CROSS LAYER SCHEDULING AND RESOURCE ALLOCATION ALGORITHMS FOR CELLULAR WIRELESS NETWORKS

by

Syed Hussain Ali

B.E., NED University of Engineering and Technology, 1993
M.Sc., King Fahd University of Petroleum and Minerals, 1998

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

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies
(Electrical and Computer Engineering)

The University Of British Columbia
(Vancouver)

October, 2008
© Syed Hussain Ali 2008
Abstract

This thesis considers the problem of cross layer scheduling and radio resource allocation of multiple users in the downlink of time-slotted and frequency-slotted cellular data networks. For these networks, opportunistic scheduling algorithms improve system performance by exploiting time variations of the radio channel. Within the broader framework of opportunistic scheduling, this thesis solves three distinct problems and proposes efficient and scalable solutions for them. First, we present novel optimal and approximate opportunistic scheduling algorithms that combine channel fluctuation and user mobility information in their decision rules. The algorithms propose the use of dynamic fairness constraints. These fairness constraints adapt according to the user mobility. The optimal algorithm is an off-line algorithm that precomputes constraint values according to a known mobility model. The approximate algorithm is an on-line algorithm that relies on the future prediction of the user mobility locations in time. We show that the use of mobility information increases channel capacity. We also provide analytical bounds on the performance of the approximate algorithm. Second, this thesis presents a new opportunistic scheduling solution that maximizes the aggregate user performance subject to certain minimum and maximum performance constraints. By constraining the performance experienced by individual users, who share a common radio downlink, to some upper bounds, it is possible to provide the system operator with a better control of radio resource allocations and service differentiation among different classes of users. The proposed solution offers better performance than existing solution under practical channel conditions. Finally, we present a dynamic subcarrier allocation solution for fractional frequency reuse in multicell orthogonal
Abstract

frequency division multiple access systems. We formulate the subcarrier allocation as an equivalent set partitioning problem and then propose an efficient hierarchical solution which first partitions subcarriers into groups and next schedules subcarriers opportunistically to users. Simulation results for three solutions illustrate the usefulness of the proposed schemes.
# Table of Contents

Abstract ................................................................. ii

Table of Contents ..................................................... iv

List of Tables ............................................................ ix

List of Figures ........................................................... x

Acknowledgements ....................................................... xiii

Dedication ............................................................... xiv

Statement of Co-Authorship .......................................... xv

1 Introduction ............................................................. 1
   1.1 System Model .................................................... 2
   1.2 Literature Review ............................................... 3
   1.3 Methodology and Motivation .................................. 6
   1.4 Contributions ................................................... 8
   1.5 Publications .................................................... 10
   1.6 Organization of Thesis ........................................ 11

Bibliography ............................................................ 15
## 2 Mobility Assisted Opportunistic Scheduling in Cellular Networks

### 2.1 Introduction

- **2.1.1 Related Work**
- **2.1.2 Contributions**
- **2.1.3 Outline**

### 2.2 Mobility and Channel Models

- **2.2.1 Finite State Semi-Markov Mobility Model**
- **2.2.2 Finite State Markov Chain Channel Model**
- **2.2.3 Hierarchical Time Scale**

### 2.3 Proposed Optimal MaOS Algorithm

- **2.3.1 Stage I: Exploiting Slow Variations**
- **2.3.2 Stage II: Exploiting Fast Variations**

### 2.4 Proposed Approximate MaOS Algorithm

### 2.5 Analysis of MaOS

- **2.5.1 Capacity Gain of Optimal Mobility Assisted Opportunistic Scheduling**
- **2.5.2 Bounds on Approximate Mobility Assisted Opportunistic Scheduling**

### 2.6 Implementation Details

- **2.6.1 Mobility Estimation and Prediction**
- **2.6.2 Estimation of Lagrange Multipliers**

### 2.7 Numerical Results of Optimal MaOS

- **2.7.1 Basic HDR Simulation Setup**
- **2.7.2 Performance Comparison**
- **2.7.3 Delay Comparison**
- **2.7.4 Effects of Parameters**
- **2.7.5 Capacity Gain**

### 2.8 Numerical Results of Approximate MaOS

- **2.8.1 Extended HDR Simulation Setup**
Table of Contents

2.8.2 Performance Comparison and Effects of Future Sample Path . . . . . . . 49
2.8.3 Robustness of Approximate MaOS Algorithm . . . . . . . . . . . . . . . 50
2.9 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
Bibliography . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

3 Throughput Constrained Opportunistic Scheduling . . . . . . . . . . . . . 65
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
3.2 Problem Formulation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69
3.3 Proposed Solution . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 71
  3.3.1 Properties of the Optimal Solution . . . . . . . . . . . . . . . . . . . . . . 71
  3.3.2 Solution Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
3.4 Performance Evaluation Methodology . . . . . . . . . . . . . . . . . . . . . . . . 75
  3.4.1 HDR Simulation Setup . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75
  3.4.2 Channel Models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75
  3.4.3 Evaluation Parameters . . . . . . . . . . . . . . . . . . . . . . . . . . . . 77
3.5 Numerical Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78
  3.5.1 Achieving Minimum Rates . . . . . . . . . . . . . . . . . . . . . . . . . . 78
  3.5.2 Achieving Minimum and Maximum Rates . . . . . . . . . . . . . . . . . 79
  3.5.3 Service Differentiation: MPG vs TCOS and MTMR . . . . . . . . . . . . 81
3.6 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
3.7 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
Bibliography . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94

4 Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA
  Networks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 97
  4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 97
        4.1.1 Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
# Table of Contents

4.1.2 Outline .................................................. 103
4.2 System Architecture and Model .................................... 103
   4.2.1 Dynamic Fractional Frequency Reused Cell Architecture .............. 103
   4.2.2 System Model ............................................. 104
   4.2.3 Channel and Data Rate Models ................................... 105
   4.2.4 Time Scale ............................................... 107
   4.2.5 Assumptions ............................................... 108
4.3 Problem Formulation ................................................. 109
4.4 Proposed Solution .................................................. 111
   4.4.1 RNC DSA Algorithm ........................................... 111
   4.4.2 BS DSA Algorithm ........................................... 114
4.5 Numerical Results ................................................. 117
4.6 Related Work ..................................................... 121
4.7 Conclusion ....................................................... 122
Bibliography ......................................................... 135

5 Conclusions and Future Work ....................................... 138
   5.1 Future Work .................................................. 141
   Bibliography ..................................................... 143

Appendices

A Proof of Theorem 1 ................................................. 145

B Proof of Theorem 2 ................................................. 146

C Proof of Proposition 1 .............................................. 148
   Bibliography ..................................................... 150
List of Tables

2.1 Frequently Used Notations ........................................ 53
2.2 Comparison of Delay Statistics ................................ 54

3.1 MPG vs. TCOS and MTMR: Average Rate, Standard Deviation of Scheduled
Data Rates (in Kbps) and Average HOL Delay (in slots) ............ 86

4.1 Estimated Distance between BS and a User in Central Cell ........ 124
4.2 Simulation parameters ............................................. 124
List of Figures

1.1 System Model for downlink cellular network. ........................................ 13
1.2 Multicell OFDMA system model. ......................................................... 14

2.1 Discrete state space mobility model and network structure: (a) three concentric rings, (b) five topological spaces, and (c) hexagonal cell structure with BS separation of 2Y. The rings or spaces are characterized by average SINR which depends on \( y_m \) (i.e., distance between BS and mobility state \( m \)). ....................... 55
2.2 Effects of future \( \tau \) and number of users \( N \) on the optimal solution of (2.10) when coefficients of the LP’s objective function are nonnegative uniformly distributed independent and identical random variables in the interval \([0,1]\). ......................... 55
2.3 Comparison of MaOS with LCS and MR: (a) average system data rate achieved for different values of \( \epsilon = \{0.99, 0.95, 0.90\} \), (b) individual data rate for \( \epsilon = 0.99 \) and (c) fairness. ................................................................. 56
2.4 MaOS performance as a function of \( \theta \) along with MR, LCS and LCS MSF performances. ................................................................. 57
2.5 Two user channel capacity using numerical and (where applicable) simulations for TD, LCS and optimum MaOS. ............................... 58
2.6 Effect of \( \tau \) on the performance of approximate MaOS, and its comparison with LCS and MR schemes. ........................................ 59
2.7 Effect of prediction errors on approximate MaOS performance. Two error scenarios are considered; first scenario restricts errors to adjacent states, and second scenario allows incorrect prediction of any state. 60

3.1 The average feasible data rate, $E(R_i)$, values for $i = 1, \ldots, 50$. Recall that $R_i$ is the feasible rate at which user $i$ can be served if scheduled in the time-slot. 87

3.2 CDF of scheduled rate for TCOS and MTMR algorithms under memoryless channel for 10, 20, 30, 40 and 50 users. Throughput is lower bounded by $C_i^l = 9.6$ Kbps. 88

3.3 TCOS and MTMR algorithms under Finite State Markov Channel for 10, 20, 30, 40 and 50 users. Throughput is lower bounded by $C_i^l = 9.6$ Kbps. (a) CDF of scheduled rate (b) average data rate $\bar{R}$ (c) HOL average delay $\bar{D}$. 89

3.4 Comparison of TCOS and MTMR under finite state Markov channel for 10, 20, 30, 40 and 50 users. Throughput is bounded within $C_i^l = 9.6$ Kbps and $C_i^u = 70.0$ Kbps. (a) CDF of scheduled rate (b) average data rate $\bar{R}$ (c) HOL average delay $\bar{D}$. 90

3.5 Average data rate (a) and corresponding $R_i$ multipliers values (b) for two users. Throughput is bounded within $C_i^l = 9.6$ Kbps and $C_i^u = 38.4$ Kbps. 91

3.6 Service differentiation TCOS vs MPG: Class A user data rates are bounded within $C_i^l = 19.2$ Kbps and $C_i^u = 80.0$ Kbps. Class B user data rates have lower bound of $C_i^l = 9.6$ Kbps and upper bound of $C_i^u = 38.4$ Kbps. 92

3.7 Service differentiation TCOS vs MPG: Class A user data rates are bounded within $C_i^l = 19.2$ Kbps and $C_i^u = 80.0$ Kbps. Class B user data rates have lower bound of $C_i^l = 9.6$ Kbps and upper bound of $C_i^u = 19.2$ Kbps. 93

4.1 Frequency band allocation in a static fractional frequency reused architecture. 125

4.2 Frequency band allocation in the proposed fractional frequency reused architecture. 125

4.3 Fractional frequency reused based cell models (a) 7-cell grid (b) 19-cell grid. 126
List of Figures

4.4 Pseudocode for RNC DSA algorithm ........................................... 127
4.5 Pseudocode for one scheduling slot of BS DSA for cell $k$ ............... 128
4.6 Traditional frequency allocation schemes (a) full frequency reuse with full interference (b) conventional sectored allocation ......................................................... 129
4.7 Grid size 7 results as a function of cell radius (a) comparison of proposed DSA with conventional schemes, (b) fraction of subcarriers in super group of proposed scheme ................................................................. 130
4.8 Grid size 19 results as a function of cell radius (a) comparison of proposed DSA with conventional schemes, (b) fraction of subcarriers in super group of proposed scheme ................................................................. 131
4.9 CDF of achieved data rates and lower bounds by the proposed and conventional schemes for cell radius of 1000 m: (a) grid size 7, (b) grid size 19. ............................................. 132
4.10 Network throughput comparison ................................................... 133
4.11 Cell edge throughput comparison for grid size=19, cell radius = 2000 m (a) data rates (b) CDF of data rates ................................................................. 134
Acknowledgements

I would like to express my profound gratitude and appreciation to my supervisor Dr. Victor C. M. Leung, for his guidance and patience throughout this research. It was because of him that the work at any point of time, never got stressful. His continuous support, advice and encouragement can never be forgotten. I would also like to thank Dr. Virkam Krishnamurthy for his collaboration and valuable advice. Thanks are also due to Dr. Ki-Dong Lee for his cooperation. I also acknowledge the valuable comments of my thesis committee members Dr. Vincent Wong and Dr. Son Vuong. I would like to thank my friends in the communication lab. Finally, I am appreciative of the Natural Sciences and Engineering Research Council of Canada, the University of British Columbia and Josephine T. Berthier for providing grants and fellowships for this research.
Dedicated to my parents
Statement of Co-Authorship

I am the first author and the main contributor of all the papers published and submitted from this thesis. Dr. Victor C.M. Leung supervised this thesis and is a co-author of all papers. Two papers published from Chapter 2 are co-authored with Dr. Vikram Krishnamurthy. He advised me to add capacity analysis and extend Dyer et al. inequality for the proposed solution. I performed the analysis and the extension. The papers published from Chapter 3 are co-authored with Dr. Ki-Dong Lee when he was a Research Associate under the supervision of Dr. Victor C. M. Leung. He reviewed the proposition proof and helped me in discussing the infeasibility results and the Lagrangian relaxation of the proposed solution. Dr. Lee is also a co-author of a paper published from Chapter 4. That paper handled two separate problems: dynamic subcarrier allocation and capacity planning. The capacity planning part was contributed by Dr. Lee which is not covered in this thesis.
Chapter 1

Introduction

In recent years, there have been active research efforts in enabling high speed data services through cellular networks. These efforts resulted in the development of standards and technologies for the third-generation (3G) of mobile systems that can support a maximum bit rate of a few Mb/s per user. Examples of such technologies are HDR (High Data Rate) [1] of cdma2000 1xEV-DO and HSDPA (High Speed Downlink Packet Access) [2] of 3GPP specification. In reality, these data rates fall short of consumer expectations of high speed data access. In order to improve the system performance further and to allow the integration of heterogeneous networks, research and development is directed toward the cross layer design [3], [4] and emerging technologies for the next generation networks (see, for example, [5], [6], [7]).

Cross layer design that combines information available in different layers of the network and makes intelligent resource allocation decisions can improve the system performance. Wireless systems generally have low performance due to the limited radio spectrum and transmission impairments. Therefore, any improvement resulting through the cross layer design helps in increasing the system capacity which results in reduced cost and an increased consumer satisfaction. The central theme of this thesis is the design of efficient resource allocation and scheduling algorithms which employ the cross layer technique of combining the information from different layers of the network in their decision rules with an objective of increasing the system capacity and the user performance. The underlying cross layer design technique combines the information from physical and medium access control layers. This introductory chapter sets the stage for the detailed discussion of the proposed algorithms. It overviews the system models along
Chapter 1. Introduction

with the related literature, research methodology, and contributions. It also provides the list of publications and an outline of the rest of the document.

1.1 System Model

This research considers the downlink resource allocation problems in the cellular data networks. Chapters 2 and 3 consider a model where multiple users time-share a single channel in a single cell. The model is based on the third generation systems like High Data Rate (HDR) [1] and High Speed Downlink Packet Access (HSDPA) [2]. The base station (BS) maintains a central scheduler which allocates the single channel in discrete time slots to the users. At a time slot, only one user can access this channel. Users periodically monitor their respective channel states and report back this information to the BS which employs adaptive modulation and coding (AMC) schemes to dynamically adjust the corresponding data rates. Users are infinitely backlogged which means that the number of users is constant and every user always contains data for transmission. Fig. 1.1 shows the model of such a network.

Chapter 4 considers an extension of the basic system model. The extension retains most of the properties of the basic model like infinitely backlogged users. Additionally, it increases the service area to include multiple cells instead of just one cell. Furthermore, there are several orthogonal subcarriers or channels available for time and frequency sharing between multiple users. This sharing of orthogonal subcarriers is called orthogonal frequency-division multiple access (OFDMA). Fig. 1.2 shows a sketch of the corresponding multicell OFDMA system model considered in this thesis. A central node, for example, radio network controller (RNC), controls the cells in the network and distributes radio resources to the individual cells. Every cell has a local scheduler running in the BS which allocates resources to the individual users.
Chapter 1. Introduction

1.2 Literature Review

One popular cross layer technique is called channel-aware or opportunistic scheduling. The scheduling schemes based on this technique exploit the time-varying and user-dependent radio channel fluctuations. The opportunistic scheduling is a general term that includes all the scheduling rules which give some sort of preference to the user with a high data rate [8].

Several opportunistic scheduling schemes have been proposed in the literature. The simplest scheme always selects the user with the maximum data rate. Such a trivial solution is often called greedy or myopic because it maximizes the system performance at the expense of fairness to the weaker users. Therefore, non-trivial opportunistic scheduling solutions try to maximize the system performance while satisfying some quality-of-service (QoS) attributes or constraints (see, for example, [9], [10], [11]). This thesis concentrates on these non-trivial opportunistic scheduling solutions.

We can broadly classify opportunistic scheduling algorithms in three categories according to the assumptions made for the system model or the type of the objective function. First, one commonly used system model is suitable for elastic traffic where flows are infinitely backlogged and the number of flows is constant. It means that each flow always contains data for transmission. For such systems, when the type of objective function is of the form

$$\max_Q \sum_{i=1}^{N} E \left( R_i 1_{\{Q(\bar{R}) = i\}} \right)$$  \hspace{1cm} (1.1)

where $R_i$ is the feasible rate for user $i$ which is defined as the rate at which user $i$ can be served if scheduled, $E \left( R_i 1_{\{Q(\bar{R})\}} \right)$ be the average scheduled data rate of user $i$ achieved by the scheduling policy $Q$, $1_{\{\cdot\}}$ is the indicator function, and $N$ users are sharing the radio channel. After extending the objective function with fairness constraints and solving the resulting constrained stochastic optimization problem, we get optimal policy of the following form [9]:

$$Q^*(\bar{R}) = \arg\max_i (\alpha^*_i R_i + v^*_i),$$
Chapter 1. Introduction

where $\alpha_i^*$ and $v_i^*$ are the Lagrangian multipliers that satisfy complementary slackness conditions of the fairness constraints. Several algorithms fall within this category, for example, [8], [9] and [10].

Second, within the infinitely backlogged model, when the objective function maximizes the product of the expected data rates, that is equivalent to the following

$$\max \sum_{i=1}^{N} \log \left( E(R_i 1\{Q_i(R) = i\}) \right)$$

[12]. The algorithms falling in this second category are called professional fair scheduling, (see for example [13], [14]).

Third, in contrast to the backlog model, other studies consider the presence of some arrival processes which feed into the user queues. The schemes falling in this category are based on Max Weight scheduling rule [15]. The objective of these schemes is to provide short-term fairness; therefore, these schemes consider time related user attributes like head-of-line packet delay [16] or queue length [17] in their decision rules.

Recently, the impact of mobility on the performance of wireless systems has become an active research area after the seminal work of Grossglauser and Tse [18] which showed that mobility increases the capacity of ad hoc mobile networks at the cost of increased delay. For cellular networks, [19] and [20] have analyzed system performance for single- and multicell scenarios respectively. According to these studies, intra- and inter-cell mobility increases the performance. This increase in performance provides an opportunity to design new cross layer resource allocation techniques where user mobility information is combined with the existing information in order to realize the capacity gain.

OFDMA builds on orthogonal frequency-division multiplexing (OFDM), which is immune to inter-symbol interference and frequency selective fading because it divides the frequency band into a group of mutually orthogonal subcarriers, each having a much lower bandwidth than the coherence bandwidth of the channel. In multi-user environments, OFDMA provides another degree of freedom by allowing dynamic assignment of subcarriers to different users at
Chapter 1. Introduction

different time instances, to take advantage of the fact that at anytime instance channel responses are different for different users at different subcarrier frequencies. Thus, dynamic subcarrier assignment (DSA) to multiple users can be employed to improve the system data rate. Since the achievable data rate is a function of the allocated power, i.e., the data rate increases with the transmit power and vice versa, it is expected that the adaptive power allocation (APA) can further improve the system data rate. The DSA and APA have attracted considerable research interests (see [21], [22], and references therein).

Most optimization-theoretic approaches to solve the DSA and APA problems assume continuous data rates and an infinite number of subcarriers in the OFDMA system (e.g., [21]). These ideal conditions render the DSA and APA problems tractable and enable the system data rate to be maximized through a greedy assignment algorithm. In the case of greedy assignments, a subcarrier is assigned to only one user that has the maximum channel gain on that subcarrier. However, this trivial solution is highly unfair because some users close to the BS can monopolize all the subcarriers while other users are starved from not receiving any assignments. Thus, it is desirable that the DSA problem should have fairness constraints on the individual users’ performance, or the objective function should be designed so that it implicitly enforces some form of fairness among users, as in [21], [23], [24].

In principle, allocating different power levels to individual subcarriers should improve performance. However, previous studies [21], [25] have shown that performance improvements are marginal over a wide range of signal-to-noise ratios (SNRs) due to the statistical effects. Therefore, a simpler solution involving DSA with equal power per subcarrier is preferred over the more complex joint DSA and APA solution.

System performance can be significantly increased if the DSA solution schemes are combined with the frequency reuse in the multicell OFDMA networks. In [26], authors propose a heuristic solution to otherwise exponentially complex optimal solution. The sub-carrier allocation by the radio network controller (RNC) is based on the presence or absence of the dominant interferer and the performance of the best user in the cell. RNC assigns subcarriers to the BS on a slower
Chapter 1. Introduction

time scale and provides its recommendations for the user allocation. The base station (BS) allocates the subcarriers to the users on a faster time scale, and in doing so, it can deviate from the RNC recommendation. In these situations, RNC algorithm is no longer optimum. On the other hand, if BS completely follows the RNC recommendations, then the system cannot benefit from the fast fading variations in the radio channel. The DSA solution presented in [27] adds fairness and does not require instantaneous channel knowledge at the super-frame level. However, the BS order of allocation at the super-frame level affects the signal-to-interference-plus-noise ratio (SINR) which in turn affects the data rate. The solution in [27] does not consider this aspect in its allocation.

