ON THE OPTIMAL FORMULATION OF
RESOURCE MANAGEMENT IN WIRELESS
NETWORKS

by

JIHUI ZHANG

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
in Computer Science

August 2005, Hong Kong

Copyright © by Jihui Zhang 2005
Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

JHUI ZHANG
ON THE OPTIMAL FORMULATION OF RESOURCE MANAGEMENT IN WIRELESS NETWORKS

by

JIHUI ZHANG

This is to certify that I have examined the above Ph.D. thesis and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the thesis examination committee have been made.

[Signature]
DR. BO LI, THESIS SUPERVISOR

[Signature]
PROF. LIONEL NI, HEAD OF DEPARTMENT

Department of Computer Science
16 August 2005
DEDICATION

To my Mom and Dad
ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my supervisor Prof. Bo Li. I thank him for leading me to this exciting research area and giving me guidance throughout my PhD studies. I am grateful for his encouragement, support, patience and inspiration all these years.

I also thank Prof. Lionel Ni, Prof. Yunhao Liu, Prof. Brahim Bensaou, Prof. Bing Zeng and Prof. Yi-Bing Lin for serving as my PhD thesis committee members and giving me valuable suggestions for improving my thesis. I also want to thank Prof. Jogesh K. Muppala, Prof. Lionel Ni, and Prof. Jelena Misic for serving as the committee members for my qualifying examinations.

Furthermore, please allow me to thank Dr. Wenwu Zhu, Dr. Qian Zhang and other employees in the Wireless and Networking Group of Microsoft Research Asia. They have provided me with valuable instruction and comments. Special thanks to Prof. Xiaodong Wang, Dr. Haitao Wu and Dr. Yunnan Wu for their kind collaboration and help in my research. I’d like to thank my teammates: Xinyan Zhang, Yao Zhao, Qi Wu, Xiwu Luo, Yang Yang, Min Li for their precious discussions. Further thanks are also to other friends I met in MSRA: Xiaoyan Sun, You Zhou, Hao Wu, Mingzhen Bao, Wings, Xi Wang, Min Liu, Wenle Wang and the rest. It has been my great pleasure to make the acquaintance of these very brilliant and gifted people.

I appreciate our department for the support I have received in various aspects relating to my PhD study. I would also like to take this opportunity to thank other professors at the HKUST for their kind help. Thanks to Ms. Connie Lau, Mr. Isaac Ma and other staff members of the department for their excellent administrative work. Special thanks to Ms. Shauna Dalton for helping me improve my writing skills throughout my PhD study.

I truly appreciate the kind help from my friends during my stay in HKUST. In particular, Yang Qin, Jiangchuan Liu, Xiaowen Chu, Ping Wu, Susu Xie, Fan
Zhang, Meixia Tao, Jun Zhang Bobby, Yufei Tao, Jinyuan Jia, Jicheng Zhao, Meng Zhang, Xiaolin Chang, Manli Zhu, Jie Yin, Jun Miao, Rui Gan, Hong Chen, Hong Chang, Qingmao Shen, Yan Zhang, An Chen, Wei Feng, Min Jiang, Gang Zeng, Yinchen Wei, Ruiduo Yang, Jimeng Sun, Jun Zhang, Jian Ma, Qiuyan Xia, Min Gao, Jinsong Han, and others too numerous to name.
# TABLE OF CONTENTS

Title Page ........................................ i
Authorization Page .............................. ii
Signature Page ................................. iii
Dedication ........................................ iv
Acknowledgments ............................... v
Table of Contents ............................... vii
List of Figures ................................. xii
List of Tables ................................. xiv
Abstract ........................................ xv

**Chapter 1  Introduction** ........................ 1
  1.1 Our Contribution .......................... 5
  1.2 Organization of the Thesis .......... 7

**Chapter 2  Resource Management in DS-CDMA Cellular Networks** ........................ 8
  2.1 Capacity in DS-CDMA ...................... 8
  2.2 Challenges in Resource Management ... 11
   2.2.1 Universal Frequency Reuse .......... 11
   2.2.2 Power Control ....................... 12
   2.2.3 Soft Handoff ......................... 12
   2.2.4 Voice Activity ....................... 13
   2.2.5 Propagation Model ................... 13
Chapter 3  Downlink Traffic Control in MC-CDMA Cellular Systems

3.1 System Model and Formulation
   3.1.1 System Model
   3.1.2 SIR Formulation

3.2 Connection Level Call Admission Control
   3.2.1 Downlink Capacity Constraint
   3.2.2 Handoff Based (HB) Reservation Scheme
   3.2.3 Connection Level Admission Control

3.3 Background Data Transmission

3.4 Performance Analysis
   3.4.1 Connection Level Services
   3.4.2 Background Data Traffic

3.5 Numerical Results

3.6 Summary

Chapter 4  Adaptive Cell Sectoring in DS-CDMA Systems

4.1 System Model and Formulation
   4.1.1 System Model

4.2 Optimal Sectoring Problem

4.3 Optimal Solution

4.4 The Cluster-Based Sectoring Algorithm
4.4.1 Cluster-based Sectoring 53
4.4.2 Angular-based Partitioning 55
4.4.3 Complexity Analysis 56
4.5 High Rate Data Services 58
4.6 Discussion 60
   4.6.1 Imperfect Sectoring 61
   4.6.2 Physical Constraints 61
   4.6.3 Imperfect Power Control and Shadowing 61
   4.6.4 Multicell Environment 62
4.7 Numerical Results 62
4.8 Summary 66

Chapter 5 Survey on Wireless Ad-hoc Networks 68
   5.1 Throughput Capacity in Wireless Ad-hoc Networks 68
      5.1.1 Capacity in Fixed Ad-hoc Networks 69
      5.1.2 Node Mobility 72
      5.1.3 Hybrid network 74
      5.1.4 High bandwidth 75
      5.1.5 Directional antennas 75
      5.1.6 Network Coding 76
      5.1.7 Multi-radio Multi-channel 76
   5.2 Energy Conservation in Wireless Ad-hoc Networks 78

Chapter 6 Mobility Assisted Routing in MANET 81
   6.1 System Model 83
      6.1.1 Network Model 83
      6.1.2 Mobility Model 85
      6.1.3 Communication Model 85
      6.1.4 Route Definition 87
   6.2 Routing in The Interference-Free Regime 87
      6.2.1 Optimal Routing without Power Control 87
      6.2.2 Optimal Routing with Power Control 90
Appendix D  Steady State Probability for Data Queue  168
Appendix E  Correctness Proof of Recurrence Relation  169
Appendix F  Cluster Integration Proof  171
Appendix G  Proof of Correctness of Fastest Routing Algorithm  172
Appendix H  Proof of Min Energy Routing Lemma in Interference-Free Regime  173
Appendix I  Proof of Min Energy Routing Lemma in Interference-Limited Regime  175
# LIST OF FIGURES

2.1 SIR model for DS-CDMA cellular system.  
2.2 User distribution and cell sectoring  
2.3 Number-based admission control  
2.4 MS admission  
3.1 Call blocking / dropping probability according to new call arrival rate \( \lambda_n = \lambda_{n,1} + \lambda_{n,2} \)  
3.2 Call blocking / dropping probability with the changing of reservation factor for \( \delta_1, (\delta_2=0.1) \)  
3.3 New call blocking (handoff dropping) according to arrival fraction of class-2 calls (\( \lambda_n=0.1 \))  
3.4 Background data packets loss probability according to the data packets arrival rate.  
3.5 Packets dropping probability according to the data packets arrival rate.  
3.6 Average data packet transmission rate (kbps) according to the packet arrival rate  
3.7 Utilization (kbps) according to the new call arrival rate  
4.1 36 users, 6 sectors. The lines represent the sectoring boundaries. The ellipse shows one of the clusters.  
4.2 180-user, 15-sector, \( \delta=1 \). (a). User distribution. The ellipse roughly includes one high-density cluster after coarse clustering, which is further partitioned into 3 sub-clusters with users between boundaries. (b) Sectoring result. The lines represent the final boundaries.  
4.3 Minimum total transmission power for a different number of sectors  
6.1 Transmission Map  
6.2 Network Setting  
6.3 The Fastest Route from Node 3 to Node 14  
6.4 The Min-Hop Route from Node 3 to Node 14  
6.5 The Minimum Energy Route from Node 3 to Node 14
6.6 Energy consumption with the increase of node mobility. 95
6.7 Energy consumption with the increase of number of nodes. 96
6.8 Problem formulation 98
6.9 CF Greedy Algorithm 100
6.10 Path Adjustment 101
6.11 Network setting 108
6.12 Individually optimal routing 109
6.13 Coordinated optimal routing 110
6.14 ME greedy routing 111
6.15 CF greedy routing 112
6.16 Iterative adjustment 113
6.17 Number of admitted packets with the increase of $\lambda$ 114
6.18 Energy consumption with the increase of $\lambda$ 115
6.19 Cumulative energy consumption with the increase of admitted packets number 116

7.1 Illustrative Example 123
7.2 8 Nodes String 134
7.3 4*4 Grid Setting 136
7.4 Network setting (30 Nodes) $T=250$ 139
7.5 Network setting (40 Nodes) $Tr = 250$ 141
# LIST OF TABLES

3.1 Traffic Model and Its Parameters 38
3.2 Traffic Parameters For Background Data Service 38
3.3 System Parameters 38
3.4 Capacity Comparison 39
4.1 List of Notations 49
4.2 Total Transmission Power. [WATT] 64
4.3 Total Transmission Power. [WATT] 64
4.4 Comparison for Different $\delta$ 65
6.1 Comparison between Fastest and Min-Hop Routing 93
6.2 ME greedy algorithm 109
7.1 Transmission Patterns for String 134
7.2 Transmission Patterns for Grid (1 Channel) 137
7.3 Time Requirement for Grid ($Tr = In = 1$) 137
7.4 Time Requirement for Large Setting (30 Nodes, $Tr = In = 250$) 139
7.5 Time Requirement for Large Setting (30 Nodes, $Tr = 250, In = 500$) 141
7.6 Time Requirement for Large Setting (40 Nodes, $Tr = 250, In = 250$) 141
ON THE OPTIMAL FORMULATION OF RESOURCE MANAGEMENT IN WIRELESS NETWORKS

by

JIHUI ZHANG

Department of Computer Science
The Hong Kong University of Science and Technology

ABSTRACT

Resource management is of great importance in wireless networks, as it can improve the scarce wireless spectrum usage, enhance the network capacity, and help to provide QoS support for various services. In this thesis, we discuss resource management in both CDMA cellular networks and wireless ad-hoc networks.

For the CDMA cellular network, we firstly present a comprehensive survey on resource management in such a system and then introduce two major aspects of work we have done so far: firstly, admission control for the downlink multi-services MC-CDMA cellular networks. In this work we quantitatively demonstrate the significant impact from specific statistical factors, which are derived from stochastic behavior of mobile subscribers and channel conditions, and design the admission strategy accordingly for connection-level traffic and the background data transmission; Secondly, adaptive cell sectoring for non-uniform traffic to minimize the
total transmission power in a CDMA cellular network. We make use of a dynamic programming approach to formulate and solve the optimal sectoring problem. In addition, to reduce the computational complexity incurred from the optimal calculation and to prevent sector boundaries going across high-density regions, we propose a cluster-based sectoring algorithm with bounded complexity under high density traffic distribution.

For the wireless ad-hoc network, we first look into the recent work on throughput capacity maximization and energy conservation in such an infrastructure-free system, where cross-layer joint optimization could be applied to achieve more efficient resource utilization. We then introduce mobility assisted routing in a mobile ad-hoc network (MANET). In this work, we explore node mobility in the search for packet delivery routes under a deterministic mobility model, while optimizing a variety of QoS criteria. We consider networks operating in both interference-free and interference-limited regimes. In the interference-free case, we present efficient polynomial time algorithms to find optimal routes. In the interference-limited case, we give a mathematical formulation of the most energy efficient routing for multiple packets, and present an efficient heuristic algorithm. Our recent work mainly focuses on joint optimization in multi-channel multi-radio wireless multi-hop networks. We formulate the optimization under a deterministic model, and we seek to minimize overall system activation time to satisfy a given end-to-end traffic demand, subject to the number of channels, the multi-access interference restriction in the system and the radio interface constraint at each node. The exact solution to such an optimization problem is prohibitively complex due to the combinatorial complexity, particularly with the deployment of multi-radio and multi-channel. We then develop an efficient column generation based approach to solve this problem.

Lastly, we discuss our future work in wireless mesh networks. Seeking efficient resource management strategies, which comply with the specific demands of this emerging area, would be the main focus of our study.
CHAPTER 1

INTRODUCTION

With the rapid development and the continuous expansion of mobile communications, and the explosive growth in the demand for new wireless services, it is expected that next generation wireless networks will eventually support a wide variety of services including voice, video, images, data or combination of these. There are two major categories of wireless systems, namely the infrastructure network, where there are pre-deployed infrastructures responsible for the communication with end mobile nodes and the intermediate information routing; and the self-organizing infrastructure-free network, where information is delivered from source to destination through multi-hop wireless peer-to-peer relays. Example networks with infrastructure support are cellular networks [69], satellite networks [23] and IEEE 802.11-based Wireless LAN operating in an infrastructure mode [19] with access points' (AP) support. In such networks, base stations (satellites or APs) are connected to a backbone network, namely an interconnected wired/wireless network or the Internet. They are deployed in a manner of covering the service area and serving as a central controller to allocate channel resources to end users. Wireless transmissions are only involved in the communication with end users. In such a system, seamless transfer among different cells, as well as service provisioning with diverse QoS guarantee for mobile users, would be the major concern. Example networks without infrastructure support can be wireless ad-hoc networks, in which mobile (or static) nodes are distributed in the system. Due to the restriction on power consumption and interference, and in order to improve system capacity, nodes may communicate with a limited number of neighboring nodes. Thus, multiple-hop relays are generally required for end-to-end packet delivery. In addition, great developments of certain hybrid versions of wireless networks, where multi-hop wireless relays together with
the central infrastructure support are involved, can be envisioned. An emerging research area which attracts great attention in recent years would be the wireless Mesh network. This consists of a powerful multi-hop wireless backbone network, normally deployed at a rooftop and working in the broadband with multi-radio supply, and a distribution of end nodes with limited capacity. Multiple short distance relays among end nodes might be traversed locally before reaching the wireless backbone network. Another example is the wireless sensor network (WSN)[2], which comprises a great number of sensor nodes that have the capability of sensing, data processing and communication, and can be used for a variety of applications (environment, health, home and the like). The sensing data would be delivered across the sensor nodes before arriving at the sink (an information center collecting the data) and sent back to the backbone network for analysis and feedback. In this thesis, our discussion mainly focuses on code division multiple access (CDMA) cellular networks and wireless ad-hoc networks.

In recently years, direct sequence code division multiple access (DS-CDMA) has emerged as the predominant wireless access technology for cellular systems to provide high speed multimedia services used in both wideband CDMA (WCDMA) and cdma2000. For example, in WCDMA networks, the air interface using 5MHz bandwidth per carrier and 3.84Mcps high chip rate can provide increased processing gain and higher receiver multipath resolution, which enables the possibility of an effective RAKE receiver and offers flexible wideband services such as wireless Internet services with the peak rate of 384 kb/s and video transmissions with data rate up to 2 Mb/s [34]. Great capacity gain and flexibility can be achieved with the deployment of a DS-CDMA radio interface. Whereas in narrowband TDMA/FDMA systems with fixed channel allocation, the capacity of each cell is time-invariant based on the specified frequency reuse pattern ensuring the co-channel interference level [69], in DS-CDMA systems, all cells share the overall frequency band and the cell capacity is interference limited. The capacity of a DS-CDMA system depends heavily on instantaneous traffic conditions in both home and neighboring cells, while the transmission quality of wireless channels
can be expressed in terms of the signal-to-interference ratio (SIR) at the receiver side by taking into account both the in-cell and out-cell interference. Due to this unique interference-limited “soft” capacity nature of the DS-CDMA system, mechanisms such as adaptive cell sectoring and voice activity monitoring could be effectively applied to improve the system capacity [28]. Despite considerable efforts and progress made in the DS-CDMA system design, the accommodation of multimedia services still poses great challenges. To fully utilize scarce resources while at the same time provide considerable QoS guarantee to diverse services, it is of great importance to work out effective resource management strategies.

Wireless Ad-hoc networks emerge as a hot topic nowadays, due to the perspective of conveniently connecting mobile (or static) nodes without any pre-implemented infrastructures. In such a system, a packet needs to be delivered from the source to the destination through multi-hop relays across the network. Those attractive properties, such as convenience, efficiency, nomadism and the like, can be realized by the ad-hoc deployment.

Co-channel interference constitutes the main constraint in wireless ad-hoc networks where communication between a transmitter-receiver pair generally restricts the simultaneous transmission on its neighboring links. Thus only links that are located out of their mutual interference range can be active at the same time. This situation deteriorates when traffic needs to traverse from source to destination through multiple hops, and concurrent transmissions of neighboring links along delivery paths are prohibited. As discussed in the seminal work “The capacity of wireless networks” [31], per node throughput capacity in a fixed ad hoc network decreases with the increasing number of nodes in the system, and thus is largely restricted by the network size. Consequently, in such a network, a major measurement is the achievable throughput capacity. A great amount of effort has been made to improve the system’s throughput capacity. Another important issue is the power/energy conservation. Power sources generally have limited or costly energy supply. In some applications, such as remote sensor deployment, energy is entirely non-renewable. Power and energy efficiency is thus
of great importance for the design and operation of the network, in order to improve system lifetime. Another reason for power conservation is that controlling transmission power may mitigate interference to the network and help improve spatial reuse, and thus increase system capacity. Thirdly, conservative usage of power may obtain better co-existence with other systems operating in the same area. According to FCC [24], UWB systems [67] operate at an ultra-high unlicensed frequency band (i.e. 3.1 to 10.6 GHz), sharing bandwidth with other existing system (e.g. IEEE 802.11a). As a result, devices are regulated with low power emission for safe coexistence with other narrow-band systems.

Generally speaking, resource management in wireless networks contains a wide scope of aspects including channel configuration, power control, connection mobility management, load balance, and so on. Specifically, channel configuration comprises of basic channel configuration (QoS requesting and QoS mapping), dynamic channel configuration (to realize bandwidth on demand for best-effort traffic), code resource management (to use orthogonal codes at best, to improve code utility, to lower inter-channel interference, or to improve system capacity) and channel allocation (to efficiently use multiple orthogonal channels for greater throughput capacity). Power control, in CDMA cellular networks including open-loop control and closed-loop control, is basically utilized to overcome the "near-far" problem and shadowing attenuation. In ad hoc network networks, it is largely applied to increase system lifetime, reduce co-channel interference and maximize spatial reuse. Connection mobility management mainly deals with how to effectively choose a handoff scheme for various types of traffic, going across different cells or a variety of systems (i.e. seamless roaming). Furthermore, load balance can be obtained through call admission control (CAC), congestion control, cell planning, packet scheduling, adaptive (multi-path) routing, and so on. In short, resource management generally aims at guaranteeing the requested QoS, improving system capacity and reducing energy/power consumption, while keeping cost as low as possible. Consequently, adaptive resource management schemes that take into account those unique characteristics and challenges of the
system, while guaranteeing QoS requirements of diverse traffic types, should be designed.

1.1 Our Contribution

The contribution of this thesis mainly lies in two areas, the CDMA cellular network and the wireless multi-hop network. In the CDMA wireless cellular network, we first consider the problem of downlink traffic control in a Multi-code Code Division Multiple Access (MC-CDMA) system, which support multiple service types with diverse QoS requirements. Prior solutions for this problem have largely focused on call admission control at the connection level while neglecting the stochastic behavior of mobile subscribers and channel conditions. We quantitatively demonstrate that these statistical factors, in particular log-normal shadowing in propagation and voice activity factors, have a significant impact on connection-level performance. Accordingly we propose a connection-level admission control scheme [94], which guarantees a long-term transmission quality for all service types, while providing priority support for handoff traffic. Furthermore, we show that conventional data services can be best handled at the packet level as background transmissions [95] through taking advantage of the statistical variations. This leads to significantly better utilization of scarce wireless spectrums.

Cell sectorization is a well-acknowledged technique for improving the overall capacity in Direct Sequence CDMA (DS-CDMA) systems and the use of adaptive antenna arrays with dynamic cell sectoring is particularly suitable for non-uniformly distributed traffic. In our work [96] [97], we firstly re-formulate cell sectoring into an optimization problem and solve it by a dynamic programming algorithm. We then show that the optimal sectoring has two major practical drawbacks: the high complexity and the poor reaction toward oscillations of users between neighboring sectors. We then present an efficient cluster-based sectoring (CS) algorithm to mitigate these two inefficiencies: firstly, the computation
complexity of the CS algorithm is much lower than that of the optimal sectoring approach. In particular under a high-density case, the complexity is bounded and does not depend on the number of users in a cell; secondly, the CS algorithm maintains an excellent property of avoiding sector boundaries frequently crossing those users located closely within short angular distances. In addition, we also investigate the support for multi-rate applications by extending the CS algorithm. Through extensive experimental study, we find that the proposed CS scheme obtains comparable performance with greatly reduced complexity when compared to the optimal solution.

In wireless ad-hoc networks, we firstly look at the mobility assisted routing in mobile ad-hoc networks (MANET). Grossglauser and Tse (2001) first demonstrate a significant improvement of throughput capacity in MANET over a stationary network under a sufficient amount of random mobility, and thus show that the variation in the link strength might be exploited for more cost-effective communications. Although previous papers mainly focus on computing asymptotic bounds under homogeneity assumptions, in our work, we take a rather different approach and explore node mobility in the search for packet delivery routes under a deterministic model, optimizing QoS criteria such as end-to-end delay and energy consumption. We consider networks operating in both interference-free and interference-limited regimes. In the interference-free case, we present efficient polynomial time algorithms to search for optimal packet delivery paths for fastest, minimum hop and most energy efficient routing. In the interference-limited regime, we incorporate power control and scheduling into the routing decision, and thus carry out cross-layer joint optimization. Our objective is to find concurrent packets' relay paths associated with exact relay instants, which minimize system-wide energy consumption under deterministic mobility pattern, traffic load and channel conditions. We derive a dynamic programming based algorithm to solve the optimization problem, and also present an effective polynomial time heuristic approach.

Lastly, we investigate the achievable performance gain of a multi-radio multi-
channel multi-hop wireless network [101]. While previous work primarily focuses on the protocol design, we research the achievable performance gain by jointly optimizing routing and scheduling in such an environment. We derive the mathematical programming formulation for the optimization, which minimizes overall system activation time in use to support the throughput requirement, subjected to multi-access interference among neighboring transmission pairs and the radio interface constraint at each node. The exact solution to such an optimization problem is prohibitively complex due to the combinatorial complexity, particularly with the deployment of multi-radio and multi-channel. We then develop a column generation based approach to solve this problem, which decomposes the original problem into sub-problems and solves them iteratively.

1.2 Organization of the Thesis

This thesis is organized as follows. In section 2, we present a survey of resource management in DS-CDMA cellular networks and point out the main challenges that need to be addressed in such a system. Then we move on to introduce our downlink traffic control scheme, which takes into account the statistical behavior of the system and ensures the long-term transmission quality for a mixture of traffic types in Section 3. An efficient adaptive cell sectoring algorithm and its extension to multi-rate applications are proposed in Section 4. Later in Section 5, we make a survey of resource management, with emphasis on throughput capacity and energy conservation, for wireless ad-hoc networks. We then investigate node mobility in the search for packet delivery routes both in the interference free regime and the interference limited regime in Section 6. The cross-layer joint routing and scheduling optimization in the multi-radio, multi-channel multi-hop wireless network is proposed in Section 7. At last in Section 8, we discuss our future work and conclude the thesis.
CHAPTER 2

RESOURCE MANAGEMENT IN DS-CDMA CELLULAR NETWORKS

Resource management generally contains a wide scope of aspects including channel configuration, power control, connection mobility management, load balance, and etc. This survey mainly focuses on some critical issues, such as call admission control, congestion control, power and rate allocation and cell planning in a DS-CDMA wireless cellular network. To put it into more detail, call admission control is devoted to the design of efficient admission algorithms at call arrival; congestion control tries to avoid or relieve congestion status throughout the system; an efficient allocation of power and rate may optimize system performance under a given system state; cell planning includes a network's construction considerations, such as bandwidth allocation between two links with adaptation to the traffic asymmetry; base station location and density planning to minimize the system cost; pilot power control for proper cell coverage; and the optimal cell sectoring issue so as to determine the direction and width of the sectors.

In this chapter, we start by describing general concepts in the DS-CDMA network, in particular the capacity constraint in Section 2.1, and discuss the main challenges in Section 2.2. We then closely examine the congestion control, rate and power control and cell planning in Section 2.3. We study the call admission control in Section 2.4, and Section 2.5 concludes the survey.

2.1 Capacity in DS-CDMA

Before we go into resource management schemes, we first take a close look at the capacity provisioning in DS-CDMA systems. In the DS-CDMA cellular network, all users share the same overall frequency band for transmissions. Each
user is assigned one or more distinct spreading codes and all these codes generally bear noise-like characteristics with very little cross-correlation to each other. The quality of communication is primarily determined by the detected signal-to-interference ratio (SIR) level at the receiver, with the generally accepted measurement on the bit energy-to-noise density ratio \( (E_b/N_0) \), expressed as \( \text{SIR} = \frac{P_t/R}{I_N/W} \), where \( R \) is the base-band information bit rate; \( W \) is the total radio frequency band for transmission, thus the corresponding spreading gain is \( G = W/R \); \( P_t \) is the received signal power for the considered spreading code and \( I_N \) denotes the total detected interference at the receiver. In order to guarantee the transmission quality for the considered service, the SIR should be maintained above a certain threshold.

To achieve a higher system capacity, one of the most efficient ways is to mitigate multiple access interference (MAI) among the mobile users. Two techniques are commonly applied [28]: 1) the cell sectoring, which deploys directional antenna arrays at base stations and splits a cell into different sectors. Generally speaking, only signals from users located within a sector are received at the corresponding antenna array. Thus the number of users that a cell can serve might be increased approximately by the same factor as the number of sectors; 2) the voice activity monitoring, which switches off the signal transmission when the user is not active, thus reducing the multi-access interference.

In order to accommodate high speed multimedia services and to support variable rate transmissions, two mechanisms can be used: Multi-code (MC) and Variable Spreading Factor (VSF) [81]. In a MC-CDMA system, all data signals over the radio interface are transmitted at a basic rate and the spreading gain \( G \) over each code channel is a constant. Multiple orthogonal spreading codes are applied simultaneously for a high speed application. For VSF transmission, each user transmits over one single code channel. A higher transmission rate is provided by varying the spreading factor \( G \) inversely with the desired data rate. Thus, in a VSF-CDMA system, users are assigned variable length codes and different power levels based on their data rates and quality of service (QoS) requirements. The
two mechanisms provide comparable performance in high speed transmissions. Consequently both are commonly deployed in DS-CDMA systems.

![Figure 2.1: SIR model for DS-CDMA cellular system.](image)

Generally speaking, the uplink and the downlink use different bandwidth regions for transmission in a DS-CDMA cellular system. Thus, they are usually considered separately for the capacity calculation. Refer to Figure 2.1, in the uplink, the received signal power at the reference base station BS0 from a specific mobile station MS\(j\) is MS\(j\)'s transmitted power \(P_j\) multiplied by the pass loss factor \(L_{0,j}\); the corresponding interference is the received power from other active mobile stations, both within the reference cell BS0 (e.g. MS\(i\)) and the out cells (e.g. MS\(k\) in BS1), plus the background thermal noise [69]. Capacity in the uplink is calculated as the maximum number of users that could be admitted while ensuring the SIR requirement for each user in any reference cell. In the downlink, a fraction of total transmitted power at the base station is dedicated to the control channels and all traffic channels share the remaining power. As shown in Figure 2.1, the reference cell BS0 allocates a fraction of its traffic channel power to an in-cell mobile user, e.g. MS\(i\); the remaining power received at MS\(i\) from its home base station BS0 and other base stations, such as BS1, appears as interference. The capacity in the downlink is thus defined as the maximum number of users that could be admitted under the constraints that the transmission quality of each
user, in terms of its SIR threshold, is guaranteed and the overall power in need does not exceed the maximum power supply at the base station.

2.2 Challenges in Resource Management

The accommodation of multimedia services with diverse QoS requirements poses many new challenges in the design of resource management strategies, such as how to efficiently allocate available resources to optimize channel utilization; how to adjust the service rate to relieve the congestion state or accommodate more traffic; how to maintain diverse QoS requirements; and how to provide priorities among different services. There are also several distinctive characteristics for the DS-CDMA system which can be explored to achieve better performance. For example, the universal frequency usage enables more efficient resource provisioning; proper rate and power allocation helps reduce the interference, thus providing higher capacity. On the other hand, these features might inevitably increase the complexity of the system design. In this section, we examine some of the issues including universal frequency reuse among cells, power control, soft handoff, voice activity factor, path loss model for a DS-CDMA system, and discuss their impact on resource management design.

2.2.1 Universal Frequency Reuse

The universal frequency reuse of a DS-CDMA system allows all cells to share the same wide frequency band. This reduces the complexity in cell planning on co-channel bandwidth allocation and potentially leads to higher capacity gain for a cell under asymmetric loading. Nevertheless, this increases the dependency among neighboring cells and incurs significant difficulty in the resource management, as more frequent coordination is required. Furthermore, because of the universal frequency reuse with interference-limited transmission, even if the cell occupancy information can be obtained, the outage, i.e., the received SIR level for an existing connection drops below a threshold, might still occur due to user mo-
bility, propagation variation, voice activity and the like. As a result, more precise control and finer tuning are generally required in the resource management.

2.2.2 Power Control

Power control in a DS-CDMA system helps to reduce excessive interference throughout the system and prolong battery life. Most existing work assumes perfect power control, i.e. in the uplink, to overcome the “near-far” problem [69], the received power from all users in the reference cell are set to be equal; in the downlink, the assigned power to each user is adjusted to achieve the exact SIR level that is required. In practical systems, power control imperfection may occur occasionally and cause certain mis-adjustment on the received power. As a result, the received power level of both the required signal and the interference might change over time. This in turn exerts additional difficulty to the power allocation and adjustment, and requires resource management strategies to incorporate a more accurate estimation on the power control imperfection.

2.2.3 Soft Handoff

Soft handoff denotes the state where a mobile transmits to and receives from more than one base station simultaneously. In the downlink, the mobile combines the signals from the connecting base stations and adds the different multipaths to reinforce the received signal, which leads to improved communication quality. In the uplink, the neighboring base stations independently decode the signals received and the best replica is selected. Soft handoff ensures smoother user communications, better communication quality during handoffs and larger cell coverage in DS-CDMA systems [89]. The tradeoff lies in a higher complexity and an additional network resource demand. As a result, soft handoff may have a great impact on the resource management design. For example, as a longer handoff transfer delay is generally endurable, it is possible for handoff traffic to be queued. Furthermore, as more base stations are involved in the communication with a
mobile user, power transmission and allocation should be carefully adjusted to maximize spectrum utilization and avoid unnecessary interference in the system. In addition, extra receivers are required for communication with each user.

2.2.4 Voice Activity

In DS-CDMA systems, cell capacity is mainly constrained by the interference from simultaneous transmissions. In such a case, the voice activity factor can be incorporated into the system design and the spectrum utilization might be significantly improved by shutting down mobile stations during their silent periods. Consequently, the system does not always need to provide full bandwidth to the admitted calls as long as their received SIR levels during active periods exceed the required threshold. However, with the partial bandwidth allocation, high outage probability could occasionally occur when most of the users are actively transmitting. The additional capacity gain should be well justified to maintain the required QoS, and congestion control is needed to combat the more varying system behavior.

2.2.5 Propagation Model

Generally speaking, the propagation model might change greatly in a variety of environments and/or during different time intervals. This imposes great difficulty as well as new chances for the system management. For example, under poor channel conditions, higher power level should be allocated to compensate for the severe path loss; however, this may cause over-provisioning of system resources and incur additional interference to other users in better channel conditions. On the other hand, by precisely tracking channel fluctuations and scheduling transmissions when the corresponding channel quality is near peak, such diversity gain might help to improve performance in terms of a higher system throughput and lower inter- and intra-cell interference [87]. Therefore, a precise loss model and accordingly better resource management strategies should be designed to
properly utilize system resources.

The issues discussed in this section pose considerable difficulties as well as opportunities for the resource management in DS-CDMA systems. The challenge here is how to design an effective resource management policy, which can accurately reflect the system behavior and optimize network performance by taking advantage of the unique characteristics of DS-CDMA systems.

2.3 Resource Management Issues

Resource management is a large scope concept. In this section, we investigate some of the major aspects, including congestion control, power and rate allocation, and cell planning. Furthermore, the cell planning considers the bandwidth allocation, base station planning, pilot power control and cell sectorization. We give a detailed discussion on call admission control in Section 2.4.

2.3.1 Congestion Control

When a system fails to find a set of power transmission levels which satisfy users' QoS requirements, congestion occurs. Even with call admission control, congestion might still happen due to some factors such as the deterioration of the wireless environment, mobility of users, users' activity factor and power control imperfection. As a result, the transmitted power and the cell interference level tend to increase and outage occurs. This incurs traffic loss and delay jitter, thus deteriorates the transmission quality. Generally speaking, when congestion occurs, three major mechanisms can be applied to relieve the congestion status: 1) to drop some on-going calls. The question is then, which set of calls to drop? The system may choose to drop call(s) in the outage condition, or to drop each existing call with pre-specified probability, or to drop call(s) that make the largest contribution to alleviate the congestion. The last scheme offers the best performance with the highest complexity [3]; 2) to decrease the transmission rate. This
could be done by either proportionally reducing the transmission rate of each user or decreasing rates to the same maximal fair SIR level [40]; 3) to reduce the number of simultaneous transmissions [54]. The transmission probability can be dynamically adjusted according to the occupancy information. To eliminate the randomness and to obtain full control over the simultaneous transmissions, cells may sequentially schedule actively transmitting periods to data services.

