers an alternative multi-quality video representation format to that of AVS described in Section 4.1.2. In SVC, a layered approach is used where a Base Layer (BL) is used for the lowest video quality representation, and multiple Enhancement Layers (ELs) are defined above it. Each EL contains

only the incremental data required to refine the video quality. The scalable layered approach permits the storage of one media file containing all the layers. This contrasts with having independent quality versions of the same media content as shown in Figure A.1. The disadvantage however, is that SVC incurs additional processing overhead to combine the layers, introducing delays.

By utilizing SVC, a video can be decoded and viewed at the lowest quality if only the BL is delivered, while additional ELs improve the quality. SVC is resource efficient as a single representation containing the BL and ELs can be delivered from the server to the BS. Then depending on the link conditions, the appropriate number of layers is transmitted to each user. Several proposals have also been made to deploy SVC over DASH [100]. Delivering multi-quality videos over a scalable format such as SVC offers unique opportunities to predictive content delivery approaches as we highlight below.

A.2 Predictive In-Network Caching

Studies of Telecom provider backhaul traffic reveal that many users request the same popular content as demonstrated in the analysis of over 370 TBs of the 3G traffic in [96]. Content caching enables the temporary storage of popular content at the edge of the network, closer to the clients [97]. This reduces the need to re-deliver content from the original server and thereby decreases the server and CN load, as well as the perceived delivery delay.

The efficiency of caching is generally determined by the cache hit ratio and the corresponding cache size. The objective is to maximize the video streaming experience while minimizing the additional costs incurred by video cache servers. Caching

was typically performed closer to the CN, but recently propositions for micro-caches located closer to the BSs have been made, as illustrated in Figure A.1. The argument for such architectures is the increased popularity of certain media content fueled by sharing on Social Networks (SNs) and content provider viewing recommendations. During network design, the problem involves optimizing the cache sizes and their geographical location. Throughout network operation, the primary challenge is to determine what content to cache (and for how long), and where to cache it in the network.

Location Awareness

Knowing the locations of users and their mobility patterns helps predict the network regions where content will most likely be requested. Figure A.1 provides a use case of this where the appropriate CN cache is selected to store the media content. At a finer level, if the exact mobility trajectory of a user is known, it may be possible to determine the media segments to cache for that user in each BS depending on the predicted time period that will be spent at each cell. Mobility trajectories of users commuting on public transportation during rush hour can also be used to optimize the caching of popular evening shows and news reports at the appropriate locations.

It is worth noting that coupling location-aware caching with SVC based delivery improves efficiency in many aspects. First, it is more storage efficient compared to having multiple quality representations as previously discussed. Secondly, hit ratios will be higher as it is very likely that at least the BL representation will be requested. And finally, it enables hierarchical media caching, where only the BL is cached at the BS, with additional ELs in the larger CN caches, as illustrated in Figure A.1. This

provides immediate access to the media stream (at a low quality), while enabling a gradual quality enhancement with acceptable delay.

Impact on Energy Consumption

By reducing the redundant data transfers between the content servers and the RAN, the effective core bandwidth of the network increases. Evidently, caching also reduces the delivery path and therefore the network related transmission energy. At the end user, efficient reception (in less time) results in lower transceiver energy, and less rebuffering delays also reduce power consumption. On the other hand, the additional power consumption resulting from expanding the architectures of video caching will need to be evaluated.

A.3 Predictive Prefetching/Content Pushing

In prefetching or content pushing, a part or the whole content is loaded into the local storage of the UE before being accessed by the user. This can be viewed as an extreme form of caching that provides seamless video streaming since the media is already preloaded. Prefetching can also distribute network load spatially and temporally by preloading anticipated user requests at the most opportune times. For instance, popular media shows accessed regularly by a user can be strategically prefetched when the user is at a geographical location where network resources are abundant. By doing so, peak congestion can be alleviated. An IEEE standard (P2200-2012) [101] has been developed to enable such mechanisms to queue content for later delivery and intelligently route and replicate content over heterogeneous networks to mobile devices with local storage.

Prefetching is most effective if the content of interest is selected by the user in advance through subscriptions to playlists and other media services. However, the primary challenge is how to proactively determine the content users are anticipated to consume. As opposed to in-network caching which is based on aggregate access probability at a geographical location, prefetching requires insights on user-specific consumption preferences. This can be achieved with user behavior modeling techniques and data mining of media consumption logs to identify usage patterns for news, specific shows, etc. As in caching, information from SNs can be utilized in prefetching strategies where only the most popular content shared by the user's closest social circles are prefetching candidates [102]. Note that when applied with SVC, only the BL may be prefetched in cases where the user access of the content is uncertain. This is illustrated in Figure A.1.

Location Awareness

Being aware of a user's regular mobility patterns facilitates efficient content pushing by proactively delivering content before a user leaves Femtocell or other small cell zones where resources are generally more abundant. Similarly, predictions of upcoming high bandwidth or low congestion networks can delay content prefetching, and thereby optimize energy consumption. With location awareness the spatio-temporal network load can also be distributed more evenly. Furthermore, coupling the users' content consumption profiles discussed earlier with their geographical locations may improve the accuracy of determining what to prefetch where.

Impact on Energy Consumption

Prefetching media streams before consumption allows the content provider to choose less congested delivery paths (e.g., through WiFi) as well as off-peak hours to transmit the content to the users. When content is transmitted in less time and/or through more energy efficient wireless interfaces, this translates to energy savings at the UE [99] and the network. Furthermore, by distributing the traffic more evenly throughout the day, congestion from peak demands is reduced. On the other hand, prefetching incurs energy waste if the user does not access the pushed content.

In Table A.1, we identify and summarize the previous discussion on the vital role of location awareness as an enabler of end-to-end media delivery. It is worth noting that to fully capitalize on location awareness, it is necessary to consider the interaction between the different layers of delivery and design cross-layer delivery strategies.

Table A.1: Summary of Location-Awareness Potential in End-to-End Media Delivery.

Video Delivery Procedure	**Implementation Challenges**	**Directions for location-aware solutions**	**Impact on Energy Consumption**
Adaptive video quality	Estimating end-to-end network conditions is difficult. Mobility introduces more challenges from rapid fluctuations in wireless link capacities and spatial load densities.	Making long-term segment quality plans based on: 1) future signal strength variations, and 2) spatial traffic distributions.	Smooth streaming reduces re-buffering delays and device energy consumption. Lowering the target quality allows energy efficient delivery when energy is scarce.
Radio resource allocation	Mobility and uneven traffic distribution cause sudden changes in the available data rates to users. Therefore, sustaining long-term QoE and fairness among users is difficult.	Rate predictions from mobility trajectories allow the BS to make efficient RA plans over multiple cells. Opportunistic long-term RA plans are made at peak channel conditions.	Designing efficient RA schemes can reduce energy by sending content in less time. RA strategies that prebuffer media to users allow the BS to enter sleep modes and save energy without impacting QoE.
In-network caching	Predicting content popularity. Optimizing the size and location of caches in the network architecture to maximize the hit ratio and reduce required storage.	Current user locations and mobility trajectories can give insight on where media is most likely to be accessed.	Reduces transport related energy consumption, but requires additional energy to power the caches.
Content prefetching/pushing	Determining the content to prefetch requires accurate user behavior modeling and data mining. Users may choose not to view the prefetched content, or view part of it.	Knowing user locations can improve prefetching accuracy by adding context. Mobility predictions allow energy efficient scheduling of content pushing.	Prefetching over WiFi and off-peak cellular network hours enables energy savings, but can incur energy waste if the content is not consumed.