Fractional frequency reuse (FFR) offers a simpler alternative to the frequency reuse problem in the multicell OFDMA networks. First time proposed for the TDMA networks in [28], it has been presented for the OFDMA networks in [29]. The subcarriers are partitioned into two groups. One group is reused in the central cell region of all cells while the other group is reused in the outer region of the cells according to the sectoring scheme. The overall frequency reuse factor is 1. In the traditional FFR scheme, subcarriers are statically allocated to the geographical regions (i.e., groups) on the basis of fixed thresholds [29]. The traditional FFR has several shortcomings. It does not consider the subcarrier’s state information in its allocation to the regions. Secondly, it divides users among groups on the basis of distance or SINR thresholds. The FFR can benefit from the channel-aware partitioning of subcarriers to the groups and their opportunistic assignment to the users.

1.3 Methodology and Motivation

The common theme of this thesis is to propose dynamic real-time algorithms that adapt to the changing conditions and use the available information intelligently in their decision rules. The proposed algorithms solve scheduling and radio resource allocation problems in multiuser cellular environments at the medium access control layer. These algorithms fall within a general
framework of constrained stochastic optimization of the system capacity, which is the sum of
the long-term expected data rates of users (1.1), under certain constraints. The constraints
represent fairness requirements. These algorithms employ cross layer techniques of combining
information from different layers of the system, for example, physical layer, medium access
layer, and user location.

In the first problem, we combine mobility information with the channel state information
in the objective function and the constraints of the optimization problem. The novelty in the
resulting solution allows us to exploit not only fading-related fast variations, but also path-loss-
related slow variations in the system capacity. In the second problem, we extend the general
framework of the constrained stochastic optimization (1.1) with the maximum bounds on the
individual data rates. Employing Lagrangian relaxation, we propose a real-time solution with
linear control variables. In the third problem, we improve the system capacity of the multicell
OFDMA network. Here, the complexity of the problem forces us to propose intelligent heuristic
solution which benefits, at one level from the average channel state information of groups of
users, and at the other level, from the instantaneous channel state information of individual
users. All the three solutions presented in this thesis are validated with extensive simulation
and numerical results.

Next, we summarize the motivation for the three research problems addressed in this thesis.
The details of the following summary is provided in respective chapters.

- As stated earlier, opportunistic scheduling exploits the time-varying channel fluctuations
due to fast fading in multi-user environments by preferring users with higher data rates.
Prior to this research, no other research collectively exploits the fast fading fluctuations
and the slow fading variations. Our proposed mobility assisted opportunistic scheduling
scheme (MaOS) optimally benefits from both of these variations and prevents long-term
starvation.

- Generally, problems based on multi-user scheduling attempt to maximize overall system
Chapter 1. Introduction

performance while satisfying lower bounds on the individual performance. Prior to this research, only one other study considered the use of the upper bounds on the individual performance [11]. Our proposed throughput constrained opportunistic scheduling (TCOS) has the capability to trade-off feasibility with throughput. This inherent attribute makes the proposed algorithm a preferred solution for realistic channel conditions.

- Fractional frequency reuse (FFR) offers a simpler solution of complex frequency reuse problem in the OFDMA multicell networks. The existing literature solves the associated set partitioning problem by statically deciding either on the SINR or on the distance between the user and the base station [29]. Our proposed FFR based DSA scheme dynamically partitions the users and subcarriers so that the overall system performance is maximized.

1.4 Contributions

This thesis solves three resource allocation problems by proposing intelligent cross layer algorithms. These algorithms provide quantitative and qualitative system improvements with minimal complexities. The intelligent algorithms revolve around opportunistic scheduling and have distinct contributions. The main contributions of these algorithms are:

- First, we formulate a semi-Markov mobility model that models user mobility among hot-spot locations within a cell. We propose an optimal mobility assisted opportunistic scheduling (MaOS) algorithm that combines user mobility information and channel fluctuations in the scheduling policy. The algorithm consists of two stages. The first stage exploits the slow time path loss variations and the second stage takes advantage of the fast time fading variations in the channel conditions by opportunistically scheduling users. The algorithm has sufficient mechanisms in place to avoid long term starvation of the users. Numerical results show that the proposed algorithm increases channel ca-
Chapter 1. Introduction

Capacity. In addition to the optimal MaOS algorithm, we also propose a computationally less intensive approximate MaOS algorithm which relies on the sample path of the user mobility. Furthermore, analytical lower bounds on the performance of the approximate MaOS are provided by generalizing the fundamental inequality of Dyer et al. [30],[31].

- Second, we formulate an opportunistic scheduling problem with minimum and maximum performance constraints. The maximum-rate constraints make the optimization problem difficult to solve in real-time. Therefore, we propose a very efficient iterative heuristic based on Lagrangian relaxation called throughput constrained opportunistic scheduling (TCOS). The proposed solution uses linear multipliers to control otherwise greedy behavior of the algorithm and constrains the solution within the feasible region. The significant aspect of the solution is its capability to trade-off feasibility with the performance. It relaxes the feasible region marginally in order to significantly improve the performance. The added advantage of the proposed solution is the ability of the system operator to have better control over resource allocation and service differentiation between different classes of users.

- Third, a dynamic sub-carrier allocation and scheduling solution is presented that employs fractional frequency re-use (FFR) concepts in the multicell multi-user OFDMA networks. Within FFR, first, we propose a new cell architecture that provides increased trunking gain by partitioning the subcarriers to geographical regions. Next, we formulate the subcarrier allocation as a set partitioning problem. Due to the exponential complexity of the optimal solution, a less expensive dynamic solution is presented that partitions the subcarriers into geographical regions, opportunistically schedules them to users and satisfies user fairness constraints. The proposed algorithm efficiently reuses subcarriers and improves the system capacity in comparison to the traditional frequency allocation schemes.
Chapter 1. Introduction

1.5 Publications

Several papers from this thesis have been published in journals and conference proceedings. In particular, the work described in Chapter 2 is published in IEEE Transactions on Mobile Computing. A version of Chapter 3 is accepted for publication in IEEE Transactions on Vehicular Technology. Some parts of Chapter 4 have been published in IEEE Wireless Communications while a detailed version is submitted to IEEE Transactions on Wireless Communications. Following is the complete publication list.


Chapter 1. Introduction

Signal Processing and Information Technology (ISSPIT), Vancouver, BC, Aug. 2006, pp. 708-713.


1.6 Organization of Thesis

This thesis is organized according to the manuscript based thesis format adopted by the Faculty of Graduate Studies, University of British Columbia\(^1\). In this format, except for the first and the last chapters, the rest of the chapters are self-contained manuscripts of research papers. Therefore in this thesis, Chapters 2 to 4 are manuscript chapters. These manuscript chapters include respective introduction, literature review, proposed solution, results and analysis, conclusions, and bibliography. Following is the organization of the thesis.

This chapter introduces research topic and reviews related literature. Furthermore, it overviews thesis contributions and provides publications list. The remaining chapters are organized as follows:

- Chapter 2 formulates a multi-user scheduling problem in the downlink of a time-slotted cellular network. The problem attempts to improve the system performance by exploiting user mobility information along with the time variations of the radio channel. We first propose a mobility model followed by the optimal and approximate mobility assisted opportunistic scheduling (MaOS) algorithms. The algorithms solve the problem by

\(^1\)http://www.grad.ubc.ca/students/thesis/index.asp?menu=002,002,000,000
Chapter 1. Introduction
dynamically adjusting fairness constraints according to the user mobility. The optimal algorithm has exponential complexity; therefore, a less expensive approximate algorithm is also presented. Additionally, we show that the use of mobility information increases the channel capacity and provide analytical bounds on the performance of the approximate algorithm.

• Chapter 3 formulates a multi-user scheduling problem with minimum and maximum throughput constraints. The problem attempts to provide system operator with better control and service differentiation among different classes of users. The chapter, proposes a dynamic algorithm, called throughput constrained opportunistic scheduling (TCOS), to optimally solve the problem in real-time. Extensive simulation results demonstrate the improved system performance and effective service differentiation of the proposed solution under realistic channel conditions.

• Chapter 4 formulates a dynamic channel allocation and scheduling problem in multi-user multi-cell OFDMA system as a set partitioning problem. The underlying objective of the problem is to improve the system capacity by employing fractional frequency reused allocation. The optimal solution would have exponential complexity; therefore, a real-time approximate algorithm is proposed which divides subcarriers according to their signal strengths and opportunistically allocates them to the users. Numerical results validate the efficiency of the proposed solution in comparison to the traditional schemes.

• Chapter 5 concludes this thesis and presents recommendations for future research.
Chapter 1. Introduction

Figure 1.1: System Model for downlink cellular network.
Chapter 1. Introduction

Figure 1.2: Multicell OFDMA system model.
Bibliography


Bibliography


Chapter 2

Mobility Assisted Opportunistic Scheduling in Cellular Networks

2.1 Introduction

We consider a downlink system model in a wireless cellular network where multiple data flows time-share a single channel. The model is inspired by third generation cellular systems such as HDR (High Data Rate) [1] and HSDPA (High Speed Downlink Packet Access) [2]. The central scheduler at the base station (BS) operates in discrete time slots and transmits to one user at a time. The channel condition, being a random time-varying and user-dependent process, affects the service rate of a respective user. Better channel condition results in higher rate, and vice versa. The channel conditions of all flows are conveyed to the scheduler as feasible rates before the start of the time slot. For this model, we consider infinitely backlogged flows where the number of flows is constant, and each flow always contains data for transmission. Such a model is suitable for elastic traffic that can tolerate variable delays. This model is introduced

in Section 1.1 and Fig. 1.1.

In the wireless context, multiple flows correspond to download queues for corresponding users, channel condition means signal-to-interference-plus-noise ratio (SINR), and service rate equals data rate. User equipment measures SINR in response to a BS pilot signal where SINR accounts for path loss and fading effects [3]. BS receives channel information in the form of feasible rate, that is, the rate at which corresponding user can be served if scheduled in the time slot. Path loss variations occur at a slow timescale due to changes in distance between transmitter and receiver, whereas fading variations occur at a fast timescale due to multipath propagation.

For the above system model, scheduling plays an important part in the system performance because it decides which user is served in each time slot. If the scheduler selects a user with a high SINR value then that user receives a high data rate. Conversely, a user selected with a low SINR value receives a low data rate. A scheduling scheme that gives preference to the users with a high data rate is called opportunistic scheduling [4]. Opportunistic scheduling takes advantage of the fast time fluctuations in the radio channel due to fading. It improves system performance by favoring users at a time when they have higher data rates. A trivial solution would be to always serve the best user, i.e., maximum rate (MR) scheduling. However, this solution leads to unfairness because a few users with excellent wireless links can starve other users. Therefore, research in this area concentrates on proposing schemes that trade overall performance with quality-of-service (QoS) requirements like fairness (see, for example, [5], [6], [7]).

In this chapter, our objective is to improve the long-term system data rate for elastic data traffic. To accomplish this objective, we exploit not only fading-related fast variations, but also path-loss-related slow variations in the feasible data rate. We present novel opportunistic scheduling schemes in which decision rules combine user’s mobility and channel information. In a macro-cellular environment with mobile user population, the average SINR of a mobile user will improve (or worsen) if the user moves toward (or away from) the BS. If the scheduler
can track and anticipate user mobility, it can adapt scheduling priorities with the help of
dynamic constraints. It can give priority to users that are in the most favorable locations, e.g.,
closer to the BS. Furthermore, adapting priorities should still satisfy user’s long-term fairness
constraints and prevent long-term starvation. The proposed *mobility assisted opportunistic
scheduling* (MaOS) algorithm achieves these objectives for infinitely backlogged elastic data
traffic. First, the algorithm adapts user priorities by dynamically adjusting fairness constraints
according to the user mobility information. In this way, it takes advantage of the slow time
variations in the radio channel. Next, it uses these priorities to opportunistically schedule the
users. Thus, it also exploits fast time fading variations.

There are two versions of the proposed algorithm: the optimal MaOS, which uses full
state space knowledge and pre-computes user priorities for all possible mobility states, and
the approximate MaOS, which relies on the future prediction of user mobility to compute user
priorities in real-time.

### 2.1.1 Related Work

**Infinite Backlog Model**

Liu et al. have proposed opportunistic scheduling algorithms that maximize the sum of the
expected user data rates and satisfy long-term fairness constraints under an infinite backlog
assumption [4],[5]. Their algorithms (abbreviated here as LCS algorithms) optimally exploit
the fast time fading variations in the feasible rate. They consider several fairness constraints,
but to keep the discussion short, we discuss only *temporal* fairness measure and present our
proposed mobility assisted solution for the same measure. We emphasize that the results and
analysis presented in this chapter are true for other measures considered in [5].

Let there be $N$ users in a cell, and $\bar{R} = (R_1, ..., R_N)$ be the feasible data rate vector at a
generic scheduling time slot. Recall that the feasible rate $R_i$ for user $i$ is defined as the rate at
which user $i$ can be served if scheduled in the time slot. Further assume that $E\left(R_i \mathbf{1}\{Q(\bar{R})\}\right)$

Chapter 2. *Mobility Assisted Opportunistic Scheduling in Cellular Networks*
is the average scheduled data rate of user $i$ achieved by the scheduling policy $Q$, where $\mathbf{1}_{\{i\}}$ is the indicator function. The overall system data rate is $E\left(R_{Q(\vec{R})}\right) = \sum_{i=1}^{N} E\left(R_{i}\mathbf{1}_{\{Q(\vec{R})=i\}}\right)$.

The *temporal* fairness measure considers time as a resource and shares it among multiple users according to the given fairness constraints. Therefore, if $P\{Q(\vec{R}) = i\}$ is the probability of scheduling user $i$ by policy $Q$ and $r_i$ is the long-term minimum temporal requirement for user $i$, where $r_i \geq 0$, $\epsilon := \sum_{i=1}^{N} r_i \leq 1$, then the LCS algorithm solves the following stochastic optimization problem:

$$\max_{Q} \quad E\left(R_{Q(\vec{R})}\right), \quad \text{s.t.} \quad P\{Q(\vec{R}) = i\} \geq r_i, \quad i = 1, 2, ..., N. \tag{2.1}$$

According to the LCS algorithm, the optimal policy is

$$Q^*(\vec{R}) = \arg\max_{i}(R_i + v_i^*), \tag{2.2}$$

where $v_i^*$ are the true Lagrange multipliers that satisfy the complementary slackness conditions of the constraints in problem (2.1). The LCS algorithm has a linear complexity of $O(N)$.

Let us consider an example to clarify physical meanings of the temporal fairness measure. Assuming that there are 10 users sharing a single channel with fairness constraints given as $r_i = 1/12$. These constraints mean that each user expects an allocation of at least 8.33 percent of the time slots for itself. This being a long-term fairness measure considers only the fraction of the total time. Thus, position of the allocated time slots or their order is not important.

**Impact of Mobility**

The positive impact of mobility on the performance of wireless systems is an active research area after [8] showed that mobility increases the capacity of ad-hoc mobile networks at the cost of increased delay.

For cellular networks, [9] and [10] have analyzed system performance for single- and multi-cell scenarios, respectively. According to these studies, intra- and inter-cell mobility increases performance. Certainly in this context, our work enforces the results of above studies that
mobility increases system performance. However, our work is significantly different than [9] and [10]. We propose a new scheduling algorithm that optimally exploits user mobility and prevents long-term starvation.

Other Models and Objective Functions

Since our emphasis is on long-term fairness, we maximize the sum of expected data rates under infinite backlog assumption. These assumptions shield the scheduling process from delay distributions and arrival process dynamics and are appropriate for elastic traffic applications like file download.

Alternatively, for delay constrained applications different objective functions or models are used. For example, proportional fair scheduling (PF) [11], [12] maximizes the product of the expected data rates which is mathematically equivalent to max \( \sum_{i=1}^{N} \log \left( E(R_i 1_{\{Q(R_i)=i\}}) \right) \) [13]. This objective function favors weak users over strong users. The PF also exploits the time-varying channel conditions and provides proportional fairness which is different than the temporal fairness considered in this chapter [4]. PF reduces delay variance among users, but in comparison to the class of algorithms that maximize the sum of the data rates PF achieves overall lower data rates [4], [7]. Recall that LCS and MaOS fall into the category of algorithms that maximize the sum of the data rates. Like LCS, PF scheduling has polynomial complexity of \( O(N) \).

Furthermore in contrast to the backlog model, some studies consider another model in which arrival processes feed queues. For this model, popular scheduling algorithms based on max-Weight scheduling rule [14] aim to establish short-term fairness. Therefore, they combine queue length [15], or head-of-line packet delay [16] with the channel information in their decision rules.
2.1.2 Contributions

Following are the main contributions of this chapter:

1. We formulate a semi-Markov mobility model which is a coarse-grained mobility model suitable for user mobility among hot-spot locations within a cell. We then propose MaOS algorithm that combines mobility and channel information in the scheduling rule. MaOS consists of two stages and like LCS considers time as a resource. In the first stage, the algorithm takes advantage of the slow time path loss variations and optimally distributes time fractions, which can be considered as priorities, among all users. In the second stage, the algorithm exploits the fast time fading fluctuations by opportunistically scheduling users according to the resulting time fractions from the first stage. We prove that the optimal MaOS algorithm increases the channel capacity.

2. Because the optimal MaOS algorithm is computationally expensive, a computationally less intensive approximate MaOS algorithm is proposed that relies on time windowing and the sample path of the user mobility.

3. We provide analytical lower bounds on the performance of the approximate MaOS. For this purpose, we generalize the fundamental inequality of Dyer, Frieze and McDiarmid [17],[18] and analyze the performance of the approximate MaOS algorithm. The original inequality works for minimization linear programs (LP) with certain conditions that makes it unsuitable for maximization problems. We extend the inequality for maximization problems and analyze the approximate MaOS with respect to the sample path and the number of users. This analysis helps us to find lower bounds on the system performance of the approximate MaOS algorithm.
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

2.1.3 Outline

Section 2.2 briefly describes mobility and channel models. Section 2.3 presents the optimal MaOS algorithm followed by the approximate MaOS algorithm in Section 2.4. Section 2.5 analyzes the optimum algorithm in terms of the channel capacity gain and proposes bounds on the approximate algorithm. Implementation details of both algorithms are covered in Section 2.6. Numerical results of the optimal MaOS algorithm and a comparison with the LCS algorithm are given in Sections 2.7. Section 2.8 compares the approximate MaOS with the LCS algorithm and Section 2.9 concludes this chapter. Frequently used notations are summarized in Table 2.1.

2.2 Mobility and Channel Models

This section presents mobility and channel models used in this chapter. For mobility model, we propose a novel finite state semi-Markov model that is suitable for mobility between hot-spot locations in a cell or wireless coverage area. This model is a coarse-grained model because it divides the cell surface into non-overlapping regions. The channel model is a popular finite state Markov chain that models Rayleigh fading process.

2.2.1 Finite State Semi-Markov Mobility Model

This chapter considers a discrete state mobility model that divides cell geography into a set of non-overlapping topological spaces $S = \{1, \ldots, M\}$ called mobility states. Fig. 2.1 shows two examples of the mobility model where states are regular concentric rings (Fig. 2.1a) [19], or irregular cell areas depending on accessibility and geography of cell (Fig. 2.1b). These states are characterized by average SINR, $\rho(m)$, where $m \in S$. User mobility among these states follows a semi-Markov process with (known) generally distributed dwell times $P_d(m)$ and mean $\bar{d}(m)$. The stationary distribution of the underlying process is found as follows [20]:

\begin{equation}
\text{25}
\end{equation}
\[ \pi(m) = \frac{\pi^e(m)d(m)}{\sum_{m' \in S} \pi^e(m')d(m')} , \]
where \( \pi^e(m) \) is the stationary distribution of the embedded Markov chain.

The above model is a coarse-grained model with discrete state space. However, user mobility occurs in continuous state space where users follow fine-grained mobility patterns like random walk or random waypoint [21]. The proposed model is adaptable to any fine-grain mobility and cell structure. For example, speed and structure of \( S \) affect dwell time distribution. The proposed model learns parameters, \( \pi^e(.) \), \( P_d(.) \) and \( \bar{d}(.) \), from simulation or actual data.

### 2.2.2 Finite State Markov Chain Channel Model

The SINR is a function of path loss, interference and fading process [3].

**Path Loss and Interference**

The path loss is a function of the distance from the BS, and the interference is a function of the distance from the neighboring cells. Because our mobility model allows user mobility over large areas, average SINR will change over time resulting in non-stationary channel model. To cope with non-stationarity and to have a simpler model that allows analysis, we employ a piece-wise stationary channel model.