To conclude, dropping on-going calls is the most straightforward way to alleviate congestion, however it is undesirable and often unbearable from users' point of view. As mentioned in [17], to simultaneously reduce the data rate of all data services yields lower throughput than to support fewer users at full rates at the same time. Thus, decreasing data rate obtains sub-optimal system utilization. On the other hand, switching off some users could lead to extra transmission jitter and delay. All these tradeoffs need careful examination.

2.3.2 Power and Rate Allocation

Provided with system parameters on the cell occupancy and channel conditions, it is important to improve system performance by properly allocating transmission rate and power. A variety of criteria might be optimized, such as to minimize power consumption, which prolongs the battery life and causes less interference; or to maximize the transmission rate, which indicates the maximized system throughput and resource utilization. A power and rate allocation scheme has been proposed for the downlink in [44] with the objective of maximizing the total number of frames transmitted during a control interval under a SIR constraint for each service type and an average transmission power limit among all base stations. The basic implication is that if a mobile terminal experiences poor channel conditions, the assigned transmission rate for this mobile should be low. The optimal allocation is formulated and the solution is obtained through an exhaustive search, with exponential complexity and degraded fairness. To cope with these two disadvantages, suboptimal algorithms are proposed where rate allocation de-
pends on the interference distribution and the power allocation depends on the traffic distribution among the cells. For the uplink [86], the optimization has been formulated to maximize the total normalized transmission rate subjected to the constraints on transmission power and the maximum allowable rate. The optimal solution yields not only higher throughput but also significant power savings. However, this can potentially lead to starvation for users under poor channel conditions, since users with better channel condition always transmit first.

2.3.3 Cell Planning

In order to reduce the network cost and maximize the utilization of scarce resources, efficient cell planning is of vital importance for service providers. Generally speaking, cell planning consists of a number of issues, such as bandwidth allocation, base station planning, pilot power control and cell sectorization. We discuss each of them in detail.

- Bandwidth allocation in the uplink and downlink

  In most systems, the uplink and the downlink use different bandwidth regions for transmission and the bandwidth allocation between the two links is symmetric. However, with the accommodation of multimedia services, a greater amount of radio spectrum is required for the downlink than the uplink. To cope with the traffic asymmetry, unbalanced bandwidth allocation is preferred and great system performance gain might be achieved by a proper assignment of bandwidth between two links according to the traffic demand.

- Base station planning

  Since the overall frequency band is shared among all active users and the capacity of each cell depends on the interference levels, proper planning of base station locations should consider not only the cell coverage but also
some other factors, such as the amount of spectrum available, estimated traffic distribution in the area and the radio channel propagation models. The main purpose of base station planning is to select the sites for base stations by taking into account the system cost, transmission quality, service coverage and so on. The optimal location of new base stations to minimize a linear combination of installation cost and total transmitted power has been investigated in [4] by taking into account traffic distribution, SIR requirements, power allocation constraints and power control mechanisms. More complex models also exist with considerations on the stochastic behavior of the system, or/and soft handoff. The optimal base station density is estimated in [33] considering power control, macroscopic selection diversity and the propagation model.

- Pilot power control

To select a proper base station to connect, a mobile station needs to measure and report the $E_b/N_0$ level of its received pilot power to the base station. The pilot power determines the cell coverage area and the average number of users within a cell. Increasing the pilot-signal power of a base station expands the coverage area of the cell, thereby increasing the number of users in the serving cell but resulting in higher intra-cell interference. Efficient pilot power control needs to balance the cell load and cell coverage area among neighboring cells, whose objective is to reduce the variation of interference, stabilize the network operation, and improve cell capacity and communication quality especially under non-uniform traffic load among the cells.

- Cell sectorization

To cope with interference, smart antennas or adaptive array processing can be utilized to enhance both energy efficiency and multiple access interference rejection capability. It has been shown in [15] [59] that the use of antenna arrays at base stations greatly increases system capacity by reduc-
ing the co-channel interference. Traditional sectoring approaches divide a cell into equal-width sectors, which has been shown to provide the same fold capacity gain under highly uniform traffic load [28]. However, for a system with “hot-spot” traffic, sectors with high-density traffic suffer high outage probability. Adaptive cell sectoring could be deployed at a base station to greatly improve the performance of such a system. In [73], the optimal cell sectoring (OS) with an objective of minimizing the total transmission power is formulated and the solution is obtained using the well-known Dijkstra’s algorithm. The direction and width of the sectors can also be adjusted according to the geographic distribution of traffic. With the observation that the sector boundaries should better cross some low-density regions in order to minimize the oscillations, and also with the objective of lower computational complexity, the authors in [97] propose a cluster-based sectoring (CS) algorithm. The complexity of CS is much lower than the OS algorithm, without any significant degradation on system performance. Specifically under the user distribution and sectoring for OS and CS shown in Figure 2.2, the CS algorithm greatly reduces the complexity of OS with only a slight increase in the total power transmitted. Furthermore, it can be observed that the sector boundaries are generally crossing low-density regions in the CS solution; while in OS, they may pass through two users very close to each other. It is expected that future dynamic cell sectoring might be designed to incorporate the stochastic nature of the system, such as user mobility, channel fading, power control and sectorization imperfection, into consideration, to better adapt to a real system.

2.4 Call Admission Control

Call admission control is one of the most important elements in the resource management, which determines whether to admit or reject a call upon its arrival. The objective is to maximize the resource utilization by admitting as many new
Figure 2.2: User distribution and cell sectoring

calls as possible while maintaining certain fairness and QoS of on-going calls. There are many call admission control schemes proposed in the literature, most of which can be classified into following three categories: number-based, SIR-based and interference-based call admission control. In this section, we introduce each of them in detail.

2.4.1 Number-based Call Admission Control

In number-based call admission control [77], the QoS requirement, in terms of an upper bound of packet error probability, is mapped into the maximum number of voice/data calls that can be simultaneously accommodated by the system, denoted by $K_v/K_d$. To reflect the capability that voice calls can tolerate higher bit error rates, $K_v$ is set to be greater than $K_d$. As shown in Figure 2.3, the admission of voice calls depends on a threshold value $T_v$. New data arrival either joins in the backlog pool or is discarded when the pool is full. To provide priority to voice calls, the admission threshold $T_v$ can be set equal to $K_v$, i.e. admitting as many voice calls as the system can support. Certain fairness can be offered to data traffic by setting $T_v$ less than $K_v$, taking into account the number of backlogged data packets. At any time slot $t$, voice calls can transmit without delay and the backlogged data packets occupy the silent period of voice calls for transmission with probability $P_t$. $P_t$ is calculated according to a function with the parameters set to be the maximum allowed simultaneous data transmissions
Kd, the current active voice calls and the number of backlogged data packets obtained from the system feedback at the beginning of each time slot.

Figure 2.3: Number-based admission control

Based on a time-invariant cell capacity assumption, the operation for number-based call admission strategies is very similar to those in narrowband systems, in which call admission only depends on the current cell loading. Such a scheme is simple for implementation and analysis. However, in a DS-CDMA system, the capacity varies with the interference level, while the number-based schemes completely ignore the soft nature of DS-CDMA capacity, so the control is generally inaccurate and non-adaptive.
2.4.2 Interference-based Call Admission

In DS-CDMA cellular networks, all users share the same wide frequency band. When a call (MS.Arrival) is admitted into a cell, it transmits to and receives information from the corresponding base station (BS0) as shown in Figure 2.4. In the uplink, the transmitted power from MS.Arrival increases the interference level on other in-cell (MS.In) and out-cell (MS.Out) users' receivers at their base stations BS0 and BS1. In the down link, the power allocated from BS0 to MS.Arrival causes additional interference to all other users (MS.In and MS.Out) in the system. The interference level depends heavily on overall system parameters. Consequently the admission of a new call can gracefully degrade the performance of all users currently in service. As a result, in a DS-CDMA system, the number of calls that can be admitted to a cell is not a fixed value, and a more reasonable measurement is the received interference power level by the receivers.

An admission control scheme based on the received interference at base stations is proposed in [74] for an uplink DS-CDMA system. Three interference margins are defined: the total interference margin (TIM), the current interference margin (CIM) and handoff interference margin (HIM). Here, TIM is the maximum acceptable link interference such that the QoS in terms of lower bound $E_b/N_0$ is guaranteed; CIM is the estimated interference level taking into account the assignment of the channel to the newly arrived call; HIM is the interference estimation by further considering the reserved channels for handoff calls. The BS interprets the current interference from its measured power strength and calculates CIM and HIM accordingly. If there is a call request and HIM<TIM, i.e., handoff arrival can be safely reserved, then the call is admitted. Otherwise, the base station checks whether it is a handoff call and CIM<TIM. If so, the base station would assign a new channel for the handoff call; otherwise the call is rejected. Compared with the number-based schemes, the interference-based call admission control better represents the inherent interference-limited nature of DS-CDMA systems at the expense of extra complexity of interference detection.
2.4.3 SIR-based Call Admission Control

Admission control based on the uplink SIR requirements is firstly proposed in [52]. A crucial measurement, namely residual capacity, is defined as the additional number of calls that can be accepted by a base station such that the system-wide outage probability does exceed a pre-defined SIR level. In the localized algorithm, the residual capacity is calculated solely based on the SIR measurement at the local base station and a call is not admitted unless the residual capacity is greater than zero. In the global algorithm, all adjacent cells’ SIR levels and residual
capacities are calculated upon a call arrival. The call is accepted if and only if the minimum of all cells’ residual capacity is greater than zero. This ensures that the admission of a new call does not affect the QoS in all surrounding cells, which is particularly important under non-uniform traffic scenario. This work is extended to support multimedia services in [39], where the uplink and the downlink are considered separately. The average SIR level for each call class is measured periodically for both links. Accordingly, for each class $i$, the system estimates the expected $\text{SIR}_{i,j}$ upon a class $j$ call arrival. Distinct admission thresholds $\Gamma_{i,j}$ are settled for admission if $\text{SIR}_{i,j} \geq \Gamma_{i,j}$, for all call classes. Priority is provided to handoff calls by setting higher SIR thresholds for new arrivals.

To better reflect stochastic behavior of a system and to guarantee a long-term transmission quality, an admission control scheme for the downlink MC-CDMA system to support multiple classes of traffic with diverse SIR requirements is proposed in [95], as with wider deployment of multimedia services, the downlink may become the bottleneck of a system. The lower-bound SIR requirements for multiple service types are formulated, incorporating the statistical factors such as voice activity and log-normal shadowing in propagation. The long-term outage probability is calculated and the capacity constraints are derived by a Gaussian approximation. Reservation based on an iteratively estimated handoff arrival rate is performed to provide priority support for handoff traffic, and thus lower dropping probability for handoff calls can be obtained. The long-term outage probability is kept at a pre-required level to ensure much stabler system performance and at the same time, gain better system utilization with QoS guarantees.

In conclusion, the number-based call admission control assumes a time-invariant cell capacity and is simple for implementation and analysis. However, it does not consider the inherent soft capacity nature of a DS-CDMA system and could lead to inaccurate results. On the other hand, SIR-based schemes best characterize the transmission quality and thus offer the QoS guarantee. However the varying system environment poses considerable difficulty in the system design and measurements. The interference-based call admission control also considers the
interference limited nature of the system. The trade-off lies in the simpler implementation of coarse measurements on quality of services.

2.5 Summary

The DS-CDMA cellular system significantly differentiates itself from the traditional narrowband TDMA/ FDMA systems, in which great capacity gains could potentially be achieved by taking advantage of its unique characteristics. There are also new challenges for the system design, and in particular for the resource management. In this survey, we discuss several critical issues in the resource management including call admission control, congestion control, rate and power allocation, and cell planning. In general, these issues are closely related. A cell planning strategy that adapts well to the traffic distribution and channel condition helps the design of call admission control schemes. A better admission policy can alleviate the pressure on the congestion control mechanism. Furthermore, an efficient rate and power allocation could be deployed to achieve optimized system performance. Perhaps greater challenges exist in the seamless integration of the different elements in real systems and adequate adaptation to the stochastic behavior of the system such as user mobility, air-interface condition variation, voice activity factors, power control imperfection and traffic conditions.
CHAPTER 3

DOWNLINK TRAFFIC CONTROL IN MC-CDMA CELLULAR SYSTEMS

In this chapter, we consider the problem of downlink traffic control in an MC-CDMA system that can accommodate multiple services with different QoS requirements. Specifically, we consider both real-time services and conventional best-effort data services, in which the real-time services require a stringent upper bound on the transmission delay that must be handled at the connection level, while the delay-tolerable data services can be queued and delivered at the background to maximize the utilization of scarce wireless spectrums. There are several distinctive features in the proposed control policy: 1) a long-term transmission quality of all services are guaranteed by ensuring upper bounds on the respective outage probability; 2) handoff arrival rates are derived using an iterative algorithm by taking into account users' mobility and cells' layout; 3) priority support is provided to real-time handoff traffic using an adaptive reservation scheme; 4) queuing is applied to the background data transmission to make full use of the wireless bandwidth. Numerical results from performance analysis indicate that the proposed scheme achieves excellent bandwidth utilization and adapts well to a wide variety of system configurations. Furthermore, we investigate the sensitivity of system capacity from various statistical factors, in particular log-normal shadowing in propagation and voice activity factors. The results demonstrate that both have a significant impact on the connection-level performance, and data queuing can help to mitigate this by taking advantage of the variations from users' activities and channel conditions. The rest of the chapter is organized as follows. Section 3.1 describes the system model and SIR formulation in the downlink MC-CDMA cellular networks. In Section 3.2, we address call
admission control for connection level services. In Section 3.3, we discuss background data transmission. We present the performance analysis and numerical results in Section 3.4 and 3.5 respectively. We conclude the chapter in Section 3.6.

3.1 System Model and Formulation

3.1.1 System Model

Considering a downlink Multi-Code Direct Sequence CDMA (MC-CDMA) cellular system, where all data signals over the radio interface are transmitted at a basic rate $R_b$ and the spreading processing gain over each code channel is a constant. To achieve a higher transmission rate, multiple spreading codes are applied simultaneously to a high speed application. A fraction $(1 - \beta)$ of the total transmitting power at the base station is dedicated to the control channel and all services in the cell share the remaining power. The radio propagation is modelled by path loss and shadowing effect [28],

$$
\zeta_{ji} = d_{ji}^{-m} L_{ji}[\sigma^2],
$$

where $d_{ji}$ is the distance from base station $BS_j (j = 0, 1, \ldots, M)$ to a given user $i$; $m$ is the path loss exponent and $L_{ji}[\sigma^2]$ is the lognormal shadowing variance. There are $K + 1$ classes of traffic ($C_1, ..., C_k, ..., C_K, C_{K+1}$) in the system, in which ($C_1, ..., C_k, ..., C_K$) are real-time traffic. All background data traffic ($C_{K+1}$) is placed in a data queue before transmission. Assume each connection level service takes on an on-off Markov model [65] such that when a call is connected, it alternates between "on" and "off" states according to the characteristic of a traffic source. Specifically when a class-$k$ ($k = 1, 2, ..., K$) call is in an "on" state, it transmits continuously at a rate $c_k^d R_b$ ($c_k^d < 1$), where $R_b$ is the rate of a basic code channel, and $c_k^d$ is the number of spreading codes assigned for each class-$k$
call in the downlink. The activity variable of a class-$k$ call is denoted by $\Psi_k$, with $P\{\Psi_k = 1\} = \alpha_k$ denoting the probability that a class-$k$ call is in the "on" state.

### 3.1.2 SIR Formulation

Under homogeneous traffic conditions, i.e., the types and amount of traffic are statistically the same, which uniformly loaded in all cells [29], the total transmission power from each base station is equal to $P_i$ in a steady state. Consider the $i$-th user of a class-$k$ call in a reference cell $BS_0$ (denoted as $U_{0ki}$). Assume base station $BS_0$ allocates a fraction $\phi_{ki}/\beta$ of total power to user $U_{0ki}$. This power is distributed equally among the $c_k^d$ code channels of $U_{0ki}$. Hence, each code channel has a power of $\phi_{ki}/\beta$. The signal to interference ratio (SIR) for $U_{0ki}$ can be expressed as,

$$\gamma_k = \left(\frac{E_b}{I}\right)_{ki} = \frac{\left(\frac{\phi_{ki}}{\beta}P_i\zeta_{0ki}\right)/R_b}{(1 - \nu)(1 - \frac{\phi_{ki}}{\beta})P_i\zeta_{0ki} + \sum_{j=1}^{M}P_i\zeta_{jki}} / W_d \quad (3.2)$$

Here, denote $M$ as the total number of neighboring base stations which interfere with the referred base station $BS_0$; $\zeta_{ki}, (j = 1, 2, \ldots, M)$ as the path loss between $BS_j$ and $U_{0ki}$; $\nu$ as the orthogonal factor in the downlink; and $W_d$, $R_b$ are the downlink spreading bandwidth and basic code rate respectively. Background noise is ignored.

### 3.2 Connection Level Call Admission Control

In this section, we describe the connection level call admission control for real-time services. To ensure a long-term transmission quality of diverse service classes, we take into account the statistic factors of the radio channel (log-normal shadowing in propagation) and mobile users (voice activity) in calculating the capacity constraint. In order to maintain a low handoff dropping probability, we
propose a handoff based (HB) adaptive reservation scheme which reserves certain amount of bandwidth exclusively for handoff calls. In addition, we use an iterative algorithm to accurately derive handoff arrival rates that adapts to current traffic conditions.

3.2.1 Downlink Capacity Constraint

To ensure a long-term transmission quality, capacity constraint needs to be calculated at each cell before call admission. According to the SIR formulation shown in (3.2), the fraction of power allocated to user $U_{ki}$ is given by,

$$f_{ki} = f_k(1 - v + \sum_{j=1}^{M} \frac{\zeta_{jki}}{\zeta_{0ki}})$$  \hspace{1cm} (3.3)

where $f_k = \frac{c_k}{\beta(G_d \gamma_k (1-v))}$, and $G_d = \frac{W_d}{R_d}$ is the downlink processing gain. Assume there are $n_k (k = 1, 2, \ldots, K)$ class-$k$ users in the cell. The outage probability constraint can be written as

$$P_{out} = P\{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \Psi_{ki} f_{ki} > 1\}$$  \hspace{1cm} (3.4)

Here, $\Psi_{ki}$ is the activity variable for class-$k$ user $i$. Let $\Phi = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \Psi_{ki} f_{ki}$. This might be approximated by a Gaussian distribution with a mean and a variance as $m_\Phi$ and $\nu_\Phi$.

The outage probability $P_{out}$ can then be obtained by calculating the tail distribution of $\Phi$,

$$P_{out} = \int_{1}^{\infty} \frac{1}{\sqrt{2\pi} \nu_\Phi} e^{-\frac{(x-m_\Phi)^2}{2\nu_\Phi}} \, dx$$  \hspace{1cm} (3.5)
The capacity region is calculated as the maximum number of users of different classes that could be concurrently accommodated in a cell, subjected to the constraint that the outage probabilities do not exceed some predefined threshold $\delta$, i.e. $P_{out} \leq \delta$. (Please refer to Appendix A for details)

### 3.2.2 Handoff Based (HB) Reservation Scheme

In order to maintain a low handoff dropping probability, certain reservation is usually necessary for handoff traffic. Furthermore, to efficiently utilize the limited capacity and to adapt to the time-variant traffic conditions, it is important to accurately estimate handoff traffic arrivals. Under a homogeneous environment, assume that class-$k$ ($k = 1, 2, \ldots, K$) new calls arrive according to the Poisson process with rate $\lambda_{n,k}$ in a cell and a call duration is exponentially distributed with the average duration time $1/\mu_{c,k}$. Following a similar procedure as in [43], we could calculate the class-$k$ call mean dwell time $1/\mu_{d,k}$ within a cell (Please refer to Appendix B for details)

\[
\mu_{d,k} = \frac{V_k L_{RCell}}{\pi A_{RCell}} = \frac{6 \cdot r_1 \cdot V_k}{\pi \cdot (3\sqrt{3}/2)r_1^2} = \frac{4V_k}{\sqrt{3}\pi r_1} \tag{3.6}
\]

where $V_k$ is the average moving speed for class-$k$ calls, $L_{RCell} = 6r_1$ is the perimeter of a regular hexagonal cell and $A_{RCell} = (3\sqrt{3}/2)r_1^2$ is the area of the hexagon. Under the homogeneous assumption, the mean departure rate of class-$k$ calls toward neighboring cells should be equal to the mean handoff arrival rate $\lambda_{h,k}$ for a reference cell, i.e.

\[
\lambda_{h,k} = \frac{\mu_{d,k}}{\mu_{d,k} + \mu_{c,k}} \cdot \{\lambda_{n,k} \cdot (1 - P_{\text{block},k}) + \lambda_{h,k} \cdot (1 - P_{\text{drop},k})\} \tag{3.7}
\]

where, $P_{\text{block},k}$ is the new call blocking probability and $P_{\text{drop},k}$ is the handoff call dropping probability for class-$k$ calls in the cell. Thus, the handoff arrival rate is,
\[
\lambda_{h,k} = \lambda_{n,k} \cdot \frac{\mu_{d,k} (1 - P_{\text{block},k})}{\mu_{d,k} \cdot P_{\text{drop},k} + \mu_{c,k}}.
\] (3.8)

According to (3.8), \(\lambda_{h,k}\) is dependent on the new call blocking probability \(P_{\text{block},k}\) and handoff dropping probability \(P_{\text{drop},k}\), which are derived from the state probabilities; on the other hand, the state information is derived from the handoff arrival rate. As a result, we calculate \(\lambda_{h,k}\) by iteration with initial values \(P_{\text{block},k} = P_{\text{drop},k} = 0\) [39].

As interrupting a call in progress is generally less desirable than occasionally blocking a new call request, it is necessary to reserve certain capacity exclusively for handoff traffic, commonly referred to as the Trunk Reservation, or Guarded Channel policy. On the other hand, in order to maximize the bandwidth utilization, a cell should accommodate as many new calls as possible. Therefore, it is important to make reservations based on handoff arrival rates. In this paper, we propose a handoff based (HB) reservation, which reserves a fraction of channels exclusively for handoff calls according to handoff arrivals.

A function \(R_{h,k} = \delta_k(\lambda_{h,k} \cdot T_k)\) is adopted as a reservation measurement for class-\(k\) handoff arrivals [38]. Here \(\lambda_{h,k}\) is class-\(k\) handoff calls’ arrival rate; \(T_k\) is the average time for a class-\(k\) call requiring connection within a cell, which is exponentially distributed with mean \(T_k = \frac{1}{(\mu_{d,k} + \mu_{c,k})}\); \(\delta_k(0 < \delta_k < 1)\) is the reservation factor for the proportion of class-\(k\) handoff calls that would be reserved, thus \(R_{h,k} = \delta_k(\frac{\lambda_{h,k}}{\mu_{d,k} + \mu_{c,k}})\). The overall reservation vector is given by

\[
R_h(s) = (R_{h,1}, R_{h,2}, ... R_{h,K}).
\]

The HB reservation scheme is easy to implement and may flexibly adapt to the changing QoS requirements by adjusting the value of reservation factor \(\delta_k\).
3.2.3 Connection Level Admission Control

Let \( n_k \) (\( 1 \leq k \leq K \)) denote the number of class-\( k \) calls in progress within a cell. The system steady-state could be written as a \( K \)-dimension row vector,

\[
s = (n_1, n_2, \ldots, n_K)
\]

Thus, for the given downlink outage probability threshold \( \delta \), the feasible state space is restricted as,

\[
S = \left\{ s : s = (n_1, n_2, \ldots, n_K) \right\} \text{ subject to }: P_{\text{out}}(s) \leq \delta_{\text{out}}
\]

where the computation of \( P_{\text{out}}(s) \) is given in Sec. 3.2.1.

By denoting \( e_i \) as a \( K \)-dimension vector with only the \( i \)-th element equals to 1, while all other elements equal to zero, and assume the system is currently in state \( s = (n_1, n_2, \ldots, n_K) \), the connection level admission control strategy can be described as follows,

1. Upon the arrival of a new class-\( k \) user, the base station checks according to (3.10) whether \( s + R_k(S) + e_k \in S \) (\( R_k(S) \), the reservation vector computed in Sec. 3.2.2). If \( s + R_k(S) + e_k \) is within the system’s admissible region, this new call is accepted; otherwise, it is blocked.

2. Upon the arrival of a handoff class-\( k \) user, the base station only checks whether \( s + e_k \) is in the feasible state \( S \). If \( s + e_k \in S \), this handoff call is accepted; otherwise, it is dropped.

When a connection level call is admitted, it would transmit whenever it has information to send with the assigned transmission rate \( c_k^a R_k \) until its termination.
3.3 Background Data Transmission

As discussed earlier, the proposed traffic control scheme tries to satisfy the call level traffic first subjected to the downlink outage probability constrains. The background data packets can be backlogged and they use the remaining available capacity to transmit. We refer this as class-\(K+1\) service. Assume that in the current cell, the inter-arrival time of class-\(K+1\) data packets is exponentially distributed with mean \(1/\lambda_{data}\) and the arrived packets would be placed in a queue with a maximum queue length \(L\).

Let \(n_{act} = (n_{act,1}, n_{act,2}, \ldots, n_{act,K})\) as the number of connection level users that are currently active. All the remaining available channels in terms of code words \(n_d\) could be utilized for the data transmission.

Generally speaking, background data traffic requires a more stringent BER (Bit Error Rate) compared to its real-time service counterparts, thus a higher SIR level needs to be guaranteed for the class-\(K+1\) traffic transmission. Assume the packet currently transmitting is destined to the \(i\)-th user in a cell. Using a similar formulation as in Section 3.2.1, the downlink SIR requirement for each class-\(K+1\) transmission code word is represented as,

\[
\gamma_{K+1} = \frac{(\phi_{K+1,i} \beta P_t \zeta_{0,K+1,i}) / R_h}{(1 - \nu)(1 - \phi_{K+1,i} \beta) P_t \zeta_{0,K+1,i} + \sum_{j=1}^{M} P_t \zeta_{j,K+1,i}) / W_d}. \tag{3.11}
\]

Thus, the fraction of power for a code word to transmit data packet is,

\[
\phi_{K+1,i} = f_d (1 - \nu + \sum_{j=1}^{M} \frac{\zeta_{j,K+1,i}}{\zeta_{0,K+1,i}}) \tag{3.12}
\]

with \(f_{K+1} = \frac{\gamma_{K+1}}{\beta(G_d + \gamma_{K+1}(1-\nu))}\).
As \( n_{act} = (n_{act,1}, n_{act,2}, \ldots, n_{act,K}) \) connection level calls are active in the cell, the outage constraint could be written as,

\[
P_{out} = P\{\sum_{k=1}^{K} \sum_{i=1}^{n_{act,k}} \phi_{ki} + \sum_{i=1}^{n_{d}} \phi_{K+1,i} > 1\} \tag{3.13}
\]

where, \( n_{d} \) is the number of code words currently used for class-\( K + 1 \) data transmission.

Again, let \( \Phi' = \sum_{k=1}^{K} \sum_{i=1}^{n_{act,k}} \phi_{ki} + \sum_{i=1}^{n_{d}} \phi_{K+1,i} \). The outage constraint \( P_{out} \) could be approximated by a Gaussian distribution and obtained through the tail distribution of \( \Phi' \),

\[
P_{out} = \int_{1}^{\infty} \frac{1}{\sqrt{2\pi\nu_{x'}}} e^{-(x-n_{df})^2/2\nu_{x'}} dx.
\]

The maximum number of code words (Please refer to Appendix C) that could be used simultaneously for data transmissions, denoted as \( n_D \), is obtained subjected to the constraint \( P_{out} \leq \delta_{out} \).

As a result, when \( n_{act} \) connection level calls are active in the cell and the data queue is not empty, \( n_D \) code words could be used simultaneously for the transmission of a head of line data packet, where the transmission rate is \( n_D \cdot R_b \) with \( R_b \) as the transmission rate for a code.

### 3.4 Performance Analysis

In the proposed scheme, time sensitive applications are given higher priority and admitted at the connection level. HB reservation is applied to reserve some channels exclusively to the handoff traffic to ensure QoS for handoff calls. Background data packets arriving at the cell are queued in a data queue first, and the head of line packet makes use of the remaining available capacity, in terms of the maximum number of code words that can be transmitted simultaneously with the current active calls in the cell, for transmission.
3.4.1 Connection Level Services

- Flow Balance Equations

Under the assumptions of Poisson arrival and exponentially distributed service and dwell time of the calls, the system could be modelled as a multidimensional birth-death process, with the stead-state given by \( s = (n_1, n_2, \ldots, n_K) \).

Accordingly, we define the following possible successor state from state \( s \) for class-\( k \) (\( 1 \leq k \leq K \)) users,

\[
\begin{align*}
  s_{k^+} &= (n_1, n_2, \ldots, n_{k-1}, n_k + 1, n_{k+1}, \ldots, n_K) \\
  s_{k^-} &= (n_1, n_2, \ldots, n_{k-1}, n_k - 1, n_{k+1}, \ldots, n_K).
\end{align*}
\]

As with the connection admission strategy we discussed in previous sections, the state transition could be triggered by any of the following events:

1. The arrival of a new class-\( k \) call, with the state transition rate \( q_{n,k} \), which would cause state transition from state \( s \) to state \( s_{k^+} \). \( q_{n,k} \) is expressed as,

\[
q_{n,k}(s) = \lambda_{n,k}, \quad \text{if } P_{out}(s + R(s) + e_k) < \delta \quad (3.14)
\]

2. The arrival of a class-\( k \) handoff call, with the state transition rate \( q_{h,k} \), which would also cause state transition from state \( s \) to state \( s_{k^+} \). \( q_{h,k} \) is expressed as,

\[
q_{h,k}(s) = \lambda_{h,k}, \quad \text{if } P_{out}(s + e_k) < \delta \quad (3.15)
\]

3. The release of an on-going call either by departure to the neighboring cells or termination of an on-going call with state transition rate \( q_{r,k} \), which would cause state transition from state \( s \) to state \( s_{k^-} \) and \( q_{r,k} \) is expressed as,
\[ q_{r,k}(s) = n_k(\mu_{d,k} + \mu_{c,k}). \quad (3.16) \]

Let \( \pi(s) \) be the stationary probability of state \( s \). The flow balance equation could be written as,

\[ \pi(s) \sum_{k=0}^{K} \{ q_{n,k}(s) + q_{h,k}(s) + q_{r,k}(s) \} \]

\[ = \sum_{k=0}^{K} I_{n_k>1} \cdot \pi(s_{k-}) \cdot \{ q_{n,k-}(s_{k-}) + q_{h,k-}(s_{k-}) \} \]

\[ + \sum_{k=0}^{K} I_{P_{out}(s+\epsilon_k)<c} \cdot \pi(s_{k+}) \cdot q_{r,k+}(s_{k+}), \quad \forall s \in S \quad (3.19) \]

\[ (3.20) \]

(Here, \( I_c \) is the indication function so that \( I_c = 1 \) when the condition \( c \) is true.) The sum of steady-state probabilities satisfies

\[ \sum_{s \in S} \pi(s) = 1. \quad (3.21) \]

- **Performance Measures**

The most important measurements for connection level services are new call blocking probabilities and handoff call dropping probabilities.

Let \( P_{\text{block},k} \) denote the blocking probability of class-\( k \) new calls. Thus,

\[ P_{\text{block},k} = \sum_{s \in S} I_{(s+R(s)+\epsilon_k) \notin S} \pi(s). \quad (3.22) \]

Let \( P_{\text{drop},k} \) denote the dropping probability of class-\( k \) handoff calls. Thus,

\[ P_{\text{drop},k} = \sum_{s \in S} I_{(s+\epsilon_k) \notin S} \pi(s). \quad (3.23) \]
Another measurement is the average total transmission rates for connection level services, which is

\[
U = \sum_{s \in S} \pi(s) \sum_{k=1}^{K} \alpha_k n_k c_k^d R_b.
\]  \hspace{1cm} (3.24)

### 3.4.2 Background Data Traffic

According to Section 3.3, \( n_D \) codes are used simultaneously to transmit the head of line data packet, i.e. the transmission rate \( R_{n_D} = n_D \cdot R_b \). Assume that all packets have a same length \( S_p \). The time required for the data transmission is \( T_{data} = \frac{S_p}{n_D \cdot R_b} \), and then the service rate of the queue is \( \mu_{data} = \frac{1}{T_{data}} \). With the assumption of Poisson arrival of data packets with mean time \( 1/\lambda_{data} \), the data service could be modelled as an \( M/D/1/L \) queue, with the steady state probability \( p'_n \) calculated in Appendix D. Here, we assume that the on-off state change of those connection level services is much slower than packet transmission in the data queue, such that the data queue would finally reach the steady state.

Thus, the average system data queue occupancy is,

\[
L_{data} = \sum_{n=1}^{L} n p'_n.
\]  \hspace{1cm} (3.25)

- Performance Measurements

1. Average transmission rate of the data packets

Assume when \( n_{act} = (I_1, I_2, ..., I_K) \) users are active in the cell, \( n_D \) code words could be applied for data transmission. In the case when the data queue is non-empty, the head of line packet would be transmitted with rate \( R_{n_D} = n_D \cdot R_b \). Thus, the average rate for background data service is,
\[ R_{BD} = \sum_{s \in S} \pi_s \cdot \sum_{I_1=0}^{n_1} \ldots \sum_{I_K=0}^{n_K} \left( \begin{array}{c} n_1 \\ I_1 \end{array} \right) \alpha_1^{I_1}(1-\alpha_1)^{n_1-I_1} \ldots \left( \begin{array}{c} n_K \\ I_K \end{array} \right) \]

\[ \alpha_K^{I_K}(1-\alpha_K)^{n_K-I_K} \cdot (1 - P_0^{n_D}) \cdot R_n, \]

where \( \alpha_K \) is the voice activity factor for class-\( k \) call; \( P_0^{n_D} \) is the probability that the data queue is empty at the time \( n_D \) codes is available.

2. Average loss rate of data packets

\[ P_{drop}^{BD} = \sum_{s \in S} \pi_s \cdot \sum_{I_1=0}^{n_1} \ldots \sum_{I_K=0}^{n_K} \left( \begin{array}{c} n_1 \\ I_1 \end{array} \right) \alpha_1^{I_1}(1-\alpha_1)^{n_1-I_1} \ldots \left( \begin{array}{c} n_K \\ I_K \end{array} \right) \]

\[ \alpha_K^{I_K}(1-\alpha_K)^{n_K-I_K} \cdot P_L^{n_D}. \]

3. Average throughput of the system

The average throughput could be expressed as the total stationary transmission rates for connection level and background data services,

\[ U_T = \sum_{s \in S} \pi_s \cdot \sum_{I_1=0}^{n_1} \ldots \sum_{I_K=0}^{n_K} \left( \begin{array}{c} n_1 \\ I_1 \end{array} \right) \alpha_1^{I_1}(1-\alpha_1)^{n_1-I_1} \ldots \left( \begin{array}{c} n_K \\ I_K \end{array} \right) \]

\[ \alpha_K^{I_K}(1-\alpha_K)^{n_K-I_K} \cdot [(1 - P_0^{n_D}) \cdot R_n + \sum_{k=1}^{K} I_k \cdot R_k], \]

where \( R_k \) is the transmission rate for class-\( k \) calls.

### 3.5 Numerical Results

The proposed call admission scheme can serve an arbitrary number of connection level classes of traffic. For simplicity, we present in this section numerical examples with only two connection level services, namely class-1 and class-2 calls.
Background data packets are queued in a data queue and make use of the remaining available code words for transmission. The traffic parameters used in this section are based on those for the spectrum calculation for IMT-2000 systems [79]. For convenience, those traffic specific parameters for call level services and background data service are listed in Table 3.1 and Table 3.2, while system specific parameters are given in Table 3.3.

<table>
<thead>
<tr>
<th>Table 3.1: Traffic Model and Its Parameters</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Rate</td>
<td>16 kbps</td>
<td>192 kbps</td>
</tr>
<tr>
<td>Activity Probability $\alpha$</td>
<td>0.5</td>
<td>0.03</td>
</tr>
<tr>
<td>Portion of New Call Arrivals</td>
<td>98%</td>
<td>2%</td>
</tr>
<tr>
<td>Average Call Holding Time $1/\mu_c$</td>
<td>120 sec</td>
<td>3000 sec</td>
</tr>
<tr>
<td>SIR Requirement</td>
<td>7db</td>
<td>9db</td>
</tr>
<tr>
<td>Average Moving Speed</td>
<td>60km/h</td>
<td>20km/h</td>
</tr>
<tr>
<td>Service Example</td>
<td>Voice</td>
<td>Medium Multimedia</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3.2: Traffic Parameters For Background Data Service</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet Length</td>
<td>1600 bits</td>
</tr>
<tr>
<td>SIR Requirement</td>
<td>10 db</td>
</tr>
<tr>
<td>Max Data Queue Length</td>
<td>20 packets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3.3: System Parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Definition/Description</td>
</tr>
<tr>
<td>$W_d$</td>
<td>Downlink spreading bandwidth width</td>
</tr>
<tr>
<td>$R_b$</td>
<td>Basic code channel rate on both uplink and downlink</td>
</tr>
<tr>
<td>$G^d$</td>
<td>Downlink spreading processing gain</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>Variation of lognormal shadowing effects</td>
</tr>
<tr>
<td>$m$</td>
<td>Transmitting path loss exponent</td>
</tr>
<tr>
<td>$\beta$</td>
<td>fraction of TTP for traffic channel</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Downlink outage probability threshold value</td>
</tr>
<tr>
<td>$r_1$</td>
<td>The radius of a cell</td>
</tr>
</tbody>
</table>

Table 3.4 illustrates the connection level capacity comparison with or without considerations on the statistical factors, including voice activity factor (VAF).
Table 3.4: Capacity Comparison

<table>
<thead>
<tr>
<th>VAF; LNS</th>
<th>Class-1 Calls Only</th>
<th>Class-2 Calls Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max: 44</td>
<td>Max: 10</td>
<td></td>
</tr>
<tr>
<td>Avg: 22 (i.e. 44*0.5)</td>
<td>Avg: 0.3 (i.e. 10*0.03)</td>
<td></td>
</tr>
<tr>
<td>No VAF; LNS</td>
<td>Max: 27</td>
<td>Max: 1</td>
</tr>
<tr>
<td>No VAF; No LNS</td>
<td>Max: 33</td>
<td>Max: 1</td>
</tr>
<tr>
<td>VAF: Voice Activity Factor LNS: Log-Normal Shadowing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and log-normal shadowing (LNS), under a homogeneous downlink cellular system. The results demonstrate that the statistical behavior of users’ activity and channel conditions can significantly impact on the number of admissible calls. For example, to ensure the required outage probability, a maximum of 44 class-1 calls are allowed, considering both the voice activity factor 0.5 and the log-normal shadowing. Thus, on average, max 22 class-1 calls could transmit in the cell, which is lower than the situation without VAF but with LNS (27) or without any statistic factor (33) considerations. Furthermore, lower VAF 0.03 for class-2 calls incurs greater variance to system behavior, thus resulting in a larger reduction of the number of admissible users, i.e. on average 0.3 class-2 calls compared to No VAF case 1 class-2 calls allowed. As a result, to ensure the QoS of connection level services, it is important to take those statistical factors into account. Also, as shown in Table 3.4, the spectrum utilization for connection level services might be low when considering those stochastic behaviors. Thus a background data queue which makes use of the remaining available capacity at this time would greatly improve the system utilization.

Figure 3.1 considers only the connection level services and shows the new call blocking and handoff call dropping probabilities according to the changing overall new call arrival rate under both with reservation (the reservation factor $(\delta_1, \delta_2) = (0.5, 0.1)$ ) and without reservation. As specified in Table 3.3, the proportion of class-2 new calls is 2% of overall new call arrivals. For the control with reservation, as capacity is reserved, a higher priority is provided to handoff
Figure 3.1: Call blocking/dropping probability according to new call arrival rate \( \lambda_n = \lambda_{n,1} + \lambda_{n,2} \)

calls. At any given arrival rate, the new call blocking probability is higher than the handoff call dropping probability for each class. Furthermore, the admission of a class-2 call with high transmission rate causes a higher downlink SIR outage probability, inherently class-1 calls have higher priority over class-2 calls and the call blocking (handoff dropping) probability is lower than that of class-2 calls as shown in the figure under the given traffic arrival proportion. For control without reservation, the call blocking and handoff dropping probability should be the same, and the curve lies between the blocking and dropping curve for control with reservation. Furthermore, with the increase of the overall new call arrival rate, both the new call blocking and handoff call dropping probability increase.

Figure 3.2 presents the connection level new call blocking and handoff call dropping probabilities according to the changing reservation factor for the class-1 call. The total arrival rate here is \( \lambda_n = 0.1 \). With the increase of factor \( \delta_1 \) (See Section 3.2.2), more bandwidths are reserved for the class-1 calls, and the handoff dropping probability is decreasing, at the same time, new call blocking proba-
Figure 3.2: Call blocking / dropping probability with the changing of reservation factor for $\delta_1, (\delta_2=0.1)$

bilities inevitably increase as shown in the figure. One thing we should mention here is that, as both classes share the same reservation space, the handoff dropping probability for class-2 calls also decreases. Such reservation space sharing is reasonable as inherently a class-1 call has higher priority over a class-2 call. If a strict separation of reservation for each class is required, we could achieve this by reserving capacity exclusively for each class as in [38].

Figure 3.3 shows the new call blocking and handoff call dropping probabilities according to the changing arrival proportion, in terms of the arrival rate, for class-2 calls (total new call arrival rate $\lambda_n=0.1$). Due to the higher data transmission rate, the greater variance in voice activity factors and the larger SIR requirement, the admission of class-2 calls may cause a higher downlink SIR outage probability. As a result, with the increasing new call arrival rate for class-2 calls, the system would endure a significant increase in call blocking and handoff dropping probabilities. Thus, the proportion of high rate calls' arrival can not be too high.

Figure 3.4 shows the background data packets dropping (by the data queue)
probability according to the changing of data packets arrival rate with (reservation factor \((\delta_1, \delta_2) = (0.5, 0.1)\) and \((1, 0.5)\)) or without reservation. The total arrival rates \(\lambda_n=0.1\). As priority is provided to connection level services by admitting class-1 and class-2 calls first and utilizing the remaining capacity for packets’ transmission, the background data traffic indeed does not affect the transmission of connection level services and the blocking/dropping probabilities for both class-1 and class-2 calls are constant. On the other hand, with the increasing packet arrival rate, there is a higher probability that a data packet encounters a full queue at its arrival and thus incurs greater packet loss probability. Furthermore, as shown in the figure, a higher reservation factor causes greater capacity to be reserved and lower bandwidth utilization. Thus, background data packets may obtain more codes for transmission and the packet loss curve is lower.

Figure 3.5 shows the packets loss probability according to the changing data packets arrival rate, where three sets of total connection level call arrival rates, \(\lambda_n =0.05, 0.1\) and 0.3, are tested and the reservation factor is \((\delta_1, \delta_2) = (0.5, 0.1)\). Generally speaking, increasing the connection level arrival rates increases
Figure 3.4: Background data packets loss probability according to the data packets arrival rate.

the connection level system utilization with decreased codes for background data transmission. As a result, higher $\lambda_n$ incurs greater packet loss probability. Again, the packet loss ratio goes up with the increase of the packet arrival rate as illustrated in Figure 3.5.

Figure 3.6 shows the average system transmission rate, considering three $\lambda_n$ settings and reservation factor (0.5, 0.1). At first, increasing the packet arrival rate increases the system utilization. When the offered packet load goes above a certain level, the system saturates and the probability that the data queue becomes empty is extremely low. Thus, all the remaining capacity after transmitting active class-1 and class-2 calls are fully utilized and the utilization does not change beyond that level and higher packets arrival rate only causes greater packet loss, as shown in the figure.

Figure 3.7 illustrates the average connection level services transmission rate and total system transmission rate including background data transmission at any time in the cell, according to the changing new call arrival rate. The utilization either without reservation or with two sets of reservation, $(\delta_1, \delta_2) = (0.5,$
Figure 3.5: Packets dropping probability according to the data packets arrival rate.

0.1) and (1, 0.5), are presented. When the new call arrival rate is low (around 0.05), the system is light loaded and thus the results are very similar. It is obvious that the control without reservation presents the highest system utilization. However, because of the increased blocking for new call arrivals, the slope of the utilization curve (No Reservation) slightly decreases as shown in the figure. When reservation is adopted, with the increase of new call arrival rate, more capacity should be reserved for in-coming handoff calls, causing the utilization curve to bend greatly. Here, the higher reservation factor (the lowest curve) results in lower system utilization. Furthermore, the utilization of background data packets' transmission is actually very high. The reason is that we take voice activity into consideration to ensure the long time transmission quality of the channels in the admission of connection level services, which greatly lower the connection level services' utilization of the system spectrum resources.
Figure 3.6: Average data packet transmission rate (kbps) according to the packet arrival rate

3.6 Summary

In this work, we study downlink traffic control for the MC-CDMA cellular networks. The key contributions in this study are: 1) we quantitatively demonstrate that several statistic parameters, especially voice activity factors and shadowing in propagation, have significant impact on the connection level performance; 2) we showed that data traffic can be best handled at the packet-level as background transmission by taking advantage of the statistical variations, in which the results show that this leads to significant improvement in bandwidth utilization.

The main points concluded from this study are: 1) the traffic control in MC-CDMA has to take into account characteristics of more realistic traffic, which can potentially lead to a significant reduction in the capacity derivation; 2) data traffic, due to its large statistical variation, should be best handled at packet level. Using call admission control at the connection level for such traffic results in a significant reduction of the number of users that can be accommodated in the system.
Figure 3.7: Utilization (kbps) according to the new call arrival rate
CHAPTER 4

ADAPTIVE CELL SECTORING IN DS-CDMA SYSTEMS

In this work, we mainly focus on the cell sectoring on the uplink in DS-CDMA cellular networks. We first formulate cell sectoring into an optimization problem and solve it with a dynamic programming algorithm. With the information on the traffic distribution and number of sectors, we can optimally design the sector boundaries such that the minimum TTP can be achieved in the systems subject to the SIR constraint for each user.

The above solution is optimal; however, in practice, it incurs high computation complexity. Hence, it does not scale well to high-density cells. We believe that, for such an adaptation system, reducing the execution time of the sectoring algorithm is also of primary importance. Furthermore, we observe that, in practice, the sector boundaries would be better across low-density regions. The reasons are as follows: Firstly, separating very close users is impractical given the difficulty in achieving precise directivity; secondly, if a boundary goes through an extremely high-density region, very frequent oscillations of users across the boundary might potentially occur, resulting in a non-stable sector partition; in other words, the adaptive sectoring algorithm needs to be executed very frequently. With these considerations, we then propose a Cluster-based Sectoring (CS) algorithm to reduce the computation complexity under high-density traffic distribution, while at the same time trying to avoid the sector boundaries across those neighboring users whose physical locations are close in angular distances. The dynamic programming technique is also applied in the CS algorithm to solve the MinTTP problem.

The basic idea in the proposed CS algorithm is to partition the users into several clusters, where the users in the same cluster are geographically nearby
in the angular distance with their neighboring users. For a high-density cluster that exceeds certain threshold, we apply an Angular-based Partitioning (AP) algorithm to further separate it into several sub-clusters to reduce the TTP. Generally speaking, the number of clusters is much smaller than the number of users, but larger than that of sectors. We then apply a dynamic programming (DP) algorithm to the clusters, not the users, to find the sector partition. Whereas the CS algorithm is not globally optimal, its time complexity is significantly reduced compared with its optimal counterpart. In particular, under the high-density traffic distribution, the main complexity from DP calculation does not depend on the number of users in the cell, making it practically feasible. We carry out detailed analysis of the complexity for the algorithm, and with extensive simulations, we find that, compared to the optimal solution, the complexity of the CS algorithm is greatly reduced. The performance degradation of the sub-optimal solution is limited and usually less than 10%, which is particularly efficient from a practical point of view.

The rest of this chapter is organized as follows. Section 4.1 describes the system model and formulates the optimal sectoring problem. In section 4.3, we use a dynamic programming algorithm to solve the MinTTP optimization. The cluster-based sectoring algorithm and detailed complexity analysis are presented in Section 4.4. In Section 4.5, we apply the algorithm to multi-rate applications in MC-CDMA system. In Section 4.6, we discuss some practical issues related to cell sectoring. We investigate the performance of our algorithm through numerical simulations in Section 4.7. Finally, we conclude the paper in Section 4.8.

4.1 System Model and Formulation

4.1.1 System Model

We mainly focus on sectoring in the uplink of a single cell DS-CDMA system, where each user specifies a minimum tolerable QoS which can be mapped into
a lower bound Signal-to-Interference Ratio (SIR) level. Adaptive antenna arrays are applied to the base station to partition the cell into multiple sectors, where each directional antenna receives signals from the users within the particular sector it serves, resulting in a spatial isolation of users in the cell. Assume ideal directivity for the adaptive antenna arrays, and thus no interference occurs between sectors. As a result, each user's receiver only experiences interference from other users within the same sector. Our objective here is to find the best partition of the cell in order to minimize the total transmission power (TTP) for the whole system, while at the same time ensuring the SIR requirement for each user. We first list the notations used in this paper in Table 1 before formally describing the MinTTP problem.

<table>
<thead>
<tr>
<th>Table 4.1: List of Notations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
</tr>
<tr>
<td>$K$</td>
</tr>
<tr>
<td>$\eta$</td>
</tr>
<tr>
<td>$\alpha_j$</td>
</tr>
<tr>
<td>$G_i$</td>
</tr>
<tr>
<td>$\chi_i$</td>
</tr>
<tr>
<td>$\chi(i)$</td>
</tr>
<tr>
<td>$\gamma_i$</td>
</tr>
<tr>
<td>$p_i h_i$</td>
</tr>
<tr>
<td>$K_i$</td>
</tr>
<tr>
<td>$K_{Cl}$</td>
</tr>
<tr>
<td>$M_i$</td>
</tr>
<tr>
<td>$\chi^C_i$</td>
</tr>
<tr>
<td>$\chi^S(i)$</td>
</tr>
<tr>
<td>$\delta$</td>
</tr>
<tr>
<td>$\tau$</td>
</tr>
<tr>
<td>$\kappa$</td>
</tr>
</tbody>
</table>
4.2 Optimal Sectoring Problem

Let $K$ be the number of users and $N$ the number of sectors partitioning the cell. The SIR requirements for user $i$ at the base station receiver can be represented by the bit energy to total interference plus thermal noise density ratio, $E_b/N_0$ [1]

\[
\left( \frac{E_b}{N_0} \right)_i = \frac{G_i p_i h_i}{\sum_{j=1, j \in X(i), j \neq i}^K \alpha_j p_j h_j + \eta} \geq \gamma_i.
\]  

(4.1)

Assume perfect power control, thus the received power at the base station from each user should be the same. A common target SIR requirement $\gamma$ is assigned to all the users in the cell, i.e. $\gamma_i = \gamma$. Further assume identical voice activity factor $\alpha_i = \alpha$ and bit rate $R_i = R$. As $G_i = W/R_i$, this results in the same processing gain $G_i = G$. The optimization problem can be formulated as follows:

\[
\min_{\theta, P} \sum_{s=1}^N \sum_{i \in X_s} p_i,
\]  

(4.2)

subject to,

\[
\frac{G p_i h_i}{\sum_{j=1, j \in X(i), j \neq i}^K \alpha_j p_j h_j + \eta} \geq \gamma_i, \quad i = 1, \ldots, K,
\]  

(4.3)

\[
P \geq 0, \quad P = (p_1, p_2, \ldots, p_K),
\]  

(4.4)

\[
\sum_{i=1}^N \theta_i = 2\pi, \quad \theta = (\theta_1, \theta_2, \ldots, \theta_N).
\]  

(4.5)

It has been shown that the system capacity is maximized when each mobile's transmitted power level is controlled in such a way that its signal arrives at the
base station with the minimum required SIR. In addition, within each sector, the received power for each user is controlled at an equal level at the base station, i.e., \( p_i h_i = p_j h_j \), where \( j \in \chi(i) \) [69]. Therefore, we can rewrite the formulation (4.2) as follows:

\[
\text{(MinTTP)} \quad \min_{\theta, \eta, \gamma} \sum_{i=1}^{N} \frac{\eta \gamma}{G - (K_i - 1)\alpha_\gamma} \sum_{j \in \chi_i} \frac{1}{h_j}
\]  

subject to,

\[
K_i < \frac{G}{\alpha \gamma} + 1, \quad \sum_{i=1}^{N} K_i = K
\]  

\[
\sum_{i=1}^{N} \theta_i = 2\pi, \quad \theta = (\theta_1, \theta_2, ..., \theta_N).
\]  

### 4.3 Optimal Solution

We present a dynamic programming (DP) algorithm to solve the above optimization problem. The DP algorithm is also the basis for the cluster-based sectoring in subsequent sections.

As discussed in the previous subsection, with the information on the \( K \)-user distribution and predefined sector number \( N \) in the cell, the optimization problem can be represented as:

\[
\min_{\theta, \eta, \gamma} \sum_{i=1}^{N} \sum_{j \in \chi_i} p_j = \min_{\theta, \eta, \gamma} \sum_{i=1}^{N} \frac{\eta \gamma}{G - (K_i - 1)\alpha_\gamma} \sum_{j \in \chi_i} \frac{1}{h_j}.
\]  

Without loss of generality, we can choose one user as the start user \( U_1 \). From this user, label the remaining users counter-clockwise from \( U_2 \) to \( U_k \), where \( k \) is
the total number of users. We set the start-partitioning point between user $U_k$ and user $U_1$, denoted as $SP = (U_k, U_1)$. Let $\varphi_{SP}(k, n)$ represent the MinTTP of $k$ users partitioned with $n$ sectors for the given start partitioning point $SP$. We have the following recurrence relation,

$$
\varphi_{SP}(k, n) = \begin{cases} 
\frac{\eta \gamma}{\alpha - (k - 1) \alpha \gamma} \sum \frac{1}{h_j}, & n = 1, 1 \leq k \leq \frac{\alpha}{\alpha \gamma} + 1 \\
\min_{n-1<i\leq k} \left\{ \varphi(i-1, n-1) + \frac{\eta \gamma}{\alpha - (k-1) \alpha \gamma} \sum_{j \in \chi(i)} \frac{1}{h_j} \right\}, & n > 1, n \leq k \leq \frac{\alpha}{\alpha \gamma} + i \\
+\infty, & Otherwise
\end{cases}
$$

(4.10)

where, $i$ is the partition point for the $n$-th sector (including $U_i$) and $\chi(i)$ is the set of consecutive users in the same sector with user $i$, i.e. $\{U_i, U_{i+1}, ..., U_k\}$. Please refer to Appendix E for the correctness proof of the recurrence function 4.10.

**Lemma 1** $\min_{1 \leq s \leq K} \varphi_{SP}(K, N)$ is the solution to the optimization problem MinTTP (4.6), where the start partitioning point $SP = \begin{cases} (U_K, U_1), & s = 1 \\
(U_{s-1}, U_s), & s > 1
\end{cases}$

The proof of the lemma is obvious by continuously changing the start partitioning boundary, one for each user, and computing the similar optimal solution $\varphi_{SP}(k, n)$ according to the $k$ consecutive start partitioning point in a cyclic way. For the system with $K$ users to be partitioned into $N$ sectors, MinTTP $= \min_{1 \leq s \leq K} \varphi(K, N)$.

The complexity of the dynamic programming algorithm of computing each $\varphi(K, N)$ is $O(NK^2)$, so the complexity of computing the optimal solution for MinTTP is $O(NK^3)$, which is the same as that of the result obtained in [73].

52
4.4 The Cluster-Based Sectoring Algorithm

In this section, we propose a Cluster-based Sectoring (CS) algorithm as a sub-optimal solution to the MinTTP problem (4.6), which might greatly reduce the complexity in the optimal solution and guide the sector boundaries across the low-density regions. We first divide the users into several clusters based on a nearby criterion, in terms of the relative angular distances between neighboring users. Here, we assume that each sector contains at least one user, and the users within a particular cluster are placed into the same sector. In addition, given the MinTTP objective and the capacity constraint $K_i < G/\alpha \gamma + 1$, we partition those "congested" clusters, i.e., clusters containing too many users, into sub-clusters with an Angular-based Partitioning algorithm and put the sub-clusters back into the overall cluster set. A dynamic programming algorithm is then applied to the whole set of clusters to obtain the MinTTP partition.

4.4.1 Cluster-based Sectoring

A cluster is a sequence of users, where each user's angular distance to its neighbor(s) is less than a predefined threshold $\delta$. For all the users in the cell, we first divide them into a set of clusters. This can be done in two steps, namely coarse clustering and angular-based partitioning. The coarse clustering algorithm is performed as follows: Firstly, arbitrary choose one user as the starting user for the first cluster. From this user, we check the remaining users in the cell one by one in a counter-clockwise manner. If the current user's angular distance to the previous one is less than $\delta$, it is added into the current cluster, i.e. the same cluster the previous user is in. Otherwise, we start a new cluster by setting the current user as the first user and repeat the above step for user additions. This procedure stops when each user in the cell has been included into one and only one cluster. Finally, if the angular distance between the starting user of the first cluster and the end user of the last cluster is less than $\delta$, we combine these two clusters into one. After the coarse clustering, the clusters containing too many
users are further partitioned by an angular-based partitioning algorithm, with details given later.

When both coarse clustering and angular-based partitioning are done, we would then obtain a set of clusters. Sectoring on the clusters, namely cluster-based sectoring, is then performed for the MinTTP problem, which can be shown as follows:

\[
\min_{\theta,P} \sum_{i=1}^{N} G - \left( \sum_{C_i \in \chi_i^C} K_{C_i} - 1 \right) \alpha \gamma \sum_{c_j \in \chi_i^C} \frac{1}{h_{C_j}}
\]

(4.11)

where \( K_{C_i} \) is the number of users in the cluster \( C_i \), \( \chi_i^C \) is the set of clusters in the sector \( i \) and \( h_{C_j} = \sum_{k \in C_j} \frac{1}{h_k} \).

A dynamic programming algorithm similar to what we use in the optimal sectoring (Section 4.3) is then designed with the recurrence relation shown as (4.12), by regarding a cluster as one particular integrated user. Correctness proof can be found in Appendix F.

\[
\varphi^c_{SP}(k, n) = \begin{cases} 
\frac{\eta \gamma}{\alpha - \left( \frac{m \eta \gamma}{K_{C_i} - 1} \alpha \gamma \sum_{j=1}^{m} h_{C_j} \right)} \sum_{j=1}^{m} h_{C_j}, & n = 1, 1 \leq \sum_{t=1}^{m} K_{C_t} < \frac{G}{\alpha \gamma} + 1 \\
\min_{n-1 < i \leq m} \left\{ \varphi^c(i - 1, n - 1) + \frac{m \eta \gamma}{\alpha - \left( \sum_{t=i}^{m} \frac{1}{h_{C_t}} \right)} \sum_{j=i}^{m} \frac{1}{h_{C_j}} \right\}, & n > 1, 1 \leq \sum_{t=i}^{m} K_{C_t} < \frac{G}{\alpha \gamma} + 1 \\
+\infty, & \text{Otherwise}
\end{cases}
\]

(4.12)

where, \( m \) is the size of the clusters, \( C_i \) is the partition point for the \( n \)-th sector (including \( C_i \)) and \( \chi^C(i) \) is set of clusters in the same sector with cluster \( C_i \). Assume a total of \( M \) clusters to be partitioned into \( N \) sectors in the system, \( \min_{1 \leq s \leq M} \varphi^c_{SP}(M, N) \) is the solution to the optimization problem (4.11), where the start partition point \( SP = \left\{ \begin{array}{ll} (C_M, C_1) & s = 1 \\
(C_{s-1}, C_s) & s > 1 \end{array} \right. \).
4.4.2 Angular-based Partitioning

When some high-density regions or "hot spots" exist in the cell, the coarse clustering method might result in very large clusters. If too many users are clogged in one cluster, thus placed into the same sector later, they would likely incur great interfere among one another, resulting in high TTP. In certain situation, the number of users in one cluster may exceed the capacity constraint \( \frac{C}{\alpha \gamma} + 1 \). It is thus necessary to further partition those large clusters into some smaller sub-clusters.

For a large cluster that needs partitioning, we first roughly decide the number of users that can be supposed in a sector with equalized sectoring [73]. The number of sub-clusters is determined by the heuristic \( m_{C_i} = \left\lfloor \frac{K_{C_i}}{K/N} + \sigma \right\rfloor \), where \( \sigma \) is a predefined threshold for adjustment. Based on user distribution information and the number of sectors, we further split the cluster by an angular-based partitioning algorithm with the objective of minimizing the total transmission power of all users in the original cluster. Note that, due to the physical constraint of directivity accuracy, it does not make too much practical sense to set two neighboring boundaries within a precision threshold. So, we only allow the sectoring boundaries to take discrete incremental angular values. In our algorithm specifically, we assume the minimum angular interval between any pair of sectoring boundaries, i.e. the precision threshold, to be \( \tau \), namely the Angular Changing Interval. Assume \( \kappa \) is the set of all feasible partitioning intervals (i.e. at least one user in each interval). Our algorithm is then applied to set \( \kappa \), instead of the set of users in the cluster, to split the original cluster into \( m_{C_i} \) sub-clusters. This can also be done by a similar dynamic programming algorithm. Note here, instead of sectoring based on the users, we set the partition on the feasible splitting intervals represented by \( \kappa \). The recurrence relation \( \varphi^A(l, n) \) is written as follows:
\[ \varphi^A(l, n) = \begin{cases} 
\frac{n}{\zeta_{-1(\kappa^C-1)\alpha\gamma}} \sum_{j=1}^{K^C} \frac{1}{h_j}, & n = 1, 1 \leq K^C < \frac{\zeta}{\alpha\gamma} + 1 \\
\min_{n-1 < i \leq l} \left\{ \varphi^A(i - 1, n - 1) + \frac{n}{\zeta_{-1(\kappa^C - 1)\alpha\gamma}} \sum_{j \in \chi_n} \frac{1}{h_j} \right\}, & n > 1, 1 \leq \kappa_n < \frac{\zeta}{\alpha\gamma} + 1 \\
+\infty, & \text{Otherwise} \end{cases} \]

(4.13)

where, \( n \) is the number of sectors; \( l \) is the size of splitting set \( \kappa \); \( K^C \) is the number of users in the original cluster; \( \kappa_n \) is the number of users in sector \( n \), i.e. number of users from splitting point \( i \) to \( l \). Later on, the sub-clusters are placed back into the overall cluster set for the cluster-based sectoring, as discussed in the previous subsection.

### 4.4.3 Complexity Analysis

In Section 4.3, it has been shown that the complexity of the optimal sectoring algorithm is \( O(NK^3) \). In this subsection we show that the complexity for the cluster-based sectoring algorithm is much lower. The computation of our algorithm can be divided into two parts:

- **Coarse clustering and angular-based partitioning.**

  The coarse clustering only involves the operation of user scanning, thus the complexity of it is simply \( O(K) \). When performing angular-based partitioning on the very large clusters, as the algorithm is applied to a cluster with sequential users, the starting partitioning point is fixed to the first user of the cluster. Let \( M' \) be the number of clusters requiring further partition, \( K_{C_i} \) be the number of users in cluster \( C_i \), and \( m_{C_i} \) be the number of sub-clusters for \( C_i \). In our algorithm, we set \( m_{C_i} = \left\lfloor \frac{K_{C_i}}{K/N} + \sigma \right\rfloor \), where \( \sigma \) is a predefined threshold for adjustment. Let \( A_{C_i} \) be the overall angular distance between the start and end user in \( C_i \). As there is at least one user
in each sector, we do not consider the angular splitting intervals without any user in between. Thus the size of $\kappa$ is $\min\{K_{C_i}, A_{C_i}/\tau\}$, and the complexity for this part is given by $\sum_{i=1}^{M'} O(m_{C_i} \cdot (\min\{K_{C_i}, A_{C_i}/\tau\})^2)$, subject to $M' \leq N$, $\sum_{i=1}^{M'} A_{C_i} \leq 2\pi$, and $\sum_{i=1}^{M'} K_{C_i} \leq K$. For the extreme case where one large cluster includes all the users, we have

$$\sum_{i=1}^{M'} O(m_{C_i} \cdot \min\{K_{C_i}, A_{C_i}/\tau\}^2) \leq O(N \cdot \min\{K, 2\pi/\tau\}^2) \leq O(NK^2).$$

- Dynamic programming on a set of clusters.

As the dynamic programming algorithm is applied directly to the clusters, the complexity can be largely reduced in the non-uniform high-density case where many users gather together closely. Let $M$ ($M < K$) be the number of clusters in the cell, the complexity to compute $\varphi^S_{SP}(M, N)$ is $O(NM^2)$, thus the MinTTP solution (i.e. $\min_{1 \leq s \leq M} \varphi^S_{SP}(M, N)$) could be obtained in $O(NM^3)$. Here, $M$ is bounded by $\{2\pi/\delta, K\}$.

Combining these two parts, the total complexity of the algorithm is

$$K + \sum_{i=1}^{M'} O(m_{C_i} \cdot \min\{K_{C_i}, A_{C_i}/\tau\}^2) + O(NM^3).$$

It is obvious that the computation effort for the optimal solution mainly comes from the cube of user number, i.e. $K^3$. Generally speaking, the complexity of the coarse clustering and angular-based partitioning (first part) can be ignored as it is one order of magnitude less than that of the cluster sectoring part. Furthermore, in the extremely high-density case, the complexity can be bounded for specified $\delta$ and $\tau$, and does not depend on the number of users. This greatly decreases the computation complexity. On the other hand, in a sparse user distribution, our algorithm gives the optimal solution, and its complexity is no higher than the user-based optimal sectoring algorithm.
4.5 High Rate Data Services

Next generation wireless cellular networks are expected to accommodate multimedia services. To provide multi-class service transmitting at variable rate in DS-CDMA systems, two main techniques are widely deployed, namely Variable Spreading Factor (VSF) and Multi-Code (MC). In a VSF-CDMA system, higher transmission rate can be obtained by lowering processing gain, thus requiring greater transmission power to ensure SIR constraint. In an MC-CDMA system, users that need higher data rates make use of multiple codes, each with a basic bit rate, for transmission in parallel. In this section, we focus on the multi-rate transmission with MC scheme and show that the previously mentioned algorithm can be easily adjusted for multiple rate transmission. Furthermore the result can also be readily extended for VSF-CDMA systems.