For the piece-wise stationary channel model, we assume that the path loss and interference remains constant within a mobility state. These values are computed at the center of the mobility state and used in the estimation of the channel model for that state. As long as a user remains in a mobility state, the user experiences channel conditions according to the stationary probability distributions for that state. Transition to another mobility state results in another channel model with different set of probability distributions. The piece-wise stationary model captures the essence of non-stationarity due to the user mobility and allows a simpler analysis compared to a fully non-stationary channel model.
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

Let mobility state $m$ be at distance $y_m$ from the BS (see Fig. 2.1a). According to a simple propagation model, also considered in [19], the path loss at distance $y_m$ from the BS is given as:

$$\Gamma(m) = \begin{cases} 
1 & \text{if } y_m \leq y_0 \\
\left(\frac{y_0}{y_m}\right)^\alpha & \text{otherwise}
\end{cases}$$

(2.3)

where $y_0$ is the maximum distance from the BS up to where full transmitted power $P_t$ is received, and $\alpha$ is the path loss exponent.

For the network structure, we consider 7-cell structure (Fig. 2.1c). The central cell forms the service area, and the other six cells act as sources of interference. The interference from the neighboring cells is approximated as [19]

$$I(m) = P_t \times \left( \Gamma(2Y - y_m) + 2\Gamma(\sqrt{(Y - y_m)^2 + 3Y^2}) + 2\Gamma(\sqrt{(Y + y_m)^2 + 3Y^2}) + \Gamma(2Y + y_m) \right).$$

(2.4)

where $P_t$ is the transmit power of the interfering cell, and $Y$ is the cell radius.

Fading Process and SINR

For the small scale fading, a flat fading channel is assumed because for systems like HDR and HSDPA, the feedback delay is relatively short compared to the fading frequency [1].

Let user $i$ be in mobility state $m$ at a generic time slot. According to the fading channel model, the received SINR behaves as follows [19]:

$$Z_i(m) = \zeta_i \frac{P_t \times \Gamma(m)}{\eta + I(m)},$$

where $\zeta_i$ is the time-dependent fading process, $P_t$ is the transmitted power, $\Gamma(m)$ and $I(m)$ are the path loss and interference experienced by the user, and $\eta$ is the background noise. Furthermore, fading process, $\zeta_i$, are independent and identically distributed (i.i.d.) copies of an exponentially distributed stationary process with unit mean.$^3$ Thus, the SINR process, $Z$,

---

$^3$ The i.i.d. and unit mean assumptions are not necessary for the validity of our results; these assumptions are considered only to simplify discussion.
also behaves as an exponentially distributed random process with mean $\rho(m) = P_t \times \Gamma(m)/(\eta + I(m))$, and, for $z \geq 0$, its distribution is given as [22]:

$$p_Z(z, m) = \frac{1}{\rho(m)} e^{-\frac{z}{\rho(m)}}.$$ 

Let $\mathcal{H} = \{Z_0, Z_1, \ldots, Z_{H-1}\}$ represent the set of discrete SINR levels in ascending order.

The fading channel takes level $h$ if the received SINR falls in the interval $[Z_h, Z_{h+1})$. The corresponding steady state probability of level $h$ in mobility state $m$ is found as [23]:

$$p(h, m) = \int_{Z_h}^{Z_{h+1}} p_Z(z, m) dz = e^{-\frac{Z_h}{\rho(m)}} - e^{-\frac{Z_{h+1}}{\rho(m)}}. \quad (2.5)$$

The transition probabilities are computed according to the procedure given in [23] and [24]. The physical layer coding and modulation schemes map fading levels to as many feasible data rates.

### 2.2.3 Hierarchical Time Scale

The overall system model employs hierarchical timescale with two levels. The higher level, referred to as mobility slot, operates on coarse-grained user mobility and has wider time slot. This time slot allows the proposed algorithm to take advantage of the slow time path loss variations in the feasible data rate. At the mobility slot level, the fast time fading variations in the feasible rate can be represented by their expected values. The lower level, referred as scheduling slot, handles scheduling decisions and has narrower time slot. The scheduling slot allows algorithm to benefit from the fast time fading fluctuations. At the scheduling slot, the slow time path loss variations can be modeled as a constant like path loss experienced at the center of the corresponding mobility state. The presence of two timescales helps to accommodate user mobility that evolves relatively slowly compared to the scheduling decisions.

**Remark on Time Index**: Channel fluctuations, feasible rates, and scheduling decisions depend on the scheduling timescale, whereas user mobility and expected feasible rates follow the mobility timescale. For simplicity of discussion, generally, we do not index variables with time.
However, time index is used for the mobility slot when considering sample path of the user mobility. For other places, the use of the time index is clear according to the context.

### 2.3 Proposed Optimal MaOS Algorithm

The proposed optimal MaOS algorithm consists of two stages. The first stage exploits the slow time path loss variations due to the user mobility and the second stage takes advantage of the fast time fading fluctuations due to the multipath propagation.

#### 2.3.1 Stage I: Exploiting Slow Variations

Recall that, at the timescale of user mobility, the fading experienced by a user in a mobility state can be abstracted by the corresponding expected feasible rate. In order to improve the system data rate, it is imperative to give preference to users when they are in the most favorable locations, i.e., having high expected feasible rates. Since the proposed algorithm considers time as a resource, if time fractions are distributed such that they maximize their product with the expected feasible rates of the corresponding users for all aggregate mobility states; then the algorithm can achieve its stated objectives of improving the system data rate. Therefore, this stage of the MaOS algorithm determines optimum time fractions that maximize the expected feasible rates of the users under the constraints of providing fairness and preventing starvation.

For \( N \) users and \( M \) mobility states per user, the set of possible aggregate states is an \( N \)-dimensional space given by \( S = S_1 \times \ldots \times S_N \), where \( \times \) denotes the Cartesian product, and \( S_i \) is the state space of user \( i \) defined in Section 2.2.1. Thus, the aggregate mobility model contains \( M^N \) states (i.e., \(|S| = M^N\)), and a state is identified by a vector \( \mathbf{m} = (m_1, \ldots, m_N) \) at a generic mobility slot. Assuming independence among users, the stationary distribution of state \( \mathbf{m} \in S \) equates to the product of the individual stationary distributions, \( \Pi(\mathbf{m}) = \prod_{i=1}^{N} \pi(m_i) \). Further, assume that the scheduler knows the above mobility model, expected feasible rates \( R(\mathbf{m}, i) \forall \mathbf{m} \) and \( i \), and can track user state accurately.
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

The proposed MaOS algorithm modifies LCS algorithm by using dynamic fairness constraints. The algorithm first computes the optimal time fraction allocations for every mobility state \( \vec{m} \) that maximize their product with the expected feasible rate summed over all states and for all users. For the notation used in (2.1) where \( r_i \) is the long-term minimum temporal requirement for user \( i \) and \( \epsilon = \sum_{i=1}^{N} r_i \leq 1 \), the proposed MaOS algorithm solves the following LP:

\[
\begin{align*}
\text{max} & \quad \sum_{\vec{m} \in \mathcal{S}} \sum_{i=1}^{N} \bar{R}(\vec{m}, i) r(\vec{m}, i), \\
\text{s.t.} & \quad \sum_{i=1}^{N} r(\vec{m}, i) \leq \Pi(\vec{m}), \quad \forall \vec{m} \in \mathcal{S}, \\
& \quad \sum_{\vec{m} \in \mathcal{S}} r(\vec{m}, i) = r_i, \quad i = 1, \ldots, N, \\
& \quad r(\vec{m}, i) \geq \frac{\theta r_i}{|\mathcal{S}|}, \quad \forall \vec{m} \in \mathcal{S}, \forall i, \text{ and } 0 \leq \theta \leq \theta_{\max},
\end{align*}
\]

where \( r(\vec{m}, i) \) are optimization variables and \( \theta \) is a parameter.

The above LP maximizes the weighted sum of expected feasible rate and complies with the stationary distribution of user mobility (2.6b), satisfies long-term temporal fairness constraints (2.6c), and prevents long-term starvation (2.6d). The time fraction \( r_i \) of (2.1) behaves as a resource, and the above LP optimally distributes it among all mobility states accessible to the user \( i \).

The rational for \( \theta > 0 \) in (2.6d) is to prevent the long-term starvation of users. When \( \theta = 0 \), the resulting solution of (2.6) is a greedy solution where users are denied access in mobility states with weak channels. This denial results in the starvation of such users. Therefore, \( \theta > 0 \) guarantees that a minimum fraction of \( r_i \), distributed over all mobility states, is assigned to every user even in mobility states with bad channels. This way MaOS algorithm avoids long-term starvation. For the feasibility of LP, \( \theta \) is upper bounded by \( \theta_{\max} = \min_{\vec{m}} (\Pi(\vec{m})) / |\mathcal{S}| / \epsilon \).

Let \( r^*(\vec{m}, i) \) maximize the objective function (2.6a). This value is a fraction of \( r_i \) because
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

of the construction of LP (2.6). For example, when $\epsilon = 1$ then $r_i$ behaves like the probability of allocation of user $i$. The resulting $r^*(\vec{m}, i)$ is the joint probability of two events: user $i$’s allocation and the occurrence of mobility state $\vec{m}$. Thus, user $i$’s allocation probability given mobility state $\vec{m}$ can be found by dividing $r^*(\vec{m}, i)$ with the corresponding probability of state.

For a more general case when $\epsilon < 1$ and considering the feasibility of (2.6b) constraint, the resulting allocation probability, identified here as normalized time fraction, $\hat{r}(,)$, is found as:

$$\hat{r}(\vec{m}, i) = \epsilon \frac{r^*(\vec{m}, i)}{\sum_{i=1}^{N} r^*(\vec{m}, i)}, \quad \forall \vec{m} \in S.$$  \hspace{1cm} (2.7)

The next stage of the algorithm uses the normalized time fractions as the dynamic fairness constraints and opportunistically schedules users.

2.3.2 Stage II: Exploiting Fast Variations

In this stage, algorithm gains from the fast time fading fluctuations in the feasible data rate. For this purpose, it uses the LCS methodology for the normalized time fractions found in the first stage. It maximizes the expectation of the scheduled (system data) rate for each mobility state as well as satisfies the respective time fraction allocation. For this objective, algorithm solves the following stochastic optimization problem after modifying problem (2.1):

$$\max_{Q} \mathbb{E}(R_Q(\bar{R}(\vec{m}))), \text{ s.t. } P\{Q(\bar{R}(\vec{m})) = i\} \geq \hat{r}(\vec{m}, i), \quad i = 1, 2, ..., N.$$  \hspace{1cm} (2.8)

The resulting problem has dynamic constraint values that are functions of $\vec{m}$. The optimal scheduling policy is found by adapting the solution (2.2) for state $\vec{m}$:

$$Q^*(\bar{R}) = \arg\max_i (R_i + v^*_i(\vec{m})), \quad \text{where } v^*_i(\vec{m}) \text{ is the true Lagrange multiplier that satisfies the corresponding constraint in (2.8).}$$  \hspace{1cm} (2.9)

Intuitively speaking, MaOS is a piece-wise LCS algorithm, that is, one for every $\vec{m}$ state.
2.4 Proposed Approximate MaOS Algorithm

The optimal MaOS algorithm has exponential complexity in the number of users because it pre-
computes time fractions for all aggregate mobility states. The approximate algorithm computes
time fractions in real-time, and relies on the future prediction of the mobility states of individ-
ual user. Unlike optimal MaOS, the approximate algorithm does not rely on the underlying
mobility model. Rather, it depends on the sample path of the user mobility.

Like optimal MaOS, the approximate algorithm also consists of two stages. First stage
gains from the slow time path loss variations. Assuming at (mobility) time $t$, the scheduler can
accurately predict the future state of all users for next $\tau$ mobility slots. Using this prediction
during the first stage, the algorithm finds time fraction values so that users close to the BS
could be given preferences. During the second stage, the algorithm opportunistically schedules
users according to the time fractions found in the first stage. The formulation of the second
stage is given in the following paragraphs.

The first stage of the approximate MaOS finds $\hat{r}(t + \Delta, i)$, $\forall i$ and $\Delta = 1, \ldots, \tau$, as a solution
to the following LP:

\[
\begin{align*}
\max & \quad \sum_{\Delta=1}^{\tau} \sum_{i=1}^{N} \bar{R}(t + \Delta, i)r(t + \Delta, i) \\
\text{s.t.} & \quad \sum_{i=1}^{N} r(t + \Delta, i) = \epsilon, \quad \forall \Delta, \\
& \quad \frac{1}{\tau} \sum_{\Delta=1}^{\tau} r(t + \Delta, i) = r_i \quad i = 1, \ldots, N.
\end{align*}
\]

(2.10a) (2.10b) (2.10c)

where $\bar{R}(t + \Delta, i)$ is the average feasible data rate of user $i$ at $t + \Delta$ (mobility) time slot in the
future in accordance with the predicted mobility state. The constraint (2.10b) ensures that for
every $\Delta$ the time fraction resource $\epsilon$ is distributed among users, and constraint (2.10c) makes
sure that for every $\tau$, the assigned fraction satisfies the long-term fairness requirement. The
above LP (2.10) is solved after every \( \tau \) mobility slots.

Let \( \hat{r}(t+\Delta, i) \) maximize the objective function of LP (2.10). In the second stage these values are used to modify and solve problem (2.1). The resulting stochastic optimization problem maximizes the expectation of the scheduled (system data) rate and satisfies the new fairness constraints, as:

\[
\max_Q \quad E \left( R_Q(\vec{R}(t+\Delta)) \right),
\]
\[
\text{s.t.} \quad P\{Q(\vec{R}(t+\Delta)) = i\} \geq \hat{r}(t+\Delta, i), \quad i = 1, 2, \ldots, N, \quad \Delta = 1, \ldots, \tau.
\]

\( \text{(2.11)} \)

**Remark:** The value of constraint constant \( \hat{r}(t+\Delta, i) \) changes only when a user makes a transition to a new mobility state. After that transition, it remains constant for several mobility time slots until the next transition. If need be, (2.10) can be modified by lower bounding \( r(t+\Delta, ) \), like (2.6d), in order to avoid possible long-term starvation.

The solution of problem (2.11) is a modified LCS policy that selects a user at every (scheduling) time slot according to the following rule:

\[
Q^*(\vec{R}) = \arg \max_i (R_i + v^*_i(t+\Delta)),
\]

\( \text{(2.12)} \)

where \( v^*_i(t+\Delta) \) is the Lagrange multiplier for the constraint active at (mobility) time \( t+\Delta \). The resulting approximate MaOS algorithm has a complexity of \( O(N\tau) \).

### 2.5 Analysis of MaOS

This section analyzes optimal and approximate MaOS algorithms. Section 2.5.1 proves that the optimal MaOS algorithm increases channel capacity. Section 2.5.2 provides bounds on the first stage (2.10) of the approximate MaOS algorithm.
2.5.1 Capacity Gain of Optimal Mobility Assisted Opportunistic Scheduling

In this section, we use numerical methods to compute the ergodic capacity for LCS and optimal MaOS algorithms in fading channel environments. The ergodic capacity is defined as [25]: “the maximum long-term achievable rate averaged over all states of the time-varying channel.” Because the channel distribution is a function of mobility state (Section 2.2.2), the ergodic capacity definition is modified as: “the long-term achievable rate that is averaged over all mobility and fading states.” Furthermore, it is assumed that the LCS and MaOS algorithms are working in the restrictive constraint regime, i.e., $\epsilon \approx 1.0$. In this regime, constraints are almost always binding; therefore, LCS cannot overschedule any user beyond its fairness requirement.

The allocation of users is spread all over the scheduling time horizon. Using this fact, first we generate the time division (TD) capacity equations of [25] for the ergodic capacity followed by similar equations for LCS and MaOS algorithms.

We only consider capacity region in two-dimension (i.e. two users sharing a common channel) because the general characteristics of the spectrum sharing methods remain the same for a large number of users as in the two user case [26]. Thus, for two users and $H$ channel conditions per user, the set of possible aggregate channel conditions is a 2-dimensional space given by $\mathbb{H}(\mathbf{m}) = \mathcal{H}_1 \times \mathcal{H}_2$, where $\times$ denotes the Cartesian product and $\mathbf{m}$ in parenthesis signifies that $\mathbb{H}$ is a function of aggregate mobility state. An aggregate channel condition is identified by a vector $\mathbf{h} = (h_1, h_2)$ and the corresponding feasible user data rate as $R(\mathbf{h}, \cdot)$. Assuming independent channels among users, the steady state distribution of state $\mathbf{h} \in \mathbb{H}(\mathbf{m})$, $P(\mathbf{h})$, is computed as a product of the individual steady state distributions. Thus, using the modified definition of ergodic capacity, we can write the TD capacity of [25] as follows:

$$C_{TD} = \bigcup_{\{r_1: 0 \leq r_1 \leq 1\}} \left( C_1 = E_{\mathbf{h}} \left[ E_{\mathbf{h}} \left[ r_1 R(\mathbf{h}, 1) \right] \right], C_2 = E_{\mathbf{h}} \left[ E_{\mathbf{h}} \left[ (1 - r_1) R(\mathbf{h}, 2) \right] \right] \right),$$

where $r_1$ is the fraction of time slots to be assigned to user 1, the remaining fraction, $1 - r_1$, to
user 2; $E_{\bar{h}}$ and $E_{\bar{m}}$ are expectations with respect to $\bar{h}$ and $\bar{m}$ respectively.

Equation (2.13) is the nonopportunistic TD capacity equation because it assigns an $r_i$ time fraction to user $i$ for all $\bar{h}$. However, LCS algorithm is opportunistic; therefore, time fraction allocation also depends on the corresponding feasible data rate with an objective of maximizing overall system rate and satisfying fairness constraints. Thus, we need to discover the fraction of times the algorithm selects a user for every state $\bar{h} \in \mathbb{H}(\bar{m})$. We identify this fraction as $\hat{\phi}(\bar{h}, i), i = 1, 2$. The corresponding LP that achieves LCS objectives for every $\bar{m}$ is:

$$\max \sum_{\bar{h} \in \mathbb{H}(\bar{m})} 2 \sum_{i=1}^2 R(\bar{h}, i) \phi(\bar{h}, i),$$

s.t. $\sum_{i=1}^2 \phi(\bar{h}, i) \leq P(\bar{h}), \forall \bar{h} \in \mathbb{H}(\bar{m})$, \hspace{1cm} (2.14b)

$$\sum_{\bar{h} \in \mathbb{H}} \phi(\bar{h}, i) = r_i, i = 1, 2, \hspace{0.5cm} r_2 = 1 - r_1,$$

$$\phi(\bar{h}, i) \geq 0, \forall \bar{h} \in \mathbb{H}(\bar{m}), i = 1, 2. \hspace{1cm} (2.14c)$$

Let $\phi^*(\bar{h}, .)$ maximize the objective function of (2.14). This value is a fraction of $r_i$; therefore, normalizing for aggregate channel condition $\bar{h}$ results:

$$\hat{\phi}(\bar{h}, i) = \frac{\phi^*(\bar{h}, i)}{\sum_{i=1}^2 \phi^*(\bar{h}, i)}, \forall \bar{h} \in \mathbb{H}(\bar{m}). \hspace{1cm} (2.15)$$

The solution of (2.14) and the subsequent computation of (2.15) results in $\hat{\phi}(\bar{h}, .)$ which are members of a set defined as $\mathcal{F}_{LCS} \doteq \{ \hat{\phi}(\bar{h}, .) : 0 \leq r_1 \leq 1 \text{ and Sol. of (2.14), (2.15) } \}$ where $\mathcal{F}_{LCS}$ is closure of $\hat{\phi}(.)$ for all values of $r_1$ and $\bar{m}$. The resulting capacity region for LCS is expressed as follows:

$$C_{LCS} = \bigcup_{\{ \hat{\phi}(.) \in \mathcal{F}_{LCS} \}} \left( C_1 = E_{\bar{m}} \left[ E_{\bar{h}} \left[ \hat{\phi}(\bar{h}, 1) R(\bar{h}, 1) \right] \right], C_2 = E_{\bar{m}} \left[ E_{\bar{h}} \left[ \hat{\phi}(\bar{h}, 2) R(\bar{h}, 2) \right] \right] \right). \hspace{1cm} (2.16)$$

Next, we extend (2.16) for the optimal MaOS. Like LCS, we need to compute time fraction allocations for a user for every $\bar{h} \in \mathbb{H}(\bar{m})$. However, this computation is different than the
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

problem of (2.14) because the fraction values now depend on the average feasible data rates in aggregate mobility states. Using notation \( \omega(.) \) in place of \( \phi(.) \), the modified LP for every \( \vec{m} \) is:

\[
\begin{align*}
\max & \sum_{\vec{h} \in \mathbb{H}(\vec{m})} \sum_{i=1}^{2} R(\vec{h}, i) \omega(\vec{h}, i), \\
\text{s.t.} & \sum_{i=1}^{2} \omega(\vec{h}, i) \leq P(\vec{h}), \quad \forall \vec{h} \in \mathbb{H}(\vec{m}), \\
& \sum_{\vec{h} \in \mathbb{H}} \omega(\vec{h}, i) = \hat{r}(\vec{m}, i), \\
& \omega(\vec{h}, i) \geq 0 \quad \forall \vec{h} \in \mathbb{H}(\vec{m}), i = 1, 2.
\end{align*}
\] (2.17a)

Comparing (2.17c) with (2.14c) shows that the values of \( \omega(.) \) depend on the solution of (2.6).