As all signals over the radio interface are transmitted at a basic rate, namely $R_b$ with one chip code, the spread processing gain over each code channel is a constant. Assume each user associates with $c_i$ codes for transmission at rate $c_i R_b$ and the same power $p_i$ is allocated to each of the $c_i$ code words. Further assume that orthogonal codes are assigned for each user’s transmission such that for one particular code channel, no self-interference experienced from the other $(c_i - 1)$ parallel code channel in use. This can be done by sub-code catenation technique [35] [53]. Again, assume identical voice activity factor $\alpha_i = \alpha$ and the same SIR requirement $\gamma_i = \gamma$ for each user. The optimal formulation can be written as:

$$\sum_{s=1}^{N} \sum_{i \in \chi_s} c_i p_i$$

subject to,

$$\frac{p_i h_i / R_b}{(\sum_{j=1, j \in \chi(i), j \neq i}^{K} \alpha c_j p_j h_j + \eta)/W} \geq \gamma, \quad i = 1, \ldots, K,$$

58
\[ P \geq 0, \quad P = (p_1, \ldots, p_K), \]  
(4.16)

\[ \sum_{i=1}^{N} \theta_i = 2\pi, \quad \theta = (\theta_1, \ldots, \theta_K). \]  
(4.17)

Let \( G = W/R_\theta \) and assume the same received power for each code in the cell, i.e. \( p_i h_i = p_j h_j \), then,

MinTTP:

\[
\min_{\theta, p} \sum_{i=1}^{N} \sum_{j \in \chi_i} \frac{\gamma \eta}{G - \gamma \cdot \alpha \cdot \sum_{k \in \chi_i, k \neq j} c_k h_j} c_j
\]  
(4.18)

subject to,

\[
\sum_{k \in \chi_i, k \neq j} c_k \leq \frac{G}{\gamma \cdot \alpha}, \quad j \in \chi_i.
\]  
(4.19)

Accordingly, the recurrence relation for the optimal DP solution can be expressed as:

\[
\varphi(k, n) = \begin{cases} 
\eta \gamma \sum_{j=1}^{k} \frac{1}{\alpha - \alpha \gamma} c_j h_j, & n = 1, \sum_{k \neq j} c_k < \frac{G}{\gamma \alpha} \\
\min_{n-1 < i \leq k} \left\{ \varphi(i - 1, n - 1) + \sum_{j \in \chi(i)} \frac{\eta \gamma}{\alpha - \alpha \gamma} \frac{c_j}{r_{i \in \chi(i), i \neq j} c_k h_j} \right\}, & 1 < n, \sum_{i \in \chi(i), i \neq j} c_t < \frac{G}{\gamma \alpha} \\
+\infty, & \text{Otherwise}
\end{cases}
\]  
(4.20)

Generally speaking, there are finite user classes \( \Delta_t \) (\( C_{\text{max}} \leq t \leq 1, C_{\text{max}} \) is the max number of classes the system supported), transmitting at different rates with the number of codes \( C^{\Delta_t} \), the recurrence relation could be further simplified as,
\[ \varphi(k, n) = \begin{cases} 
\eta \gamma \sum_{t=1}^{C_{\text{max}}} \frac{1}{G - \alpha \gamma} \frac{c_j}{c_{\text{max}}} \left( \sum_{t'=1}^{c_{\text{max}}} N_{t'}^{\Delta t'} \sum_{j \in \Delta t'} \frac{c_j}{h_{t_j}} \right), \\
1 = n, \sum_{t'=1}^{C_{\text{max}}} N_{t'}^{\Delta t'} \sum_{j \in \Delta t} \frac{c_j}{h_{t_j}} \left( \sum_{t'=1}^{c_{\text{max}}} N_{t'}^{\Delta t'} \sum_{j \in \Delta t} \frac{c_j}{h_{t_j}} \right), \\
1 < n, \sum_{t'=1}^{C_{\text{max}}} N_{t'}^{\Delta t'} \sum_{j \in \Delta t} \frac{c_j}{h_{t_j}} \left( \sum_{t'=1}^{c_{\text{max}}} N_{t'}^{\Delta t'} \sum_{j \in \Delta t} \frac{c_j}{h_{t_j}} \right), \\
+\infty, \\
\end{cases} \]

where, \( N_{t'}^{\Delta t'} \) is the total number of Class-\( t' \) users in the cell; \( N_{t'}^{\Delta t'} \) is the number of Class-\( t' \) users in the sector \( i \). The sub-optimal cluster-based sectoring algorithm which partitions clusters into sectors may be applied as follows,

\[ \varphi^c(m, n) = \begin{cases} 
\eta \gamma \sum_{t=1}^{C_{\text{max}}} \frac{1}{G - \alpha \gamma} \frac{m}{c_{\text{max}}} \sum_{r=1}^{m} \frac{c_j}{h_{t_j}} \\
1 = n, \sum_{t'=1}^{C_{\text{max}}} N_{t'}^{\Delta t'} \sum_{j \in \Delta t} \frac{c_j}{h_{t_j}} \\
1 < n, \sum_{t'=1}^{C_{\text{max}}} N_{t'}^{\Delta t'} \sum_{j \in \Delta t} \frac{c_j}{h_{t_j}} \\
+\infty, \\
\end{cases} \]

where, \( N_{C_t}^{\Delta t'} \) is the number of Class-\( t' \) users in cluster \( C_t \) and \( \frac{1}{h_{t_j}} = \sum_{k \in C_j} h_{t} \).

### 4.6 Discussion

In the previous sections, we consider a single cell DS-CDMA system with perfect power control, cell sectoring, and static radio propagation. To cope with more realistic systems, several practical issues need to be considered.
4.6.1 Imperfect Sectoring

In the ideal case, a directional antenna only receives signals from the users within the sector it covers. In reality, there are always some overlaps of antenna patterns, incurring cross-interference among neighboring sectors. In other words, the antenna responsible for the current sector may receive signals from those users located in the overlap region of the adjacent sectors. To suppress the additional interference, the transmission power for each user should be increased, resulting in a higher MinTTP. This is another driven force to design the sector boundaries across low density regions. To be precise, we could take into account those mobiles in the overlap region as additional interference.

4.6.2 Physical Constraints

In practice, we might have to consider a transmission power constraint for mobile users, i.e. the power consumption for any user less than a threshold; and spreading angle constraint for antenna patterns, i.e. the sector width within an admissible interval. To ensure the maximum transmission power for each mobile user, one item $p_i \leq P_{\text{max}}$ should be considered in the MinTTP formulation. Similarly, $\theta_{\text{min}} < \theta_i < \theta_{\text{max}}$ helps ensure the sector width constraint.

4.6.3 Imperfect Power Control and Shadowing

A commonly accepted model for imperfect power control and shadowing is the Log-normal approximation, with power control error $L[\sigma_i^2]$ and shadowing effect $L[\sigma^2]$. Here, the variable $10 \cdot \log_{10}(L[\sigma_2^2(\sigma^2)])$ obey Gaussian distribution of $N(0, \sigma_2^2(\sigma^2))$.

Thus, the SIR requirement can be revised according to some statistical constraint on the outage probability, i.e.,

$$\Pr\left\{ \frac{L_i[\sigma_i^2] \cdot p_i (h_i \cdot L_i[\sigma^2])/R_b}{\left( \sum_{j=1,j\in X(i),j\neq i}^{K} \alpha_c p_j (h_j L_j[\sigma^2]) + \eta \right)/W} \geq \gamma_i \right\} \leq \Gamma_{\text{outage}}$$
where, $\Gamma_{outage}$ is the acceptable SIR outage probability ensuring transmission quality and $p_i(h_i \cdot L_i(\sigma^2)) = p_j(h_jL_j(\sigma^2))$.

The outage constraint might be approximated by certain distribution to gain transmission power for each user and optimize MinTTP accordingly.

Though the diversity in channel strength could improve the performance through scheduling transmissions with peak channel quality, due to the difficulties in precisely tracking the channel fluctuations and in providing QoS and fairness to the end users, in practice, we have to average out the randomness with decreased performance.

4.6.4 Multicell Environment

In the previous algorithm, we consider only the single cell system, ignoring the impact of neighboring cells. To extend our result to the multicell environment, we have to take into account the out-of-cell interference [73]. Assume full knowledge of the user distribution in the interfering cells; we could incorporate some parameters indicating the out-of-cell interference in the SIR formulation and derive similar dynamic programming algorithms for the MinTTP optimization.

To conclude, the practical concerns listed above may greatly degrade system performance with lower achievable cell capacity and higher transmission power consumption, thus presenting more challenges to precise parameter prediction and adjustment.

4.7 Numerical Results

In this section, we investigate the performance of our cluster-based sectoring algorithm through simulations. For comparison, we also implement other two partition algorithms, namely, Optimal Sectoring and Fixed Sectoring. For simplicity, in the later part of the section, we denote OS as the optimal MinTTP result obtained from section 4.1, CS (Clustered Sectoring) as the MinTTP result
of the cluster-based sectoring algorithm described in Section 4.3 and FS as the TTP of the traditional fixed sectoring, where all the sectors in a cell are of the same size and fixed orientation.

In our experiment, we use the following parameter settings: processing gain, \( G = 64 \); SIR requirement \( \gamma = 7 \) dB; voice activity \( \alpha = 0.8 \); thermal noise power \( \eta = 1e-12 \); angular changing interval in a cluster for the angular-based partition algorithm \( \tau = 2\pi/360 \) (reflect the accuracy of antennas). The channel gain \( h \) is calculated with a simplified function, modelling path loss for outdoor to indoor and a pedestrian test environment, shown in [78].

We first consider a user distribution shown in Figure 4.1. The results in Table 4.2 indicate that, both OS and CS exhibit much better performance than FS (more than 40% power savings). Compared with OS, the CS solution gives limited degradation (In both 5-sector and 6-sector cases, less than 1%). The sector partitioning results for OS and CS are also shown in Figure 4.1. The sector boundaries are generally crossing low-density regions in the CS solution, while in OS, they may pass through two users very close to each other. In addition, generally speaking, when larger \( \delta \) is chosen, the boundaries more likely cross two users far away. Furthermore, the computational complexity of CS is greatly reduced compared to OS. When we set the angular value for clustering to be \( \delta = 2\pi/360 \), the 36 users are partitioned into 19 clusters; when \( \delta = 2\pi/36 \), only 12 clusters are left (This is obvious as the larger the interval, the more users are likely to be placed into one cluster, resulting in less clusters). Thus, approximately only (i.e. 14.7% as the complexity is about \( O(NM^3) \)) computing effort is required for \( \delta = 2\pi/360 \) case and \((12/36)^3 \) (i.e. 3.7%) for \( \delta = 2\pi/36 \) compared to OS solution (The computing for angular-based partitioning has not been included, as it is much lower than the effort on final cluster-based sectoring). In this case, higher angular distance for clustering (\( \delta \)) helps to save great amount of computation effort while the boundary locations are reasonably better with the cost of little performance degradation. However, this is not always the case. For the "hot spots" case, too large \( \delta \) results in very large clusters for further
angular-based partitioning (In the worst case this might result in only one cluster including all users). This may greatly degrade the system performance and lose the advantage of clustering to partition through those users relatively far away. Parameters should be carefully chosen and adjusted.

<table>
<thead>
<tr>
<th>PARTITION SCHEME</th>
<th>TOTAL TRANSMISSION POWER (WATT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5-SECTOR</td>
</tr>
<tr>
<td>TTP(OS)</td>
<td>4.97</td>
</tr>
<tr>
<td>TTP(CS) $\delta=2\pi/360$</td>
<td>4.98</td>
</tr>
<tr>
<td></td>
<td>$\delta=2\pi/36$</td>
</tr>
<tr>
<td>TTP(FS)</td>
<td>27.9</td>
</tr>
</tbody>
</table>

A higher density user distribution is illustrated in Figure 4.2. Here, 180 users are partitioned into 15 sectors. An extremely “hot spot” (roughly in the area of the ellipse) exists in this example, where 32 users are placed into one cluster after the coarse clustering, thus exceeding the capacity constraint of a sector. We apply the angular-based partitioning algorithm to the large cluster to divide it into sub-clusters, with the objective of the local MinTTP (in this example, 3 sub-clusters for this large cluster, shown in Figure 4.2(a)). In Table 4.3, we also show the MinTTP results for a different partitioning algorithm for the 15-sector case and 18-sector case. The MinTTP of CS is still quite close to the OS solution (less than 5% in both 15-sector and 18-sector cases). Under this user-distribution, FS is not feasible because some sector(s) for FS partitioning exceed(s) the capacity constraint.

<table>
<thead>
<tr>
<th>PARTITION SCHEME</th>
<th>TOTAL TRANSMISSION POWER (WATT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15-SECTOR</td>
</tr>
<tr>
<td>TTP(OS)</td>
<td>44.6</td>
</tr>
<tr>
<td>TTP(CS)</td>
<td>46.7</td>
</tr>
<tr>
<td>TTP(FS)</td>
<td>NF</td>
</tr>
</tbody>
</table>
Figure 4.1: 36 users, 6 sectors. The lines represent the sectoring boundaries. The ellipse shows one of the clusters.

Now we investigate the result with varying $\delta$ setting. As explained earlier, by increasing $\delta$, the number of clusters decreases, resulting in much faster computation. This is shown in table 4.4 as the percentage of the computing effort for CS over OS. At the same time, the performance degradation percentages are also calculated. In this case, the directivity property cannot be well maintained as most of clusters require further partitioning with angular-based sectoring.

<table>
<thead>
<tr>
<th>PARTITION SCHEME</th>
<th>TOTAL TRANSMISSION POWER (WATT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15-SECTOR</td>
</tr>
<tr>
<td>TTP(OS)</td>
<td>44.6</td>
</tr>
<tr>
<td>TTP(CS)</td>
<td>46.7</td>
</tr>
<tr>
<td>TTP(FS)</td>
<td>NF</td>
</tr>
</tbody>
</table>

To investigate the impact of the number of available sectors ($N$), we vary $N$ from 15 to 23 for the above 180-user configuration (Figure 4.2), where $\delta = 2\pi/360$. 

65
Figure 4.2: 180-user, 15-sector, δ=1. (a). User distribution. The ellipse roughly includes one high-density cluster after coarse clustering, which is further partitioned into 3 sub-clusters with users between boundaries. (b) Sectoring result. The lines represent the final boundaries.

The corresponding MinTTP results for OS and CS are shown in Figure 4.3. From the experiment, we observe that, similar to the OS scheme, the MinTTP results for the CS algorithm also decrease with the increased sectoring number. In general, the gap between OS and CS solutions is less than 10%, and when a large number of sectors are used, the MinTTP achieved by CS is nearly optimal. Furthermore, as shown in the figure, the power saving trend slows down with the growth of N, suggesting careful adjustment on tradeoffs between the antenna cost and performance gain.

4.8 Summary

In this work, we have studied the optimization problem of adaptive cell sectoring to minimize the total transmission power (TTP), while at the same time satisfying the QoS (i.e., lower bound SIR requirements) in DS-CDMA systems. As discussed in the section 4.3, the complexity of optimal solutions is \(O(NM^3)\), and thus suffers long computation time in the high-density cells. With the consideration of reducing system complexity and also the observation that it is better for the sector boundaries across low-density regions, we propose a cluster-based sectoring algorithm. A later section 4.4.3 on complexity analysis shows that our
Figure 4.3: Minimum total transmission power for a different number of sectors

A cluster-based algorithm may achieve great computation gain. Specifically for the high-density case, the complexity is bounded by the selected angular parameters and independent on the number of users in a cell. Furthermore, the proposed algorithms can be easily extended to accommodate multi-rate services and adjusted for more realistic systems.

Numerical results show that our cluster-based sectoring algorithm achieves comparable performance with the optimal sectoring algorithm; while at the same time, it achieves prominent savings in the computational efforts in terms of generally much lower complexity than that of the optimal algorithm. From the sectoring result, we could also find that the boundaries for our cluster-based sectoring algorithm are generally across those neighboring users relatively far away, which is beneficial from a practical point of view.
CHAPTER 5

SURVEY ON WIRELESS AD-HOC NETWORKS

A wireless ad-hoc network is a collection of fixed or mobile communication devices (nodes) that communicate without any infrastructure support. In such a network, each node is capable of transmitting to or receiving from other nodes and a packet could be delivered from source to destination directly or through multi-hop relays across the network. Due to the interference-limited nature in the wireless network, communication between a transmitter-receiver node pair generally constrains the transmission on neighboring links at the same time. A receiver can receive the data from the desired sender only if the interference from other transmitters in the neighborhood is sufficiently small. This interference-limited effect greatly differentiates the wireless ad-hoc network from its wired counterpart and exhibits great difficulty in the capacity derivation. Resource management in such a multi-hop wireless network also involves a large variety of aspects, similar to cellular systems. In this chapter, we mainly focus on two major issues: one is the capacity provisioning and the other one is the energy/power allocation. For the former, we talk about the theoretical bound of throughput capacity, as well as the possible improvement under different scenarios. For the latter, we introduce existing research efforts to design energy/power conserving algorithms. In the following sections, we introduce them in detail.

5.1 Throughput Capacity in Wireless Ad-hoc Networks

In the seminal work "The capacity of wireless networks", Gupta and Kumar [31] demonstrate a throughput capacity of $\Theta(1/\sqrt{|V|\log|V|})$ ($|V|$ is the number of
randomly located nodes) per node in a fixed ad hoc network, which decreases with the number of nodes in the system. This somewhat depressing result invokes great interest in computing possible performance improvements, either under different scenarios, such as mobility setting [30] [7] [66] [25] and hybrid networks with base stations’ support [47] [55]; or by exploiting more resources, such as high bandwidth supply [64], directional antenna deployment [92] and network coding [27]. The corresponding asymptotic bounds are derived under statistical assumptions.

Apart from the asymptotic work, a great number of papers [37] [46] [48] [70] [83] also devote to the investigation of the optimal performance gain through cross-layer joint routing, scheduling and power control, under a deterministic model. Here, routing determines the paths along which a flow is routed from source to destination; scheduling coordinates the concurrent transmissions of independent users in order to eliminate strong interference; and power control determines the admissible power vector for concurrent transmissions.

5.1.1 Capacity in Fixed Ad-hoc Networks

Gupta and Kumar in [31] demonstrate that per node throughput of $O(1/\sqrt{(n)})$ or $O(1/\sqrt{(N \log N)})$ bits/sec is achievable for arbitrary or random networks. To prove the upper bound, the authors notice that according to the interference model, each transmission should occupy a separate area with a radius proportional to transmission distance $tr(n)$. Thus, allowable simultaneous transmissions are $1/tr(n)^2$ and overall capacity of the network is at most $T/tr(n)^2$. Further assume the mean S-D distance equals to $L$, then the number of hops is $L/tr(n)$. Denote the throughput as $r(n)$, then total generated traffic in the network is $nr(n)L/tr(n)^2$. According to the connectivity requirement $tr(n) > \sqrt{\log n/n}$, $r(n) \leq c\sqrt{T/n\log n}$ (c is a constant). To prove the lower bound, they choose specific MAC and routing schemes. Specifically, the area is divided into voronoi cells, where each cell contains a disk of area $100 \cdot \log n/n$ (with radius $r$) and is contained
within a circle of radius $2r$. Under such a setting, it can be proved that each cell contains at least one node with high probability. By choosing the transmission range $= 8r$, the number of interfering cells is bounded by a constant, depending on the interference model. Thus, available capacity in the network is proportional to $T$. Packet routing is done by drawing a straight line from source to destination and relaying the traffic through each intersected cell. Then the traffic carried in each cell is proportional to the number of straight lines passing through each cell, which is less or equal to $c \cdot r(n) \sqrt{n \log n}$. As a result, in a stationary multi-hop network, when the network size becomes large, the per-node throughput goes to zero. The capacity constraint of such a fixed ad-hoc network is mainly two fold: the co-channel interference among neighboring transmission links, where interference in an area around a receiver excludes other transmissions within it; and multiple hops a packet may have to traverse for end-to-end delivery, where as the number of hops increases, much capacity is devoted to packet relay. Additional insights show that throughput is maximized when one constrains its transmission power to reduce the interference region, thus allowing the scheduling of non-interfering concurrent transmissions.

Later, authors in [51] set up simulations to evaluate the interactions of 802.11 MAC with ad hoc forwarding, and their effects on ad hoc network capacity. They simulate out the constant factor in Gupta’s capacity bound with several regular configurations under 802.11.

Capacity constraints under a deterministic setting have also been studied extensively. Following the model of Gupta and Kumar, authors in [48] design a deterministic approach to achieve the throughput bound. They divide the unit area into multiple squares, by drawing equally spaced lines horizontally and vertically. Thus, those links located in squares sufficiently far away could simultaneously transmit without interference. The authors make use of the idea of “$k \times k$ permutation routing” in parallel and distributed computing for the ad-hoc scheduling and routing, and prove its capability to obtain the achievable throughput for general configurations.
Toumpis et al.'s work [83] investigates the capacity region, which describes the set of achievable rate combinations under variable transmission protocols. They name a scheme as the set of feasible simultaneous transmissions. Accordingly, they fill in the Basic Rate Matrix the supportable simultaneous rate at any instance. The capacity of the network is defined as the weighted sum (time-scheduling) of the basic rate matrices under the transmission protocol in use.

Jain et al.'s paper [37] investigates optimal throughput gain, given the network configuration and workload demand. They use a conflict graph to model the multi-access interference among concurrent transmissions. Their idea is similar to [83] by time-sharing the concurrent transmissions while working out more precise linear programming formulation to maximize the capacity. To model the capacity graph, they map all the transmission links as a node and connect any two nodes that interfere with each other. Thus, the maps of achievable simultaneous transmissions are calculated by solving the Independent Set in the graph. Later, they derive the lower and upper bound for the problem.

Kodialam and Nandagopal [46] develop conditions for the achievable rate and propose approximation algorithms with a performance guarantee for a static multi-hop wireless network. Specifically in this paper, they consider only the primary interference constraint on nodes such that they can transmit to and receive from at most one node at a time, while ignoring the interference of simultaneous transmissions among node pairs in the vicinity. They model and solve the scheduling problem as an edge coloring problem. They formulate routing as a linear programming problem, and use the primal-dual algorithm to derive efficient approximation algorithms for the joint routing and scheduling problem.

According to these studies, co-channel interference and the multi-hop relays packets may need to transverse, constitute the major constraints of a wireless multi-hop network, and the throughput capacity decreases quickly with the increase of network size. In view of this, a great amount of effort has been dedicated to the capacity improvement in such a network. In the later sections, we discuss
them in detail.

5.1.2 Node Mobility

Due to node mobility, a mobile wireless network experiences random variations in channel strengths of wireless links, which makes communication in such a network challenging. In the work "mobility increases the capacity of ad hoc wireless networks" [30], Grossglauser and Tse investigate the effect of mobility on the scaling rate of capacity as the number of nodes increases in wireless ad hoc networks. They apply a two-hop relay model, where a packet may go directly from its source to destination or through one intermediate relay, and any transmission occurs among its closest neighbors. They show that a rate of $\Theta(1)$ per user is achievable, which is a great improvement from the rate $\Theta(1/\sqrt{|V| \log |V|})$ ($|V|$ is the number of nodes) per user in a fixed ad hoc network [31]. The basic conclusion is that rapidly changing link strengths can be exploited when deciding routes for the packets, in order to achieve potentially more cost-effective communications. However, this $\Theta(1)$ rate is obtained under at least two idealistic assumptions [30]. Firstly, it is assumed that users move significantly random during the time range of communication sessions. Secondly, the paper focuses on the performance metric of long-term throughput, without taking the end-to-end delay into account. As a result, the delay experienced by these packets under the two-hop transmission strategy could be unbearably large.

Using the same statistical model as in [30], subsequent works [7] [66] incorporate delay constraints into their work. In [66], the authors follow the "one relay" strategy, while focusing on the capacity derivation under the delay constraint. They perform calculations of following probabilities in sequence: firstly, the probability of close range transmission, i.e. the relay node meets the destination within the delay constraint; secondly, the probability of capture, i.e. no other relays are within the interfering range. To achieve the bound, they use the diversity coding, i.e. to add redundant information in the packet, split it into
many blocks and send to destination through different relay nodes. Consequently, the packet could be correctly received with parts of the blocks received. They claim that for delay less than a threshold, the impact of mobility is very small and also, such critical delay is almost independent on the number of nodes in the system. The paper [7] assumes fixed source-destination pairs and proposes a relay scheme, which may achieve throughput that is a poly-logarithmic factor off from the optimal. The basic idea was to choose mobile relays that move toward destinations and thus relay packets closer and closer to their destination.

Authors in later work [25] determine the asymptotic throughput-delay trade-off for both fixed (Gupta's) and mobile (Grossglauser's) ad hoc networks. For the fixed network model, they show the trade-off, $D(n) = \Theta(nT(n))$ ($T(n)$ and $D(n)$ are the throughput and delay). For the mobile setting, $D(n) = \theta(n^{1/2}/v(n))$, ($v$ is the velocity of the nodes). To prove the result in the fixed case, they divide the unit torus into cells of equal size $a(n)$, as the number of interfering cells is a constant and the number of S-D lines crossing a cell is $O(n\sqrt{a(n)})$. For the mobile case, they model a queue formed at a relay node for each S-D pair as a GI/GI/1-FCFS. They realize that the delay of mobile networks can be decreased by increasing the size of the "neighborhood" of each node. Thus, they propose a scheme, which considers node mobility, while making use of the model in the fixed network. The scheme suggests either directly sending packet to D if D locates in one of the neighbor cells, or to a relay on the straight line connecting T-D (T, transmitter). This scheme achieves the same throughput delay tradeoff as in the fixed network.

Previous work focuses on computing asymptotic bounds under the assumption of homogeneity of node behavior. Following simple routing rules, they derive statistical performance gains in terms of throughput and delay for a mobile ad hoc network. However, in a real setting, the movement of a node does not obey such significant randomness and a currently satisfying relay node might fail later. Consequently, for a given packet, it is hard to guarantee its in-time delivery. In view of this, the authors adopt a deterministic model in [99] [100], where no
statistical assumptions are made with regard to either the mobility pattern or the traffic demand. They derive algorithms to search for packet delivery routes, which optimize QoS criteria such as end-to-end delay and energy consumption, by taking into account node mobility. They consider networks operating in both the interference-free regime [99] and the interference-limited regime [100]. The optimal performance serves as an upper bound for the practical scenario where future information about the mobility pattern may not be precisely predicted, and offers insights into understanding system tradeoffs.

5.1.3 Hybrid network

Deployment of extra base stations (BSs) in an ad hoc network, where those BSs connected by high-bandwidth networks are responsible for faraway delivery, might largely relieve congestion from multi-hop routing. In this case, as part of the traffic load, which requires a large number of relays along the path from source to destination, is re-directed through the infrastructure network, the system-wide interference could be greatly reduced.

Under assumptions that the number of ad hoc nodes per access point (AP) or base station (BS) is bounded, the nodes (and AP/BSs) transmit at constant rate with fixed range and the BSs themselves are connected, Kozat et al. [47] declare the throughput capacity per node of $O(W/\log(N))$ ($W$ is total bandwidth and $N$ is the number of nodes) is achievable, where positions of both base stations and ad-hoc nodes are randomly selected and the number of base stations scales linearly with the number of ad hoc nodes.

Liu et al. [55] adopt a routing scheme for a hybrid network. Specifically, if the destination locates within $k$ nearest neighboring cells from the sender, ad-hoc packet relay mode (i.e. the packets are delivered through the ad-hoc nodes) is used; otherwise, packets are delivered through BSs. As a result, they split the whole system into two parts, i.e. the multi-hop ad-hoc relay and the infrastructure relay, and separately calculate the combined throughput capacity.
of the two modes. Here, the capacity derivation for the ad-hoc mode strictly follows Gupta's work [31]. Results show that when the number of base stations grows asymptotically faster than $\sqrt{n}$ ($n$ is the number of ad hoc nodes in the system), the throughput capacity increases linearly with the number of BSs.

In Zemlianov's work [93], the authors also look at the per user throughput in a hybrid ad hoc network. Two crucial values $\sqrt{n/\log n}$ and $n/\log n$ are derived and they conclude that when the number of base stations $m$ is approximately less or equal to the first value, the impact from infrastructure deployment can be ignored; when the $m$ approximately greater the later value, further increasing the infrastructure nodes does not improve throughput; only when the AP/BS deployment is between two extremes, the additional bandwidth offered by the infrastructure support could be effectively shared among ad-hoc nodes, and the capacity improvement from infrastructure deployment is comparable to the ad-hoc mode relaying.

### 5.1.4 High bandwidth

Negi et al. [64] show that in a system with low spectral efficiency, i.e. $\frac{P}{NW} \leq 1$ ($N$ is the Gaussian noise power spectral density), the uniform throughput capacity per-node may increase with the number of nodes in the system. The main idea is that, when the bandwidth goes to infinity, the interference from concurrent transmissions could be negligible; at the same time, with the increase of node density, routes may go through shorter distance hops to achieve a higher transmission rate along the route. Consequently, the later improvement comes from power control.

### 5.1.5 Directional antennas

Yi et al. [92] show that the deployment of directional antenna reduces the interference area, thus increasing the throughput by a factor related to antenna angles at both transmitter and receiver sides. Specifically, let $T_a$ be the trans-
mitter angle and $R_a$ be the receiver angle, the capacity gain for an arbitrary (or random) network is $\sqrt{\left(\frac{\frac{2\pi}{R_a}}{\frac{4\pi}{R_a}}\right)}$ for the directional transmission with omnidirectional reception, $\sqrt{\left(\frac{\frac{2\pi}{R_a}}{\frac{4\pi}{R_a}}\right)}$ for the omnidirectional transmission with directional reception, and $\frac{2\pi}{\sqrt{R_a R_o}} \left(\frac{\frac{4\pi^2}{R_a R_o}}{\frac{2\pi}{R_a}}\right)$ for the directional transmission and directional reception.

### 5.1.6 Network Coding

Gastpar and Vetterli in [27] adopt a relay traffic pattern, where one source-destination pair is considered and remaining nodes act purely as relays. They investigate the precise contribution of relays to capacity. With the point-to-point coding model, the achievable rate is $O(1)$ only. However, they show that when arbitrarily complex network coding is allowed, throughput capacity of $\Theta(\log n)$ bits per second is achievable.

### 5.1.7 Multi-radio Multi-channel

According to the IEEE 802.11 standard, 802.11b specifies 11 channels, 3 of which are orthogonal channels that can work simultaneously within a neighborhood with negligible inter-channel interference; and 802.11a provides 13 orthogonal channels [5]. In practice, channelization has already been deployed in the 802.11 based wireless LAN operating at infrastructure mode, where the neighboring access points (AP) are assigned different channels to avoid interference and thus to improve the throughput. Recent development in the “community wireless networks” provides feasible solutions to build large-scale high-performance multihop wireless backhaul systems [14]. In such networks, most of the nodes are near stationary and have continuous power supplies. Furthermore, more than one network interface card (NIC), referred to as radio here, can be installed on some powerful nodes. Each radio is designed with a capability of switching over multiple orthogonal channels and performing transmission or reception on
any particular channel at a time. Through the efficient coordination of multiple radios in utilizing multiple orthogonal channels, significant benefits in terms of higher system throughput or lower system operating time can be achieved [6].

The centralized control with multi-channel supply is considered in [70]. They propose a set of centralized channel assignment and routing algorithms for a multi-channel network and use adaptive mechanisms to iteratively improve the system performance. Multiple MAC protocols working under the multi-channel environment have also been designed recently. Some work [60] [61] assume multiple NICs equipped at each node to sense all channels concurrently; Some proposals use two radios [36] [90], one serving the control channel and the other one switching across all the other channels for sending data; some are specifically designed for frequency hopping spread spectrum (FHSS) wireless cards, such as Hop Reservation Multiple Access (HRMA) protocol [82]; and recent work mainly focuses on utilizing only one transceiver switching along multiple channels, where particular mechanisms are carried out to negotiate a channel for data transmission [5] [76]. A recent paper [21] proposes a routing protocol in community wireless networks with multi-radio supply. The protocol incorporates a routing metric into the selection of a high-throughput path, where the metric accounts for the loss rate and channel bandwidth of each link forming the path.

In the work [101], the authors investigate the achievable performance gain by jointly optimizing routing and scheduling in a stationary multi-radio multi-channel multi-hop network. They derive the mathematical programming formulation for the optimization, which minimizes overall system activation time to support a given traffic load. To deal with the high combinatorial complexity, they develop an efficient column generation based approach to solve this problem, and derive close upper and lower bounds for it. The details are given in a separate chapter later.
5.2 Energy Conservation in Wireless Ad-hoc Networks

In a multi-hop wireless network, energy/power conservation is of great importance not only to increase a device's lifetime, but also to avoid excessive interference and gain improved coexistence with other systems [22]. It is obvious that when the nodes are either costly or in short power supply, energy efficiency is vital for the lifetime improvement. In some extreme cases, energy is entirely nonrenewable and is thus an overriding constraint for the design and operation of a network. Furthermore, as multi-hop relays are involved in packet delivery in an ad hoc network, decreasing the transmission power results in lower interference to other nodes and allows better spatial re-use on concurrent transmissions. In addition, in some systems such as Ultra-Wideband (UWB) networks, there is strict restriction on the power emission in the network to avoid too much interference with other systems working in the overlapping frequency band.

The problem of multi-hop routing with energy conservation has been studied extensively in recent years [9] [20] [41] [63] [71] [72] [75]. These efforts mainly focus on designing efficient protocols for packet level energy conserved routing.

Singh and Woo [75] develop power-aware cost metrics and propose to minimize these costs for energy conserving routing. Roduplu and Meng [71] propose a distributed protocol to compute the minimum energy route. Based on this, in their later work [72], they discuss the maximization of throughput (bits) subject to the constrained total energy (Joule) consumption in the network, assuming that the concurrent communications are non-interfering. Mobile systems are also considered in [71], where the focus is on computing and maintaining the most energy efficient topology, for each fixed configuration of node locations.

Kar et al. [41] develop an online algorithm to achieve potentially high throughput capacity with the constraint on each node’s energy supply. This is performed through searching for the shortest path for each source destination pair with the link weight set as a function of the energy required to transmit on the link and
the fraction of energy the transmitter already used.

These papers deal with the packet relay along the multi-hop network, subject to certain optimization criteria. As no delay constraint is exerted for the packet delivery, they do not explicitly consider the MAC scheduling for interference avoidance.

Energy efficient joint routing, scheduling and power control with co-channel interference concern are addressed in later work. Bharia and Kodialam in [9] investigate the most energy efficient communication over point to point links at a given rate. They consider the primary interference constraint, i.e. a node can transmit to and receive from at most one node at a time, while ignoring interference among simultaneous transmissions in a vicinity. They formulate the problem as an optimization problem with non-linear objective function and non-linear constraints. They design a polynomial time 3-approximation algorithm to solve the problem with performance guarantee.