Let \( \omega^*(\vec{h}, .) \) maximize the objective function of (2.17), which is a fraction of \( \hat{r}(\vec{m}, .) \). The following normalization computes the time fraction allocation of user \( i \) in channel condition \( \vec{h} \),

\[
\hat{\omega}(\vec{h}, i) = \frac{\omega^*(\vec{h}, i)}{\sum_{i=1}^{2} \omega^*(\vec{h}, i)}, \quad \forall \vec{h} \in \mathbb{H}(\vec{m}).
\] (2.18)

The time fraction allocations \( \hat{\omega}(\vec{h}, .) \) are members of a set \( \mathcal{F}_{MaOS} \triangleq \{ \hat{\omega}(\vec{h}, .) : 0 \leq r_1 \leq 1 \text{ and Sol. of (2.17), (2.18)} \} \). Thus, the resulting capacity region for MaOS is expressed as

\[
C_{MaOS} = \bigcup_{\{\hat{\omega}(\vec{h}, .)\} \in \mathcal{F}_{MaOS}} \left( C_1 = E_{\vec{m}} \left[ E_{\vec{h}} [\hat{\omega}(\vec{h}, 1) R(\vec{h}, 1)] \right], C_2 = E_{\vec{m}} \left[ E_{\vec{h}} [\hat{\omega}(\vec{h}, 2) R(\vec{h}, 2)] \right] \right).
\] (2.19)

The following theorem, proof of which is given in Appendix A, formalizes the relationship of channel capacities of optimum MaOS and LCS algorithms.

**Theorem 1.** The channel capacity of MaOS algorithm (2.19) is greater or equal to the channel capacity of LCS algorithm (2.16), i.e., \( C_{MaOS} \geq C_{LCS} \).
2.5.2 Bounds on Approximate Mobility Assisted Opportunistic Scheduling

Recall that the first stage of the approximate MaOS algorithm distributes time fractions among $N$ users according to the sample paths of their mobility. For this purpose, the algorithm solves LP (2.10) after every $\tau$ mobility time intervals where $\tau$ is the length of the predicted time window. The coefficients of decision variables, $\bar{R}(t + \Delta, \cdot)$, are expected feasible rates for users at (mobility) time $t + \Delta$. These coefficients are random processes which are functions of user mobility and cell characteristics. Thus, the resulting LP (2.10) is a stochastic LP with random rewards.

In contrast with MaOS, an opportunistic scheduling algorithm like LCS does not consider user mobility. Therefore, it does not take advantage of the slow time variations in the radio channel. Although, it considers fast time channel variations in its scheduling decision, it also tries to satisfy long-term fairness constraints. Because user mobility takes place on a slower timescale than scheduling decisions, the LCS algorithm may end up satisfying portions of the fairness requirements on per mobility time slot. This behavior is common in the restrictive constraint regime, that is, when $\epsilon \approx 1$. Thus, the LCS behavior can be approximately modeled as a resource distribution system that assigns $r_i$ fraction of time to user $i$ at every mobility slot. If we like to write an equivalent objective function similar to (2.10) for the LCS model, we will consider coefficients to be constant and equal to the time average of the expected feasible rates. By abuse of notation, we use $E_{\tau}[\bar{R}(\tau, i)]$ to represent the time-average of the expected feasible rate of user $i$ for $\tau$ mobility slots. The resulting objective function values are

$$ u = \tau \sum_{i=1}^{N} E_{\tau}[\bar{R}(\tau, i)]r_i $$

where $u$ gives the lower bound on the objective function of (2.10).

For the approximate MaOS, it is interesting to find the effects of the length of the sample path $\tau$ and the number of users $N$ on the quality of the solution of (2.10). For this purpose, we extend the fundamental inequality of Dyer, Frieze, and McDiarmid (identified here as DFM),
which gives upper bounds on the expected optimal cost of a minimization LP with random coefficients [17], [18]. Our extension allows the determination of lower bounds on the expected optimal value of a maximization problem. In particular, the extension provides a relationship between the LCS model’s objective function (2.20) and the optimal value of (2.10).

In this section, we briefly summarize the DFM result and then present our extension. For notational convenience, the general result relies on the notation of standard form LP as follows:

\[
\begin{align*}
\min \text{ or max} \ & \ u = \sum_{j=1}^{J} c_j x_j, \\
\text{s. t.} \ & \ \sum_{j=1}^{J} a_{kj} x_j = b_k, \quad k = 1, 2, ..., K, \\
& \quad x_j \geq 0, \quad j = 1, 2, ..., J,
\end{align*}
\]  

(2.21a) 

(2.21b) 

(2.21c)

where \(x_j\)'s are optimization variables. The coefficients \(c_j\) are independent nonnegative random variables, but constraint coefficients \(a_{kj}\) and right hand side \(b_k\) are fixed constants. There are \(J\) decision variables and \(K\) equality constraints.

According to the DFM result [17], [18], if it is possible to find \(\beta, 0 < \beta \leq 1\) such that

\[E(c_j \mid c_j \geq l) \geq E(c_j) + \beta l,\]

then, for the minimization LP (2.21), the following inequality gives upper bound on the expected optimal objective function \(E(u^*)\):

\[\beta E(u^*) \leq \sum_{j=1}^{J} \hat{x}_j E(c_j),\]  

(2.22)

where \(\hat{x}_1, \hat{x}_2, ..., \hat{x}_J\) is any fixed feasible solution.

In its present form, the inequality of (2.22) is not suitable for maximization problems because of two reasons. First, it requires positive coefficients. Second, the dual of maximization problem has deterministic coefficients. Therefore, dual problem does not fall within the DFM scope,
which requires random coefficients. The following theorem extends the DFM inequality for the maximization problem (2.21). The proof of this theorem is given in Appendix B.

**Theorem 2:** Suppose $c_j$, $1 \leq j \leq J$, are independent nonnegative random variables in the interval $[0,1]$. Suppose for all $l > 0$ and $P(c_j \leq l) > 0$, we have

$$
E(c_j | c_j \leq l) \leq \beta l.
$$

Let $\hat{x}_1, \hat{x}_2, ..., \hat{x}_J$ be any fixed feasible solution, and $u^*$ be the value of the optimal solution to the maximization LP (2.21), then we have:

$$
\sum_{j \notin B_e} E(c_j)\hat{x}_j + \beta \sum_{j \in B_e} E(c_j)\hat{x}_j \leq \beta E(u^*).
$$

(2.24)

where $B_e$ is feasible bases.

In order to translate the above result to our problem (2.10), we assume that $\bar{R}(\tau, .)$’s are normalized average data rates in the interval $[0,1]$, where, for convenience we use $\bar{R}(\tau, .)$ in place of $\bar{R}(t + \Delta, .)$. Furthermore, we observe that $J = N\tau$, size of $B_e = K$ and $K = N + \tau$. Thus, the first summation on the left side of (2.24) contains $J - K$ terms. Increasing $N$ and $\tau$ results in a linear increase in $K$ and a quadratic increase in $J$. Furthermore, $\beta < 1$; therefore, the relative contribution of the second sum on the left side of (2.24) decreases when $N$ and $\tau$ increase. Thus, we can ignore this term at high values of $N$ and $\tau$. The remaining first sum on the left side can be approximated to the lower bound of (2.20). Moreover, if we assume that $\bar{R}(\tau, i)$ are i.i.d. and $\epsilon = \sum_{i=1}^{N} \tau_i = 1$, then (2.24) reduces to

$$
E_\tau[\bar{R}(\tau, .)] \leq \beta E(u^*)/\tau.
$$

(2.25)

As an example, we consider a special case when $\bar{R}(\tau, .)$ are uniform i.i.d. random variables in the interval $[0,1]$, then the $E_\tau[\bar{R}(\tau, .)] = 1/2$, and $\beta = 1/2$. It means that in order to satisfy (2.25), the optimal value per mobility slot of (2.10), which turns out to be $E(u^*)/\tau$, should approach 1.
The numerical results validate the above argument. Fig. 2.2 plots the expected optimal value per mobility slot of (2.10) when the coefficients of the decision variables are uniform i.i.d. random variables. By increasing $\tau$ and $N$, the expected optimal value per mobility slot approaches the ideal value of 1. This means that, if a large number of mobile users ($N$) are sharing a channel and a sufficiently large window size ($\tau$) is employed, then approximate MaOS can achieve performance close to the maximum possible data rate.

In the real case, $\bar{R}(\tau,.)$ would not be uniform i.i.d. random variables. However, for any general distribution, it is possible to find a $\beta$ that satisfies (2.23). Thus, (2.25) can still be used to provide a lower bound on the optimum solution.

## 2.6 Implementation Details

This section provides implementation details of the MaOS algorithms. It first identifies some of the candidate mobility estimation and prediction techniques from the literature. Next, it describes a stochastic approximation technique for the estimation of Lagrange multipliers.

### 2.6.1 Mobility Estimation and Prediction

The optimal MaOS algorithm only needs to estimate the current mobility state of every user. It is possible to correctly estimate user mobility state because the mobility model considers a large area within a cell as a state. A user anywhere in that area is considered to be in that state. Global Positioning System (GPS), or monitoring the respective radio signal strength, and alternative technologies like angle of arrival (AoA) and time delay of arrival (TDoA) may be used for location estimation [27]. Furthermore, the presence of two timescales helps in filtering out errors in the estimation process.

The approximate MaOS algorithm also needs to predict the mobility state in the future. We propose following alternative techniques.

1. Straight line movement witnessed on highways mostly follows deterministic mobility with
constant speed and velocity [28]. Therefore, the future trajectory can be accurately predicted.

2. For stochastic user mobility, the approximate MaOS algorithm can employ a hierarchical mobility model. The higher layer of this model will have discrete-state space like the one presented in this chapter (Section 2.2.1). The lower layer of the model will have a continuous-state space modeled by some linear dynamical system. This continuous-state space model will help to monitor not only the state but also fine-grained location information in the form of coordinates, velocity, and acceleration [29],[30]. The fine-grained mobility information will help in predicting the future discrete states to be visited by a user. For example, [29] showed fairly accurate prediction of the next state several seconds ahead of time. With the help of fingerprinting of the road network, the future prediction of mobility states can be further improved.

It is expected that the estimation and prediction of user mobility results in communication and processing overhead. The communication overhead could be minimal if system relies on the radio signal strength based mobility estimation and prediction schemes because the radio signal strength is already being measured by the system for channel state information. Therefore, the information gathered for the channel state and used for finding the appropriate modulation and coding scheme could also be used in the estimation of user mobility. The processing overhead will be mostly handled by the base station where sufficient processing power could be easily available because, unlike mobile station, base station’s processing is not limited by the battery capacity.

2.6.2 Estimation of Lagrange Multipliers

Like LCS, the optimal and approximate MaOS algorithms also use a stochastic approximation technique for estimation of Lagrange multipliers. The optimal MaOS sets the initial value of $v(.) = 0 \forall i$ and $\bar{m}$ and maintains a database for every mobility state $\bar{m}$, where it holds the most
recent estimated values of \( v(\cdot) \) and supporting parameters used in its computation. Whenever the system enters into mobility state \( \vec{m} \), the algorithm retrieves the stored values from the database. As long as the system remains in state \( \vec{m} \), it updates the values of \( v(\cdot) \) for every scheduling slot \( k + 1 \) using the following computation [4]:

\[
v_{i}^{k+1}(\vec{m}) = \max \left( v_{i}^{k}(\vec{m}) - \delta \left( 1_{\{Q^{k}(\vec{R}(\vec{m})=i}\}} - \hat{r}(\vec{m}, i) \right), 0 \right). \tag{2.26}
\]

We use a small constant \( \delta \) as the step size because the corresponding LCS algorithm, being indifferent to the user mobility information, views the piecewise stationary channel as a nonstationary process. For nonstationary processes, the stochastic approximation technique recommends a constant step size [4]. The presence of a database to store the values of \( v(\cdot) \) and supporting parameters for future use and the difference in timescales between mobility transitions and scheduling decisions guarantees the saturation of (2.26).

The approximate MaOS cannot store the values of \( v(\cdot) \) into a database because it lacks the complete knowledge of the state space. Therefore, whenever a new state is visited, the algorithm resets \( v(\cdot) \) and the supporting parameters. Next it uses an expression similar to (2.26) for subsequent scheduling and mobility slots until next mobility transition. The timescale difference between mobility transitions and scheduling may allow approximate MaOS to achieve saturation.

### 2.7 Numerical Results of Optimal MaOS

In this section, we report our simulation results for optimal MaOS algorithms for the HDR data set. First, we report implementation details for the simulation setup, followed by the performance and the channel capacity results for the MaOS and LCS algorithms. We also compare these algorithms with the MR scheduling that always selects the best user.
2.7.1 Basic HDR Simulation Setup

A hexagonal cell that has the maximum coverage distance of 5 km from the central BS to a mobile user forms the service area. The service area is divided into two concentric rings \((m = \{1, 2\})\). The innermost ring covers an area up to 3 km from the BS. The BS’s maximum output power is 15 W, and 80 percent of this power is used for the shared data channel [31]. The system provides HDR data service with the feasible data rates of \(\{2,457.6, 1,843.2, 1,228.8, 921.6, 614.4, 307.2, 204.8, 153.6, 76.8, 38.4\}\) Kbps, and the corresponding SINRs of \(\{9.5, 7.2, 3.0, 1.3, -1.0, -4.0, -5.7, -6.5, -8.5, -9.5, -12.5\}\) dB [1].

The average SINR \(\rho(m)\) at the center of the rings is found from the path loss model of (2.3) for \(\alpha = 4, y_0 = 1\) km and \(\eta = 1\) W. The steady-state channel distribution is determined using (2.5) for a speed of 60 km/h. The resulting expected feasible rate values are 1,438 Kbps in the inner ring and 116 Kbps in the outer ring. The scheduling decisions are made at every 1.67 ms.

For the two concentric rings, the coarse-grained mobility model is learned through the simulation of a user following a fine-grained mobility model. For our simulation, we considered random-walk mobility [21] with a constant speed of 60 km/h to model fine-grained user movement. Travel intervals are randomly distributed between 8 to 12 minutes. After completing one travel interval, the user selects a new direction randomly from 0 to \(2\pi\). When the user moves out of the service area, it wraps around and reenters the service area. Mobility slots are of 1s duration. We emphasize that the choice of identical mobility and channel distributions is only for the ease of simulation. The proposed algorithm can handle nonidentical cases equally well.

The following results are for eight users in the system with infinite backlog. Half of the users require (minimum) fairness requirements of \(\epsilon/6\) and other half expect at least \(\epsilon/12\) fraction. In (2.26), we set \(\delta = 0.05\).
2.7.2 Performance Comparison

Fig. 2.3(a) plots the system data rate (normalized with respect to the maximum achievable data rate of 2,457.6 Kbps) for the LCS and MaOS algorithms for $\epsilon = \{0.99, 0.95, 0.90\}$. For completeness, this figure also reports the performance of the MR algorithm, which is a greedy algorithm without any regard to fairness. Because, here, we are concentrating on performance, we set $\theta = 0$ in (2.6).

The results show that MaOS performs better than LCS because it exploits not only the fast time fading fluctuations but also the slow time path loss variations in the radio channel. First, through the solution of (2.6), MaOS finds optimum user priorities, in the form of time fractions that act as fairness constraints, to take advantage of the slow time path loss fluctuations. These priorities are such that users with strong channels, that is, those in the inner ring at any given time, are preferred over weak users, that is, those in the outer ring at the same time. Next, it applies these priorities opportunistically through the solution of (2.8) to gain from the fast time fading variations. This overscheduling of strong users allows MaOS to avoid them when they transition to weak mobility states. Similarly, weak users at any given time do not remain weak indefinitely. They move closer to the BS, and that is when MaOS compensates them for their loss. The net result is that all users get substantially data rates higher than or equal to LCS as shown in Fig. 2.3(b). This figure compares the average data rates achieved by all eight users under the MR, LCS and MaOS algorithms. Because the first four users have higher fairness requirements, their data rates are higher than the remaining four users.

On the other hand, LCS suffers performance losses in order to support constant fairness constraints. These constant fairness constraints force the LCS algorithm to schedule users when they are in the unfavorable location. Although the channel information is considered in scheduling, but this information only allows it to benefit from the fast time fading while slow time path loss variations remain untapped. Thus, LCS ends up satisfying fairness constraints for users when they have weak channels.
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

The performance improvements seen in case of MaOS have been achieved without sacrificing the long-term fairness requirements, as shown in Fig. 2.3(c). This figure compares the temporal fairness requirements for all users to their actual allocations performed by the scheduling algorithms. The MR, being a greedy algorithm, failed to support the required fairness measure. It gave equal access to all users, not because of any conscious decision on the part of the algorithm, but because of identical mobility and channel distributions used in our simulation. The other algorithms considered in the comparison, namely, MaOS and LCS, are able to satisfy the minimum fairness requirements of $r_{1,4} = \epsilon/6$ and $r_{5,8} = \epsilon/12$.

Returning to the Fig. 2.3(a), we observe that increasing the value of $\epsilon$ significantly increases the performance gap between MaOS and LCS. Although the MaOS performance does not change significantly, the LCS performance decreases with the increase of $\epsilon$. We observed that, for $\epsilon$ values of 0.90, 0.95 and 0.99, MaOS performs better than the LCS algorithm by a margin of 11.4, 16.8 and 23.2 percent, respectively. The performance improvement of MaOS in comparison with LCS is expected to be significant in the restrictive constraint regime when $\epsilon$ is close to 1, for example, $\epsilon = 0.99$. According to [5], in the restrictive constraint regime, the LCS algorithm has less opportunity to improve the system’s performance. This loss of opportunity is due to the fact that almost all constraints are binding for all time durations; therefore, the scheduler cannot over schedule any user. When a user has weak channel for long duration of time, the binding constraint forces the scheduling of that user to satisfy its fairness requirement. On the other hand, MaOS handles such a user by scheduling it more than its long-term share when it is in favorable locations. This extra scheduling saves MaOS from scheduling that user when it is in unfavorable locations resulting in the performance improvement.

When $\epsilon$ decreases, fewer constraints are binding, and LCS can over schedule strong users more than their required share. Thus, when these users are in unfavorable locations, LCS algorithm can avoid them for a longer duration of time. This is the reason for the reduction in the performance gap between the MaOS and LCS algorithms when $\epsilon$ decreases.
2.7.3 Delay Comparison

In the previous section, MaOS performed significantly better than LCS algorithm, but the improvement is at the cost of adverse delay statistics. The results presented there are for $\theta = 0$. For these results, the first batch of users, that is, users belonging to the set $\{1, 2, 3, 4\}$, experienced a head-of-line (HOL) packet delay of 8.59 s and a variance of 235680 s$^2$. Here, the HOL delay is defined as the time gap between two consecutive allocations. The LCS algorithm faced almost equal expected delay of 8.69 s, but its variance was only 34728 s$^2$. The expected HOL delay is indifferent to both algorithms because of the temporal fairness measure that requires a certain minimum fraction of time for users.

The delay variance is significantly high for MaOS algorithm. This high variance is because of the way MaOS distributes user priorities. When $\theta = 0$, MaOS completely avoids users in the unfavorable locations as long as there are other users in the favorable locations. Therefore, users in unfavorable locations can be starved of service. To overcome long-term starvation, the proposed algorithm provides a mechanism of ensuring minimum resource allocation to users even when they are in the most unfavorable locations so that their connection remains active and higher layer applications can perform reasonably. The mechanism is to use a nonzero value of $\theta$ in (2.6). When $\theta > 0$, MaOS lower bounds user preferences, (that is, fairness), by the constraint (2.6d). The lower bound considerably decreases the MaOS delay variance, as shown in Table 2.2 for different values of $\theta$. For example, even when $\theta$ is very small, like 0.1, the delay variance decreases to 2366 s$^2$. This value is even better than the corresponding LCS value. By increasing $\theta$, we can reduce delay variance for MaOS. This pattern is observed for the second set of users, $\{5, 6, 7, 8\}$, as well. For the data set considered in Table 2.2, $\theta_{\text{max}}$ is 0.4.

The decrease in MaOS delay variance is due to the way it computes (2.26) and uses Lagrange multipliers in (2.9). The algorithm associates Lagrange multipliers to the aggregate mobility state $\bar{m}$. Therefore, it is able to comply with the minimum fairness bounds for $\theta > 0$ case for every mobility state resulting in lower variance values. On the other hand, the LCS algorithm
does not associate corresponding Lagrange multipliers to the mobility states. It computes and uses them for the whole time horizon. Though it is able to provide asymptotic fairness, it does not guarantee strict fairness on a much reduced timescale like mobility state. If LCS is also forced to comply with the fairness for every mobility state \( \vec{m} \), then we identify the resulting algorithm as LCS mobility state fair (MSF). For this compliance, LCS MSF associates Lagrange multipliers with \( \vec{m} \) but retains the static constraints of LCS. With this modification, LCS MSF provides the best delay variance among the competition (see Table 2.2). For completeness, we also report the corresponding values for the MR algorithm. The MR algorithm does not differentiate users according to their fairness requirements. Therefore, all users experienced almost comparable values.

The decrease in delay variance achieved by MaOS for \( \theta > 0 \) and LCS MSF are not without loss. The loss is in the form of reduction in the average system data rate. Figure 2.4 shows the average data rates for MR, LCS, LCS MSF and MaOS. Recall that \( \theta \) affects only MaOS algorithm.

### 2.7.4 Effects of Parameters

In another experiment, we change system parameters to observe their impact on the performance of the proposed scheme. In this new experiment, every cell has a radius of 2000 m and \( y_0 = 500 \) m. The mobility model consists of three mobility states of concentric rings. The ring radii in increasing order are 800, 1400 and 2000 m. For fine-grain mobility model, users drive in straight lines with constant speeds. There were 6 users in the system with fairness requirements of \( r_{1,...,3} = \epsilon/9 \) and \( r_{4,...,6} = \epsilon/4.5 \). The other model parameters are identical to those reported in Section 2.7.1.

The average data rate of each algorithm in normalized units is as follows: MR (0.82), LCS (0.56), MaOS (0.78 for \( \theta = 0 \)), LCS MSF (0.44). MaOS is able to significantly improve performance in comparison to the LCS algorithm. In general, we have observed that, for macro-
cellular structures where free-space path loss variations significantly affect the channel quality
and users experience those variations, MaOS can provide increased performance.

2.7.5 Capacity Gain

Using the mobility and channel parameters in Section 2.7.1, we plot the capacity region for two
users using (2.13), (2.16) and (2.19) in Fig. 2.5. This figure also contains simulated capacity
regions generated by the LCS and optimal MaOS algorithms for $\epsilon = 0.99$ and $\theta = 0$. The
simulated and numerical results validate that MaOS provides a larger capacity region than the
LCS algorithm.

2.8 Numerical Results of Approximate MaOS

This section provides simulation results for the approximate MaOS algorithm. After describing
the simulation setup, it reports performance comparisons and carries out robustness analysis
of the algorithm.

2.8.1 Extended HDR Simulation Setup

For the performance analysis of the optimal MaOS algorithm, a two or three-state mobility
model was considered because of the exponential complexity in the number of states. However,
for the approximate MaOS, we extend the model to a multiple-state model. The extended
model has several mobility states in accordance with the average SINR values of the HDR data
set.

We extend the topology shown in Fig. 2.1(a) with multiple concentric rings. In the extended
model, each ring corresponds to a SINR level in the HDR data set; the ring order, from the
innermost to the outermost, follows the descending order of the SINR levels. As there are eleven
SINR levels in the HDR data set (Section 2.7.1), if the cell radius $Y$ permits, then there can
be eleven possible rings on the cell surface. A user at a distance $y$ from the BS, where $y \leq Y$,
will fall in ring $m$ if

\[
\begin{cases}
    \rho(1) \leq \frac{P_t \Gamma(y)}{\eta + I(y)} & m = 1, \\
    \rho(m) \leq \frac{P_t \Gamma(y)}{\eta + I(y)} < \rho(m - 1) & m = 2, ..., 11.
\end{cases}
\] (2.27)

For the eleven SINR levels in the HDR system, we compute the corresponding outer radii using (2.3),(2.4), and (2.27). The outer radii in increasing order are \{1, 070, 1, 222, 1, 556, 1, 715, 1, 957, 2, 325, 2, 562, 2, 682, 3, 005, 3, 181, 3, 764\} meters. These values are found for the following system parameters: $y_0 = 1$ km, $\alpha = 4$, $Y = 4$ km, $\eta = 1$ W, $P_t = 15$ W and $P = 0.8P_t$ [31]. The steady state probability of channel conditions are found according to (2.5).

There are 12 users in the cell, and they follow random walk mobility model [21]. The mobility model parameters are identical to the one given in Section 2.7.1 except user speed. For one experiment users move with a speed of 60 km/h, and for the other experiment, they follow a speed of 30 km/h. The fairness requirements for each user are $\epsilon/12$ where $\epsilon = 0.99$.

### 2.8.2 Performance Comparison and Effects of Future Sample Path

Fig. 2.6 plots the system data rate for the MR, MaOS and LCS algorithms as a function of $\tau$ in (2.10). Except MaOS, all the other algorithms are independent of $\tau$. From this figure, it is clear that an increase in the sample path size ($\tau$) improves the data rate of the approximate MaOS algorithm. This is in line with the theoretical result of Section 2.5.2.

However, for small values of $\tau$, approximate MaOS can perform worse than LCS. For small $\tau$, users do not cover long distances within the duration of $\tau$, therefore, distance change does not significantly affects the data rate. Thus, the chances of performance improvement by MaOS are limited in this case. On top of that, because of short time duration, the algorithm makes several quick changes in the constraints of (2.11). These changes make it difficult for the approximate MaOS to find true Lagrange multipliers. Therefore, in order to satisfy fairness, the algorithm ends up reducing overall performance. Although, if the algorithm knows Lagrange multiplier values before hand than it can recover some of this loss. For example, Fig. 2.6 also plots the
MaOS algorithm performance with known (estimated) values of Lagrange multipliers. These values were computed in an earlier simulation for the exact user mobility and channel statistics. With known multiplier values, approximate MaOS achieved comparable performance to LCS when \( \tau \approx 120 \text{ s} \), which is better than the original value of \( \tau \approx 180 \text{ s} \). With the increase of \( \tau \), MaOS performs better because larger time window allows it to take advantage of the path loss variations in the feasible rates. The LCS and MR algorithms are independent of the sample path size.

The increase in speed from 30 km/h to 60 km/h increases the performance of all algorithms because frequent changes in radio channel provide more opportunity to improve the data rate.

### 2.8.3 Robustness of Approximate MaOS Algorithm

Future prediction of mobility states is crucial for the success of the MaOS algorithm. In this section, we simulate the effects of prediction errors on the performance of the algorithm. We study the approximate MaOS for the simulation setup of Section 2.8.1 under two scenarios.

The first scenario assumes that errors are independent and only restricted to the adjacent rings. In this scenario, predictor makes an error with some probability \( \gamma \). When it makes an error, it erroneously predicts one of the adjacent rings with equal probability. For situations where there is only one adjacent ring, that is, the correct ring is either the innermost or the outermost ring, the predictor selects the only adjacent ring.

The second scenario assumes that the errors are independent but once an error is made with probability \( \gamma \), the predictor can select any ring with equal probability. This is an extreme scenario as most of the prediction errors in real situations occur close to the actual location.

The MaOS performance is shown in Fig. 2.7 for both scenarios as a function of the probability of error, \( \gamma \), for future prediction of 800s. When \( \gamma = 0 \), the predictor is ideal with no errors. From this figure, we see that for the first case, MaOS experiences graceful performance loss as \( \gamma \) increases. Even when all the rings are erroneously replaced by their neighbors, that is, \( \gamma = 1 \),
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

the MaOS algorithm achieved a performance of 0.51 which is higher than the LCS algorithm. The adjacent cell errors do not cause serious dent in the performance of the algorithm because the expected data rates of adjacent rings do not differ significantly. For the second scenario, the performance loss is considerably higher because of the extreme types of errors. Even in this case, performance loss is graceful, and it remains higher than the corresponding LCS for $\gamma$ as high as 0.5, that is, when half of the rings are incorrectly predicted and replaced by any other ring with equal probability.

The prediction errors do not result in fairness violations, as temporal fairness measure considers fraction of time assignment which remains independent of errors.

2.9 Conclusion

This chapter proposes algorithms that combine channel and mobility information in the scheduling of downlink data packets in cellular networks. The proposed MaOS algorithms improve on the LCS algorithm in [5] by dynamically adapting long-term fairness constraints according to the mobility information of a user. The MaOS algorithms compute constraint values such that the scheduler gives preference to users that are in the most favorable locations with high expected feasible rates. This extra scheduling helps the scheduler to avoid such users when they enter mobility states with bad channels. The optimal MaOS algorithm precomputes constraint values for all states according to a discrete-state-space model. The approximate MaOS algorithm computes constraint values according to the future prediction of mobility states. With the help of the extended DFM inequality, we prove that lengthening the future prediction and increasing the number of users in the network can achieve maximum possible performance. The extended inequality is also capable of handling more general cases and provides lower bounds on the performance of the approximate MaOS. Simulation results show that the proposed algorithms perform better than LCS and satisfy the fairness constraints. Moreover, it is observed that the improvement is significant in the restrictive constraint regime. The improvement is
observed without increasing the long-term expected delay, though delay variance increases. The proposed algorithm provides a mechanism to control the value of delay variance. The use of mobility information in opportunistic scheduling also increases channel capacity. The performance improvements are at the expense of complex algorithms that require additional information in the form of mobility state estimation and prediction.
<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>number of users</td>
</tr>
<tr>
<td>$Q$</td>
<td>scheduling policy</td>
</tr>
<tr>
<td>$R_i$</td>
<td>feasible rate (FR) for $i$ &amp; $\bar{R}$ is FR vector</td>
</tr>
<tr>
<td>$1_{{Q(.)=i}}$</td>
<td>indicator function (=1 if $i$ selected)</td>
</tr>
<tr>
<td>$E(R_{Q(\bar{R})})$</td>
<td>average system data rate</td>
</tr>
<tr>
<td>$P_i{Q(.)=i}$</td>
<td>probability of scheduling user $i$</td>
</tr>
<tr>
<td>$E(R_i1_{{.}})$</td>
<td>average scheduled data rate of user $i$</td>
</tr>
<tr>
<td>$\alpha, \theta, \beta, \delta$</td>
<td>constants, see (2.3), (2.6), (2.23), and (2.26) respectively</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>probability of prediction error</td>
</tr>
<tr>
<td>$r_i, \epsilon'$</td>
<td>time fraction allocation, $\epsilon' := \sum_{i=1}^{N} r_i$</td>
</tr>
<tr>
<td>$v_*'$</td>
<td>Lagrange multiplier</td>
</tr>
<tr>
<td>$\mathcal{S}, m$</td>
<td>set of mobility states , $m \in \mathcal{S}$</td>
</tr>
<tr>
<td>$\mathcal{S}, \bar{m}$</td>
<td>stationary distribution of vector $\bar{m}$ and state $m$</td>
</tr>
<tr>
<td>$\mathcal{Z}(\cdot)$</td>
<td>SINR</td>
</tr>
<tr>
<td>$\rho(m)$</td>
<td>mean SINR in mobility state $m$</td>
</tr>
<tr>
<td>$\Gamma(m)$</td>
<td>path loss in mobility state $m \in \mathcal{S}$</td>
</tr>
<tr>
<td>$P_t, P_I$</td>
<td>transmit and interference power respectively</td>
</tr>
<tr>
<td>$I(m)$</td>
<td>interference in mobility state $m \in \mathcal{S}$</td>
</tr>
<tr>
<td>$Y$</td>
<td>cell radius</td>
</tr>
<tr>
<td>$\mathcal{H}$</td>
<td>ordered SINR levels</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>fading process</td>
</tr>
<tr>
<td>$p_Z(.)$</td>
<td>distribution of SINR</td>
</tr>
<tr>
<td>$p_r(.)$</td>
<td>steady state distribution of channel</td>
</tr>
<tr>
<td>$\bar{R}(\cdot)$</td>
<td>expected feasible rate for a user</td>
</tr>
<tr>
<td>$r^*(\cdot)$</td>
<td>optimal sol. of (2.6)</td>
</tr>
<tr>
<td>$\hat{r}(\cdot)$</td>
<td>normalized time fractions (2.7)</td>
</tr>
<tr>
<td>$\tau, \Delta$</td>
<td>mobility time window &amp; $\Delta = 1, ..., \tau$</td>
</tr>
<tr>
<td>$\mathbb{H}$</td>
<td>aggregate channel conditions $\mathbb{H} = \mathcal{H}_1 \times \mathcal{H}_2$</td>
</tr>
<tr>
<td>$\bar{h}$</td>
<td>aggregate channel condition vector $\bar{h} \in \mathbb{H}$</td>
</tr>
<tr>
<td>$E_{\bar{m}[,]}, E_{\bar{h}[,]}$</td>
<td>expectation wrt to $\bar{m}$ and $\bar{h}$ respectively</td>
</tr>
<tr>
<td>$C_x^*$</td>
<td>ergodic capacity where $x = {\text{TD, LCS, MaOS}}$</td>
</tr>
<tr>
<td>$\phi_x^<em>(\cdot), \omega_x^</em>(\cdot)$</td>
<td>optimal sol. of (2.14) &amp; (2.17) respectively</td>
</tr>
<tr>
<td>$\phi_\cdot(\cdot), \hat{\omega}(\cdot)$</td>
<td>normalized fractions, see (2.15) and (2.18)</td>
</tr>
<tr>
<td>$E_{\tau}[R(.)]$</td>
<td>time-average of expected feasible rate</td>
</tr>
<tr>
<td>$u^*$</td>
<td>optimum value of LP (2.21)</td>
</tr>
<tr>
<td>$c_j, x_j$</td>
<td>coefficients &amp; decision variables of (2.21)</td>
</tr>
<tr>
<td>$\hat{x}_j$</td>
<td>any feasible solution of (2.21)</td>
</tr>
<tr>
<td>$\mathbb{B}_e$</td>
<td>feasible bases of (2.21)</td>
</tr>
<tr>
<td>$K, J$</td>
<td>number of constraints and variables in (2.21)</td>
</tr>
</tbody>
</table>
### Table 2.2: Comparison of Delay Statistics

<table>
<thead>
<tr>
<th>Users</th>
<th>Expected Delay (s)</th>
<th>Delay Variance (s²)</th>
<th>Delay Variance (s²) of MaOS as a function of θ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LCS</td>
<td>MaOS</td>
<td>LCS MSF</td>
</tr>
<tr>
<td>1 → 4</td>
<td>8.69</td>
<td>8.59</td>
<td>8.43</td>
</tr>
<tr>
<td>5 → 8</td>
<td>17.10</td>
<td>17.44</td>
<td>18.03</td>
</tr>
</tbody>
</table>
Chapter 2. Mobility Assisted Opportunistic Scheduling in Cellular Networks

Figure 2.1: Discrete state space mobility model and network structure: (a) three concentric rings, (b) five topological spaces, and (c) hexagonal cell structure with BS separation of $2Y$. The rings or spaces are characterized by average SINR which depends on $y_m$ (i.e., distance between BS and mobility state $m$).

Figure 2.2: Effects of future $\tau$ and number of users $N$ on the optimal solution of (2.10) when coefficients of the LP’s objective function are nonnegative uniformly distributed independent and identical random variables in the interval $[0, 1]$. 
Figure 2.3: Comparison of MaOS with LCS and MR: (a) average system data rate achieved for different values of $\epsilon = \{0.99, 0.95, 0.90\}$, (b) individual data rate for $\epsilon = 0.99$ and (c) fairness.
Figure 2.4: MaOS performance as a function of $\theta$ along with MR, LCS and LCS MSF performances.
Figure 2.5: Two user channel capacity using numerical and (where applicable) simulations for TD, LCS and optimum MaOS.
Figure 2.6: Effect of $\tau$ on the performance of approximate MaOS, and its comparison with LCS and MR schemes.
Figure 2.7: Effect of prediction errors on approximate MaOS performance. Two error scenarios are considered; first scenario restricts errors to adjacent states, and second scenario allows incorrect prediction of any state.
Bibliography


Chapter 3

Throughput Constrained
Opportunistic Scheduling

3.1 Introduction

Recently, several opportunistic scheduling algorithms have been proposed for sharing time-slotted downlink channels among multiple users in cellular data networks ([1], [2], [3] and refs. therein). With the help of channel information at the base station (BS), and adaptive modulation and coding, these networks adapt each individual user’s feasible transmission data rate in accordance with the respective wireless channel condition; a better channel condition results in a higher data rate, and vice versa. Examples of such networks are HDR (High Data Rate) [4] of cdma2000 1xEV-DO, and HSDPA (High Speed Downlink Packet Access) [5] of 3GPP. In these systems, a scheduler resides at the BS, and only one user can access the shared channel at a given time-slot. Fig. 1.1 shows the model of such a network. For these networks, opportunistic scheduling improves performance because it schedules a time-slot to a user when the users performance (i.e., data rate) is among the highest.

Liu, Chong and Shroff have proposed an opportunistic scheduling framework for infinite

---

Chapter 3. Throughput Constrained Opportunistic Scheduling

backlog traffic demand [1],[2]. The infinite backlog model considers a static user population and infinite size queues. The proposed framework maximizes the overall performance of the system, sum of the expected throughput, under certain minimum quality of service (QoS) requirements for individual users. These requirements are formulated, as temporal, utilitarian, or minimum-performance guarantee constraints. The first two types of formulations offer relative fairness constraints where a minimum required fraction of resource, e.g., time for temporal fairness or system performance for utilitarian fairness, is enforced for each user. The third formulation enforces absolute requirements on an individual users performance. In this formulation, the scheduler satisfies the minimum data rate requirement for every user. This formulation offers direct service guarantees to users. In the subsequent discussion, we consider the minimum-performance guarantee (MPG) formulation, and identify the corresponding solution in [2] as the MPG solution. This solution scales the instantaneous data rate of every user to compensate for violation of the minimum performance constraint and selects the user with the maximum scaled instantaneous data rate in every slot. By properly setting the scaling parameters, the resulting algorithm gives the optimal performance while satisfying the minimum-performance constraints.

When minimum-performance constraints for all users are satisfied, the MPG solution behaves as a greedy algorithm and selects the user with the best channel condition. There could be situations when the best user may not actually use the assigned opportunity since the user may not have much traffic to receive from base station or the user may face buffer overflow at the receiver side. The greedy behavior could result in data rate variations among the same class of users. For example, if one user has a very strong channel while other users in the same class have weak channels most of the time, then the strong user will have considerably higher data rate than the rest. Furthermore, the actual data rates that users achieve depend on the presence of other users in the network.

Because of the above reasons, we extend the opportunistic scheduling framework with additional constraints that upper-bound the data rates of individual users. We retain the objective
function and lower bounds on the individual data rates as in [2] where the objective function maximizes the overall performance of the system which is the sum of the expected individual data rates. The resulting formulation maximizes the system rate and constrains individual rates within respective minimum and maximum bounds. The proposed solution technique is identified as Throughput Constrained Opportunistic Scheduling (TCOS).

The maximum-rate constraints, first introduced in [3], will lower the overall system performance; therefore, their presence at first seems counterintuitive. However, there are several reasons why these constraints are useful. We extend some of the reasons given in [3] as follows:

1. The upper bound will be an incentive for the users of a (cheap) low-rate service to upgrade to an (expensive) high-rate service. Thus, they provide a tool for system operators to increase their revenue.

2. Without the upper bounds, the QoS given to a user will be greatly influenced by the number of users in the network. When only one user is active, then all the system data rate is available to that user. This data rate will decrease as more and more users become active. Thus, there will be considerable variations in the QoS among different users. The upper bound on data rate will decrease this variation.

3. The presence of a performance upper bound allows better service differentiation between different classes of users. As we will see in Section 3.5.3, the proposed TCOS solution provides better isolation among different classes of users than the MPG solution.

4. As seen in Section 3.5.3, the upper bounds provide a feasible method of reducing the average head-of-line delay experienced by users in an opportunistic scheduling systems with infinite backlog traffic demand. Without upper bounds, the only method available to the system operator is to increase the minimum performance for users. However, increasing minimum performance may result in an infeasible solution or a decrease in the accommodation capacity of the system.
Chapter 3. Throughput Constrained Opportunistic Scheduling

The problem of maximizing data rate subject to minimum and maximum constraints has been considered previously in [3]. The resulting algorithm that solves the same problem we are considering in this chapter is identified as *Maximum Throughput with Minimum/Maximum Rates* (MTMR) in [3]. Like MPG, this algorithm also uses a scaling term with instantaneous data rate for every user. The scaling term is an exponential term with a controlling parameter called the token counter, which has a positive drift when the respective user receives a lower rate than the minimum bound and a negative drift when the user receives a higher rate than the maximum bound. As we will see in Section 3.5, though the MTMR algorithm is optimal, its performance is inferior to TCOS under more realistic radio channel conditions.

The proposed TCOS scheme has improved features compared to the MTMR scheme of [3]. Firstly, TCOS retains the generalized framework of [2]. Thus, the solution presented here can be easily extended to the other fairness measures in the framework, namely, temporal and utilitarian fairness. Secondly, the proposed solution is independent of the wireless channel characteristics. It is likewise suitable for channels with or without memory. Moreover, the proposed scheme is equally useful for non-stationary wireless channels.

We generalize the framework for opportunistic scheduling with both minimum-rate and maximum-rate constraints. The maximum-rate constraints make the optimization problem difficult to solve in real-time. As a computationally inexpensive solution method, we propose a very efficient iterative heuristic based on Lagrangian relaxation. The proposed solution uses linear multipliers to control otherwise greedy behavior of the algorithm and constrains the solution within the feasible region. The significant aspect of the solution is its capability to trade-off feasibility with throughput. It relaxes the feasible region marginally in order to significantly improve performance. Using simulations under different channel conditions, we compare the proposed TCOS solution with MTMR and MPG solutions.

Furthermore, we show the usefulness of upper bounds on performance through simulations. The upper-bounds provide improved service differentiation between different classes of users. For the problem considered in this chapter, the upper bounds also provide a feasible method of
reducing average packet delay.

The remainder of this chapter is organized as follows. In Section 3.2, we present the problem formulation and review the MPG and MTMR solution schemes. Section 3.3 presents the proposed TCOS scheme. Following our evaluation methodology in Section 3.4, simulation results of the proposed scheme in comparison with MTMR and MPG schemes are given in Section 3.5. Section 3.6 covers related models and scheduling schemes found in the literature and Section 3.7 concludes this chapter.