Cruz and Santhanam in [20] introduce a link scheduling and power allocation scheme to minimize the total average power consumption, subject to the minimum average data rate demand, peak transmission power constraint and the multi-access interference in a stationary wireless network. They claim that the optimal power consumption on each link is a convex function of the transmission rate, and accordingly adopt a shortest path algorithm to guide routing.

Needly et. al. [63] propose a joint routing and power allocation scheme for time varying (due to channel state variation/ user mobility) wireless networks. By handling the queues at each node, the proposed strategy stabilizes the network and achieves bounded average delay under statistical conditions.

In the following two chapters, we investigate the cross-layer optimization in wireless multi-hop networks under a deterministic setting, where routing, scheduling and power control can be jointly considered. Specifically, two kinds of systems, namely the mobile ad-hoc network where mobility is exploited in the packet routing and the multi-radio multi-channel multi-hop network where radios are capable
of switching over different channels to avoid co-channel interference, are intensively investigated.
CHAPTER 6

MOBILITY ASSISTED ROUTING IN MANET

In this chapter, we take node mobility into account for designing an optimal routing scheme in mobile ad hoc wireless networks (MANET) operating in both the interference-free regime and the interference-limited regime. Specifically, in the *interference-free regime*, each transmission from a certain node is assumed to be unaffected by potential concurrent transmissions from other nodes. Representative systems operating in the (near) interference-free regime are those characterized by a relatively wide bandwidth, or a relatively large number of available channels, used to accommodate a relatively small amount of end-to-end traffic. The effect of interference becomes negligible in such systems since concurrent traffic at any time may be arranged in different orthogonal channels. One example is the Ultra-Wide Bandwidth (UWB) systems [64][67]. Yet another example is sensor networks, which are mission-driven wireless ad hoc networks comprising a large number of light-weight, low-energy, and cheap sensor nodes. Although the bandwidth might not be high in a sensor network, the traffic load is typically very low. Thus, the number of available channels is relatively large with respect to the amount of traffic demands. In contrast to these systems, many ad hoc wireless networks operate in the *interference-limited regime*, where interference imposes constraints on feasible concurrent transmissions [31]. In such a case, the problem involves joint optimization on routing in the network layer and scheduling in the link layer.

Understanding the fundamental relationships among throughput, delay, interference and power control is a very challenging task. In this work, we assume a deterministic mobility model with full knowledge of the mobility pattern over
the entire communication time range. Although future information might not be available in practice, this enables us to derive bounds for more realistic systems and gain some insights into the fundamental tradeoffs. Since this work targets systems where mobility affects routing, we assume that the allowable delay is comparable with the time-scale of node mobility [7][66].

In the interference-free regime, we propose efficient algorithms for calculating the optimal packet delivery routes. The obtained optimal performance serves as an upper bound for the practical scenario where future information about the mobility pattern may not be precisely predicted. Two kinds of systems, either with or without power control, are considered here. For a system without power control capability, the transmission power at each hop is fixed and two nodes can directly communicate only when they fall into their mutual transmission range. In such a case we consider minimizing the time and energy consumption required for end-to-end packet delivery. In other words, with the former criterion, we seek the fastest route; with the latter criterion, we seek the route with the minimum number of hops, since the energy consumed at each hop is a constant. We propose hop-expansion based algorithms that carry out the computations inductively over the number of hops, to find both the fastest and the minimum-hop routes. When power control is allowed in the system, any transmitter may flexibly control its communication range by adjusting its transmission power. In this context, we consider minimizing the total energy consumed at the multiple hops, while satisfying certain end-to-end delay bound. We propose a layered algorithm that performs the computations inductively over the discrete time periods, to find the minimum energy path subject to delay constraints.

In the interference-limited regime, multi-access interference should be explicitly considered and the concurrent transmissions can coexist only if the SINR requirements can be guaranteed. We consider the objective of minimizing the total energy consumption for multiple packets delivery, subject to the end-to-end delay constraint and SINR requirements for concurrent transmissions in a power controlled system. This problem can be formulated with a directed graph, and
solved by a dynamic programming algorithm, which performs the computations inductively over the discrete time periods. The resulting worst case complexity for the optimal solution is high, as it increases exponentially with the number of packets for coordinate delivery. To reduce the complexity, we propose an efficient heuristic approach, where the packets are admitted one after another in a greedy way, i.e. the addition of a path does not incur the modification of previously packed routes. For each packet in service, the heuristic involves a shortest path calculation in a directed graph, which minimizes the cumulative link costs along the path. Due to the interference limited nature of multi-access transmissions, the interference from any transmission link on previously admitted paths might constrain later packets' usage of neighboring links for concurrent transmissions. Refer to Figure 6.1, transmission from node '1' to '2', i.e. link 1 → 2, can not simultaneously coexist with link 3 → 4, as the strong interference from node '3' may prohibit the correct reception of node '2' from node '1'. To take this effect into account and to prevent aggressive usage of those links that have significant impact on the system, we define a variable 'impact factor' for each active link, which denotes the fraction of active links that are disabled by the addition of a particular link. Here, active links are transmission links that can be used for packet relay before the routing of the current packet. Consequently, we incorporate both parameters, i.e. the energy requirement and the impact factor, into the cost function. In addition, we also consider ways to iteratively adjust existing paths for a better performance. The algorithms would be illustrated in detail in the following sections.

6.1 System Model

6.1.1 Network Model

We consider a mobile wireless network, where all mobile nodes are equipped with identical communication devices such that each node may act as transmitter or receiver as needed and the nodes are cooperating on the packets delivery. Let
Figure 6.1: Transmission Map

$N$ denote the set of nodes in the network, which are labelled 1, \ldots, $|N|$. We assume a constant transmission rate and fixed packet length, such that the time for each packet to travel any one hop is a constant, $\tau$ (seconds). We adopt a discrete time system, where the overall time range is represented by uniformly spaced time-ticks $t = 1, 2, \ldots, T$, with inter-tick duration $\tau$ seconds, the same as one hop packet relay. System decisions are made globally over a total of $T$ intervals, with length $T \cdot \tau$ seconds. Furthermore, let $K$ denote the set of packets to be delivered in the current system. Each packet $k \in K$ from source $S_k$ to destination $D_k$ is associated with two time ticks $T^s_k$ and $T^e_k$, where $T^s_k$ is packet $k$'s generating time at the source and $T^e_k$ is its deadline by which the packet must be delivered to the destination. As a result, each packet could be represented by a tuple $(S_k, D_k, T^s_k, T^e_k) \in N \times N \times T \times T$. Here, different from [25], we define delay as the time it takes for a packet to reach the destination after it becomes available at the source.
6.1.2 Mobility Model

For simplicity, we adopt a quasi-static mobility model where the locations of the nodes can only change at the beginning of each discrete time interval $t$ ($t \in T$). An arbitrary mobility pattern is assumed for each node. During time $t$, the position of node $i$ is $(x_i(t), y_i(t))$ and the collection of node locations is called a map and denoted by $M(t) \equiv \{(x_i(t), y_i(t)) , i = 1, \ldots, |T|\}$. That is, we assume that the nodes stay at the locations $M(t)$ for $\tau$ seconds, then instantaneously jump to new locations $M(t + 1)$ at the beginning of the next time interval.

6.1.3 Communication Model

In a non-interfering system without power control capability, a fixed transmission range $R$ is defined such that two nodes can communicate when their relative distance $d \leq R$. In a system with power control, it might be possible for direct communication between any two nodes by adjusting the transmission power. A minimum signal to interference plus noise ratio ($\text{SINR}$) level is required at each receiver for transmission to be successful. Assume $C$ communication pairs are now active, the $\text{SINR}$ requirement at the $i$-th receiver can be expressed as

$$\text{SINR}_i = \frac{P_i g_{ii}}{\sum_{j \neq i} P_j g_{ji} + \sigma^2} \geq \gamma_i,$$  \hspace{1cm} (6.1)

where $P_i$ is the transmission power at the $i$-th transmitter $\text{Tr}_i$, $g_{ji}$ is the channel path loss from the $j$-th transmitter $\text{Tr}_j$ to the $i$-th receiver $\text{Re}_i$, $\gamma_i$ is the $\text{SINR}$ requirement for the $i$-th ($i \leq C$) transmission pair and $\sigma^2$ is the background white Gaussian noise level.

Let $F$ denote the square matrix with each entry

$$F_{ji} = \begin{cases} 
\frac{g_{ji}}{g_{jj}} \gamma_j, & \text{if } i \neq j \\
0, & \text{if } i = j 
\end{cases}$$  \hspace{1cm} (6.2)
Further let

$$
u = \left[ \frac{\sigma_1^2 \gamma_1}{g_{11}}, \ldots, \frac{\sigma_C^2 \gamma_C}{g_{CC}} \right]^T$$

(6.3)

$$P = [P_1, \ldots, P_C]^T$$

(6.4)

The SINR requirements (6.1) could be written as

$$(I - F)P \geq u.$$  

(6.5)

According to [32], if the Perron-Frobenius eigenvalues of $F$, $\lambda = \max |\lambda_i| \geq 1$ ($i = 1, \ldots, C$), then no feasible power vector exists. If and only if $\lambda < 1$, a minimum power allocation satisfying the SINR requirements with the given concurrent transmission pairs can be calculated as

$$P = (I - F)^{-1}u.$$  

(6.6)

Since it takes one unit of time for a packet to be relayed at one hop, the one-hop energy consumption vector is $P \cdot \tau$ and the total energy requirement for concurrent transmissions during the interval is

$$\varepsilon(\cup_{1 \leq i \leq C} T_{ri} \rightarrow R_{ei}) = \sum_{1 \leq i \leq C} (P_i) \cdot \tau.$$  

(6.7)

Specifically in an interference-free system, where interference from concurrent transmissions could be ignored,

$$\text{SINR}_k = \frac{P_k g_{kk}}{\sigma^2} \geq \gamma_k,$$

(6.8)

the power consumption simplifies to,

$$P_k \geq \frac{\gamma_k \sigma^2}{g_{kk}}.$$  

(6.9)
6.1.4 Route Definition

A route/path is calculated to deliver one packet from its source node to the destination through multiple hops’ relay within the delay constraint. In this paper, an immediate transmission, or a hop, from node $x$ to $y$, is denoted by $x \rightarrow y$. We use notation $u \xrightarrow{p} v$ to denote a route $p$ from node $u$ to node $v$, as a concatenation of hops, e.g., $p \equiv v_1 \rightarrow \ldots \rightarrow v_d, v_1 = u, v_d = v$. Since the system is dynamic, specifying a route $p$ also requires specifying the time intervals during which each constituent transmission occurs. We use $t_{x \rightarrow y}$ to denote an assignment of time interval along with the transmission $x \rightarrow y$. In the sequel, a route $p$ would refer to a concatenation of hops, together with an assignment of the time periods for each constituent hop.

6.2 Routing in The Interference-Free Regime

In this section, we investigate the optimization on the end-to-end packets delivery for an interference-free system, where computing the optimal route for a packet could be decoupled from other packets.

6.2.1 Optimal Routing without Power Control

We consider a packet with delay constraint $[T^s, T^c]$ delivered from $S$ to $D$ in a system without power control. Two optimization objectives, minimizing the end-to-end delivery time and minimizing the total power consumption, are investigated. The resulting optimal paths for both objectives can be obtained with the hop-expansion based algorithms, which perform computations inductively over the number of allowed hops, with polynomial time complexity.

In the optimizations below, we assume global knowledge of the maps over time, $M(t) \equiv \{(x_i(t), y_i(t)), i = 1, \ldots, |N|, t = 1, \ldots, T\}$. Define a function $f(u, v, t)$, which returns the first time period $t_f$ after $t$, during which node $u$ and node $v$ can directly communicate, i.e., $d_{uv}(t_f) \leq R$, where $R$ is the maximum
transmission range. This function sufficiently summarizes all information about
the mobility pattern that would be used in the algorithms.

**Fastest Path**

Define $\pi(p)$ as the time instant for the last hop in path $p$. Minimizing the end-
to-end delivery time can be formally expressed as,

$$p^*_t(S \rightarrow D) \equiv \arg \min_{p: S \rightarrow D} \pi(p). \quad (6.10)$$

Introduce

$$p_t(S \rightarrow v, h) \equiv \arg \min_{p: S \rightarrow v, |p| \leq h} \pi(p), \quad (6.11)$$

which gives the fastest path $S \xrightarrow{p} v$ with less than or equal to $h$ hops. Initially,
set

$$p_t(S \rightarrow v, 1) = S \rightarrow v, \quad \text{if } f(S, v, T^* - 1) \leq T^* \quad (6.12)$$

In other words, direct transmission is possible if $S$ and $v$ ever fall into their mutual
transmission range within the delay constraint.

Since a route with loops can always be shortened by removing loops, we have

$$p^*_t(S \rightarrow v) = p_t(S \rightarrow v, +\infty) = p_t(S \rightarrow v, |N| - 1). \quad (6.13)$$

The proposed algorithm proceeds in $|N| - 2$ steps of “hop expansion”, where in the
$h$’s expansion, we compute $\{p_t(S \rightarrow v, h + 1), v \in N\}$ from $\{p_t(S \rightarrow v, h), v \in N\}$.

Specifically, given $\{p_t(S \rightarrow v, h), v \in N\}$, we only need to consider paths $S \xrightarrow{p} v$ with exactly $h + 1$ hops, since only these paths could possibly provide further
improvement. Partition a path with $h + 1$ hops as $S \xrightarrow{p_1} u \rightarrow v$, where $|p_1| = h$.
Since $p_t(S \rightarrow u, h)$ is the fastest route with less than or equal to $h$ hops, the
minimum delivery time among all paths of the form $S \xrightarrow{p_1} u \rightarrow v$ can be achieved as
$p_1 = p_t(S \rightarrow u, h)$. This observation leads to a procedure to compute $p_t(S \rightarrow v, h + 1)$: check if there exists a path of the form $S \xrightarrow{p_1} u \rightarrow v, p_1 = p_t(S \rightarrow u, h)$, which
is faster than $p_t(S \rightarrow v, h)$.

88
A detailed implementation is described in Algorithm 1 (Please refer to Appendix G for proof), which computes the fastest routes from a given source node $S$ to all other nodes with worst case complexity $O(|N|^3)$.

**Algorithm 1 Fastest Routing Algorithm**

```plaintext
for $v \in N$ do
    $T_{\text{min}}(v) = f(S, v, T_s - 1)$;
    $\text{PrevNode}(v) = \text{NULL}$;
end for

for $h = 1$ to $|N| - 2$ do
    $T_{\text{min old}} = T_{\text{min}}$;
    for $u \in N \setminus \{S\}$ do
        for $v \in N \setminus \{S\}$ do
            $t = f(u, v, T_{\text{min old}}(u))$;
            if $t < T_{\text{min}}(v)$ and $t \leq T_s$ then
                $T_{\text{min}}(v) = t$;
                $\text{PrevNode}(v) = u$;
            end if
        end for
    end for
end for
```

**Minimum Hop Path**

Define $|p|$ as the number of hops in path $p$. Since the transmission power at each hop is a constant, minimizing the total energy consumption is equivalent to minimizing the number of hops for a packet delivery, which can be formally expressed as,

$$p^*_h(S \rightarrow D) = \arg \min_{p : S \rightarrow D} |p|.$$  \hfill (6.14)

Introduce

$$p_h(S \rightarrow v, h) = \arg \min_{p : S \rightarrow v, |p| \leq h} |p|,$$  \hfill (6.15)

which gives the minimum-hop path $S \xrightarrow{h} v$ with less than or equal to $h$ hops. The procedure for calculating the minimum-hop path is conceptually similar to the search for the fastest path. Indeed, Algorithm 1 can be slightly revised for the minimum hop routing. For simplicity, we only give a brief description here.
When searching for the minimum-hop path, we initially record all nodes that are directly reachable from \( S \). These one hop transmissions are the optimal paths for the corresponding nodes. Subsequently, after each step of hop expansion for the fastest routing, we identify nodes that become reachable from \( S \) during the current hop expansion, and record the corresponding fastest path as the minimum hop path. This procedure continues until all nodes are recorded or no hop expansion is applicable. This results in the fastest minimum-hop paths.

### 6.2.2 Optimal Routing with Power Control

In a system with power control capability in the interference-free regime, any two nodes may directly communicate by adjusting the transmission power. We consider the minimizing the total energy consumption along the route \( S \xrightarrow{p} D \), subject to certain delay constraint \([T^*, T^*]\) for end-to-end packet delivery.

Let \( e(p) \) denote the total energy usage over all constituent hops of path \( p \). The minimum energy consumption path, with the associated each hop's transmission time within \([T^*, T^*]\), could be formally expressed as,

\[
\text{\( p^*_e(S \xrightarrow{} D) \equiv \arg \min_{\substack{p: S P D}} e(p), \)} \tag{6.16}
\]

Introduce

\[
\text{\( p_e(S \xrightarrow{} v, t) \equiv \arg \min_{\substack{p: S P v, \pi(p) \leq t}} e(p), \)} \tag{6.17}
\]

which gives the minimum energy path from \( S \) to \( v \) by time \( t \), with the corresponding minimum energy consumption \( \{E(S \xrightarrow{} v, t), S, v \in N\} \). Denote \( e(i \to j, t) \) as the lowest energy required for one-hop relay from \( i \) to \( j \) during the \( t \)-th time period. For consistency, we let \( e(i \to i, t) = 0 \), indicating that no energy is consumed if node \( i \) holds the packet during time interval \( t \).

**Lemma 2**

*The minimum energy required to deliver one packet from \( S \) to \( v \) by time \( t \) \((T^* < \)
\[ t \leq T^e, \] is
\[ E(S \rightarrow v, t) = \min_{u \in N} \{ E(S \rightarrow u, t - 1) + e(u \rightarrow v, t) \}. \] (6.18)

Let \( u^* \) be the node achieving the minimization above. An optimal path for \((S, v)\) is

\[ p_e(S \rightarrow v, t) \equiv \begin{cases} 
  p_e(S \rightarrow u^*, t - 1), & u^* = v \\
  S \rightarrow v, & u^* = S \\
  p_e(S \rightarrow u^*, t - 1) \oplus u^* \rightarrow v, & u^* \neq v, u^* \neq S
\end{cases} \] (6.19)

where \( \oplus \) is an operation which concatenates two sub-paths into a path.

According to Lemma 2 (Please refer to Appendix H for proof), the minimum power route can be obtained by a layered optimization that performs the computations inductively over the discrete time periods. Initially, set \( E(S \rightarrow v, T^*) = e(S \rightarrow v, T^*) \) and \( p_e(S \rightarrow v, T^*) = S \rightarrow v \). For time period \( t \), we first calculate \( \{ e(u \rightarrow v, t), u, v \in V, T^* < t \leq T^e \} \) with (6.9). Then update the optimal paths and the corresponding minimum energy values, according to (6.19) and (6.18). The procedure continues until the delay bound \( t = T^e \).

As a result, the worst case complexity for computing the minimum energy path from a source node to all other nodes is \( O(T^e |N|^2) \), where \(|N|\) is the total number of nodes.

### 6.2.3 Simulation Results in Interference-free System

Simulations have been conducted on a mobile network configuration, shown in Figure 6.2(a). There are 20 mobile nodes, uniformly located in a square region of size 1000m \( \times \) 1000m. Each arrow denotes the trajectory of a node: the tail and the head of the arrow are the starting and ending positions, respectively. These starting and ending positions are randomly generated. There are overall \( N = 100 \) discrete time intervals. Each node is assumed to move at a constant speed toward the ending position, over the considered time interval \([1, N]\). Furthermore, during
each discrete time interval, the locations of the nodes do not change. Extensive experiments have been done in our work, however, for clarity in illustration, only a small set of results are presented in this section.

For an interference-free system without power control capability, the transmission range is set to be $R = 150m$, meaning that two nodes can directly transmit packets if and only if they fall within their mutual communication distance $R$. The delay constraints for all packets considered in the interference-free regime are $[T_s, T_e] = [1, T]$.

Figure 6.3 and Figure 6.4 show the fastest and minimum-hop route for packet delivery from node 3 to node 14. Each sub-figure captures a snapshot of the nodes’ locations, along with one hop of packet relay, denoted by an arrow. The associated time period for each hop relay is indicated in the title. As a result, transmissions along the path tend to occur at the instant when two nodes first get into the mutual transmission range or immediately after the previous delivery. As expected, the overall minimum-hop routes requires less hops (3 hops) compared with the fastest route (6 hops), at a cost of longer delivery time, 61 versus 40 intervals. Further note that if the network is static and all the nodes are stationary with the original map $M(1)$, it is not possible for node 3 to connect
Figure 6.3: The Fastest Route from Node 3 to Node 14

<table>
<thead>
<tr>
<th>Table 6.1: Comparison between Fastest and Min-Hop Routing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Relay Hops</strong></td>
</tr>
<tr>
<td>Fastest Routing</td>
</tr>
<tr>
<td>Min-Hop Routing</td>
</tr>
</tbody>
</table>

with and transmit a packet to node 14.

A more general comparison between Fastest Routing and Minimum Hop Routing is presented in Table 6.1. The average number of hops and time intervals required among all pairs of nodes for both schemes are shown in the table. Accordingly, an average of 35.6% energy saving could be obtained by minimum hop routing with 28.6% more delivery time consumed compared with the fastest routing scheme.

Figure 6.5 shows the minimum energy route from node 3 to node 14 in an interference-free system with power control. It can be observed that energy savings are generally achieved by adopting less hop and short-range transmissions. According to the simulation results averaged over all pairs of nodes, the minimum energy consumption decreases monotonically as the allowed delay increases,
Figure 6.4: The Min-Hop Route from Node 3 to Node 14

demonstrating the tradeoff between the end-to-end delay bound and the minimum energy consumption.

In Figure 6.6, we examine the energy consumption with the increase of node mobility, which maps to the distance that the nodes can traverse during the packet delay bound. The first column denotes the result for a static setting, with the nodes' locations the same as map M(1), the starting locations (Figure 6.2(b)), for the mobile setting (Figure 6.2(a)); later columns show the cases where the nodes traverse 1/4, 1/2 and the full distances on the Figure 6.2(a) within the delay constraint. We calculate the averaged energy consumption over all mobile nodes. As expected, with the increase on the node mobility, better relay nodes and channel conditions might be encountered in the packet delivery, resulting in lower energy consumption.

We also investigate the performance with the increase of node density in the
network. The results are shown in Figure 6.7. We average the minimum energy value for packets delivery over all source-destination combinations of the first five mobile nodes. As shown in the figure, with the increase of number of users in the network, more nodes are involved in the relay, resulting in better relay paths and sound transmission links, with reduced energy consumption.

### 6.3 Routing in The Interference-Limited Regime

When the system works in an interference-limited regime, transmission on one link constrains the capacity on neighboring links at the same time. In this case,
the multiple packets must be coordinated in the delivery, and cross-layer joint routing, scheduling and power control should be considered.

6.3.1 Minimum energy routing: optimal solution

In this section, we first present an algorithm to calculate the coordinate multiple packet delivery paths in a power controlled mobile ad hoc network. We take into account node mobility for global optimization, and the objective is to minimize the system-wide energy consumption, with the constraints on the packet delay and the SINR requirement. As packets are relayed during each time interval between any two nodes along the system time-line, we could formulate the problem with a directed graph. Later, we introduce a dynamic programming algorithm on the optimization.

Problem formulation

Consider a directed graph \( G = (GV, GE) \) shown in Figure 6.8. The vertex set \( GV = \{v_t^i \mid 1 \leq t \leq |T| + 1, 1 \leq i \leq |N|\} \) denotes the set of mobile nodes at the beginning of each time interval, labelled as \( (t,i) \), thus \( |GV| = |N| \times (|T| + 1) \).
Lemma 3  The minimum energy requirement for multiple packets \( k \in K \) delivery from \( S_k \) to \( D_k \) by time \( t \) \((T_{\text{min}} \leq t \leq T_{\text{max}})\) can be calculated as,

\[
E(\bigcup_{k \in K} \mathcal{N}(t) | S_k \rightarrow D_k, t) = \min_{u \in U} \{E(\bigcup_{a \in \mathcal{A}(t)} S_a \rightarrow u \cup \bigcup_{f \in \mathcal{F}(t)} S_f \rightarrow D_f, t - 1) + \varepsilon(\bigcup_{a \in \mathcal{A}(t)} v_{u_a}^{t} \rightarrow v_{D_a}^{t+1} \cup \bigcup_{s \in \mathcal{S}(t)} s_s \rightarrow v_{D_s}^{t+1})\}
\]

where \( U = \{(u_k), k \in \mathcal{A}(t), \forall u_k \in N\} \), is the set of union on intermediate nodes. Let \( u^* = (u_k^*) \) be the nodes, which achieve the optimization above. The optimal relay path \( p^k_\varepsilon \) for packet \( k \in K/\mathcal{N}(t) \) by time \( t \) is

\[
p^k_\varepsilon(S_k \rightarrow D_k, t) \equiv \begin{cases} 
 v_{S_k}^t \rightarrow v_{D_k}^{t+1}, & u_k^* = S_k \\
 p^k_\varepsilon(S_k \rightarrow D_k, t - 1), & u_k^* = D_k \\
 p^k_\varepsilon(S_k \rightarrow u_s^*, t - 1) \oplus v_{u_s^*}^{t+1} \rightarrow v_{D_k}^{t+1}, & u_k^* \neq D_k, u_k^* \neq S_k
\end{cases}
\]

where \( \oplus \) is an operation to concatenate two sub-paths into a path.

The directed edge set \( GE = \{e_{i,j}^t (1 \leq t \leq |T|, 1 \leq i \leq |N|, 1 \leq j \leq |N|)\} \), connecting from \( v_i^t \) to \( v_{j}^{t+1} \), denotes all the links during each time interval. We define, \( \{e_{i,j}^t, i \neq j\} \) as 'relay link', which represents the relay from node \( i \) to node \( j \) at \( M(t) \); and \( \{e_{i,i}^t, i = j\} \) as 'hold link', which means that node \( i \) holds the packet and no relay happens during the interval \( t \). We associate each edge \( e_{i,j}^t \) with a weight \( w_{i,j}^t = g_{i,j}^t \), which can be the path loss on the link \( i \rightarrow j \) during time \( t \), labelled as \( w(t, i, j) \) in the figure. As a result, \(|GE| = |T| * |N|^2\). This 'transmission graph' \( G \) sufficiently summarizes all the system information over the time range.

For a packet \( k \in K \) represented by the tuple \((S_k, D_k, T_k^S, T_k^D)\), the source vertex is \( v_{S_k}^{T_k^S} \), simplified as \( s_k \); and the corresponding destination vertex is \( v_{D_k}^{T_k^D+1} \), simplified as \( d_k \) on the transmission graph \( G \). Assume parallel paths \( \{p_k (k \in K)\} \) exist and denote the path set \( p = (p_1, p_2, \ldots, p_K) \). In Figure 6.8, these paths consist of a collection of relay/hold links along the time-line. For the set of relay links \( \{e_{i,j}^t \in p_k, i \neq j, k \in K\} \) concurrently transmitting during time interval \( t \) \((t \in T)\), the energy requirement can be calculated from (6.7) in Section 6.1.3 and represented by

\[
\varepsilon_t = \varepsilon(\bigcup_{\{e_{i,j}^t \in p_k, k \in K, i \neq j\}} v_i^t \rightarrow v_j^{t+1}), \quad (6.21)
\]
where, \( \cup \) denotes the union of the relay links. Consequently, the system-wide energy usage for the path set \( p \) is

\[
\epsilon(p) = \sum_{t \in T} \epsilon_t.
\]  

(6.22)

Our objective is to calculate the optimal path set \( p_e = (p_1, p_2, ..., p_K) = (\cup_{k \in K} S_k \rightarrow D_k) \) for the set of packets \( \{k \in K\} \) from source vertex \( s_k \) to destination \( d_k \), with the lowest overall energy consumption, formally expressed as

\[
p_e(\cup_{k \in K} S_k \rightarrow D_k) \equiv \min_{P: \{\cup_{s_k P_k d_k}\}} \epsilon(p),
\]  

(6.23)

with energy consumption \( E_e(\cup_{k \in K} S_k \rightarrow D_k) \).
Dynamic programming solution

In this section, we propose a dynamic programming algorithm to calculate the optimal path set $p_k$, provided with transmission graph $G$ and the packet set $K$. For simplicity, we record the sequence of relay links, i.e., $e_{i,j}^k (i \neq j)$ along each path. This sufficiently presents the packet delivery routes, together with the assignment of time periods for the constituent relays.

To begin with, we define the following matrices: for $\forall t \in T$, $SC(t)$ (starting packets), denotes the set of packets $k \in K$ with $T_k^s = t$; $AC(t)$ (active packets), denotes the set of packets with $T_k^s < t \leq T_k^e$; $FC(t)$ (finished packets), denotes the set of packets with $T_k^e < t$; $NC(t)$ (non-starting packets), denotes the set of packets with $T_k^s > t$. Furthermore, let $T_{min} = \min_{k \in K} T_k^s$ and $T_{max} = \max_{k \in K} T_k^e$.

According to Lemma 3 (Please refer to Appendix I for proof), the minimum energy relay paths for the packet set $K$ can be obtained by the dynamic programming algorithm that performs computations inductively over discrete time periods. Initially, set

$$E(\bigcup_{k \in SC(T_{min})} S_k \rightarrow \nu_k, \ T_{min}) = \varepsilon(\bigcup_{k \in SC(T_{min})} S_k \rightarrow \nu_k^{T_{min}+1})$$

for $\forall \nu_k \in N$ and the corresponding $p_k^b(S_k \rightarrow \nu_k, \ T_{min}) = s_k \rightarrow \nu_k^{T_{min}+1}$. For the later time periods $T_{min} < t \leq T_{max}$, we calculate the energy consumption $\varepsilon$ in (6.21) for all possible arrangement on path updates at $M(t)$. Specifically, $\forall u \in U$, $U = \{ (\bigcup_{k \in AC(t) \cup SC(t)} u_k), \ \forall u_k \in N \}$, we calculate $\varepsilon(\bigcup_{u \in AC(t)} v_{u_a}^t \rightarrow v_{u_a}^{t+1} \cup \bigcup_{s \in SC(t)} s_s \rightarrow v_{s_u}^{t+1}), \ \forall u_a, \nu_s \in N$; update the optimal path $p_k^b(D_k \rightarrow \nu_k, t), k \in AC(t) \cup SC(t)$ and the corresponding minimum energy values according to Lemma 2. The procedure continues until $t = T_{max}$, with the optimal value

$$E_{\varepsilon}(\bigcup_{k \in K} S_k \rightarrow \nu_k) = E(\bigcup_{k \in K} S_k \rightarrow \nu_k, \ T_{max}).$$

For each time interval, the algorithm involves a minimization over all possible updates on any $k$-node combinations, prepared for $k = |SC| + |AC|$ starting/active packets’ intermediate relay paths. The computation for updates on one $k$-node
set is at most $O(|N|^k)$ and the number of combinations is $O(|N|^k)$. Thus, the computation effort needed at each map is at most $O(|N|^{2|K|})$ and the worst case complexity for the optimization is $O(|N|^{2|K|} \cdot |T|)$, where $|N|$ is the number of nodes in the network, $|K|$ is the number of packets and $|T|$ is the total system intervals. The obtained optimal result serves as an energy consumption lower bound on the packets’ routing, provided that mobility is taken into account for global optimization. However, due to the high complexity for the computation, i.e. the exponential factor on $|K|$, the algorithm is not suitable for a large problem set with a great number of packet demands. Thus we introduce an efficient heuristic algorithm in the next section.
6.3.2 Heuristic Approach

According to the optimization illustrated in the last section, the complexity of the algorithm increases exponentially with the number of packets to be delivered. In this section, we introduce an efficient heuristic approach with much reduced complexity. The objective here is to achieve high throughput while maintaining relatively low energy consumption.

The general idea is to admit the packets one after another in a greedy manner. Specifically, for each packet in service, the algorithm performs three steps as shown in Figure 6.9, namely Initialization, Path Selection and Admission. The initialization step generates the active graph $AG$ and associates each link with a cost, which is calculated from a function capturing both the extra energy requirement and the quantified impact to the network. Here, $AG$ denotes a transmission graph whose links can safely coexist with the relay links of previously admitted paths, and thus be used for packet relay. We name these relay links 'Active Links'. The path selection step finds the minimum cost path in $AG$ for the current packet delivery. If no feasible path exists for the packet, it is rejected; otherwise, the minimum cost path is assigned to the packet in the admission step. Furthermore, those relay links along the path and all other links that fail to coexist with the relay links are eliminated from $AG$, resulting in a reduced graph for later packets’ admission. This procedure continues until all packets are examined. As
the packets are served one by one, another advantage for the proposed algorithm is that, it can flexibly deal with the overloaded traffic demand by admitting the packets according to the sequence of arrival or priorities, whereas the optimal algorithm may fail to present any feasible solution.

Noticeably, as the algorithm performs in a greedy manner, once a path is settled, it subsequently does not change during the process of admission. The relay links along these paths cause interference to other links in the vicinity. Consequently, previously admitted paths might restrain the later packets' routing and affect the system's performance. Later in this section, we introduce an iterative adjustment scheme (Figure 6.10), which deletes relay links along the previously admitted path, and re-calculates a path with lower cumulative extra energy consumption.

The cost function based greedy (CF) algorithm

For clarity of illustration, we first define some notations. Let packets be indexed in the order they are served; $s_k/d_k$ denotes the source/destination vertex of the packet $k$, $k \in K$ in the transmission graph $G$, with the information tuple $(S_k, D_k, T_k^s, T_k^e)$.

Denote $\mathcal{Y}$ as the set of packets already admitted in the system with the corresponding delivery paths $\{p_r \ (r \in \mathcal{Y})\}$. These paths can be viewed as a collection of concurrent transmissions along the maps. To admit one more relay link in a map, all the existing transmissions of the same map need to increase their power level to satisfy the SINR requirement. Thus, we obtain an evaluation of adding one more link in each map $M(t)$, by subtracting the total required energy before the addition of this link from the total energy required after the addition of this link. Specifically, denote $e'_r = \{e'_{s,e} \in p_r, s \neq e, \forall r \in \mathcal{Y}\}$ as the set of relay links in use at $M(t)$. The extra energy requirement for adding a link
\( e_{i,j}^t \) to transmit concurrently with \( e_{i}^t \), can be calculated as
\[
\varepsilon_{i,j}^t = \varepsilon( \cup_{e_s^t \in e_i^t} v_{s}^{t+1} \cup v_{j}^{t} \rightarrow v_{j}^{t+1} ) \\
- \varepsilon( \cup_{e_s^t \in e_i^t} v_{s}^{t} \rightarrow v_{j}^{t+1} ).
\]

For consistency, \( \varepsilon_{i,i}^t = 0 \), indicating that there is no extra energy consumption if node \( i \) holds the packet during \( t \). Furthermore, we define \( \varepsilon_A \) as the extra energy requirement averaged over all active relay links in \( AG \).

With the activation of one relay in the current \( AG \), some neighboring active relay links of the same map might be disabled, due to the multi-access interference. Let \( \Lambda_A^t \) be the whole set of active links, which can coexist with a previously admitted link set \( e_{i}^t \) at \( M(t) \); and \( \Lambda_{i,j}^t \subseteq \Lambda_A^t \) denotes the set of active relay links which fail in concurrently transmitting with one extra relay link \( \{ e_{i,j}^t \in \Lambda_A^t, i \neq j \} \) and \( e_{i}^t \) at \( M(t) \). For consistency, we count the relay link \( e_{i,j}^t \) itself in \( \Lambda_{i,j}^t \). We define the 'impact factor'
\[
\alpha_{i,j}^t = \frac{|\Lambda_{i,j}^t|}{|\Lambda_A^t|}, \quad (6.24)
\]
which is the fraction of active transmission links that become infeasible after adding the relay link \( e_{i,j}^t \). Thus \( 0 < \alpha_{i,j}^t \leq 1 \) quantifies the impact of \( e_{i,j}^t \) to the system at \( M(t) \). The procedure for admitting a packet \( k \in K \) is,

1. Initialization: Generate \( AG \) and associate with costs.
   
   In the transmission graph \( G \), we eliminate the links \( \{ e_{i,j}^t \in GE, e_{i,j}^t = \infty \} \).
   
   This forms a reduced active graph \( AG \). Associate each relay link \( e_{i,j}^t (i \neq j) \) in \( AG \) with a cost, calculated with the function
   
   \[
   c_{i,j}^t = e_{i,j}^t + \lambda * \varepsilon_A * \alpha_{i,j}^t, \quad (6.25)
   \]
   
   where \( \lambda \geq 0 \) is a constant. Further set \( c_{i,i}^t = 0 \).

2. Path Selection: Search for the minimum cost path.
   
   In \( AG \) with link cost \( c_{i,j}^t \), find the minimum-cost path \( p_k^t \) for packet \( k \), where the cost is additive over the multiple relay links along \( p_k^t \). This could be
formulated and solved by an algorithm for single source shortest path search in the acyclic graph [18].

3. Admission: Decide the admission and record the path.

If \( p_k^* \) doesn’t exist, packet \( k \) is rejected; otherwise, record the relay path for packet \( k \). Specifically, Insert \( k \) into the admitted packet set, \( T = T + \{k\} \). Update relay link set \( e'_T = e'_T + \{e_{i,j}^t\}, \forall e_{i,j}^t \in p_k^*, i \neq j \). Eliminate the relay links from current active graph \( AG \), where \( GE = GE - \{e_{i,j}^t\}, e_{i,j}^t \in p_k^* \); and set \( G = AG \).

According to the algorithm stated above, a packet is routed on the minimum cost path, which minimizes the cumulative cost \( c_{i,j}^t \) for the links \( e_{i,j}^t, t \in T \) along the path, where \( c_{i,j}^t \) is calculated from a function considering two major parameters, i.e. the extra energy requirement \( \epsilon_{i,j}^t \) and the fraction of affected links \( \alpha_{i,j}^t \) at \( M(t) \).

The extra energy requirement at \( M(t) \) is calculated with consideration to the mutual interference among pairs of transmissions \( e_{i,j}^t \cup e_{i,j}^t \). If the transmission pairs are too close to each other, mutual interference can be extremely high and congestion occurs. In such a case, it is non-feasible to serve this extra link \( e_{i,j}^t \), even though there is no power limit in the system. Thus, \( \epsilon_{i,j}^t \) depends heavily on the concurrent transmission pairs and might be greatly higher than the individually calculated energy demand on each relay link, where no multi-access interference is considered.

Furthermore, with mobility globally taken into account in the path selection, the packets tend to compete for the relay links in sound condition. Due to the greedy nature of the path selection, a relay link on the admitted path does not change. This link causes interference to other concurrent transmissions and might prohibit later packets’ usage of relay links in its vicinity. To choose a link that gives relatively low impact to other relay links at the same map, we incorporate the impact factor \( \alpha_{i,j}^t, i \neq j \) into the cost function, \( c_{i,j}^t = \epsilon_{i,j}^t + \lambda \epsilon_A \alpha_{i,j}^t \). Here, \( \epsilon_A \)
is incorporated for quantitative adjustment; \( \lambda \geq 0 \), such that the cost for adding the relay link is monotonically increasing with the impact factor \( \alpha_{ij}^{t} \). Note that a critical constant \( \lambda \) should be pre-defined: if \( \lambda = 0 \), the algorithm comes to be the most straightforward greedy minimum energy (ME) routing for each packet; To set a higher \( \lambda \) is to put more emphasis on the link’s impact to the network. According to the cost function, the impact factor is considered in a cumulative way, as a path causes interference along all its relay links. Generally speaking, if there are two feasible paths with comparable energy consumption and impact factor at each relay link, it is better to use the path with fewer relays, as more relays affect more maps.

The cost function captures the information on both the extra energy consumption for adding the relay link and its impact on the system at that instance. The shortest path algorithm determines a path that tries to utilize links in sound condition, i.e. links with short transmission distance and low mutual interference; while at the same time, provides room for later packets by occupying fewer hops with low impact to other active relay links.

Compared with the optimal scheme, the greedy algorithm better captures the demand for on-line routing that requires lower complexity and serves packets according to the arrival sequence or the priority. The computation complexity for calculating the impact factor for one active relay link is at most \( O(|N|^2) \), thus the worst case complexity for initialization step is \( O(|T| \cdot |N|^4) \). Furthermore, the route selection step is determined by the shortest path computation, which is \( O(|T| \cdot |N|^2) \) for a directed graph. Thus, the overall greedy algorithm computes the set of packet delivery paths with worst case complexity \( O(|K| \cdot |T| \cdot |N|^4) \).

**Iterative Adjustment**

To further reduce the energy consumption in the system, we make the iterative adjustment as follows: in each round of iteration, we sequentially scan all the packets \( \gamma \in \mathcal{Y} \). Assume the admitted path for \( \gamma \) is \( p_\gamma \). We subtract the packet
from the admission set, i.e. \( Y = Y - \{ \gamma \} \). Eliminate all the relay links along \( p_{\gamma} \) from the current transmission set, i.e. \( e_{t}^{i} = e_{t}^{i} - \{ e_{i,j}^{t} \} \), \( \forall e_{i,j}^{t} \in p_{\gamma} \). Re-format the active graph and re-calculate the shortest path for packet \( \gamma \), following a similar procedure as illustrated in Section 6.3.2. Here, for the purpose of reducing energy consumption, unlike the cost function we use above, in the adjustment, we directly minimize the cumulative extra energy consumption, i.e. \( c_{i,j}^{t} = c_{i,j}^{t}, \) thus resulting in the minimum energy path. If a lower energy consumption path \( p'_{\gamma} \) can be found, we update the path with \( p'_{\gamma} \). Furthermore, the re-scheduling of the relays for a packet might reduce the system-wide interference and present room for more packet admission. Thus, for those packets that are rejected by the CF greedy algorithm, we could re-calculate the minimum energy path for them. The iterative sequential adjusting/packing procedure continues until the equilibrium reaches, i.e. no packet could obtain a lower energy consumption path by individually adjusting its relays.

Discussion

**Power Constraints:** In some systems, there is strict limitation on the power emission at each node. Such restriction ensures better coexistence with other systems working in the same area and prolongs the battery life. The algorithm proposed in this paper can be easily adjusted to deal with this by incorporating the power limitation in minimum power vector calculation (6.7). As a result, the high-cost links are eliminated from later usage, which restricts the energy consumption at each node.

**Complexity Heuristic Approach:** According to the analysis for the CF greedy algorithm in Section 6.3.2, main computation efforts are devoted to the calculation of the impact factor \( \alpha_{i,j}^{t} \). To reduce the complexity, we could apply a slightly different parameter, which measures the fraction of interfering nodes. Let \( \Lambda_{i,j}^{t} \) denote the set of nodes in the interfering region for the active link \( e_{i,j}^{t} \) at time \( t \). Suppose node \( i \) and \( j \) located at \( X_{i}^{t} \) and \( X_{j}^{t} \) in \( M(t) \). \( \Lambda_{i,j}^{t} \) includes all nodes \( n \)
located at $X^t_n$ which satisfy

$$|X_n - X_i| \leq (1 + \Delta)|X_j - X_i|$$  \hspace{1cm} (6.26)

where $\Delta$ is a constant, quantifying the interference region. Further denote $\Lambda^t_A$ as the set of nodes, which can be actively in use before the addition of relay link $e^t_{i,j}$. We could define the impact factor

$$\alpha^t_{i,j} = |\Lambda^t_{i,j}|/|\Lambda^t_A|,$$

which denotes the fraction of nodes interfered with by the addition of transmission $e^t_{i,j}$. Basically, the addition of a link impacts on its neighboring links and limits their usage by other packets. Thus, the measurement on the interfering links is more precise than that of the interfering nodes.

### 6.3.3 Simulation Results for the Interference-Limited Regime

Simulations have been conducted on a mobile network configuration. As illustrated in Figure 6.11, the $N$ mobile nodes uniformly locate in a square region of size $100 \times 100$. Each arrow denotes the trajectory of a node: the tail and the head of the arrow are the starting and ending positions respectively; and these starting and ending positions are randomly generated. Each node is assumed to move at a constant speed toward the ending position, within the system time interval $[1, T]$ and, during each discrete time interval, the the node locations do not change.

We first present results for a small setting, as shown in Figure 6.11 (left). Specifically, there are 6 mobile nodes, and overall $T = 5$ discrete time intervals.

We consider 4 packets routing with the $(S, D, T^*, T^*)$ information as $f1 = (2, 4, 1, 4)$, $f2 = (3, 6, 1, 4)$, $f3 = (3, 5, 2, 5)$ and $f4 = (4, 3, 2, 5)$. The individually calculated optimal paths for the four packets are shown in Figure 6.12, where the impact from any other concurrent transmission is not considered. It shows that optimal relays may occur during the same interval: for example at time $t = 3$,
Figure 6.11: Network setting

$f_1$ tend to use relay $N_2 \rightarrow N_4$, $f_2$ $N_3 \rightarrow N_4$ and $f_4$ $N_4 \rightarrow N_3$, with relatively sound channel condition; in some cases, the packets even tend to share the same channel, e.g. link between $N_3$ and $N_4$ for $f_2$ and $f_4$. In an interference-limited system, such concurrent transmissions might be infeasible or may cause very high energy consumption, and thus have to be re-scheduled. Figure 6.13 shows the coordinate optimal routing, where packet $f_1$ is re-scheduled to be delivered directly at \( t = 4 \) from $N_2$ to $N_4$, and thus to make room for $f_4$’s delivery at \( t = 3 \), when the channel between $N_4$ and $N_3$ is good at \( t = 3 \). The optimal paths for $f_3$ is the same as its individually calculated counterpart and $f_2$ is re-scheduled to relay from $N_3$ to $N_2$ first and then to $N_6$. Though transmissions $2 \rightarrow 6$ and $3 \rightarrow 5$ occur at the same instant $t = 2$, their mutual interference is sufficiently low. As a result, the packets’ routing is globally coordinated for the energy usage minimization.

In Figure 6.14, we show the result for the straightforward minimum energy (ME) algorithm, where only 3 packets can be admitted. The reason is that the greedy behavior of 2-hop path selection for $f_2$ and $f_3$ prohibits further relays along the maps. The result for the proposed heuristic algorithm that takes into account the impact on the network ($\lambda = 0.5$) is shown in Figure 6.15. In this case, all four packets are admitted. In order to avoid too many interfering links, each packet tends to use direct delivery. Later, Figure 6.16 shows the result after the iterative adjustment. To decrease the energy usage, $f_2$ uses the two-hop relay again and $f_4$ is rescheduled to deliver at $t = 4$ with greatly lower energy consumption. The
energy consumption for the cost function based (CF) heuristic approach is 2.72J, then after adjustment it is 0.62J, which is comparable to the optimal solution 0.48J.

We also study the average system performance in a uniformly distributed 10-node network with overloaded traffic demand, shown in Figure 6.11 (right). The total time $T = 10$. The results are averaged over 40 traffic settings, each with 30 packet requests. The source-destination pair for each packet is randomly generated, with the delay constraint $[1, 10]$. In Figure 6.17 and Figure 6.18, we
show the effect of the parameter $\lambda$ on the average number of packets that can be admitted by the system and the corresponding average energy consumption. Two sets of results are illustrated, the upper sub-figure is for the mobile network setting and the lower one is for the static setting, where the node locations are the same as map $M(1)$, i.e. the starting position for the mobile setting. For clarity, in Figure 6.17, 'Greedy' column denotes the result for the cost function based (CF) greedy algorithm, and 'AdjustMP' column for the iterative adjustment with more packets addition; in Figure 6.18, the denotations for 'Greedy' and 'AdjustMP' are the same as in Figure 6.17, while 'Adjust' columns refer to the results for the iterative adjustment with lower energy consumption only. Results show that, firstly, higher $\lambda$ may increase the throughput with greater energy consumption. However, when $\lambda$ goes too high, such as greater than 1 as shown in the figure, the improvement in throughput becomes marginal, while the energy requirement increases significantly. As a result, we need to carefully study the tradeoffs and properly adjust the const $\lambda$; Secondly, from the 'Adjust' column, great energy saving can be observed with the iterative adjustment. Furthermore, according to the 'AdjustMP' column, even though the energy consumption tends to surge
when the system works near the throughput margin, after the iterative (with more packets admission) process, the average total energy consumption decreases, although the average number of admitted packets increases. This demonstrates the efficiency of the iterative adjustment process. Lastly, as expected, mobile setting performs much better than its static counterpart with a greater number of packets admitted and lower energy consumed. The reason is that with the mobility of nodes taken into account globally, better relay nodes and channel conditions might be encountered during the process of packet delivery.

In table 6.2, we show the average throughput and energy consumption of the straightforward minimum energy (ME) algorithm. Again, results for both mobile (column 2) and static network (column 3) are given, where the mobile setting achieves much higher throughput (16 over 12) with slightly lower energy consumption. Compared with the results shown in Figure 6.18 and 6.17, it is
Figure 6.15: CF greedy routing

obvious that our CF heuristic approach performs much better than the ME greedy algorithm. Particularly with moderate const $\lambda$, the CF heuristic approach can achieve higher throughput with much lower energy consumption. Simulation results also indicate that with greater emphasis on the impact factor by setting higher $\lambda$, packets tend to use less hop paths, due to the cumulative impact effect in the cost function.

In Figure 6.19, we take a close look at the cumulative energy requirement with the increase of the number of admitted packets in the mobile system. We compare the results among the minimum energy greedy packing (ME) and the cost function based greedy algorithm (CF), with three different $\lambda$ setting, (0.01, 1 and 100). For clarity, the ‘y’ axis is represented in logarithm. As shown in the figure, when the system is light-loaded, the energy requirement for the most greedy ME algorithm is the lowest. The reason is that the CF approach tends
to avoid long hop paths, at the cost of higher energy usage. With more packets admitted, the increase in energy consumption for the ME scheme is steep; while its CF counterparts are comparatively moderate. This is because the not-so-greedy behavior of CF scheme provides room for later packet relay by occupying the links with less impact on the system. Thus, as shown in the figure, after some packets' admission, the energy consumption for CF scheme with moderate \( \lambda \) becomes lower than ME algorithm. One thing to be noted here is that when the system works near saturation, some jitters occur. This comes from the fact that we calculate the averaged value over all the settings. Generally speaking, adding more relays causes a surge in energy consumption when the system becomes congested. As the packet sets are generated randomly, for a set with relatively low throughput, it might contribute a lot to the average energy value near the throughput margin. Then, for the average value of the next packet in the figure, it might be omitted.
Figure 6.17: Number of admitted packets with the increase of $\lambda$

from the average calculation, resulting in an energy consumption jitter shown in the figure. Note that very high $\lambda$ causes very high energy consumption, thus suggesting careful adjustment on the critical const $\lambda$.

6.4 Summary

Node mobility incurs rapid variations in channel strength, which could be exploited for possibly more cost-effective communications. In such a network, it is appealing to take advantage of node mobility for fast packet delivery, low energy consumption and high throughput. This work investigates the relationships among several key parameters including throughout, delay, energy consumption and interference in the context of routing in a mobile multi-hop network.
Figure 6.18: Energy consumption with the increase of $\lambda$

In this chapter, we first investigate optimal routing in the interference-free regime, where a packet’s delivery route could be optimally selected, while ignoring interference from other simultaneous transmissions. Two kinds of systems, both with and without power control, are considered. For a system without power control, the transmitted power per hop is a constant and each transmitter has a fixed transmission range. We propose hop-expansion based algorithm to minimize the packet delivery time and minimize the total energy usage. In a system with power control, the transmission ranges may be flexibly adjusted. The optimization objective is to minimize the total energy usage for one packet delivery from its source to the destination, subject to certain end-to-end delay bound. This is calculated by the layered optimization.
Figure 6.19: Cumulative energy consumption with the increase of admitted packets number

Results in the interference-free regime demonstrate the tradeoff between the delivery time and energy consumption. To deliver a packet as fast as possible, the information tends to be transmitted when two nodes first fall into the radio range or relayed immediately after the previous transmission. On the other hand, to deliver a packet with minimum energy consumption, a potential transmitter may hold the information first and transmit to the receiver when they are close to each other.

Later we extend the minimum energy scheme to the interference limited regime. We investigate the energy efficient packet routing in a multi-hop wireless network, where mobility is taken into account by adopting a deterministic model. We consider the objective of minimizing the energy consumption for packet delivery, subject to the packet delay constraint and SINR requirement among concurrent transmissions. This can be formulated and solved by a dynamic program-
ming algorithm. Furthermore, we also present a heuristic approach, which admits packets in a greedy manner. The heuristic approach only involves the shortest path computation, thus can better scale to the network size and the on-line traffic demand.

Simulation results indicate that, with mobility globally taken into account, the performance can be greatly improved over a wide range of network settings. Generally speaking, in a wireless ad hoc network, to deliver a packet in the most energy-efficient manner, multiple hops relay may be involved and a sender might better hold the packet first, and transmit when the relay link is in a sound channel condition. When interference among concurrent transmissions can be ignored, to take advantage of node mobility, multiple packets might use nearby concurrent transmissions or even share exactly the same transmission link at the same instant for packet relay. However, in an interference-limited system, these concurrent transmissions have to be globally adjusted or re-scheduled to satisfy the multi-access interference constraint and keep relatively low energy consumption. Specifically, to ensure high throughput and low energy usage, we should preferably choose those transmission links with relatively low energy requirement while at the same time, impact very little on the system, such that multiple packets may efficiently share the resources. In this case, the packets may also tend to choose less-hop relay paths. In other words, the system performance would be greatly improved if the packets' delivery can be arranged through few relays, each with relatively low energy consumption and little impact to the system.

The results obtained in this study is under a deterministic mobility model, which provides upper bounds on optimal routing by taking advantage of node mobility under different QoS schemes. Such an approach might not be practically feasible. Our on-going work is examining more realistic scenarios, including several commonly adopted mobility patterns.
CHAPTER 7

JOINT ROUTING AND SCHEDULING IN MULTI-RADIO MULTI-CHANNEL MULTI-HOP WIRELESS NETWORKS

Co-channel interference and multi-hop relays constitute main performance constraints of a general wireless network [31]. Due to co-channel interference, only links that are located out of their mutual interference range can transmit at the same time. When traffic needs to transverse multiple hops from source to destination and concurrent transmissions of neighboring links are prohibited along delivery paths, the throughput in the system is largely degraded. Consequently, routing (i.e. selection of multi-hop relays) and scheduling (i.e. activation of concurrent transmissions) are closely related while greatly affecting each other, and they should be jointly considered for the performance optimization in a general wireless network. Think of a system deployed with multiple radios and multiple channels, and allow a radio interface to switch across the channels as needed, more concurrent transmissions might be enabled over those previously conflicting links. This offers the potential of substantially improving the system performance in the network, yet it provides new challenges in the design of more efficient scheduling and routing algorithm, which is the focus of our research in this paper. Previous work in the multi-radio multi-channel network have largely focused on the design of link level protocols [5] [36] [60] [61] [76] [82] [90]. There are also works addressing routing for such a network [21] [70]. In this work, we focus on the derivation of maximum performance gain through cross-layer joint routing and scheduling. We consider a stationary multi-hop wireless network with possible multiple radio interfaces installed in high capacity nodes, which complies well with the “community wireless network” application. We investigate the achievable performance in
a multi-radio multi-channel network, given system information such as the channel supply, the traffic load, the number of radio interfaces in each node, and the network/channel status. Specifically, the performance is measured by the minimal total system activation time in use to satisfy the specific traffic demand. As the arrangement of concurrently transmitting links along the system time-line and the selection of end-to-end paths may have strong impact on each other, we jointly consider routing and scheduling in such a multi-radio multi-channel static network by treating the corresponding (large-scale) combinatorial optimization problem. The new multi-radio multi-channel scenario incurs much higher computational complexity in the optimization, which complicates the problem and asks for the design of more efficient algorithms to handle the problem. This largely differentiates our work from previous efforts on cross-layer joint optimization [20] [37] [46].

Specifically in our work, under the constraints of co-channel multi-access interference and the number of radios deployed at each node, we use the notion 'concurrent transmission pattern' to model a feasible link schedule with multi-radio multi-channel supply, which is a collection of transmission rates associated with the links that can be simultaneously supported in the network. By activating a variety of concurrent transmission patterns throughout different time slots, we can obtain a set of capacity graphs, in which each link indicates the maximum rate supply between a pair of connecting nodes. Routing under such a capacity graph can be formulated and solved as a multi-commodity network flow problem. We present a mathematical formulation for the joint optimization, which turns out to be a Mixed Integer Programming (MIP) problem that is well-known to be NP-hard. We then re-formulate the optimization problem based on concurrent transmission patterns. Since the number of possible concurrent transmission patterns increases exponentially with the number of nodes, links, and channels due to the combinatorial nature, it is impossible to enumerate them. We further adopt a column generation (CG) based approach [62] to solve the problem. The CG approach decomposes the original problem into two sub-problems and solves
them iteratively, with possible improvement during each iteration. The resulting sub-optimal algorithm for joint routing and scheduling yields efficient selection and arrangement of the current transmission patterns to support the traffic demand and gives an upper bound on system activation time requirements.

The remainder of the chapter is organized as follows. In the next Section 7.1, we present the MIP formulation for the joint routing and scheduling problem and a re-formulation based on concurrent transmission patterns. The column generation based approach is introduced in the subsequent Section 7.2. We investigate results under different network configurations in Section 7.3, and conclude the paper in Section 7.4.

7.1 Problem Formulation

7.1.1 System Description

We consider a set of $N$ nodes, labelled $\{1, 2, ..., N\}$, which are arbitrarily distributed in a wireless network. Each node is equipped with one or more wireless interface cards, referred to as radio in this work. We denote the number of radios in each node $i$ as $\gamma_i$, $i = 1, 2, ..., N$. Assume $K$ orthogonal channels are applicable in the system, with no inter-channel interference. We consider the system with channel switching capability, such that a radio can dynamically switch across the different channels. However, at any instant, it can transmit to or receive from at most one other radio in its vicinity, and both the transmitter and the receiver should tune into the same channel to communicate.

Assume no power control is adopted at the nodes and each node $i$, $i = 1, 2, ..., N$ has a transmission range $T_i$. Denote the distance between nodes $i$ and $j$ as $d_{i,j}$. Then node $i$ can transmit to node $j$ only when $d_{i,j} \leq T_i$. In that case, there can be a directed transmission from $i$ to $j$, denoted as link $(i, j)$. Consequently, the set of feasible transmission links can be represented by

$$\mathcal{E} = \{(i, j), d_{i,j} \leq T_i, 1 \leq i, j \leq N, i \neq j\}. \quad (7.1)$$

120
Further denote the transmission rate at an active link \((i, j)\), working on one channel, as \(r_{i,j}\). Here, we assume all channels share similar properties, in terms of the same transmission range and transmission rate at each link, and the channel conditions are stationary. Furthermore, for any individual channel in use, the mutual interference among concurrent transmissions in the vicinity is the main constraint in link scheduling. In this work, we adopt a protocol model\(^1\) where two links \((i, j)\) and \((u, v)\) can be simultaneously active on the same channel if and only if the corresponding receivers \(j\) and \(v\) locate out of the mutual interference range of the transmitters. That is, \(d_{i,v} > (1 + \delta) \cdot T_i \& d_{u,j} > (1 + \delta) \cdot T_u\), where \(\delta \geq 0\) is a small positive constant. This condition also implies the constraint that, on any particular channel, each node can transmit to or receive from at most one other node at a time. If the above condition holds, we say the two links are independent. Accordingly, we could define the following matrix \(F(\|\mathcal{E}\|, \|\mathcal{E}\|)\) to summarize the independence status between any pair of links, where \(\forall (i, j), (u, w) \in \mathcal{E}\)

\[
F[(i, j), (u, w)] = \begin{cases} 
1, & \text{if } (i, j) (u, w) \text{ are independent} \\
0, & \text{otherwise.}
\end{cases}
\]

The system we are looking into operates in a synchronous time-slotted manner. In each time slot, the central network can select a schedule, which is the concurrent transmission on a set of independent links, and this selected schedule would be active throughout the slot. Further assume the arrangement of schedules is periodic. Within each period, there are \(T\) time slots, labelled as \(t = \{1, 2, \ldots, T\}\), and the network would switch over different schedules along them. The time-scale of a period is relatively small, compared with the system time range for traffic delivery. Thus in the long term, the network can be viewed as transmitting smoothly with the link capacity obtained by the combination of link rates from different schedules along the \(T\) time slots.

Let the traffic demand be a collection of \(M\) end-to-end sessions,

\[
S_m = (s_m, d_m, R_m), \ s_m, d_m = 1, 2, \ldots N, \ R_m > 0, \ m = 1, \ldots M.
\]

\(^1\)The formation can be slightly modified to accommodate more accurate physical model in our work [10].
Specifically for the session $S_m$, a source node $s_m$ would have a rate demand $R_m$ to a destination node $d_m$. For clarity, we assume that the source nodes always have data to send and the destinations are always ready to accept data.

We assume a global mechanism exists, which regulates the nodes transmitting/receiving on a specific channel within any slot. Our objective is to provide the bounds on the system performance in a multi-radio multi-channel network. As we consider the application of a wireless community network with relatively stable topology and traffic demand, such global agreement might be achieved system-wide.

### 7.1.2 Problem statement

Our objective is to minimize the system activation time, i.e. the total number of active time slots in use, to deliver the assigned traffic load over the multi-radio multi-channel multi-hop wireless network described as above. We assume that packet transmission at an individual node can be perfectly scheduled by an omniscient and omnipotent central entity. Thus, we do not consider issues such as MAC contention. Our formulation can be easily extended to accommodate other objectives, such as to maximize the minimum throughput for all source destination pairs.

Assume multi-path routing is allowed for the traffic delivery of multiple sessions with different demands. This turns out to be a multi-commodity routing problem in a general wired system. However, due to the interference-limited nature in wireless networks, the capacity of a link depends heavily on the underlying scheduling policy, resulting in a variety of feasible routing solutions. To achieve minimum time slots in use, we need to jointly optimize routing and scheduling for the traffic delivery. Specifically, for the scheduling part, feasible co-channel concurrent transmissions in the system consist of several isolated (independent) links. By combining such concurrent transmission links over multiple channels subject to the radio interface constraint at each node, we could obtain concurrent
transmission patterns (CTPs). Taking the benefit of time-sharing scheduling of different CTPs over the time slots, we could achieve a variety of capacity graphs, associated with capacity constraint on each link. Multi-commodity routing can thus be performed on the underlying capacity graph to find the optimal paths.

We illustrate the idea through a simple example in Figure 7.1 and show how to achieve our objective by taking routing and scheduling jointly in the formulation. Assume that all links have an identical throughput of one unit on each channel and the total time slot is $T = 10$. We consider the traffic demand $R = 0.2$ unit from source node N1 to destination node N4. In other words, N1 must actively sending out traffic for 2 slots every period. The directional edges in the figure represent all feasible transmission links, where each link denotes a transmitter-receiver pair located within their mutual transmission range. For simplicity, the interference range for each node is assumed to be the same as the transmission range. According to our system model, any two links in this figure would interfere with each other when transmitting on the same channel, due to the co-channel interference. That is, on any given channel, only one link can be active at a time. We assume two channels, namely $c1$ and $c2$, are deployed in the system, and each node is equipped with one radio interface.

According to the figure, to deliver traffic from N1 to N4, four possible paths might be chosen from, i.e. $(N1 \rightarrow N2 \rightarrow N4), (N1 \rightarrow N2 \rightarrow N3 \rightarrow N4), $
(\(N1 \rightarrow N3 \rightarrow N4\)) and (\(N1 \rightarrow N3 \rightarrow N2 \rightarrow N4\)). When multi-path routing is allowed, we could use any combination of them. However, the link capacity in a wireless network can not be obtained as a priori, thus we can not tell which routing scheme is optimal before looking into the underlying scheduling policy. Assume that we decide to route traffic through the path (\(N1 \rightarrow N2 \rightarrow N4\)). Due to the mutual interference between links (\(N1 \rightarrow N2\)) and (\(N2 \rightarrow N4\)), and the one radio constraint at N2 (i.e. N2 can only transmit or receive at a given time), the two links along the path must be scheduled at different time slots. Thus the minimum number of time slots required for the traffic delivery is 4, through separately scheduling each link to be active for two slots on either channel.

For this example, the optimal routing from N1 to N4 indeed contains two paths, i.e. \(N1 \rightarrow N2 \rightarrow N4\) and \(N1 \rightarrow N3 \rightarrow N4\), each burdening half of the demand. The corresponding scheduling is, 1) activating both links (\(N1 \rightarrow N2\)) and (\(N3 \rightarrow N4\)) at the same time on different channels. This scheduling pattern occupies one time slot every period; 2) similarly, activating links (\(N1 \rightarrow N2\)) and (\(N3 \rightarrow N4\)) concurrently on different channels for one slot time. To comply with the notations we define later, specifically in this case, two CTPs are applied in the scheduling. They are, Pattern 1: link (\(N1 \rightarrow N2\)) on channel c1/c2 and link (\(N3 \rightarrow N4\)) on channel c2/c1; Pattern 2: link (\(N1 \rightarrow N3\)) on channel c1/c2 and link (\(N2 \rightarrow N4\)) on channel c2/c1. Through a time-sharing combination of these two patterns, one for each slot time, the optimal capacity graph has the capacity constraint (i.e. long-term link rate) 0.1 on each of the four links. From this example, we could observe that the end-to-end performance heavily depends on the link scheduling along delivery paths. Thus, to optimize the system performance, routing and scheduling must be jointly considered in a multi-hop network. Detailed formulations of concurrent transmission patterns, capacity graphs, as well as joint routing and scheduling optimization, are introduced in the following sections.
7.1.3 Optimal Formulation

We show in the following the optimal formulation of allocating the time slots to the transmission links to be active on different radios and channels, to deliver the specific traffic load. We define $f_{i,j}^m$ as the amount of flow on link $(i,j)$, $(i,j) \in \mathcal{E}$ for session $m$ and further define the following indicator variables:

$$y_t = \begin{cases} 1, & \text{if time slot } t \text{ is in use;} \\ 0, & \text{otherwise.} \end{cases} \quad (7.2)$$

$$v_{i,j}^{t,k} = \begin{cases} 1, & \text{if link } (i,j) \text{ at slot } t \text{ active on channel } k; \\ 0, & \text{otherwise.} \end{cases} \quad (7.3)$$

The optimal formulation (OF) based on the slot usage can be shown as,

$$[\text{OF SLOT}] \quad \min \sum_{t \in T} y_t \quad (7.4)$$

$$\sum_{j \in N \setminus \{i,j\} \in \mathcal{E}} f_{i,j}^m - \sum_{j \in N \setminus \{i,j\} \in \mathcal{E}} f_{j,i}^m = 0, \quad \forall i \in N - \{s^m, d^m\}, m = 1...M \quad (7.5)$$

$$\sum_{j \in N \setminus \{s^m,j\} \in \mathcal{E}} f_{s^m,j}^m - \sum_{j \in N \setminus \{j,s^m\} \in \mathcal{E}} f_{j,s^m}^m = R^m, \quad \forall m = 1...M \quad (7.5)$$

$$\sum_{m=1}^{M} f_{i,j}^m \leq \frac{1}{T} \sum_{t \in T} \sum_{k=1}^{K} v_{i,j}^{t,k} \cdot r_{i,j}, \quad \forall (i,j) \in \mathcal{E} \quad (7.6)$$

$$v_{i,j}^{t,k} \leq y_t, \quad \forall (i,j) \in \mathcal{E}, t \in T \quad (7.7)$$

$$x_{n}^{t,k} = \sum_{(n,j) \in \mathcal{E}} v_{n,j}^{t,k} + \sum_{(i,n) \in \mathcal{E}} v_{i,n}^{t,k}, \quad \forall k \in K, n \in N, t \in T \quad (7.8)$$
\[
\sum_{k=1}^{K} x_{n}^{t,k} \leq \gamma_n, \forall n \in N
\]  
(7.9)

\[
v_{i,j}^{t,k} + v_{u,w}^{t,k} \leq 1 + F[(i, j), (u, w)], \forall (i, j), (u, w) \in \mathcal{E}, k \in K
\]  
(7.10)

\[
y_t = \{0, 1\}, f_{i,j}^m \geq 0, v_{i,j}^{k,t} = \{0, 1\}, \forall (i, j) \in \mathcal{E}, k \in K, t \in T, m = 1, 2,...M
\]  
(7.11)

Here, \(x_{n}^{t,k}\) indicates whether node \(n\) is transmitting or receiving on channel \(k\). The objective function (7.4) minimizes the total number of time slots that should be actively in use. Equation (7.5) and (7.5) are the constraints for the multi-commodity routing. Equation (7.6) exerts capacity constraints on the set of links. Later equations (7.8) (7.9) and (7.10) ensure the constraints for possible concurrent transmissions. In the following, we discuss some important concepts in detail.