### 3.2 Problem Formulation

In this section, we present the MPG problem and related solution method. We formulate a generalized form of the MPG problem and discuss the MTMR solution method to that problem.

Opportunistic scheduling with MPG works as follows. Let $C^l_i$ be the performance lower bound, i.e., the minimum long-term data rate requirement, and let $N$ be the number of users in the cell. Further assume that $\bar{R} = (R_1, \cdots, R_N)$ is a feasible rate vector at a generic scheduling time-slot where $R_i$ is a feasible rate at which user $i$ can be served if scheduled in the time-slot. This feasible rate vector is known to the BS based on the channel state feedback for all the users in the cell (see Fig. 1.1). The MPG solution solves the following optimization problem:

\[
(P_1) := \max_Q E \left( R_{Q(\bar{R})} \right), \\
\text{s.t. } E \left( R_i 1_{\{Q(\bar{R}) = i\}} \right) \geq C^l_i, \quad i = 1, 2, \cdots, N,
\]  

where $Q$ is a scheduling policy that satisfies the long-term performance requirements, $E \left( R_i 1_{\{Q(\bar{R}) = i\}} \right)$ is the average scheduled data rate of user $i$ using policy $Q$, $1_{\{\cdot\}}$ is the indicator function, and $C^l_i \geq 0$. The overall system data rate is $E \left( R_{Q(\bar{R})} \right) = \sum_{i=1}^N E \left( R_i 1_{\{Q(\bar{R}) = i\}} \right)$.

The following policy $Q^*$ is optimal for MPG [2]:

\[
Q^*(\bar{R}) = \arg\max_i (\alpha^*_i R_i),
\]  

69
Chapter 3. Throughput Constrained Opportunistic Scheduling

where \( \alpha_i^* \) is a scaling parameter that is chosen so that \( \min(\alpha_i^*) = 1 \) and \( E \left( R_i^1 \mathbf{1}_{Q^*(\vec{R})=i} \right) \geq C_i^l \) for any user \( i \), and if \( E \left( R_i^1 \mathbf{1}_{Q^*(\vec{R})=i} \right) > C_i^l \), then \( \alpha_i^* = 1 \). The parameters \( \alpha^* \)'s scale the instantaneous feasible rates of the users.

In this chapter, we generalize the above framework to the case where scheduled data rates are also upper-bounded by the maximum-rate constraints. The maximum-rate constraints for opportunistic scheduling are introduced in [3]. More specifically, we consider the problem \((P_1)\) with additional constraints as follows:

\[
(P_2) := \max_{Q} \quad E \left( R_{Q(\vec{R})} \right), \tag{3.3a}
\]

\[
\text{s.t.} \quad E \left( R_i^1 \mathbf{1}_{Q(\vec{R})=i} \right) \geq C_i^l, \quad i = 1, 2, \cdots, N, \tag{3.3b}
\]

\[
E \left( R_i^1 \mathbf{1}_{Q(\vec{R})=i} \right) \leq C_i^u, \quad i = 1, 2, \cdots, N, \tag{3.3c}
\]

where \( C_i^u \) is the performance upper bound, i.e., the maximum long-term data rate requirement for user \( i \), and \( C_i^u > C_i^l, \forall i \). Problem \((P_2)\) appears similar to problem \((P_1)\) though there are substantial difference in term of the solution complexity. The difference is basically caused by introducing maximum-rate constraints, which are in nature functionally different from minimum-rate constraints.

As stated earlier, the MTMR algorithm optimally solves problem \((P_2)\). It serves the flows according to the following rule,

\[
Q^*(\vec{R}) = \arg\max_i \quad (e^{a_i T_i} R_i), \tag{3.4}
\]

where \( T_i \) is a token counter for user \( i \), and \( a_i > 0 \) is a parameter. The solution updates the token counter at the \((k+1)^{th}\) slot as follows:

\[
T_i(k+1) = T_i(k) + TR_i - R_i^1 \mathbf{1}_{Q(\vec{R})=i}, \tag{3.5}
\]

where \( TR_i = C_i^l \) if \( T_i(k) \geq 0 \) and \( TR_i = C_i^u \) if \( T_i(k) < 0 \); \( R_i^1 \mathbf{1}_{Q(\vec{R})=i} = R_i \) if user \( i \) is scheduled in time slot \( k \) and \( R_i^1 \mathbf{1}_{Q(\vec{R})=i} = 0 \) otherwise. The token counter has a positive drift when
flow $i$ receives less than $C_i^l$ service rate and a negative drift when flow $i$ receives more than $C_i^u$ service rate. If the scheduled rate of user $i$ falls within $C_i^l$ and $C_i^u$ bounds, then $T_i$ stays close to 0 [3]. Results reported in [3] validate that the MTMR is able to strictly enforce the lower and upper bounds on the user data rates.

### 3.3 Proposed Solution

#### 3.3.1 Properties of the Optimal Solution

In this section we solve ($P_2$) by the proposed TCOS algorithm. Before describing the proposed solution, we will like to provide some insights on the significance of the constraints in ($P_2$) and their treatment in the proposed solution. The lower bound constraints of (3.3b) are critical for the connection quality. Several applications require some minimum data rate to perform reasonably. However, the upper bound constraints of (3.3c) are a tool in the hands of system operator to tune the level of service. They provide increased isolation between users, fairer distribution of resources and a way to increase revenues. However, some violation of these bounds do not result in connection loss or service disruption. The only drawback of such an event is the allocation of some radio resources to users which are not entitled for that allocation. Therefore, we will like to propose a flexible solution that achieves lower bounds and enforces upper bounds with occasional violation. Thus, the proposed solution provides necessary tool to the system operator to achieve service differentiation and fairness without losing substantial system performance by strictly enforcing upper bounds.

For the proposed solution, we assume that the system is capable of providing a feasible solution, i.e., $C_i^l \leq E \left( R_i 1_{(Q(\bar{R})=i)} \right), \forall i$. Then, the following policy $Q^*$ is optimal for problem ($P_2$):

$$Q^*(\bar{R}) = \arg \max_i \left( (\alpha_i^* - \gamma_i^*) R_i \right),$$

where $\alpha_i^*$ and $\gamma_i^*$ are chosen so that:
1. \( \min_i (\alpha_i^*) = 1, \min_i (\gamma_i^*) = 0, \) and \( \max_i (\gamma_i^*) = 1, \)

2. \( E\left( R_i 1_{\{Q_i^* (\vec{R}) = i\}} \right) \geq C_i^l \) for all \( i, \)

3. \( E\left( R_i 1_{\{Q_i^* (\vec{R}) = i\}} \right) \leq C_i^u \) for all \( i, \)

4. For any user \( i, \) if \( E\left( R_i 1_{\{Q_i^* (\vec{R}) = i\}} \right) > C_i^l, \) then \( \alpha_i^* = 1, \) and, if \( E\left( R_i 1_{\{Q_i^* (\vec{R}) = i\}} \right) < C_i^u, \)
   then \( \gamma_i^* = 0. \)

In the following proposition, we prove the optimality of policy (3.6).

**Proposition 1**: The policy \( Q^* \) defined in (3.6) is the optimal solution to the problem \((P_2)\). It maximizes the system performance and satisfies the long-term minimum and maximum performance constraints.

**Proof**: See Appendix for proof.

In (3.6), \( \gamma_i^* \)'s and \( \alpha_i^* \)'s are scaling parameters for the maximum-rate and minimum-rate constraints, respectively. These parameters provide linear control to otherwise greedy algorithm. If user \( i \) is unlucky and experiences a weak wireless channel, then \( \alpha_i^* > 1, \) which scales the performance score in (3.6), causing more resources to be allocated to user \( i. \) Thus, its data rate will improve. Once the user satisfies the minimum-rate constraint, then the corresponding \( \alpha_i^* \) will be 1, and no further scaling is needed.

As stated earlier, \( \gamma_i^* \) compensates for any violation of the maximum-rate bound. For users whose scheduled data rates are less than the maximum-rate bound, the corresponding \( \gamma_i^* \) values will be equal to 0. As a result, the associated constraints, which are non-binding constraints, do not play any role in resource allocation for the associated users. On the other hand, if the wireless channel of user \( i \) is in such a good condition that the scheduled data rate is higher than \( C_i^u, \) then \( \gamma_i^* \) is increased such that it approaches 1, which scales down the corresponding performance value in (3.6) for that user; therefore, the user will be scheduled less often. When the user has a very strong wireless channel and the scheduled rate is likely to remain higher than the upper bound, then a value of 1 is set for the corresponding \( \gamma_i^* \). For such a case,
the corresponding $\alpha_i^*$ would already be 1 because $C_i^u > C_i^l$. Thus, according to (3.6), the performance score of that user will be zero; therefore, the user will not be considered for scheduling until the performance score is returned to a value lower than the maximum-rate constraint.

### 3.3.2 Solution Procedure

We use the standard Lagrange multiplier method to solve problem $(P_2)$. The method typically includes two steps: forming the Lagrangian dual function and iterative search for improving the dual solution. In the case of linear programming problems, there is no duality gap, which is the difference between the primal optimal solution and the dual optimal solution. However, in the case that the problem is non-linear, we often have non-zero duality gap. In other words, it is often the case that the dual solution is not a feasible solution to the primal problem. To resolve this kind of matter, the dual solution is used to find the associated primal solution. By doing this, we can obtain an upper bound (the best objective value of the dual solution found so far in the procedure) and a lower bound (the best objective value of the primal solution found) even if we do not yet reach the exact optimal solution. These two bounds can be used as the performance index how close our solution approaches the true optimum so far.

The size of duality gap depends on the type of problem and is unpredictable. In this chapter, we manipulate the standard Lagrange multiplier method to trade-off computational efficiency and a little bit of violation of the constraints of the primal problem. We show by extensive simulation in a later section that the violation occurs only for the maximum performance constraints. Note that the violation of minimum performance constraints is critical and should not occur since it can drop ongoing connections whereas violation of maximum performance constraints may cause a little more resources to be allocated to some users and is not so critical. We also show by experiments that the violations very rarely occur and the degree of violation is very small.
Chapter 3. Throughput Constrained Opportunistic Scheduling

The true values of $\alpha_i^*$ and $\gamma_i^*$ can be found using the technique of stochastic approximation considered in [1],[2]. Our algorithm employs the following rules to update these parameters in scheduling slot $k + 1$ based on the parameter values in slot $k$,

$$
\alpha_i^{k+1} = \max \left[ \alpha_i^k - \epsilon \left\{ E \left( R_i \mathbf{1}_{(Q(R_i)=i)} \right) - C_i^u \right\}, 1 \right]
$$

$$
\gamma_i^{k+1} = \max \left[ \min \left\{ \gamma_i^k - \epsilon' \left\{ C_i^u - E \left( R_i \mathbf{1}_{(Q(R_i)=i)} \right) \right\}, 0 \right\}, 1 \right]
$$

where $\epsilon$ and $\epsilon'$ are small positive real numbers that act as the step sizes for learning the true multipliers for the minimum and maximum constraints, respectively. Generally, a decreasing step size is recommended for stationary cases in order to facilitate the saturation of the estimated values. For non-stationary cases, constant step sizes are more appropriate [2]. In this work, we recommend constant values for $\epsilon'$ because of their association with the maximum-rate constraint. We elaborate on the benefits for a constant $\epsilon'$ in the following.

We are maximizing the overall system performance while restricting the maximum performance of each user; therefore, for some users, it is required that the upper-bound scaling parameters, $\gamma_i$’s, be flexible. These users will have high data rates close to the upper bound. For example, when user $i$’s data rate increases to $C_i^u$, then the corresponding $\gamma_i$ should approach 1. This value of $\gamma_i$ would reduce user $i$’s chance of being scheduled and decrease its data rate. The decrease in data rate should lower $\gamma_i$ so that user $i$ regains a fair chance of being selected for scheduling. Therefore, a flexible algorithm is required in order to make $\gamma$ values adaptable to the changing conditions. This adaptive behavior is only possible with a fixed step size. An added advantage of a fixed step size is that our algorithm is also suitable for non-stationary wireless channel conditions [2].

For $\epsilon$, we also recommend a small positive value in order to make the TCOS algorithm adaptable to changing channel and network behaviors. Otherwise, it becomes difficult to find a decreasing $\epsilon$ that maintains a balance between adapting and saturating the algorithm. The balancing act depends on several factors including the number of users in the network and
their channel statistics. Therefore, an engineering solution is to use a small positive value to avoid the difficulty in tuning for an appropriate value, and to make the proposed solution more suitable for non-stationary channels.

### 3.4 Performance Evaluation Methodology

In this section, we present our simulation methodology. We employ HDR data rates and signal-to-noise ratio (SNR) levels for the radio channel states, and consider discrete-state Rayleigh fading channel models.

#### 3.4.1 HDR Simulation Setup

Like [3], we consider the feasible data rates of HDR: \{2457.6, 1843.2, 1228.8, 921.6, 614.4, 307.2, 153.6, 76.8, 38.4\} Kbps with corresponding SNR threshold: \{9.5, 7.2, 3.0, 1.3, -1.0, -4.0, -6.5, -9.5, -12.5\} dB. We simplify our implementation by ignoring the following two features of the HDR system: contiguous slot assignments at low data rates, and the use of forward error-correcting (FEC) codes. Our implementation independently selects a user at every time slot according to the policy, and we ignore the effects of FEC codes on the system performance. As found in [3], both of these features have insignificant effects on the results.

#### 3.4.2 Channel Models

We consider two types of discrete-state Rayleigh fading channel models: finite state memoryless and finite state Markov channels (FSMC) [6]. The FSMC model is more realistic as that channel has memory that helps in modeling deep fade situations. This model is a popular model used in the literature for flat fading channels [7]. For both types of channel models, the steady state distribution of SNR (Z) is assumed to be exponential with average \( \rho \); for \( z \geq 0 \), it is given as [8]:

\[
p_Z(z) = \frac{1}{\rho} e^{-\frac{z}{\rho}}.
\]  

(3.8)
In our simulation setup, the average SNR, $\rho$, is a user-dependent parameter that is uniformly distributed in the interval $(0, 5]$. This random selection of $\rho$ emulates variable locations and fading conditions for the users.

Let $\mathcal{H} = \{Z_0, Z_1, \cdots, Z_{H-1}\}$ be the set of discrete signal levels in ascending order. The fading channel is in state $h$ if the received signal is in the interval $[Z_h, Z_{h+1})$. The steady state probability of state $h$ is given as [6]:

$$p(h) = \int_{Z_h}^{Z_{h+1}} p_Z(z) dz = e^{-\frac{Z_h}{\rho}} - e^{-\frac{Z_{h+1}}{\rho}}.$$  \hfill (3.9)

The resulting expected values of the feasible data rates, $E(R_i)$, for each user in our simulation setup, are shown in Fig. 3.1. These values are obtained from the respective expectations of the input channel process for each user.

For the FSMC model, we also compute a transition probability matrix. Here, we assume that the SNR remains constant during a time-slot, only immediate neighboring states are directly accessible from a state when making transitions between time-slots, and mobile users move with a speed of 60 Km/hr. The transition probabilities for Rayleigh fading are computed as follows [6],[7].

Let $S_h$ be the number of transmitted symbols per second for fading state $h$. $\bar{S} = \sum_h S_h p(h)$ is the average rate of transmission in symbols per second. Using this overall average, we can compute the average number of symbols transmitted during the time-slots when the channel is in state $h$, i.e., $\bar{S}(h) = \bar{S} p(h)$. The adjacent transition probabilities are approximated as:

$$t_{h,h+1} \approx \frac{N_h}{\bar{S}(h)}, \quad h = 0, 1, \cdots, H - 2$$ \hfill (3.10)

and

$$t_{h,h-1} \approx \frac{N_h}{\bar{S}(h)}, \quad h = 1, 2, \cdots, H - 1,$$ \hfill (3.11)

where $N_h, h = 1, 2, \cdots, H - 1$, is the expected number of times per second the received SNR passes downward across the threshold $Z_h$. $N_h$ is computed as follows:

$$N_h = \sqrt{\frac{2\pi Z_h}{\rho} f_m e^{-\frac{Z_h}{\rho}}}$$
where \( f_m \) denotes the mobility-induced Doppler spread. The self-loop probabilities are found as:

\[
t_{0,0} = 1 - t_{0,1}, \quad t_{H-1,H-1} = 1 - t_{H-1,H-2}
\]

and

\[
t_{h,h} = 1 - t_{h,h-1} - t_{h,h+1}, \quad h = 1, 2, \ldots, H - 2.
\]

### 3.4.3 Evaluation Parameters

The following describes the parameters used in our performance evaluations and comparisons. We obtain the cumulative distribution function (CDF) of the average data rates among the users at the end of the simulation. We also compare the simple average of all users rates, i.e.,

\[
\bar{R} = \frac{1}{N} \sum_{i=1}^{N} E(R_i).
\]

The third parameter for comparison is the average head-of-line (HOL) packet delay, which is defined as \( \bar{D} = \frac{1}{N} \sum_{i=1}^{N} (1/\theta_i) \), where \( \theta_i \) is the fraction of time that user \( i \) is allocated the channel. Generally, \( \bar{D} \) follows \( O(N) \) [9]. No analytical method can be found to compute the fraction of time allocation for problems (\( P_1 \)) and (\( P_2 \)). Numerical computation is the only method of estimating \( \theta \) under a simplified assumption of continuous data rates as in [10]. We also compare the control terms of the TCOS and MTMR algorithms. These control terms are the multipliers of \( R_i \) in (3.4) and (3.6), i.e., \( e^{a_iT_i} \) and \( \alpha_i^* - \gamma_i^* \), respectively.

For all the algorithms considered in the performance comparisons, time is measured in slots. Since the data rate considered are that for HDR system, therefore, a slot translates into 1.67 msec. For TCOS and MPG, values of \( C_i^L \)'s, \( C_i^u \)'s, and \( E(R_i1_{(Q_i(R)\approx i)}) \) are measured in normalized bits per slot, where normalization is with respect to the maximum data rate of 2457.6Kbps. Thus, the normalized data rate related values have a range between 0 to 1. We convert the final results back to the true data rates (in Kbps) for ease of comparisons. The values of \( \epsilon \) and \( \epsilon' \) in (3.7) are 0.02 and 0.1, respectively. For the MTMR algorithm, as in [3], all the rate related terms are measured in bits per slot. The value of \( a_i \) in (3.4) is \( 6.25 \times 10^{-5} \) for all users [3].
Like in [3], for simulation we also set the lower bounds for data rate equal to $1.2 \times C_l^i$.

### 3.5 Numerical Results

In Section 3.5.1, we compare the MTMR and TCOS algorithms without upper bound constraints, first under memoryless channels and then under FSMC. Next, we compare these two algorithms when upper bound constraints are applied and results are summarized in Section 3.5.2. In Section 3.5.3, we compare the MPG with TCOS and MTMR algorithms to show the significant role of upper bounds for service differentiation among different classes of users.

#### 3.5.1 Achieving Minimum Rates

Our first experiment is to compare the MTMR and TCOS algorithms without upper bound constraints. It means that the resulting algorithms solve problem $(P_1)$. Since we are not considering upper bounds on data rates in the first experiment, the TCOS algorithm (3.6) behaves as the MPG algorithm (3.2) because $\gamma_i^* = 0 \ \forall i$ and for all times. Both algorithms are compared first under memoryless channels and later under FSMC. The minimum bound for all users are set as $C_l^i = 9.6$ Kbps. We run traces of both algorithms on identical users and channel realizations for 1200 seconds.

Fig. 3.2 plots the CDF of the average data rate of individual users on a logarithmic scale for 10, 20, 30, 40, 50 users under memoryless channels. It is observed that both algorithms are able to satisfy the lower performance constraints for all user populations. Increasing the number of users decreases individual user rates. The resulting average data rates of both algorithms are almost equal. The corresponding rates for the user populations given above are $\{188, 106, 75, 54, 43\}$ Kbps, respectively. However, TCOS exhibits a higher HOL average delays than MTMR: $\{23.9, 56.0, 62.9, 82.6, 92.6\}$ for TCOS against $\{21.6, 48.8, 54.7, 70.2, 77.2\}$ for MTMR, for the respective user populations. The reason for the increased delay is given later in this subsection.
Fig. 3.3 reports the results of the above experiment under FSMC. It is observed that TCOS performs better than MTMR. The reason for its better performance is that TCOS (3.6) uses a linear control term to enforce bounds. Therefore, when a user is in a deep fade and its scheduled data rate is less than the minimum bound, then the user’s score gradually rises. The gradual rise does not force the TCOS to immediately serve such a user. However, MTMR (3.4) uses an exponential control term in its decision rule; therefore, the score of such a user rises exponentially and forces MTMR to immediately schedule that user. Thus, MTMR ends up scheduling users more often in their deep fade conditions. This results in MTMR giving lower average data rates - more than 200% lower than TCOS for 40 and 50 users (see Fig. 3.3(b)). However, this immediate scheduling results in lower HOL delays for MTMR as shown in Fig. 3.3(c)). The difference in delay decreases as the number of users increases in the system.

### 3.5.2 Achieving Minimum and Maximum Rates

In this sub-section we compare TCOS and MTMR algorithms with minimum and maximum-rate constraints under FSMC. The results presented in the previous section show that both algorithms are almost identical in performance for memoryless channels. Therefore, we do not consider memoryless channels in this experiment and concentrate only on the more realistic FSMC model. The minimum and maximum bounds for all users are set as $C^l_i = 9.6$ Kbps, and $C^u_i = 70.0$ Kbps.