- Concurrent Transmission Pattern

Due to the co-channel interference, a feasible arrangement of concurrent transmissions on one channel consists of a set of spatially isolated non-interfering links. To take multi-radio multi-channel supply into consideration, a concurrent transmission pattern (CTP) is defined as a feasible schedule of transmission rates on different links at a time.

Formally, any CTP \(p\) can be represented by a set of length \(|\mathcal{E}|\), where each item corresponds to a link \((i, j) \in \mathcal{E}\) and denotes the feasible transmission rate on \((i, j)\), written as \(p = \{a_{i,j}^p, (i, j) \in \mathcal{E}\}\). Such a CTP can be obtained through the combination of concurrent transmissions over multiple channels, while maintaining the radio constraint at each node.

To be precise, the elements for a CTP can be calculated as
\[ a_{i,j} = \sum_{k=1}^{K} v_{i,j}^k r_{i,j}, \quad \forall (i,j) \in \mathcal{E}, \quad (7.12) \]

where variables \( v_{i,j}^k \) satisfy the constraints (7.8) (7.9) and (7.10). Here we ignore the superscript \( t \) in the equations.

Specifically, equation (7.8) represents the relationship of the link usage to link radio usage mapping. If a link \((i,j)\) is active on the channel \(k\), both end nodes \(i\) and \(j\) must dedicate one radio to serve the communication. As a node can communicate with at most one other node on a channel at a time, an indicator variable \( x_n^k \) is sufficient to summarize the status of node \(n\) on channel \(k\) for any transmission pattern, where

\[ x_n^k = \begin{cases} 1, & \text{if node } n \text{ transmits/receives on channel } k; \\ 0, & \text{otherwise.} \end{cases} \quad (7.13) \]

Equation (7.9) ensures that the radio in use does not exceed the radio equipment at each node and equation (7.10) shows that if two links are concurrently active on a channel, they should be independent of each other.

- Capacity Graph

Denote \( \mathcal{P} \) as the collection of all feasible CTPs, where each pattern \( p = \{a_{i,j}^p, (i,j) \in \mathcal{E}\} \). We define a 'capacity graph' as the set of capacity constraints, i.e. the maximum transmission rate, associated with each link. This can be obtained by a time-sharing activation of different concurrent transmission patterns along the \(T\) time slots, where the activation of the link \((i,j)\) for one slot can contribute \( \frac{1}{T} \cdot r_{i,j} \) throughput to the link. Thus, the elements of a feasible capacity graph can be represented by,

\[ c_{i,j} = \frac{1}{T} \sum_{t \in T} \sum_{k=1}^{K} v_{i,j}^k \cdot r_{i,j}, \quad \forall (i,j) \in \mathcal{E}. \]
Denote $f^m_{ij}$ as the amount of session $m$'s flow on link $(i, j)$, $(i, j) \in \mathcal{E}$, the equation (7.6) ensures that the total amount of traffic from different sessions passing through a link never exceeds the capacity of the link, where the capacity constraint is calculated by the combination of the active CTPs as illustrated above.

- Multi-commodity Routing

Provided with the underlying capacity graph by the time-sharing scheduling of the CTPs, the multi-commodity routing algorithm can be performed to obtain optimal delivery paths. The objective is to minimize the total number of active time slots in use, to satisfy the set of end-to-end traffic sessions $S_m$. In the optimal formulation [OF SLOT] above, equation (7.5) states the flow conservation requirement. That is, for each traffic session going through any node except for the source and destination, the amount of incoming flow is equal to the out-going flow. Equation (7.5) shows the traffic demand at the source nodes.

### 7.1.4 Pattern-Based Re-Formulation

The [OF SLOT] formulation is designed according to the system requirement, which turns out to be a Mixed Integer Programming (MIP) problem. As a NP-hard problem, it is hard to solve to optimality within a reasonable time, even for a moderate system size. If we have the knowledge of all concurrent transmission patterns (CTP) as a priori, the optimization problem can be reformulated based on the set of CTP. In this section, we perform such reformulation, such that an efficient column generation (CG) approach can be applied.

Let $\mathcal{P}$ be the whole set of CTPs and define each CTP as $p = \{a^p_{ij}\}, p \in \mathcal{P}$. Further denote $f^m_{ij}$ as the amount of flow on the link $(i, j)$, $(i, j) \in \mathcal{E}$ from session $m$; $x_p$ as the number of time slots that would be allocated to the pattern $p$. The joint routing and scheduling optimization which minimizes total system time slots in use can be reformulated as follows:
\[
[\text{OF\_PATTERN}] \quad \min \sum_{p \in \mathcal{P}} x_p \quad (7.14)
\]

\[
\sum_{j \in N: (i,j) \in \mathcal{E}} f_{i,j}^m - \sum_{j \in N: (j,i) \in \mathcal{E}} f_{j,i}^m = 0, \quad \forall i \in N - \{s_m, d_m\}, m = 1...M \quad (7.15)
\]

\[
\sum_{j \in N: (s_m,j) \in \mathcal{E}} f_{s_m,j}^m - \sum_{j \in N: (j,s_m) \in \mathcal{E}} f_{j,s_m}^m = R_m, \quad \forall m = 1...M \quad (7.16)
\]

\[
\frac{1}{T} \sum_{p \in \mathcal{P}} x_p \cdot a_{i,j}^p - \sum_{m=1}^{M} f_{i,j}^m \geq 0, \quad \forall (i,j) \in \mathcal{E} \quad (7.17)
\]

\[
f_{i,j} \geq 0, x_p = \{0, 1, ..., T\}, \quad \forall (i,j) \in \mathcal{E}, p \in \mathcal{P}, m = 1, 2...M \quad (7.18)
\]

Here, the objective function (7.14) is for the cumulative time slots minimization. Note here, if \(\min \sum_{p \in \mathcal{P}} x_p \leq T\), then end-to-end traffic demands can be routed through feasible scheduling of CTPs in the system. Again, equation (7.15) ensures the flow conservation, equation (7.16) guarantees the traffic demand at source nodes, and equation (7.17) exerts the capacity constraint at each link.

In this case, to obtain the optimization, we need to construct the whole set of CTPs \(\mathcal{P}\) as a priori, which incurs major difficulty in solving the problem [OF\_PATTERN]. According to [37], even to derive all feasible schedules in a single channel wireless network is an NP-hard problem. In our problem, as the number of feasible concurrent transmission patterns increases exponentially with the number of links, radios and channels involved, set \(\mathcal{P}\) could be extremely large. To handle such complexity, we adopt the column generation technique [62] to reformulate and solve [OF\_PATTERN]. The details are given in the following Section 7.2.

129
7.2 Column Generation Based Approach

The column generation (CG) approach helps to reduce the complexity of constructing a whole set of columns, by effectively selecting columns that make improvements for the optimization. Specifically, it decomposes a linear programming (LP) problem into two parts, i.e. a master problem and a sub-problem. The master problem starts with a small subset of columns, namely the Basis, to solve the optimization problem. To check the optimality of the original problem under the current Basis, a sub-problem is solved. If the current solution is non-optimal, the sub-problem identifies a new column that might make improvements to the original optimization. This column is expanded to the Basis, and the master problem is re-calculated based on the enlarged column set. Therefore, in order to find the optimal solution, the algorithm alternates between the master problem and the sub-problem, and a new column is selected in each round. The procedure of the iteration does not stop until the master problem contains all columns that contribute to the optimality of the original problem. Generally speaking, the set of columns that are selected for optimization is relatively small compared to the whole set of columns in the original problem, which demonstrates the efficiency of the CG approach.

Specifically in our column generation based approach, we consider the LP relaxation of the [OF_PATTERN] problem. The content of a ‘column’, which would be figured out by the sub-problem, corresponds to exactly one concurrent transmission pattern (CTP) discussed in Section 7.1.3. Our CG master problem performs the LP optimization, and the sub-problem generates such CTP related columns one by one to improve the objective result. As in this work, a relatively small set of CTPs contributes to the optimization, the CG approach avoids generating the whole set of feasible CTPs by leaving out most of them during the procedure.

In our CG approach, the master problem is indeed the LP relaxation of the [OF_PATTERN] problem, while taking into account a subset of CTPs \( P_0 \subseteq P \).
Initially, $\mathcal{P}_0$ can be constructed as a set of basic transmission patterns with the form of $\{e_{i,j}, \ (i,j) \in \mathcal{E}\}$. For $e_{i,j}$, only the item corresponds to the link $(i,j)$ equals to $r_{i,j}$, i.e. one radio’s transmission rate on $(i,j)$, and all other items equal to 0, representing only one active link at a time.

The master problem [MASTER] can be formulated as,

$$\text{[MASTER]} \quad \min \sum_{p \in \mathcal{P}_0} x_p$$

$$\sum_{j \in N : (i,j) \in \mathcal{E}} f^m_{i,j} - \sum_{j \in N : (j,i) \in \mathcal{E}} f^m_{j,i} = 0, \quad \forall i \in N - \{s_m, d_m\}, \ m = 1\ldots M \quad (7.20)$$

$$\sum_{j \in N : (s_m,j) \in \mathcal{E}} f^m_{s_m,j} - \sum_{j \in N : (j,s_m) \in \mathcal{E}} f^m_{j,s_m} = R_m, \quad \forall m = 1\ldots M \quad (7.21)$$

$$\frac{1}{T} \sum_{p \in \mathcal{P}_0} x_p \cdot a^p_{i,j} - \sum_{m=1}^{M} f^m_{i,j} \geq 0, \quad \forall (i,j) \in \mathcal{E} \quad (7.22)$$

$$f^m_{i,j} \geq 0, \ 0 \leq x_p \leq T, \ \forall (i,j) \in \mathcal{E}, \ p \in \mathcal{P}_0, \ m = 1,2\ldots M \quad (7.23)$$

After solving the master problem, we should verify the optimality and possibly decide a new column to join in the Basis, i.e. the column set of the [MASTER] problem, for another round of computation. Specifically, we tend to find a column, corresponding to a CTP $p \ (p \in \mathcal{P} \setminus \mathcal{P}_0)$, whose related variable $x_p$ leads to the most negative reduced cost among all the remaining columns. Denote $\xi_p$ as the reduced cost of the column with the respective CTP $p$, and $\{\tilde{y}_{i,j}, \ (i,j) \in \mathcal{E}\}$ as the optimal dual variables of the link demand constraints (7.22) in the [MASTER] problem. Thus,

$$\xi_p = 1 - \sum_{(i,j) \in \mathcal{E}} \tilde{y}_{i,j} \cdot a^p_{i,j}$$

$$131$$
The minimum reduced cost in turn corresponds to a CTP \( \hat{p} \) which gives the largest value of \( \sum_{(i,j) \in \mathcal{E}} \bar{y}_{i,j} \cdot \alpha_{i,j}^p \), i.e.

\[
\min_{p \in \mathcal{P} \setminus \mathcal{P}_0} \xi_p = 1 - \max_{p \in \mathcal{P} \setminus \mathcal{P}_0} \sum_{(i,j) \in \mathcal{E}} \bar{y}_{i,j} \cdot \alpha_{i,j}^p
\]

Consequently, a sub-problem [SUB] is solved for selecting the new column with minimum reduced cost, to enter the Basis. This can be formulated as follows,

\[
\text{[SUB]} \quad \max \sum_{(i,j) \in \mathcal{E}} \bar{y}_{i,j} \cdot (a_{i,j} = \sum_{k=1}^{K} v_{i,j}^k \cdot r_{i,j})
\]

\[
\sum_{k=1}^{K} (\sum_{(n,j) \in \mathcal{E}} v_{n,j}^k + \sum_{(i,n) \in \mathcal{E}} v_{i,n}^k) \leq \gamma_n, \quad \forall n \in N
\]

\[
v_{i,j}^k + v_{u,w}^k \leq 1 + F[(i,j), (u,w)], \quad \forall (i,j), (u,w) \in \mathcal{E}, k = 1...K
\]

\[
v_{i,j}^k = \{0, 1\}, \quad \forall (i,j) \in \mathcal{E}, k = 1...K
\]

If the solution to the [SUB] problem results in a negative reduced cost, [MASTER] is re-optimized according to the newly expanded Basis with its corresponding CTP set \( \mathcal{P}_0 = \mathcal{P}_0 \cup \mathcal{P} \), and the iterative computation continues. Otherwise, the [MASTER] problem already obtains the optimal solution to the LP relaxation of [OF_PATTERN], and the CG procedure terminates. The LP relaxation gives a lower bound to the original [OF_PATTERN] problem, which can be considered as the optimization in a continuous time system, or a system with an infinite number of time slots in a period. In this case, each CTP can be scheduled to be active for any fraction of a period as needed. This gives a lower bound of [OF_PATTERN] problem. To further achieve a feasible solution to the [OF_PATTERN] problem, we take the advantage of the whole set of columns obtained from LP relaxation and substitute them back to [OF_PATTERN] for the optimal solution with integer constraints. This serves as an achievable upper bound of the problem.
7.3 Numerical Results

In this section, we use numerical results to verify the effectiveness of our work. Specifically, we investigate 3 different configurations. The first one is a simple string topology containing 8 nodes. We would like to demonstrate the performance gain from efficient scheduling along the delivery path. In addition, we would present preliminary insights on how multi-radio multi-channel work and what the extreme is for an improvement in performance. We then look at a 4*4 regular Grid. We perform joint routing and scheduling using our CG based approach and illustrate details in the schedule of concurrent transmission patterns and the corresponding optimal multi-path selection. We compare results with the Shortest Path routing with optimal scheduling scheme and show the gain from the joint optimization. Results for multi-radio deployment to only two end nodes are also presented. Remaining results relate to a large setting. We show the trend of improvement from channel and radio addition when some system parameters are adjusted. We also consider the case with single shortest path routing (with optimal scheduling) to demonstrate the efficiency of joint routing and scheduling. Given the traffic demand between source-destination pairs, this demand can be mapped to the rate requirement on a set of links. We slightly revise our formulation to fit this shortest path with optimal scheduling problem and compare results with the joint optimization. More insights are provided later.

Specifically in our experiment, we normalize the transmission rate at each active link, working on one channel as a unit constant. For simplicity, we consider the traffic load for each session to be 0.1. We use the label ‘IP-Optimal’ to denote results from MIP [OF SLOT]/[OF PATTERN] optimization; ‘LP’ to denote results from the LP relaxation of [OF PATTERN], calculated by the CG approach; and ‘IP-CG’ to denote results for the optimal solution of [OF PATTERN] problem, based on the set of CTPs obtained from the above CG calculation for the LP relaxation of [OF PATTERN]. As we mentioned before, LP gives optimal results for a continuous time system, which serves as the lower bound for our slotted time
system. In order to show reasonable results, when we obtain non-integer values from LP, we enlarge it to the nearest integer. On the other hand, the IP-CG offers a feasible and in most cases, efficient solution to the joint routing and scheduling problem, which serves as an upper bound for the [OF_PATTERN] problem.

### 7.3.1 String Topology

We first look at a simple string topology illustrated in Figure 7.2, where the distance between neighboring nodes is normalized to 1 and the transmission range is thus settled to connect adjacent nodes only. The interference range is the same as the transmission range. We consider the traffic from N1 to N8 with 7 hops. There are \( T = 10 \) time slots in a period and this indeed asks for one slot’s traffic flow between two ends every period.

When the channel number is equal to 1 (Ch=1), we need 3 time slots to satisfy the traffic demand by arranging those non-interfering links, such as \([1, 2)\ (4, 5),\ (7, 8)]\), to work together. When 2 channels (Ch=2) are deployed, the minimum active time slots are 2, where those non-overlapping interfering links such as \([(1, 2)\ (3, 4)]\) can be active, concurrently on different channels. Detailed arrangement

\[
\begin{array}{|c|c|c|}
\hline
\text{Ch=1} & \text{Active Links} & \text{Time} \\
\hline
\text{Pattern 1} & (2, 3), (5, 6) & 1 \text{ slot} \\
\text{Pattern 2} & (1, 2), (4, 5), (7, 8) & 1 \text{ slot} \\
\text{Pattern 3} & (3, 4), (6, 7) & 1 \text{ slot} \\
\hline
\text{Ch=2} & \text{Active Links} & \text{Time} \\
\hline
\text{Pattern 1} & \text{Ch1: (1, 2), (5, 6)} & 1 \text{ slot} \\
& \text{Ch2: (3, 4), (7, 8)} & \\
\hline
\text{Pattern 2} & \text{Ch1: (2, 3), (6, 7)} & 1 \text{ slot} \\
& \text{Ch2: (4, 5)} & \\
\hline
\end{array}
\]
of CTPs can be found in Table 7.1. Under this setting, no improvements can be made by adding more channels. This comes from the fact that, each link only interferes with links at most two hops away. Due to the radio limit, at most 2 non-overlapping interfering links can be arranged at the same time with the maximum channel requirement 2. In this case, the total system time is distributed to the traffic scheduling from/to some 'bottleneck nodes' (i.e. nodes that require maximum processing time. For example, 2 time slots must be dedicated to N2, where one slot for receiving and another one for delivering out the traffic) and the active time slots must be no less than the number of slots to schedule traffic on those bottleneck nodes. In this special example, IP-Optimal, LP and IP-CG converge to the same value.

Preliminary results show that two conditions should be satisfied in order to benefit from the multi-channel deployment. Specifically, there are non-overlapping interfering links in the vicinity and the links contribute to the traffic delivery. The limits of multi-channel addition come from the traffic scheduling on some 'bottleneck' nodes, where all the system time must be dedicated to the receiving and transmitting on these nodes.

We also investigate the performance improvement from more radio interfaces' deployment at each node. When the number of radios at each node is 2 (r=2) and channel number is 2 (Ch=2), the minimum number of time slots in use is still 2. This is due to the fact that 3 interfering links exist in a vicinity and they must be scheduled on different channels or at different time slots. In this case, the bottleneck is the link scheduling. When the system has 3 channels (Ch=3), IP-Optimal=1, which is the best this system can achieve.

Consequently, with more radios deployed, those links from/into the same node might be scheduled at the same time. Generally speaking, the scale increase of system resources, in terms of the number of radios and channels, causes the scale decrease of a system time requirement\(^2\). Although it is the case in a continuous

\(^2\)This can be considered as having multiple copies of base systems working together.
time system, this does not always hold in a slotted time system as illustrated here, due to its discrete nature.

7.3.2 Grid Setting

In this section, we show the results for a 4*4 regular grid setting illustrated in 7.3, where the set of nodes locate at \( (X, Y) = (0.7 \times m, 0.7 \times n), \ m, n = 0, 1, 2, 3 \). We assume the transmission range \( Tr = 1 \) and the interference range equals the transmission range. The corresponding feasible transmission links are also illustrated in the figure by directed edges.
Table 7.2: Transmission Patterns for Grid (1 Channel)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Active Links</th>
<th>Time Slots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern 1</td>
<td>(1, 2), (8, 12), (10, 15)</td>
<td>5</td>
</tr>
<tr>
<td>Pattern 2</td>
<td>(2, 3), (5, 10), (12, 16)</td>
<td>5</td>
</tr>
<tr>
<td>Pattern 4</td>
<td>(1, 5), (3, 8), (15, 16)</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 7.3: Time Requirement for Grid ($Tr = In = 1$)

<table>
<thead>
<tr>
<th></th>
<th>Ch=1</th>
<th>Ch=2</th>
<th>Ch=3</th>
<th>Ch=4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LP</td>
<td>IP-CG</td>
<td>LP</td>
<td>IP-CG</td>
</tr>
<tr>
<td>Joint-SR</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>SP-SR</td>
<td>30</td>
<td>30</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Joint-MR</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>SP-MR</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

We investigate the joint routing and scheduling for one source destination (S-D) pair 1 → 16 with the traffic demand 0.1. If we assume total time slots the same as the string topology, this may require only one slot activation time for the end-to-end flow in each period and imply single path routing for traffic. Consequently, in this example, we consider a more reasonable setting with $T = 100$. Due to the larger network size and greater number of time slots involved, it is hard for us to obtain IP-optimal within an acceptable time. Instead, we calculate the LP (Lower Bound) and IP-CG (Upper Bound) for the joint routing and scheduling optimization, and make comparisons among them under a variety of system settings.

To compute LP and IP-CG, our column generation sub-routine iteratively figures out the sequence of CTPs that might achieve lower time slots requirement. When only one channel exists in the system (Ch=1), according to the computation, LP and IP-CG give the same number of time slots (i.e. 15) in use, demonstrating the optimality of IP-CG.

Table 7.2 lists the transmission patterns, i.e. the collection of simultaneously active links, as well as the number of time slots in use for each pattern, to
achieve the IP-CG with one channel. Consequently, the flow from node 1 to node 16 is split into two sub-flows, namely $1 \rightarrow 2 \rightarrow 3 \rightarrow 8 \rightarrow 12 \rightarrow 16$ and $1 \rightarrow 5 \rightarrow 10 \rightarrow 15 \rightarrow 16$ and $1 \rightarrow 5 \rightarrow 10 \rightarrow 15 \rightarrow 16$, each burdens half of the load. According to the computation, LP and IP-CG give the same number of time slots (i.e. 15) in use, demonstrating the optimality of IP-CG. To compare, if the traffic goes though the shortest path $1 \rightarrow 6 \rightarrow 11 \rightarrow 16$, no spatial reuse can be applied as the three links interfere with each other. Thus, we have to separately schedule the links along the path and the minimum total time slots required are 30. As a result, when multi-path routing is applied, the traffic load is balanced over the network. By taking full advantage of spatial reuse of non-interfering links and the channel differentiated arrangement of interfering links on multiple paths, we could avoid creating those high system time requirement intermediate bottleneck nodes. This demonstrates the great performance gain from Joint Routing and Scheduling over the shortest path routing scheme.

Table 7.3 shows LP and IP-CG results for the joint multi-path routing and scheduling in the above regular grid network. The first two lines (Joint-SR and SP-SR) is obtained when only one radio is equipped at each node, where Joint-SR denotes results from joint optimization and SP-SR for the shortest path routing with optimal scheduling. We can see that for Joint-SR, when the number of channels increases to 3 (Ch=3), no improvement can be made with more channels’ addition as all the time slots are dedicated to bottleneck nodes (i.e. N1/N16 serving as source/destination to send/receive all traffic). At this stage, both LP and IP-CG converge to the same minimum time requirement, i.e. 10 slots, as 10 slots’ traffic must be sent out every period. The last two lines show the results when more radios are equipped in the system. Instead of adding an extra radio to all nodes in the network, which is costly, we select only N1 and N16 to have one extra radio. We ignore the result with Ch=1 as the extra radio can not function when the system has only one channel. The LP and IP-CG for Ch=2 are equal, illustrating the optimality of IP-CG. Great improvement is achieved with the addition of only two more radios in the system. Indeed, even when all
nodes in the system are equipped with 2 radios, the minimum time requirement with Ch=2 is also 8 slots, which is the same as our case. This demonstrates the efficiency of applying radios to only some of the nodes. In addition, according to the results, joint optimization achieves much better performance than the general shortest path routing.

In a continuous time system, varying the traffic load by the same scale factor does not affect the traffic routing and the corresponding scheduling\(^3\). Again, in the time slotted system, this may not be the case due to the discrete nature.

7.3.3 Large Setting

We look into a “wireless community network” as illustrated in Figure 7.4 (left), there are 30 nodes uniformly distributed in a square region of size 1000\(m\) x 1000\(m\).

\(^3\)This can be obtained by the sensitivity analysis for the simplex method
The transmission range $Tr = 25$ and is equal to the interference range. The set of feasible transmission links are shown in 7.4 (right), where each link is bidirectional. There are two crossing traffic sessions, i.e. 13 → 1 and 2 → 6, each with rate requirement 0.1 in the system. Due to the large network size and greater number of time slots we are looking into, we can not afford to compute [OF_SLOT] to optimality within a reasonable time. Thus, we apply the CG approach and obtain LP and IP-CG accordingly. Results with different channel supplies are given in Table 7.4, where the labels in use share similar meaning as in the previous grid setting. Again, for comparison, both results for joint optimization and shortest path (SP) routing with optimal scheduling under the same traffic load with single or multiple radios' supply are presented. With more channels deployed in the system, a greater number of concurrent transmissions can be arranged, resulting in a lower time requirement. As expected, joint routing, scheduling with multi-path selection exhibits much better performance than the shortest path solution. The reason lies in that cross-layer joint optimization makes full use of the spatial separation of the links and schedules concurrent transmission to the largest extend, whereas fewer links are involved in SP routing thus hard for channel re-use, and the multiple shortest paths may tend to go through some bottleneck “short” links, which may create bottleneck nodes with extremely high system time requirement and place traffic scheduling in further difficulty. Again we pick up several nodes (5 in this case), namely 1, 2, 6, 13, 15, to be equipped with one extra radio. This complies with the design of “wireless community network”, where only some nodes may be capable of deploying multiple radios. As shown in the results, increasing the radio supply at only a small number of nodes allows great performance improvement.

We increase the interference range of the above setting to $In = 500$ and show the results in Table 7.6. Due to the greater interference range, the concurrently active links should stand far apart, i.e. more links need to remain silent when one is transmitting, resulting in the reduced number of feasible CTPs. Thus, total time slots in use would increase compared with the case that allows a shorter
Table 7.5: Time Requirement for Large Setting (30 Nodes, $Tr = 250, In = 500$)

<table>
<thead>
<tr>
<th></th>
<th>Ch=1</th>
<th>Ch=2</th>
<th>Ch=3</th>
<th>Ch=4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LP</td>
<td>IP-CG</td>
<td>LP</td>
<td>IP-CG</td>
</tr>
<tr>
<td>Joint-SR</td>
<td>70</td>
<td>70</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>SP-SR</td>
<td>80</td>
<td>80</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Joint-MR</td>
<td>-</td>
<td>-</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>SP-MR</td>
<td>-</td>
<td>-</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

Figure 7.5: Network setting (40 Nodes) $Tr = 250$

Table 7.6: Time Requirement for Large Setting (40 Nodes, $Tr = 250, In = 250$)

<table>
<thead>
<tr>
<th></th>
<th>Ch=1</th>
<th>Ch=2</th>
<th>Ch=3</th>
<th>Ch=4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LP</td>
<td>IP-CG</td>
<td>LP</td>
<td>IP-CG</td>
</tr>
<tr>
<td>Joint-SR</td>
<td>21</td>
<td>22</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>SP-SR</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Joint-MR</td>
<td>-</td>
<td>-</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>SP-MR</td>
<td>-</td>
<td>-</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

interference range $In = 250$. At the same time, as the number of interfering links in a neighborhood increases, the achievement from the addition of channels becomes more significant. As the channel number increases, time requirements for two (interference) settings become closer until the system time lower bound is achieved. Consequently, the impact from a larger interference range can be mitigated by adopting more channels into the system.
In the last example, we increase the density of nodes (see Figure 7.5) and compare results between Joint Routing/Scheduling and Shortest Path Routing. There are 10 more nodes (total 40), where $Tr = 250$ and $In = 250$. Compared with its 30 nodes’ counterpart, as more nodes are supplied as relays, better routes might be selected when the system is allowed to use joint optimization, resulting in improved system usage time. Similar to previous examples, as more links locate within their mutual interference range, greater improvement can be achieved through using multiple channels in the system. Furthermore, for the later cases, as the links become dense, fewer bottleneck links exist and better joint routes might be scheduled, while making full use of link re-use factor. The performance gap between the SP and Joint approach increases.

Furthermore, according to the previous results with $r = 1$ at each node, the greatest performance achievement can be found when the channel number changes from 1 to 2. With the increasing of channels, the performance improvement becomes marginal and no further progress can be made. This comes from the fact that each radio can tune to only one channel at a time. Thus, the system time requirement should be no less than the time to route the total traffic in and out of each node. If the total system time is distributed to schedule the traffic from/to some bottleneck nodes, performance can not be further improved by multi-channel deployment. This situation can be made even worse when the nodes are sparsely located and transmission range is short. In such cases, the number of non-overlapping interfering links is small and it is more possible to create bottleneck links and nodes (e.g. bridges and articulation points), where several traffic must pass through.

7.4 Summary

In this chapter, we investigate joint routing and scheduling optimization in a multi-hop network with multiple radios’ and orthogonal channels’ supply. Given the traffic load and network configurations, our objective is to minimize the over-
all system activation time to satisfy the end-to-end rate demand, subject to system resource limitations such as the available channels and the number of radios at each node, and the wireless multi-access constraint. To avoid the high complexity in optimization, we adopt a column generation (CG) based approach, which generates concurrent transmission patterns (CTP) with performance improvement one by one as needed. Accordingly, we could obtain very close upper bound and lower bound for the joint optimization problem. As the set of patterns involved in progress making is relatively small, compared to the whole set of patterns, the column generation approach can be very efficient. In summary, our contributions in this paper are as follows: firstly, we study the performance bound by joint routing and scheduling optimization in a multi-radio multi-channel multi-hop network; secondly, we apply an efficient column generation approach to derive the upper and lower bounds for the optimization, with extremely high combinatorial complexity.

Results from extensive experiments show that in most cases, upper and lower bounds derived in our work are the same or very close to each other, demonstrating the efficiency of the proposed algorithm. Compared with shortest path routing with optimal scheduling, great performance gain can be achieved by jointly considering multi-path routing and scheduling in the optimization. Also, joint multi-path routing can make better use of multiple channels in a multi-radio multi-channel network. Furthermore, the benefit of multi-radio multi-channel deployment largely comes from arranging those co-channel interfering links to transmit concurrently on different channels. When the interference range and the node density increase, more interfering links exist in the neighborhood and higher performance gain can be achieved from multi-radio and multi-channel deployment. Due to the radio constraints, no further improvement can be obtained from channel addition when all system time are dedicated to several bottleneck nodes.
CHAPTER 8

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This thesis studies resource management techniques in wireless networks. In this chapter, we summarize the main research results we have carried out so far and discuss the possible future research directions.

8.1 Summary of the Thesis

In this thesis, we firstly present a comprehensive survey on resource management in a CDMA cellular network. We examine the basic mechanism in capacity provisioning, main challenges and related issues including congestion control, rate and power allocation, and cell planning, with an emphasis on the call admission control. We then move on to introduce our work on admission control for downlink multi-services MC-CDMA cellular networks. Whereas previous solutions mainly focus on the admission control at the connection level while neglecting stochastic behaviors of mobile subscribers and channel conditions, in our work we quantitatively demonstrate the significant impact from those statistical factors, in particular log-normal shadowing in propagation and voice activity factors, and design the admission strategy accordingly for connection-level traffic. Furthermore, by taking advantage of those statistical variations, we could arrange delay tolerable data traffic at the packet level as the background transmission to improve the spectrum resource unitization.

We then investigate adaptive cell sectoring for non-uniform traffic in a CDMA cellular network to minimize total transmission power. Previous work transforms optimal cell sectoring into a shortest path problem and solves it with Dijkstra's
algorithm. In our work, we make use of a dynamic programming approach to formulate and solve the optimal sectoring problem in a more straightforward manner, and avoid the overhead of graphic mapping. In addition, to reduce the computational complexity incurred from the optimization and to prevent sector boundaries crossing high-density regions, we propose a cluster-based sectoring algorithm, which achieves bounded complexity under high density traffic. Later the algorithm is expanded to cope with multi-rate applications.

For the wireless ad-hoc network, we first present a survey of recent work on throughput capacity improvement and energy conservation in such a system without infrastructure support. It can be observed that cross-layer joint routing, scheduling and power control can be exploited for better resource utilization. We then introduce the major work we have done in this area. The first one is the mobility assisted routing in mobile ad-hoc networks (MANET). In this work, we explore node mobility in the search for packet delivery routes, optimizing QoS criteria such as end-to-end delay and energy consumption, assuming a deterministic mobility model. Our focus in this study is to obtain insights on inherent relationships among throughput, delay and power consumption in the context of routing. We consider networks operating in both the interference-free regime and interference-limited regime. In the interference-free case, we present efficient polynomial time algorithms that find optimal packet delivery paths for fastest, minimum hop and most energy efficient routing. In the interference-limited case, we first give a dynamic programming formulation of the most energy efficient routing for multiple packets. Because of the high complexity introduced in the optimization, we then present an efficient heuristic algorithm for the problem.

Our recent work tries to derive the performance bound of a multi-channel multi-radio wireless multi-hop network. While previous work has largely focused on the protocol design, we investigate the achievable performance gain by jointly optimizing routing and scheduling under a deterministic model. We derive the mathematical formulation for the optimization, which minimizes overall system activation time to satisfy a given end-to-end traffic demand, subject to the
number of channels' limitation, the multi-access interference among neighboring transmission pairs in the system, and the radio interface constraint at each node. The exact solution to such an optimization problem is prohibitively complex due to the combinatorial complexity, particularly after introducing multi-radio and multi-channel into the system. We then develop a column generation based approach to solve this problem, which decomposes the original problem into sub-problems and solve them iteratively, and derives very effective management on system resources.

8.2 Future Research Directions

Apart from the on-going work in previously mentioned areas, we would like to do research on a promising emerging technology, namely Mesh Networking in the future. As a special combination of both ad-hoc and infrastructure wireless networks, mesh networking can provide new perspective as well as great challenges to the capacity provisioning and the corresponding resource management policy design.

8.2.1 Mesh Networks

A wireless mesh network is a fully wireless network that employs multihop communications to forward traffic to and from wired Internet entry points. Mesh network is an alternative to purely wired infrastructure networks or purely mobile wireless ad hoc networks [14] [42], in that it builds on a mixture of fixed and mobile nodes interconnected via wireless links to form a multihop network. Different from flat ad hoc networks, a mesh network introduces a hierarchy in the network architecture. It implements dedicated nodes (namely, wireless routers) to communicate with each other and provide broadband wireless services from user to user or from user to those wired Internet entry points. As a result, such wireless routers form a wireless backbone, which provides multihop connectivity between nomadic users and wired gateways. Wireless mesh help the deployment
of community networks and provide mobile user with a low cost, high bandwidth and seamless multihop inter-connection service.

A lot of industrial efforts have been dedicated to the development of commercial wireless mesh solutions [84] [68] for both indoor and outdoor environments, such as Tropos, LocustWorld, MeshNetwork, Radiant Networks, and the like. Other research projects [58] are also going on, including MIT Roofnet and Nokia Rooftop (802.11 hot-spot operators.). Recently, Microsoft Research has organized a Mesh Networking Summit [57], providing a forum in which to discuss the mesh networking benefits and challenges. In addition, IEEE Working Groups are also making efforts to provide mesh networking extensions to their standards, 802.11s (Wi-Fi technology), 802.15.5, 802.16.a and 802.20. Through which, they are trying to establish and define the requirements for mesh networking in wireless personal area networks (WPANs), WLANs and wireless metropolitan area networks (WAMANs), Mobile broadband wireless access (MBWA). Similar to what Wi-Fi Alliance provides to IEEE 802.11 standard for WLANs, WiMAX forum [88] has been established to ensure and improve the interoperability of equipments from different manufacturerers, and thus facilitate the deployment of broadband wireless mesh networks based on the 802.16 standard.

A lot of benefits can be derived from the deployment of wireless mesh networking: Firstly, the cost of installing and maintaining a wired backhaul connection infrastructure, such as wired APs connecting to the Internet, can be greatly reduced by replacing them with wireless routers; Secondly, mesh networking enables large scale deployment, due to the easy placement and movement of the wireless routers (Access points); Thirdly, mesh networking may guarantee better reliability, when redundant paths are provided under multi-path routing; Fourthly, two tiered hybrid routing protocol can be designed to exploit node heterogeneity. As the wireless routers, together with the channels among them, are relatively stable, for traffic delivery within the backbone involving only wireless routers, the routing scheme can be proactive, i.e. calculated before the real traffic demand comes. On the other hand, the routing involving mobile users could be reactive
to reflect user mobility and channel varying dynamics. Also we could incorporate two operating modes, namely ad hoc mode and AP mode, into the same network, where traffic to the backbone can be delivered directly to the wireless router (AP mode) or through multiple hops by other peer nodes (Ad-hoc mode); Fifthly, the pure wireless distribution system with peer to peer networking allows self-management; Lastly, advanced technology can be easily applied to the mesh network to help further increase the network capacity. According to the standards, multiple orthogonal channels can be adopted simultaneously within the system, where each wireless mesh router may be equipped with multiple radios [6] to work cooperatively in different channels. Furthermore, as wireless routers generally have high capability, directional antenna and MIMO technologies can easily be deployed in the system.

Although new perspectives can be envisioned, mesh networking poses new challenges to the resource management in such a system. One of the major problems is the scalability of both the network architecture and protocols in building a multihop wireless backhaul network. New scalable and distributed scheduling, MAC, and routing protocols [21] have to be designed to manage the data traffic. These algorithms must be aware of the characteristics of the physical channels, which leads to the cross-layer design. Secondly, a backhaul network is obliged to support a huge amount of traffic, thus capacity of a multihop wireless network becomes a de facto constraint. Research efforts have been made to improve the capacity of wireless mesh networks by exploiting such alternative approaches as multiple radio interfaces, multiple-input multiple-output (MIMO) techniques, beamforming antennas, and opportunistic channel selection. New scalable and opportunistic networking functions, which exploit a variety of diversity such as multi-user diversity (i.e. users under different channel conditions/quality), spatial diversity (by smart antennas) and channel diversity (by multiple channels), should be designed for both scheduling and routing. Thirdly, in resource management, it is an essential requirement for the backhaul network to provide QoS guarantee to the users. We should ensure a fair share of system resources allo-
cated to the users in the network, no matter how far the user is from the Internet entry point (i.e. independent of the spatial location) [26]. Coordinated multi-hop resource management policy must be developed to achieve high performance while preserving a system-wide notion of fairness. To conclude, resource management schemes should do their utmost to improve network throughput while compromising with system-wide fairness and provide a prompt reaction to variations in system capacity due to changes in traffic patterns, channel conditions and contention.
REFERENCES


150


151


[57] Microsoft Mesh Networking Summit 2004
http://research.microsoft.com/meshsummit/techprogram.aspx


156


APPENDIX A

CONNECTION LEVEL CAPACITY CONSTRAINT

For the given user $U_{0ki}$ (i.e. $i$-th user of a class-$k$ call in the reference cell of $BS_0$), the corresponding signal to interference ratio (SIR) is,

$$\gamma_k = \left( \frac{E_b}{I} \right)_{ki} = \frac{\left( \frac{\phi_k}{c_k^2} \beta P_t \zeta_{0ki} \right) / R_b}{(1 - v)(1 - \frac{\phi_k}{c_k^2} \beta) P_t \zeta_{0ki} + \sum_{j=1}^{M} P_t \zeta_{jki}} / W_d$$  \hspace{1cm} (A.1)

Where, $\frac{\phi_k}{c_k^2} \beta$ is the power allocated by $BS_0$ to each code of $U_{0ki}$. $v$ is the orthogonal factor in the downlink and $W_d$ $R_b$ are the downlink spreading bandwidth and basic code rate. $\zeta_{jki} = d^{-m}_{jki} L_{jki}[\sigma^2]$ is the pass loss between $BS_0$ and $U_{0ki}$. After some manipulations, we have,

$$\phi_{ki} = f_k(1 - v + \sum_{j=1}^{M} \frac{\zeta_{jki}}{\zeta_{0ki}})$$  \hspace{1cm} (A.2)

where $f_k = \frac{c_k^{2}\gamma_k}{\beta(G_d + \gamma_k(1-v))}$, and $G_d = \frac{W_d}{R_b}$.

The outage constraint, i.e. the required power exceeds the power supply at $BS_0$, could be written as:

$$P_{out} = P\{ \sum_{k=1}^{K} \sum_{i=1}^{n_k} \Psi_{ki} \phi_{ki} > 1 \}$$  \hspace{1cm} (A.3)
Let $\Phi = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \Psi_{ki} \phi_{ki}$ and $T = 1 - \nu + \sum_{j=1}^{M} \left( \frac{d_{jki}}{d_{0ki}} \right)^{-m} \frac{L_{jki}}{L_{0ki}}$. The outage constraint can be approximated by the Gaussian approximation with mean and variance as follows,

$$m_{\Phi} = \sum_{k=1}^{K} n_k \alpha_k f_k m_T \quad (A.4)$$

$$v_{\Phi} = \sum_{k=1}^{K} f_k^2 n_k \alpha_k \{ \nu_T + (1 - \alpha_k) \cdot m_T^2 \} \quad (A.5)$$

where, $m_T$ and $\nu_T$ is the mean and variance value of $T$.

$$m_T = 1 - \nu + M \cdot E_j \cdot E(L(2\sigma^2)) \quad (A.6)$$

$$\nu_T = \frac{E_j^2 \cdot \text{Var}(\sum_{j=1}^{M} \frac{L_{jki}}{L_{0ki}})}{} \quad (A.7)$$

here, $E(L(2\sigma^2)) = \exp(\sigma^2/h^2)$, $h = 10 \log_{10} e$ is the mean value of Log-Normal distribution with the variance $2\sigma^2$ for the normal random variable. The outage $P_{out}$ could be obtained by calculating the tail probability of $\Phi$,

$$P_{out} = \int_{1}^{\infty} \frac{1}{\sqrt{2\pi v_{\Phi}}} e^{-\frac{(x-m_{\Phi})^2}{2v_{\Phi}}} \, dx \quad (A.8)$$

Here, for simplicity, we assume an “average location” for each mobile with $E_j$ as the mean value of $(d_{jki}/d_{0ki})^{-m}$, which is approximately 0.0474 [102], with $j$ as the most nearest cell to the considered cell 0. In this work, we only consider the impact from the six nearest neighbors.

For a given outage threshold $\delta_{out}$, there exists a value $a$, $Q(a) = \delta_{out}$, where is the integral over the tail of a Gaussian distribution $f \sim N(0, 1)$. Then, we have the following relation,
\[ 1 - m_\phi = a\sqrt{v_\phi} \quad (A.9) \]

Given the vector for other class \((C_j, j \neq k)\) user numbers in the cell, \(n^k = (n_1, \ldots, n_{k-1}, n_{k+1}, \ldots n_K)\), by solving for \(n_k\) in the above equation, we could find the capacity constraint for \(C_k\) calls, denoted as function \(\max(n^k, k)\), which satisfy the constraint for outage threshold, \(P_{out} \leq \delta_{out}\), as follows,

\[ An_k^2 - Bn_k + C = 0 \]

\[ A = (\alpha_k f_k m_T)^2 \]

\[ B = 2(1 - \sum_{i=1}^{k-1} n_i \alpha_i f_i m_T - \sum_{i=k+1}^{K} n_i \alpha_i f_i m_T)\alpha_k f_k m_T + a^2 \alpha_k f_k^2 (v_T + (1 - \alpha_k) m_T) \]

\[ C = (1 - \sum_{i=1}^{k-1} n_i \alpha_i f_i m_T - \sum_{i=k+1}^{K} n_i \alpha_i f_i m_T)^2 - a^2 \sum_{i=1}^{k-1} n_i \alpha_i f_i^2 (v_i + (1 - \alpha_i) m_T^2) - \]

\[ \sum_{i=k+1}^{K} n_i \alpha_i f_i^2 (v_i + (1 - \alpha_i) m_T^2) \]

\[ n_k = \max(n^2, k) = \frac{1}{2A} [B - \sqrt{B^2 - 4AC}] \quad (A.10) \]

which is the maximum number of \(C_k\) calls that could be admitted to the cell given the existence \((n_1, \ldots, n_{k-1}, n_{k+1}, \ldots, n_K)\) of other class calls.
APPENDIX B

DEPARTURE RATE CALCULATION

Assume a hexagon cellular system with uniform user traffic distribution within each cell. If the average moving speed for the class-$k$ services is $V_k$, the boundary-crossing rate will be [56]

$$\lambda_{h,k} = \frac{\rho_k V_k L}{\pi}$$  \hspace{1cm} (B.1)

where $L = 6r_1$ is the perimeter length of the hexagon cells, and $\rho_k$ is the per unit user density within cell, which is calculated as[43],

$$\rho_k = \frac{\lambda_{n,k}(1 - P_{\text{block},k}) + \lambda_{h,k}(1 - P_{\text{drop},k})}{\mu_{d,k} + \mu_{c,k}} \cdot \frac{1}{A_{\text{Cell}}}$$  \hspace{1cm} (B.2)

where $A_{\text{Cell}}$ is the area of a cell; $P_{\text{block},k}$ is the class-$k$ new call blocking probability; $P_{\text{drop},k}$ is the class-$k$ handoff call dropping probability; $1/\mu_{d,k}, 1/\mu_{c,k}$ is the class-$k$ average dwell and call holding time both with exponential distribution. $\lambda_{n,k}, \lambda_{h,k}$ are the class-$k$ new call and handoff call arrival rates.

Assume homogeneous traffic for the overall system. The mean departure rate across the cell boundary should be equal to mean handoff call arrival rate to the cell.

$$\lambda_{h,k} = \frac{\mu_{d,k}}{\mu_{d,k} + \mu_{c,k}} \cdot \{\lambda_{n,k} \cdot (1 - P_{\text{block},k}) + \lambda_{h,k} \cdot (1 - P_{\text{drop},k})\}$$  \hspace{1cm} (B.3)

Combine (B.1), (B.2), (B.4) together, the mean dwell time $1/\mu_{d,k}$ within a regular hexagonal cell may be approximated as,
\[ \mu_{d,k} = \frac{V_k L_{RC_{ell}}}{\pi A_{RC_{ell}}} = \frac{6 \cdot r_1 \cdot V_k}{\pi \cdot (3\sqrt{3}/2) r_1^2} = \frac{4V_k}{\sqrt{3}\pi r_1} \] (B.4)
APPENDIX C

BACKGROUND DATA PACKETS TRANSMISSION

The downlink SIR requirement for each class-$K + 1$ background data packet transmission code word is represented as,

$$\gamma_{K+1} = \frac{(\phi_{K+1,i}\beta P_i\zeta_{0,K+1,i}) / R_b}{(1 - \nu)(1 - \phi_{K+1,i}\beta) P_i\zeta_{0,K+1,i} + \sum_{j=1}^{M} P_j\zeta_{j,K+1,i}) / W_d}$$  \hspace{1cm} (C.1)

Here, $\phi_{K+1,i}\beta$ is the fraction of power allocated to one code of the $i$-th (class-$K + 1$) user. $\zeta_{j,K+1,i}$ is the pass loss from $BS_j$ to the $i$-th user. Thus, the fraction of power for a class-$K + 1$ code word is,

$$\phi_{K+1,i} = f_{K+1}(1 - \nu + \sum_{j=1}^{M} \frac{\zeta_{j,K+1,i}}{\zeta_{0,K+1,i}})$$ \hspace{1cm} (C.2)

with $f_{K+1} = \frac{\gamma_{K+1}}{\beta(\zeta_{0} + \gamma_{K+1}(1 - \nu))}$.

Assume $n_{act} = (n_{act,1}, n_{act,2}, ... n_{act,K})$ connection level calls active,

$$P_{out} = P\{\sum_{k=1}^{K} \sum_{i=1}^{n_{act,k}} \phi_{ki} + \sum_{i=1}^{n_d} \phi_{K+1,i} > 1\}$$ \hspace{1cm} (C.3)

where, $n_d$ is number of code words currently used for class-$K + 1$ data transmission.
Again, let $\Phi' = \sum_{k=1}^{K} \sum_{i=1}^{n_{\text{act},k}} \Phi_{ki} + \sum_{i=1}^{n_{d}} \Phi_{K+1,i}$, $T_i = 1 - v + \sum_{j=1}^{M} \left( \frac{d_{jki}}{d_{0ki}} \right) - m \frac{L_{jki}}{L_{0ki}}$, and assume an "average location" for each mobile who receive the data packets. The outage constraint $P_{\text{out}}$ can be approximated by the Gaussian approximation with mean and variance as,

$$m_{\Phi} = (\sum_{k=1}^{K} n_{\text{act},k}f_k + n_d f_{K+1}) m_T$$

$$v_{\Phi} = (\sum_{k=1}^{K} f_k^2 n_{\text{act},k} + f_{K+1}^2 n_d) v_T$$

and obtained by calculating the tail probability of $\Phi$,

$$P_{\text{out}} = \int_{1}^{\infty} \frac{1}{\sqrt{2\pi} v_{\Phi'}} e^{-\frac{(x-m_{\Phi'})^2}{2v_{\Phi'}}} dx \quad (C.4)$$

Again, the maximum code words that could be used simultaneously for data transmission, i.e. Max $n_d$, is obtained subject to the constraint $P_{\text{out}} \leq \delta_{\text{out}}$. Follow the manipulation in Appendix A,

$$n_d = \frac{1}{2m_T f_{K+1}} [(2f_{K+1}m_T(1 - m_T \sum_{k=1}^{K} n_{\text{act},k}f_k) + a^2 f_{K+1}^2 v_T)$$

$$+ \sqrt{(2f_{K+1}m_T(1 - m_T \sum_{k=1}^{K} n_{\text{act},k}f_k))^2 - 4m_T^2 f_{K+1}^2 (1 - m_T \sum_{k=1}^{K} n_{\text{act},k}f_k)^2 - a^2 \sum_{k=1}^{K} f_k^2 n_{\text{act},k} v_T} ] \quad (C.5)$$

167
APPENDIX D

STEADY STATE PROBABILITY FOR DATA QUEUE

With the assumption of fixed service rate of the data queue $\mu_{data} = 1/T_{data}$, exponentially distributed time interval for data packets arrival with mean $1/\lambda_{data}$ and maximum data queue length constraint $L$, the data service is modelled as an M/D/1/L queue, with $L - 1$ as the maximum waiting queue length. Denote $\rho = \lambda_{data} \cdot T_{serve}$. The state probability for the M/D/1 data queue could be shown to be [29]:

\begin{align*}
p_0 &= 1 - \rho \\
p_1 &= (1 - \rho)(e^\rho - 1) \\
\vdots &\quad \\
p_n &= (1 - \rho) \sum_{k=1}^{n} (-1)^{n-k} e^{k\rho} \left[ \frac{(k\rho)^{n-k}}{(n-k)!} + \frac{(k\rho)^{n-k-1}}{(n-k-1)!} \right] \\
\end{align*}

When finite queue length $L$ is considered, the steady-state probability of $i$ packets in the queue at a departure point, say $\{\pi_i\}$ for M/D/1/L is equal to,

\begin{equation}
\pi_i = \frac{1}{\sum_{j=0}^{L-1} p_j} p_i \tag{D.2}
\end{equation}

let $p'_n$ denotes the steady state probability that $n$ packets in the queue at arbitrary point in time, (i.e. the probability that an arriving packet finds a system with $n$ packets), it can be expressed as,

\begin{align*}
p'_0 &= \frac{\pi_0}{\pi_0 + \rho} \\
p'_n &= \frac{\pi_n}{\pi_0 + \rho} \\
P'_L &= \frac{n-1 + p'_0}{\rho} = 1 - \frac{1}{\pi_0 + \rho}
\end{align*}

168
APPENDIX E

CORRECTNESS PROOF OF RECURRANCE RELATION

Lemma: Recurrence relation (4.10) is correct.

\[
\varphi_{SP}(k, n) = \begin{cases} 
\frac{\eta \gamma}{G - (k-1)\alpha \gamma} \sum_{j \in \chi_i} \frac{1}{h_j}, & n = 1, 1 \leq k < \frac{G}{\alpha \gamma} + 1 \\
\min_{1 \leq i \leq k} \left\{ \varphi(i - 1, n - 1) + \frac{\eta \gamma}{G - (k-1)\alpha \gamma} \sum_{j \in \chi(i)} \frac{1}{h_j} \right\}, & n > 1, n \leq k < \frac{G}{\alpha \gamma} + i \\
+\infty, & \text{Otherwise}
\end{cases}
\]

Here, \( \varphi_{SP}(k, n) \) represent the MinTTP of \( k \) users partitioned with \( n \) sectors for the given start partitioning point \( SP \). \( i \) is the partition point for the \( n \)-th sector (including \( U_i \)) and \( \chi(i) \) is the set of consecutive users in the same sector with user \( i \), i.e. \( \{U_i, U_{i+1}, ..., U_k\} \).

Proof: The equation can be proved by induction.

1. \( n = 1, 1 \leq k < \frac{G}{\alpha \gamma} + 1 \)

   The \( k \) users contribute only for one sector, then \( \varphi(k, n) = \varphi(k, 1) \)

   \[
   = \frac{\eta \gamma}{G - (k-1)\alpha \gamma} \sum_{j \in \chi_i} \frac{1}{h_j} = \frac{\eta \gamma}{G - (k-1)\alpha \gamma} \sum_{j=1}^{k} \frac{1}{h_j}
   \]

2. \( n = 1, 1 \leq k < \frac{G}{\alpha \gamma} + i \)

   Let \( \varphi(i - 1, n - 1) \) be the optimal solution with \( i - 1 \) consecutive users split into \( n - 1 \) sectors. Assume the partitioning point for the last sector \( n \) is set
to be $i$. The optimal solution $\varphi(k, n)$ takes place if and only if the solution to previous $i - 1$ users is optimal with $n - 1$ sectors, i.e. $\varphi(i - 1, n - 1)$ [10]. Then for this case, the MinTTP is $\varphi(i - 1, n - 1) + \frac{\eta \gamma}{\sigma - (k - i) \alpha \gamma} \sum_{j \in \chi(i)} \frac{1}{h_j}$, (the last item is the TTP for $n$-th sector). Consider all the feasible placement of the partition point $i$ for the $n$-th sector and extract the minimum solution, i.e. \[
min_{n-1 < i \leq k} \left\{ \varphi(i - 1, n - 1) + \frac{\eta \gamma}{\sigma - (k - i) \alpha \gamma} \sum_{j \in \chi(i)} \frac{1}{h_j} \right\}, \text{ here assume at least one user in a sector. This is the optimal solution for } \varphi(k, n). \]

3. $n < 1$, $k < n$, $k > \frac{\sigma}{\alpha \gamma} + i$ all are non-feasible settings and thus $\varphi(k, n)$ is set to $+\infty$. 

170
APPENDIX F

CLUSTER INTEGRATION PROOF

Lemma: A cluster derived from Section 4.4.1 can be regarded as one particular integrated user to be used in the (4.11).

Proof: According to Section 4.3, the total transmission power for the system can be written as $\sum_{i=1}^{N} \frac{m}{G-(K_{i}+1)\alpha \gamma} \sum_{j \in \chi_{i}} \frac{1}{h_{ij}}$, where $K_{i}$ is the number of users in sector $i$, $\chi_{i}$ is the set of consecutive users in the same sector with user $i$. As mentioned above, the users in one cluster should be put into the same sector. Assume cluster $j$ with $k$ users $C_{j} = U_{1}, U_{2}, \ldots, U_{k}$ is placed into sector $i$, then the TTP contributed by $C_{j}$ is $\frac{m}{G-(K_{i}+1)\alpha \gamma} \sum_{s=1}^{k} \frac{1}{h_{us}}$. Thus, $C_{j}$ can be viewed as a super-user to have number factor as $k$ to contribute in the $K_{i}$ place and path loss $\frac{1}{h_{C_{j}}}$ which results in (4.11).

\[
\begin{align*}
= & \sum_{s=1}^{i-1} \frac{m}{G-(K_{s}+1)\alpha \gamma} \sum_{j \in \chi_{s}} \frac{1}{h_{ij}} + \sum_{s=i+1}^{N-1} \frac{m}{G-(K_{s}+1)\alpha \gamma} \sum_{j \in \chi_{s}} \frac{1}{h_{ij}} + \sum_{j \in S_{i+1}} \frac{1}{h_{ij}} + \\
& \frac{m}{G-(k-1)\alpha \gamma} \sum_{j \in S_{i2}} \frac{1}{h_{ij}} \\
= & \sum_{s=1}^{i-1} \frac{m}{G-(K_{s}+1)\alpha \gamma} \sum_{j \in \chi_{s}} \frac{1}{h_{ij}} + \sum_{s=i+1}^{N-1} \frac{m}{G-(K_{s}+1)\alpha \gamma} \sum_{j \in \chi_{s}} \frac{1}{h_{ij}} + \sum_{j \in S_{i1}} \frac{1}{h_{ij}} + \frac{m}{G-(k-1)\alpha \gamma} \sum_{j \in S_{i2}} \frac{1}{h_{ij}} \\
= & \text{MinTTP for N-1 cell sectoring}
\end{align*}
\]
APPENDIX G

PROOF OF CORRECTNESS OF FASTEST ROUTING ALGORITHM

In Algorithm 1, at the beginning of the main for-loop (the one associated with \( h \)), \( T_{\text{min}}(v), v \in V \), give the minimum delivery time from \( S \) to \( v \) found up to the current hop expansion step \( h \). The inner for-loops enumerate all paths of the form \( S \xrightarrow{p_t} u \rightarrow v, p_1 = p_t(S \rightarrow u, h) \), for possible update of \( T_{\text{min}}(v), v \in V \). Furthermore, whenever \( T_{\text{min}}(v) \) decreases, \( \text{PrevNode}(v) \) changes accordingly, always keeping track of the node that delivered the packet to \( v \) along the fastest route found thus far. Hence, at the end of the algorithm, the end-to-end fastest route can be constructed by backtracking with information recorded in the array \( \text{PrevNode} \).
APPENDIX H

PROOF OF MIN ENERGY ROUTING LEMMA IN INTERFERENCE-FREE REGIME

Lemma 2: The minimum energy required to deliver one packet from $S$ to $v$ by time $t$ ($T^s < t \leq T^e$), is

$$E(S\rightarrow v, \ t) = \min_{u \in N} \{E(S\rightarrow u, \ t-1) + e(u \rightarrow v, \ t)\}. \quad (H.1)$$

Let $u^*$ be the node achieving the minimization above. An optimal path for $(S,v)$ is

$$p_e(S\rightarrow v, \ t) \equiv \begin{cases} p_e(S \rightarrow u^*, \ t-1), & u^* = v \\ S \rightarrow v, & u^* = S \\ p_e(S \rightarrow u^*, \ t-1) \oplus u^* \rightarrow v, & u^* \neq v, \ u^* \neq S \end{cases}$$

where $\oplus$ is an operation which concatenates two sub-paths into a path.

The proof can be done by induction over the discrete time serials.

Initially $n = T_s$, by the first time interval $T_s$, only direct transmission from $S$ to $v$ is allowed. $E(S\rightarrow v, \ T_s) = e(S \rightarrow v, \ T_s)$ and $p_e(S\rightarrow v, \ T_s) \equiv S \rightarrow v$. Here, $e(i \rightarrow j, \ t)$ is the lowest energy requirement for one-hop relay from $i$ to $j$ during the $t$-th time period. True.

Assume by time $\{n-1, T_s < n \leq T_e\}$, following (6.18) and (6.19), the optimal path from $S$ to $v$ can be calculated as $\{p_e(S \rightarrow v, \ n-1), v \in V\}$ with energy consumption $E(S\rightarrow v, \ n-1)$.

Consider the optimal path from $S$ to $v$ by time $t$, i.e. $p_e(S \rightarrow v, \ n)$. Let $u^*$ be the last relay node, such that $p_e(S \rightarrow v, \ n) = S \rightarrow u^* \rightarrow v$.

If, $n_{u^* \rightarrow v} \leq n - 1$, the optimal path has been recorded by time $n - 1$. As a result, $p_e(S \rightarrow v, \ n) \equiv p_e(S \rightarrow v, \ n - 1)$ and $E(S\rightarrow v, \ n) = E(S\rightarrow v, \ n - 1)$. 173
If, \( n_{u \rightarrow v} = n \),

\[
\min_{u \in V} \{ E(S \leadsto u, n-1) + e(u \rightarrow v, n) \} \leq E(S \leadsto u^*, n-1) + e(u^* \rightarrow v, n)
\]

\[
\leq \epsilon(p_e(S \leadsto v, n)) = E(S \leadsto v, n),
\]

Thus, the minimum energy consumption path can be obtained from (6.18) and (6.19).

This completes the proof and as a result, \( p_e(S \leadsto v, n) \) would be the optimal sub-path for any path that goes through \( S \leadsto v \) in time \([T_s, n]\) and \( v \leadsto D \) in time \([n, T_e]\).
APPENDIX I

PROOF OF MIN ENERGY ROUTING
LEMMMA IN INTERFERENCE-LIMITED
REGIME

Lemma 3: The minimum energy requirement for multiple packets \( k \in K \) delivery from \( S_k \) to \( D_k \) by time \( t \) (\( T_{min} \leq t \leq T_{max} \)) can be calculated as,

\[
E(\cup_{k \in K/N(t)} S_k \rightarrow D_k, t) = \min_{u \in U} \{ E(\cup_{a \in AC(t)} S_a \rightarrow u_a \cup_{f \in FC(t)} S_f \rightarrow D_f, t - 1) \\
+ \varepsilon(\cup_{a \in AC(t)} v_{ua}^{t} \rightarrow v_{Da}^{t+1} \cup_{s \in SC(t)} s_a \rightarrow v_{Ds}^{t+1}) \}
\]

where \( U = \{(\cup_{u_k}, k \in AC(t), u_k \in N)\} \), is the set of union on intermediate nodes. Let \( u* = (\cup_{u_k^*}) \) be the nodes, which achieve the optimization above. The optimal relay path \( p^*_k \) for packet \( k \in K/N(t) \) by time \( t \) is

\[
p^*_k(S_k \rightarrow D_k, t) \equiv \begin{cases} 
  v_{S_k}^{t} \rightarrow v_{D_k}^{t+1}, & u_k^* = S_k \\
  p^*_k(S_k \rightarrow D_k, t - 1), & u_k^* = D_k \\
  p^*_k(S_k \rightarrow u_k^*, t - 1) \oplus v_{u_k^*}^{t} \rightarrow v_{D_k}^{t+1}, & u_k^* \neq D_k, u_k^* \neq S_k
\end{cases}
\]

where \( \oplus \) is an operation to concatenate two sub-paths into a path.

**Proof:** The proof can be done by induction over the discrete time serials.

Initially \( t = T_{min} \), by the first time interval \( T_{min} \), only direct transmission from \( s_k \) to \( v_{D_k}^{t+1} \), \( k \in SC(t) \) is allowed.

\[
E(\cup_{k \in SC(t)} S_k \rightarrow D_k, t) = \varepsilon_{\cup_{k \in SC(t)}} (\cup_{s_k \rightarrow v_{D_k}^{t+1}}),
\]

and \( p_e = \cup_{k \in SC(t)} p^*_k(S_k \rightarrow D_k, t) = \cup_{k \in SC(t)} s_k \rightarrow v_{D_k}^{t+1} \). True.

Assume by time \( \{t - 1, T_{min} < t \leq T_{max}\} \), following (6.20), for any \( u \in U = \{(\cup_{u_k}, k \in AC(t), u_k \in N)\} \), the set of optimal paths for \( AC(t) \cup FC(t) \) can be
calculated as,
\[
    p_e = (\bigcup_{a \in \text{AC}(t)} p^e_a(S_a \leadsto u_a, \ t - 1)) \\
    \bigcup_{f \in \text{FC}(t)} p^f_e(S_f \leadsto D_f, \ t - 1)).
\]

with total energy consumption
\[
    E(\bigcup_{a \in \text{AC}(t)} S_a \leadsto u_a \bigcup_{f \in \text{FC}(t)} S_f \leadsto D_f, \ t - 1).
\]

Consider the set of optimal paths \( p_e \) from \( S_k \) to \( D_k \), \( k \in K/NC(t) \) by time \( t \). Let \( u^* = (\bigcup u^*_k) \), \( k \in \text{AC}(t), u^*_k \in \nu^t \) be their last relay nodes. Further let \( t_{u^* \rightarrow D} = (\bigcup_{k \in \text{AC}(t)} t_{u^*_k \rightarrow D_k}) \) be the set of time intervals for the last relays.

If \( \max\{t_{u^* \rightarrow D}\} \leq t - 1 \), the optimal path for \( k \in \text{AC}(t) \cup \text{FC}(t) \) has been recorded by time \( t - 1 \). As a result, \( p^e_k(S_k \leadsto D_k, \ t) \equiv p^e_k(S_k \leadsto D_k, \ t - 1), \forall k \in \text{AC}(t) \cup \text{FC}(t) \) and \( p^e_k(S_k \leadsto D_k, \ t) = s_k \rightarrow v^t_{D_k}, \forall k \in \text{SC}(t) \) with
\[
    E(\bigcup_{k \in K/NC(t)} S_k \leadsto D_k, \ t) \\
    = E(\bigcup_{a \in \text{AC}(t)} S_a \leadsto D_a \cup_{f \in \text{FC}(t)} S_f \leadsto D_f, \ t - 1) \\
    + \varepsilon(\bigcup_{k \in \text{SC}(t)} s_k \rightarrow v^t_{D_k}).
\]

Otherwise, let \( \Phi(t) \subseteq \text{AC}(t) \) be the set of packets with \( t_{u^*_k \rightarrow D_k} = t \), and \( U = \{(\bigcup u_k), \ k \in \text{AC}(t), \ \forall u_k \in N\}, \)
\[
    \min_{u \in U} \{E(\bigcup_{a \in \text{AC}(t)} S_a \leadsto u_a \cup_{f \in \text{FC}(t)} S_f \leadsto D_f, \ t - 1) \} \\
    + \varepsilon(\bigcup_{u \in \text{AC}(t)} v_{u_a}^{t-1} \rightarrow v^t_{D_a} \cup_{s \in \text{SC}(t)} s_s \rightarrow v^t_{D_s}) \\
    \leq \{E(\bigcup_{a \in \text{AC}(t)} S_a \leadsto u^*_a \cup_{f \in \text{FC}(t)} S_f \leadsto D_f, \ t - 1) \} \\
    + \varepsilon(\bigcup_{\phi \in \Phi(t)} v_{u^*_\phi}^{t-1} \rightarrow v^t_{D_\phi} \cup_{s \in \text{SC}(t)} s_s \rightarrow v^t_{D_s}) \\
    \leq \varepsilon(p^*_e) \\
    = E(\bigcup_{k \in K/NC(t)} S_k \leadsto D_k, \ t),
\]

As a result, the minimum energy consumption path can be obtained from (6.20).