Fig. 3.4(a) plots the cumulative distribution function (CDF) of the scheduled data rates for both algorithms on a logarithmic scale for 10, 20, 30, 40 and 50 users. From this plot, it is clear that the scheduled data rates of almost all users fall within the limits of 9.6 and 70.0Kbps. There is a small probability that the TCOS algorithm violates the performance upper bound. This violation can be attributed to the fact that the multiplier algorithm (3.7) may produce an infeasible solution to the primal problem, which is resolved by employing some projection methods [11]. Since the violation occurs very rarely, we don’t have any projection procedure
in the solution method to gain more computational efficiency. For example, for 50 user case reported in Fig. 3.4(a), only 2 users violated 70 Kbps upper bound and the maximum violation was not more than 1.2 Kbps. In addition, it is tolerable to allow a small occurrence of infeasible solutions that violate the upper bound only, for resource allocations in practical systems.

Figs. 3.4(a) and (b) present that TCOS performs better than MTMR in terms of the average data rates. For example, MTMR achieved data rates of \{59, 42, 28, 15, 11\} Kbps for configurations with 10, 20, 30, 40, 50 users. For the corresponding user configurations, TCOS achieves \{70, 65, 58, 48, 35\} Kbps, corresponding to improvements of \{18, 54, 30, 220, 218\}% over MTMR. These improvements are at the expense of slightly higher HOL delays, as shown in Fig. 3.4(c). The reason for the performance improvement is due to the linear control term in the decision rule that allows the algorithm to skip users in deep fade situations.

To clarify the difference in behavior of TCOS compared to that of MTMR, we consider the behaviors of two users, a weak and a strong user, during the simulations. For this comparison, we select user 9 and user 16. From Fig. 3.1, it is clear that user 9 has a weak channel and user 16 has a strong channel. Fig. 3.5 plots average data rates and control terms, i.e., the $R_i$ multipliers in (3.4) and (3.6), over simulation runs of 1200 seconds under FSMC, with lower and upper bounds of $C_i^L = 9.6$ and $C_i^U = 38.4$ Kbps respectively. From Fig. 3.4(a), it is clear that for the MTMR algorithm no user have been able to achieve the upper bound of 70 Kbps. Therefore, we lower upper bound $C_i^U$ from 70.0 Kbps to 38.4 Kbps so that the upper bound is also binding for the MTMR algorithm and we can compare the behavior of both algorithms for two users, one weak user who is barely able to achieve the lower bound and the other strong user who is restricted by the algorithms from violating the upper bound. According to Fig. 3.5(a), the TCOS data rates have more oscillations around the bounds than the MTMR rates. This behavior is due to the linear control term $(\alpha_i^* - \gamma_i^*)$ used in the TCOS algorithm (3.6), which helps to take advantage of the high channel states. Thus, high data rates are achieved with less number of assignments. For example, the spikes in the data rate of user 16 increase its rate slightly more than the upper bound in the beginning. The corresponding control term
gradually tapers these spikes, and in this process, for more than 60% of the time, the control term is such that user 16 is not considered in the scheduling process, i.e., $\alpha_{16}^{\ast} - \gamma_{16}^{\ast} = 0$ (see Fig. 3.5(b)). This allows TCOS to concentrate on other users and increase the overall system performance. Likewise for user 9, initially the resulting data rate becomes lower than the lower bound. However gradually, it recovers and remains higher than the bound. The corresponding control term increases when lower bounds are violated and gradually smoothes out as bounds are satisfied (see Fig. 3.5(b)). This results in an overall improved data rate albeit at a cost of higher HOL delays. On the other hand, MTMR maintains steady state data rates close to the respective bounds for both users, even though in order to maintain these rates, the corresponding control multipliers ($e^{\alpha_{i}T_{i}}$) oscillate and make sudden transitions (see Fig. 3.5(b)). This results in overall lower data rates for MTMR but improved HOL delays as reported earlier.

As stated earlier, one of the motivations for placing upper bounds on performance is to reduce average HOL delay. Comparing the HOL performance of TCOS with and without upper bounds in Fig. 3.3(c) and Fig. 3.4(c) respectively, we observe that the average delay decreases when TCOS has upper bounds. This reduction is due to the fact that with upper bounds some radio resource (e.g., time in TDMA systems) is saved from greedy allocations to the best users and allocated more evenly among other users. The reduction in average HOL delay improves overall QoS for users and is another reason for placing an upper bound on the individual performance.

### 3.5.3 Service Differentiation: MPG vs TCOS and MTMR

The objective of this experiment is to provide further justifications for the application of upper bounds on user performance. We will show that the use of performance upper bounds allows better isolations between different classes of users, by comparing the TCOS and MTMR algorithms, which allow upper and lower bounds, with the MPG algorithm, which only allows lower bounds. The main comparison is between MPG and TCOS algorithms while the results
for MTMR are presented for completeness sake. The comparison is performed under memoryless channel condition so that MTMR can also satisfy upper bounds for user performance. Furthermore, under memoryless channel, the algorithm performances are independent of the underlying channel models. Thus, the resulting comparisons helps us in understanding the affects of upper bounds on service differentiation.

We consider 40 users in the network. These users are evenly divided into two classes: class A for high rate users and class B for low rate users. For class A users, the minimum- and maximum-rate bounds are $C^l_i = 19.2$ Kbps and $C^u_i = 80.0$ Kbps, respectively, and for class B users, these bounds are $C^l_i = 9.6$ Kbps and $C^u_i = 38.4$ Kbps, respectively.

In Fig. 3.6, we plot the CDFs of scheduled data rates for all three algorithms and for both classes of users. Table 3.1 lists the corresponding average data rates and standard deviation of rates achieved by the algorithms. As expected, MPG achieves a higher system performance, and some users in both classes are able to get data rates exceeding 80.0 Kbps. This behavior is due to the fact that MPG does not have an upper bound on the performance. However, the average data rate of class A users is only 53.1 Kbps. On the other hand, the same class of users are able to get an average data rate of 66.7 Kbps with the TCOS algorithm. TCOS is able to provide this higher data rate because it does not provide excess service to class B users when they have achieved their performance upper bound. Thus, TCOS spends more time in serving class A users. Furthermore, with TCOS, class B users and their wireless channel conditions do not significantly affect the data rates of class A users.

In order to show that the presence of performance upper bounds also reduces variations in the QoS, we compare between these algorithms the standard deviations in data rates among users in the same class. Class A users exhibit a deviation of 22.8 Kbps when served by TCOS, and a deviation of 38.5 Kbps when served by the MPG algorithm. Thus, TCOS results in reduced variations in QoS. A similar reduction in data rate deviation is seen for class B users as well.

As expected for memoryless channels, MTMR performs almost similar to the TCOS. There-
fore, almost identical CDF plots to that of TCOS are achieved by MTMR for both classes of users. This behavior is also evident from the results in Table 3.1 and in line with the results of Section 3.5.1.

In order to see how the separation between the lower and upper bounds affects service differentiation for the TCOS and MTMR algorithm, we further perform the above experiment with different bounds. The resulting CDFs of data rates for these algorithms are shown in Fig. 3.7. Here, class A users have a wider separation between the lower and upper bounds than that of class B users. Class A users are bounded within 19.2 and 80.0 Kbps while class B users have bounds of 9.6 and 19.2 Kbps. Since the lower bounds of both classes are the same as the previous experiment, MPG remains independent of these changes and performs exactly identical to Fig. 3.6. TCOS and MTMR on the other hand provide a more stringent service differentiation as evident from the average data rate and standard deviation statistics given in Table 3.1. This shows that TCOS and MTMR are able to provide a higher average data rate to class A users with a lower standard deviation compared to the MPG algorithm.

Table 3.1 also reports the average HOL delay experienced by both classes of users for all the three algorithms. The experimental results in Fig. 3.6 present that TCOS achieves lower delays of 34.4 and 85.0 slots respectively for class A and class B users while MPG achieves an average delay of 58.3 and 93.4 slots. This result demonstrates that we can improve HOL delay for opportunistic scheduling systems by using upper bounds. If the MPG algorithm needs to lower delay for class A users then it has to increase the lower bounds of these users as far as the system is not infeasible.

Continuing with the results in Table 3.1, we see that TCOS provides a flexible HOL delay trade-off between high-rate and low-rate service users. Thus TCOS achieves an improved average delay for class A users by lowering the upper bounds of class B users in Fig. 3.7. This change results in an improved HOL delay of 28.1 slots for class A users. For memoryless channels, MTMR achieves comparable results to TCOS.
3.6 Related Work

In this chapter, we focus on improving the system data rate; therefore, we have selected an infinite backlog queue model that assumes all queues are saturated. The advantage of the infinite backlog model lies in the fact that it shields the scheduling process from the arrival process dynamics. Additionally, this model is appropriate for elastic traffic applications like best-effort file download or web browsing. However, delay-constrained applications like multimedia download often consider another model in which arrival processes feed queues. For this model, one class of popular scheduling schemes are based on max-Weight scheduling rule [12]. These schemes aim to establish better control over delay distributions. Therefore, they combine queue length [13], or head-of-line packet delay [14], [15], [16] with the channel state information in their decision rules. These solutions are stable in the sense that they bound queue lengths or delays if arrival processes fall within the performance region of the system. In [17], the authors considered a more general case and allow some of the queues to be unstable. More recently, [18] generalized these schemes for situations where imperfect queue length information is available.

Alternatively, proportional fair scheduling (PF) [19], [20], to a lesser degree, is useful for delay-sensitive applications. The PF scheme is independent of the arrival process dynamics. It maximizes the product of the expected data rates which is mathematically equivalent to \( \max \sum_{i=1}^{N} \log (E(R_i | Q(R_i) = 1)) \) [21]. This objective function favors weak users over strong users. The proportional fairness is different from the fairness considered in this chapter. PF reduces delay variance among users, but in comparison to the class of algorithms that maximize the sum of the data rates PF achieves overall lower data rates [1], [3].

3.7 Conclusion

In this chapter, we have presented an opportunistic scheduling algorithm that supports upper and lower bounds on individual user performance. The proposed algorithm, identified as
throughput constrained opportunistic scheduling (TCOS), allows better isolation among different classes of users. It also reduces variations in the quality of service experienced by the users, so that the presence of other users in the network do not adversely affect the quality of service provided to existing users. Furthermore, the presence of upper bounds also help in reducing average delay when compared to opportunistic scheduling without upper bounds. In comparison with MTMR, an existing algorithm that serves similar objectives, TCOS performed equally well under memoryless channel conditions. For channels with memory like FSMC, results show that TCOS performs better than the MTMR algorithm. The improvement in performance comes due to the ability of TCOS to trade-off feasibility with performance. In general, MTMR yields a better delay performance and more steady data rates. Thus, MTMR is suitable for services demanding better delay and delay jitter requirements, such as real-time conversational and streaming services. On the other hand, TCOS generally achieves better data rates under realistic channel conditions, at the expense of increased delays. Thus TCOS is more suitable for applications that are more delay tolerant, like web browsing and file transfer.

Acknowledgment

The author thanks Dr. M. Andrews for useful communication on the MTMR algorithm.
Table 3.1: MPG vs. TCOS and MTMR: Average Rate, Standard Deviation of Scheduled Data Rates (in Kbps) and Average HOL Delay (in slots)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Attributes</th>
<th>Class A Users</th>
<th>Class B Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPG</td>
<td>Avg. Rate</td>
<td>53.1</td>
<td>47.8</td>
</tr>
<tr>
<td></td>
<td>Std. Dev</td>
<td>38.6</td>
<td>49.3</td>
</tr>
<tr>
<td></td>
<td>Avg. HOL Delay</td>
<td>58.3</td>
<td>93.4</td>
</tr>
<tr>
<td>TCOS (Fig. 3.6)</td>
<td>Avg. Rate</td>
<td>66.7</td>
<td>28.4</td>
</tr>
<tr>
<td></td>
<td>Std. Dev</td>
<td>22.8</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td>Avg. HOL Delay</td>
<td>34.4</td>
<td>85.0</td>
</tr>
<tr>
<td>TCOS (Fig. 3.7)</td>
<td>Avg. Rate</td>
<td>71.6</td>
<td>17.9</td>
</tr>
<tr>
<td></td>
<td>Std. Dev</td>
<td>18.2</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Avg. HOL Delay</td>
<td>28.1</td>
<td>110.7</td>
</tr>
<tr>
<td>MTMR (Fig. 3.6)</td>
<td>Avg. Rate</td>
<td>64.5</td>
<td>28.9</td>
</tr>
<tr>
<td></td>
<td>Std. Dev</td>
<td>20.3</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>Avg. HOL Delay</td>
<td>33.1</td>
<td>77.7</td>
</tr>
<tr>
<td>MTMR (Fig. 3.7)</td>
<td>Avg. Rate</td>
<td>71.7</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>Std. Dev</td>
<td>18.2</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Avg. HOL Delay</td>
<td>27.9</td>
<td>111.2</td>
</tr>
</tbody>
</table>
Figure 3.1: The average feasible data rate, $E(R_i)$, values for $i = 1, ..., 50$. Recall that $R_i$ is the feasible rate at which user $i$ can be served if scheduled in the time-slot.
Figure 3.2: CDF of scheduled rate for TCOS and MTMR algorithms under memoryless channel for 10, 20, 30, 40 and 50 users. Throughput is lower bounded by $C_i^l = 9.6$ Kbps.
Figure 3.3: TCOS and MTMR algorithms under Finite State Markov Channel for 10, 20, 30, 40 and 50 users. Throughput is lower bounded by $C_i^l = 9.6$Kbps. (a) CDF of scheduled rate (b) average data rate $\bar{R}$ (c) HOL average delay $\bar{D}$. 
Figure 3.4: Comparison of TCOS and MTMR under finite state Markov channel for 10, 20, 30, 40 and 50 users. Throughput is bounded within $C^l_i = 9.6$ Kbps and $C^u_i = 70.0$ Kbps. (a) CDF of scheduled rate (b) average data rate $\bar{R}$ (c) HOL average delay $\bar{D}$. 
Figure 3.5: Average data rate (a) and corresponding $R_i$ multipliers values (b) for two users. Throughput is bounded within $C_i^l = 9.6$Kbps and $C_i^u = 38.4$Kbps.
Chapter 3. Throughput Constrained Opportunistic Scheduling

Figure 3.6: Service differentiation TCOS vs MPG: Class A user data rates are bounded within $C^l_i = 19.2\text{Kbps}$ and $C^u_i = 80.0\text{Kbps}$. Class B user data rates have lower bound of $C^l_i = 9.6\text{Kbps}$ and upper bound of $C^u_i = 38.4\text{Kbps}$. 
Figure 3.7: Service differentiation TCOS vs MPG: Class A user data rates are bounded within $C^l_i = 19.2$Kbps and $C^u_i = 80.0$Kbps. Class B user data rates have lower bound of $C^l_i = 9.6$Kbps and upper bound of $C^u_i = 19.2$Kbps.
Bibliography


Chapter 4

Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks\(^5\)

4.1 Introduction

Orthogonal frequency division multiplexing (OFDM) is a popular multi-carrier modulation scheme. It provides immunity to intersymbol interference and frequency selective fading. It gets this immunity as it divides the frequency band into a group of mutually orthogonal subcarriers, each having a much lower bandwidth than the coherence bandwidth of the channel \([1]\). In multi-user environment, multiple access of OFDM can be achieved by employing a Time Division Multiple Access (OFDM-TDMA) or Code Division Multiple Access (OFDM-CDMA). For both of these multiple access schemes, a user transmits over the entire spectrum which leads to the

lowering of performance as the user may experience deep fades and narrow band interference [2].

As an alternative, OFDM subcarriers can be time and frequency multiplexed among multiusers. Orthogonal frequency division multiple access (OFDMA) allows this type of multiplexing. It forms traffic channels which are based on one or a cluster of OFDM subcarriers. Every user is assigned a subset of traffic channels at any given time. Thus, OFDMA provides a degree of freedom by allowing dynamic assignment of channels/subcarriers\(^6\) to different users at different time instances, to take advantage of the fact that at any time instance, channel responses are different for different users at different subcarrier frequencies [3]. These features have resulted in the adoption of OFDM/OFDMA technology for high-speed wireless communication systems. For example, digital video broadcasting (DVB-T), wireless LAN (IEEE 802.11a), and fixed and mobile wireless metropolitan area networks (IEEE 802.16 and 802.16e) [4], [5]. OFDMA is also the preferred downlink physical layer specification for the 3GPP’s Evolved Universal Terrestrial Radio Access (E-UTRA) [6].

Dynamic subcarrier assignments (DSA) to multiple users can improve the system data rate of an OFDMA system [7]. This improvement is due to the multiuser diversity gain as the channel characteristics for different users are independent of one another. Thus, the subcarriers under deep fade for one user at a given time may not be in deep fade for other users. Therefore, each subcarrier can have good channel response for some users in a multiuser environment. In systems where adaptive modulation and coding (AMC) techniques are employed, better channel response results in higher data rates. Furthermore, it is expected that adaptive power allocation (APA) can further improve the system data rate. However, in this chapter, we only consider the downlink DSA problem in the multicell OFDMA networks.

The problem of DSA for single cell OFDMA systems has been studied in detail. However, DSA for multicell OFDMA environments with frequency reuse has not been investigated extensively. The existing solutions for traditional cellular networks based on TDMA or CDMA

\(^6\) The terms subcarriers and channels are interchangeably used in this chapter.
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

technologies cannot be applied to the OFDMA networks due to several reasons, some of which are listed in [2]. First, the existing solutions employ binary decision variables and fixed *signal-to-interference-noise ratio* (SINR) thresholds to decide the allocation of a channel to a cell. These traditional schemes are suitable for homogeneous applications like voice. Employing AMC results in variable data rates for modern applications; therefore, fixed SINR thresholds based schemes are not valid. Second, channels are frequency selective and the data rate requirements of modern applications are variable. Third, finding an optimum solution is difficult because an OFDM system has a large number of channels and for multicell networks the problem size increases manifolds. Fourth, the measurement and control overhead involved in gathering the interference information in a central scheduler like radio network controller (RNC) is huge. Last, a completely distributed scheme will not be able to handle unevenly distributed load.

Authors in [2] address the multicell DSA with frequency reuse in OFDMA networks. Their solution is a semi-distributed resource allocation where allocation decisions are divided between the radio network controller (RNC) and the base station (BS). The RNC algorithm runs at the super-frame level, coordinates the inter-cell interference, and assigns subcarriers to BSs. It also recommends user for every subcarrier. The BS algorithm runs at the frame level (user packet level) and performs subcarrier allocation according to the channel and traffic information. It is possible that the BS does not follow RNC recommendations. According to [2], in this case RNC algorithm does not remain optimal for that frame. Furthermore, when all users in a BS have traffic to send then the BS algorithm follows RNC recommendations. In this case, the BS has reduced capability of exploiting multiuser diversity gain. The solution presented in [8] adds fairness and does not require instantaneous channel knowledge at the super-frame level. However, the BS order of allocation at the super-frame level affects the signal-to-interference-plus-noise ratio (SINR) which in turn affects the data rate. The solution in [8] does not consider this aspect in its allocation.

Generally, the cell boundary users in a cellular network receive low data rates because of increased path loss and inter-cell interference. This problem of low cell edge bitrate has recently
got considerable attention for the OFDMA systems due to the advances in new standards such as IEEE 802.16e [5] and E-UTRA [6]. In particular, for E-UTRA, inter-cell interference co-ordination is considered for the downlink [9]. Several frequency reuse schemes have been suggested. Notable among those are soft frequency reuse [10] and partial frequency reuse [11]. However, [12] reported high performance loss due to the frequency reuse schemes and suggested a frequency reuse factor of 1. Our proposed scheme, in this chapter, falls within the broader recommendation of [12]. It dynamically assigns subcarriers to different regions, allocates them to different users and maintains a frequency reuse of 1. Which subcarriers are assigned to a user depend on the SINR of the subcarrier and the fairness requirement of the user.

Fractional frequency reuse (FFR) offers a simpler alternative to the frequency reuse problem in the multicell OFDMA networks [13]. The FFR scheme statically partitions the cell surface into two distinct geographical regions. One region consists of the inner cell area and the other includes the cell edge areas. The set of users present in the inner cell area are identified as super group, whereas the users located in the cell edge areas are called regular group. The regular group users are further partitioned into sectors. The set of subcarriers to service these groups of users are also called super and regular groups of subcarriers. We identify the original FFR as static FFR scheme. The boundary between the super and regular groups is statically set on the basis of the distance from the serving BS or the SINR threshold. Similarly, the set of subcarriers, to serve each group, are selected randomly and their numbers are proportional to the ratio of the users in the geographical area to the total number of users in the cell. The set of subcarriers assigned to the groups are reused in other adjacent cells. Fig. 4.1 shows the partitioning of cells into groups and corresponding frequency allocation of the static FFR scheme for 3 cells. The cell edge areas of each cell are divided into 3 sectors. The set of subcarriers assigned to the super group are reused in every cell’s inner areas. The regular group subcarriers are divided into sectors and reused in the outer regions of the corresponding sectors of all the adjacent cells. Thus, sector A subcarriers are repeated in sector A of all the cells. Thus, the frequency reuse factor of the static FFR is 1. The super group subcarriers face higher interference from
the adjacent cells. However, these subcarriers are available for all users in the area covered by the super group; therefore, they offer an increased trunking gain which is a function of the number of users sharing a radio resource. The regular group users have lower interference from the adjacent cells due to the sector based distribution of the subcarriers. The lower interference allows increased data rates. However, sectorization results in reducing the trunking gain as only a handful of users, falling within the outer region of the sector, can use these subcarriers.

The static FFR scheme has several shortcomings. First, it divides users in two groups on the basis of static distance or SINR thresholds. This partitioning reduces the trunking gain of each group because only a fraction of the total cell population is part of a group. Second, it partitions the available subcarriers randomly into the groups. The number of subcarriers assigned to a group is proportional to the number of users in the group. The subcarrier partitioning does not consider their radio channel states. We feel that FFR could benefit if only subcarriers are partitioned into groups and users in a cell are virtual members of both the groups. This way the users will be able to get access to the subcarriers of both groups and result in increased trunking gain. Furthermore, FFR can also benefit if subcarriers are partitioned on the basis of whether their presence in a particular group, that is, super group or a sector within the regular group, benefits the overall system performance or not. The motivation for dynamically distributing subcarriers into groups comes from the fact that a user may experience different SINR values on different subcarriers due to the subcarrier specific multipath and interference profiles. Therefore, it is possible for the network to find a set of subcarriers that perform better for a user if they are part of the super group instead of the regular group. Similarly, the network can find another set which provides a better gain for the same user if that set is allocated to a sector within the regular group instead of the super group. Thus, the dynamic subcarrier partitioning according to the long term channel state can improve the overall system performance. The proposed DSA scheme overcomes both of these shortcomings of the static FFR system.

The objective of this research is to improve the long-term system data rate of a down-
link OFDMA multicell network by intelligently distributing and allocating radio resources first among geographical locations of cells and later, within a cell, among users. For this purpose, we propose a dynamic FFR cell architecture where subcarriers are dynamically partitioned among geographical locations by the RNC DSA. The BS schedules those subcarriers to the users opportunistically. The combined solution is based on the cross layer technique of channel-aware allocation and scheduling of resources.

4.1.1 Contributions

The following are the main contributions of this chapter.

1. We propose a novel fractional frequency reused cell architecture where subcarriers are dynamically partitioned into super and regular groups. This partitioning is based on the popular cross layer technique of channel-aware resource allocation. On the other hand, users are not partitioned into physical groups. Their membership in these groups is fuzzy as no rigid boundaries like distance or SINR thresholds are used. The actual membership is left to be decided by the BS scheduler. The fuzzy boundaries result in the increased trunking efficiency of the frequency allocation scheme.

2. We formulate an optimization problem for the combined RNC and BS DSAs in a fractional frequency reused environment. The problem is based on set partitioning problems in combinatorial optimization discipline. The problem maximizes the aggregate system data rate by allocating subcarriers to users. It satisfies channel allocation and minimum data rate constraints.

3. Since determining an optimal allocation of subcarriers has exponential complexity; therefore, we propose an heuristic scheme that solves the combined RNC and BS DSA problem in two steps. The first step runs at the RNC and assigns subcarriers to virtual groups and satisfies the minimum rate requirements of the corresponding users. The second step
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

runs at every BS controlled by the RNC and operates at the scheduling slot level. The BS DSA algorithm schedules subcarriers to the users opportunistically, whereas satisfying their minimum performance constraints.

The proposed solution is extremely simple with linear complexity and achieves a frequency reuse factor of 1. It provides increased performance in comparison to the existing frequency allocation schemes.

4.1.2 Outline

This chapter is organized as follows. Section 4.2 provides details of the system and channel models along with the assumptions considered in this chapter. The dynamic FFR cell architecture is described in subsection 4.2.1. Section 4.3 formulates the combined DSA problem in the multicell FFR OFDMA networks. Section 4.4 presents the proposed solution scheme and Section 4.5 provides its numerical results. Related work is reported in Section 4.6 and chapter concludes in Section 4.7.

4.2 System Architecture and Model

4.2.1 Dynamic Fractional Frequency Reused Cell Architecture

Unlike the static FFR, where users and subcarriers are partitioned into two groups, namely super and regular groups, the proposed dynamic FFR only partitions subcarriers into physical groups. Among users, there are no rigid boundaries between super and regular groups. Both user groups cover the whole cell surface. Therefore, all users of a cell are virtual members of both groups. Their physical membership in a particular group or both groups is determined by the scheduling algorithm at the BS and it can change from one scheduling slot to another. Fig. 4.2 shows the cell partitioning and the frequency band allocation of the proposed FFR architecture. We see from this figure that the super group (that is, the circle) covers the whole
cell and the subcarriers allocated to it can be assigned to any user in the cell. The regular group also covers the whole cell surface, and as an example, it is further partitioned into 3 sectors. The subcarriers assigned to a sector are available for scheduling to the inner and the cell edge users falling within the boundaries of the sector.

The subcarriers are distributed between the groups and the sectors belonging to the regular group by the proposed RNC DSA algorithm. The subcarriers assigned to the super group and the sectors within the regular group are orthogonal. The distribution of subcarriers to a geographical region is performed dynamically considering their average channel states. Furthermore, the distribution decisions are made so that the overall system performance increases and the performance requirements for users belonging to the respective region are satisfied. By allowing overlap between super and regular group users, the trunking efficiency of the scheduling algorithm operating at the BS level also increases.

In this chapter, we consider 7 and 19-cell grids with 3 sectors per cell\(^7\) (see, Fig. 4.3).

4.2.2 System Model

We consider an OFDMA cellular network that consists of \(K\) cells. The service area consists of all \(K\) cells where each cell is serviced by a base station (BS). There is a central node in the form of radio network controller (RNC) that manages the \(K\) BSs. Every BS has a local scheduler. Fig. 1.2 plots a sketch of such a model.

Let \(\Upsilon_k\) denote the set of and \(|\Upsilon_k| = M_k\) be the number of users in a cell \(k\). The total number of users are \(M_c = \sum_{k=1}^{K} M_k\) and \(\bigcup_{k=1}^{K} \Upsilon_k\) is the set of all users in the network. In the proposed dynamic FFR architecture, super and regular groups cover the whole cell area. The regular group divides the cell area into sectors. Let us assume that each cell is partitioned into \(L\) sectors where \(l\) identifies a particular sector. Further assume, that \(\Upsilon_k^l\) denote the set of users and \(|\Upsilon_k^l| = M_k^l\) is the number of users in a sector \(l\) of a cell \(k\). Thus, \(\Upsilon_k = \bigcup_{l=1}^{L} \Upsilon_k^l\) and

---

\(^7\)The choice of 7 and 19-cell grids and 3 sectors per cell is due to the fact that these cell models are commonly used in the literature. We want to emphasize that our proposed solution is capable of handling any cell model.
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

\[ M_k = \sum_{i=1}^{L} M_k^i \]

Furthermore, \( M^l = \sum_{k=1}^{K} M_k^l \) represents the total number of users in the \( l^{th} \) sector of all cells.

Let \( N \) be the number of subcarriers available which represents the radio resource that is shared among the users in the network. Let us assume that \( C_{sup} \) and \( C_{reg} \) be the set of subcarriers assigned to the super and regular groups, respectively. The regular group subcarriers are further divided into sectors. Let \( C_{reg}^l \) be the set of subcarriers allocated to a sector \( l \) where \( C_{reg} = \bigcup_{l=1}^{L} C_{reg}^l \). Because the proposed architecture does not allow reuse of a subcarrier within a cell, these sets are orthogonal, that is, \( C_{sup} \cap C_{reg} = \{\} \) and \( \bigcap_{l=1}^{L} C_{reg}^l = \{\} \).

Further assume, \( Q_{sup} \) and \( Q_{reg}^l \) be the set of interferers experienced by a user in the super group and a sector \( l \) of the regular group, respectively. For a 19-cell grid, there will be 18 interferers experienced by the super group users. On the other hand, there will be 7 interferers for the regular group users. For example in Fig. 4.3(b), the set of interferers for a user in the sector A of cell 1 includes cells numbered \{5, 6, 13, 14, 15, 16, 17\}.

### 4.2.3 Channel and Data Rate Models

Channel model consists of path loss, shadowing, fast fading and interference from the neighboring cells.

The ITU’s path loss model for the vehicular test environment [14] is a function of the distance from the serving BS \( r \), carrier frequency \( f \) and the base station antenna height \( \triangle h_b \). For general cases, path loss at a distance \( r \), \( \Gamma(r) \), is expressed as

\[
\Gamma(r) = \left[ 40(1 - 4 \times 10^{-3} \triangle h_b) \right] \log_{10} r - 18 \log_{10} \triangle h_b + 21 \log_{10} f + 80 \quad \text{dB}
\]

which reduces to the following

\[
\Gamma(r) = 128.1 + 37.6 \log_{10} r \quad \text{dB} \quad (4.1)
\]

for \( f = 2000 \text{ MHz} \) and \( \triangle h_b = 15 \text{ meters} \). In the above, \( r \) is in km and the path loss, \( \Gamma(r) \), is in dB.
For shadowing, we consider a correlated model. A user $i$ from the serving BS $k$ at a discrete time instant $\tau$ will experience the shadowing effects as follows:

$$\begin{align*}
X_{i,k}(d) &= 2^{-d/d_{corr}}, \\
S_{i,k,\tau} &= X_{i,k}(d)S_{i,k,\tau-1} + \sqrt{1 - X_{i,k}(d)^2} N(0, \delta),
\end{align*}$$

(4.2)

where, $X(d)$ is a normalized autocorrelation function of shadowing at a distance $d$ from the previous calculation of shadowing, $N(0, \delta)$ is a normal random variable with mean 0 and a standard deviation of $\delta$, and $S_{i,k,\tau}$ represents the shadowing value. For simplifying notation, in the subsequent discussion, we remove the $\tau$ subscript from the shadowing value. It is understood by the context that the variable $S_{i,k}$ represents the shadowing value at a certain discrete time instant. From (4.1) and (4.2), it is clear that path loss and shadowing are independent of subcarriers.

This chapter considers Rayleigh distributed fast fading which is modeled by the Jakes spectrum [15]. The fading is carrier specific and the fading gain for a user $i$ on a subcarrier $j$ is $\zeta_{i,j}$.

Consequently, combining path loss, shadowing and fast fading results in the following channel gain experienced by a user $i$ on a subcarrier $j$ when receiving signal from a BS $k$:

$$G_{i,j,k} = 10^{-\frac{r_{i,k}(r)}{10}} \times S_{i,k} \times \zeta_{i,j}$$

(4.3)

The resulting SINR is expressed as:

$$\text{SINR}_{i,j} = \frac{G_{i,j,k}P_{i,j,k}}{N_0 \Delta f + \sum_{q=1}^{Q} G_{i,j,q}P_{i,j,q}}$$

(4.4)

where $N_0$ is the noise power spectral density, $\Delta f$ is the subcarrier spacing, $Q$ is the set of interferers, $G_{i,j,}$ and $P_{i,j,}$ are the channel gain and the transmission power, respectively. SINR values can be computed for all subcarriers and for all users in the super and regular group settings. When SINR is computed for the super group, then $Q = Q_{sup}$, and when it is found for sector $l$ of the regular group, then $Q = Q_{reg}^l$. 

106
The computation of channel gain, that is, $G_{i,j,k}$ and $G_{i,j,q}$ in (4.4), requires an estimate of distance $r$ between user $i$ and transmitting BS. In case of $G_{i,j,k}$ the transmitting BS is the serving BS, and for $G_{i,j,q}$ the transmitting BS’s are the interfering BSs. For the grids shown in Fig. 4.3, Table 4.1 provides the estimated distances between transmitting BSs and a user located in the sector A of the central cell. It is assumed that the user is located on a segment between the central BS and the neighboring BS.

Employing continuous rate adaptation, the SINR found in (4.4) can be mapped to the data rate as follows:

\[
R_{i,j} = \Delta f \log_2(1 + \lambda\text{SINR}_{i,j})
\]

(4.5)

where, $\lambda$ is a constant related to the target bit error rate (BER) as $\lambda = -\frac{1.5}{\ln(5\text{BER})}$ [16], $\Delta f$ is the subcarrier spacing and $R_{i,j}$ is the achievable data rate by the $i^{th}$ user and the $j^{th}$ subcarrier pair.

In the following, a subcarrier $j$ which falls within the regular group will have $R_{i,j}^{\text{reg}}$ achievable data rate for user $i$. Similarly, $R_{i,j}^{\text{sup}}$ identifies the achievable data rate of subcarrier $j$ in the super group. Furthermore, the RNC algorithm employs the average ($R_{i,j}^{\text{sup}}, R_{i,j}^{\text{reg}}$) whereas the BS uses the instantaneous values ($R_{i,j,t}^{\text{sup}}, R_{i,j,t}^{\text{reg}}$) of these achievable rates.

4.2.4 Time Scale

Our proposed solution consists of two DSA algorithms; one runs at the RNC level and divides subcarriers into the super and different sectors of the regular group. The other runs at the BS level and schedules subcarriers to the users. Therefore, we consider two time scales. The BS decisions are made at the scheduling slot level. A number of slots form a frame and a superframe is constructed by combining a number of adjacent frames. The RNC DSA operates at the super-frame level. Like Chapters 2 and 3, here too, we consider that the scheduling decisions are made at the slot level instead of the packet or frame level.
4.2.5 Assumptions

In this sub-section, we list the set of assumptions that are critical in the problem formulation and its proposed solution.

- Due to the complexity of the joint DSA and APA in the multicell OFDMA networks, we will like to simplify the problem by assuming that the power is equally distributed among the subcarriers. By this assumption, the joint problem reduces to only a DSA problem. This is a very common assumption considered in several studies [2], [13], [17]. The variations in SINR results in the transmission rate adjustments using the adaptive modulation and coding. Therefore, the computation of SINR in (4.4) uses constant values for the power, that is, \( P_{i,j,k} = P_{i,j,q} = P/N \ \forall \ i,j,q \), where network has \( N \) subcarriers, and \( P \) is the total BS transmission power.

- The user equipment has the capability of estimating the SINR for all subcarriers. Estimation of SINR values and forwarding it back to the BS is part of several standards developed for wireless communications [4], [18], [19].

In the proposed algorithm, both RNC and BS DSA’s require SINR values. These values are mapped to the achievable data rates by (4.5) and used in the RNC and BS DSA. The requirement for the RNC DSA is that the average achievable rates are made available to the RNC just before the execution of the RNC DSA which happens at the start of the super-frame. The second requirement for the RNC DSA is that these values are found for the super and regular group settings. Thus, there are \( 2 \times N \times M_c \) values estimated and collected for every execution of the RNC algorithm. The BS DSA requires the instantaneous achievable rates values for every scheduling slot. Unlike RNC, BS only needs these values for the group that subcarrier belongs to. Thus, if a subcarrier has been assigned to the super group then BS only needs the super group values for that subcarrier. Similarly, if another subcarrier is assigned to a sector of the regular group then the BS
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

needs rate information for the regular group for that subcarrier.

- A subcarrier is only assigned to one user in a cell. Thus, there is no intra-cell interference. However, all the subcarriers assigned to a cell are reused in the adjacent cells. Therefore, frequency reuse factor is 1 in the proposed scheme.

- As stated earlier, the BS DSA schedules the subcarriers to the users. The scheduling algorithm assumes that all the users serviced by a given BS always have data for transmission. This assumption results in the infinitely backlogged traffic model, just like in Chapters 2 and 3. In situations where the RNC DSA makes recommendations about the subcarrier assignment to the individual users, like in [2], and the infinite backlogged model is assumed, then the BS DSA has limited freedom of allocating subcarriers to the users. In this situation, the BS DSA becomes RNC centric algorithm and cannot take full advantage of the fast fading variations in radio channels. However, our proposed solution has the capability to take full advantage of the fast fading variations even in the presence of infinitely backlogged traffic demand. It gets this capability due to the two aspects of the proposed solution. First, the proposed RNC DSA does not recommend subcarrier assignments of users. Instead, it divides the subcarriers to the groups and leaves the BS DSA to assign them to the users. Second, the BS DSA employs opportunistic scheduling. Opportunistic scheduling has the capability of benefiting from the multiuser diversity due to the fast fading variations in radio channels.

4.3 Problem Formulation

In this section we formulate the dynamic subcarrier assignment problem for the downlink of the OFDMA multicell networks. The formulated problem is a representation of the joint RNC and BS DSA objectives.

The DSA objective is to maximize the system data rate while satisfying lower data rate
requirements of users. Additionally, problem satisfies some other constraints related to the subcarrier assignments like a subcarrier could be assigned to only one user in a cell, that is, no intra-cell interference is allowed, and the available subcarriers are reused in all cells of the network.

Let $x_{i,j}^{sup}$ be the binary decision variable, that is, $x_{i,j}^{sup} \in \{0,1\}$. When this variable is 1, it signals that the subcarrier $j$ is assigned to the user $i$ and belongs to the super group of subcarriers. When its complement $x_{i,j}^{sup} = 1$, it signals that the subcarrier $j$ is assigned to the user $i$ and it falls in the regular group of subcarriers. The super and regular group subcarriers are orthogonal.

The combined DSA solves the following binary integer program for every time slot $t$:

\[
\begin{align*}
\max_{x_{i,j}^{sup}} & \quad \sum_{j=1}^{N} \sum_{i=1}^{M_c} x_{i,j}^{sup} R_{i,j,t}^{sup} + \sum_{j=1}^{N} \sum_{i=1}^{M_c} x_{i,j}^{sup} R_{i,j,t}^{reg}, \\
\text{s.t.} & \quad \sum_{i \in \Upsilon_k} x_{i,j}^{sup} \in \{0,1\}, \quad k = 1 \cdots K, \quad j = 1 \cdots N, \\
& \quad \sum_{i=1}^{M_c} x_{i,j}^{sup} \in \{0,K\}, \quad j = 1 \cdots N, \\
& \quad \sum_{i=1}^{M_c} x_{i,j}^{sup} \in \{0,K\}, \quad j = 1 \cdots N, \\
& \quad \sum_{j=1}^{N} x_{i,j}^{sup} R_{i,j,t}^{sup} + x_{i,j}^{sup} R_{i,j,t}^{reg} \geq C_i, \quad i = 1 \cdots M_c.
\end{align*}
\]  

where $C_i$ be the lower bounds on data rates for user $i$.

The objective function (4.6a) maximizes the weighted sum of the instantaneous achievable data rates for all users and all subcarriers in the network. The weights, $x_{i,j}^{sup}$ and $x_{i,j}^{sup}$, are binary decision variables and represent the allocation of subcarriers to the users. The constraint (4.6b) enforces that a subcarrier can be assigned to only one user in a cell, and constraint (4.6c) satisfies that if a subcarrier $j$ is assigned to the super group in one cell then this subcarrier should be
reused in all the cells. Likewise, (4.6d) adds similar reuse constraints for the regular group subcarriers. The constraint (4.6e) enforces that the subcarriers assigned to a user results in equal or more data rate than the lower bound for the user.

4.4 Proposed Solution

Finding optimum solution of (4.6) is exponentially complex because the problem falls within the realm of set partitioning problems which are NP hard. Secondly, the solution requires the knowledge of the instantaneous achievable rates for all the user subcarrier pairs in all cells at a central location. This requirement adds an extremely high overhead that would nullify any data rate gains made by the DSA. Therefore, we decompose the joint problem into two parts. The first part is solved by a central location, like RNC, which computes the membership of subcarriers in the super group or a sector within the regular group. For decision making, RNC DSA requires average achievable rates information for all the user subcarrier pairs in the super and regular group settings. The second part operates at the BS level and allocates subcarriers to the users. The BS DSA requires the instantaneous data rate information for all the user subcarrier pairs. In the following sub-sections, we present the proposed solution to solve these problems.

4.4.1 RNC DSA Algorithm

The RNC algorithm distributes all the available subcarriers among geographical regions within a cell. In the proposed dynamic FFR architecture, the geographical regions include the super group and sectors of the regular group. The distribution of the subcarriers to the geographical regions is replicated for all the cells in the grid. The algorithm employs greedy approach where the assignments are made so that they result in increased system throughput for the whole grid. Furthermore, the algorithm considers the minimum data rate requirements of the group of users so that the assigned subcarriers are able to satisfy the load in the respective geographical
regions. Therefore, algorithm has a sense of fairness while distributing resources at the RNC level. The RNC algorithm pseudocode is presented in Fig. 4.4 where \( i, j, l \) and \( k \) are iterators for users, subcarriers, sectors and cells, respectively. In the following discussion, reference to step number is to the line numbers given in Fig. 4.4.

The algorithm requires, as input, average achievable rates for all users on all subcarriers in the super and regular group environments, that is, \( R_{i,j}^{sup}, R_{i,j}^{reg} \), and the minimum data rate requirements \( C_i \) for all users.

As an output, algorithm returns binary vector \( \mathbf{X}^{sup} = \{ x_{i,j}^{sup} | j = 1 \cdots N \} \), and a binary matrix \( \mathbf{X}^{reg} = \{ x_{l,j}^{l,reg} | l = 1 \cdots L, j = 1 \cdots N \} \). A value of 1 for \( x_{i,j}^{sup} \) signals that the subcarrier \( j \) is assigned to the super group, whereas a value of 1 for \( x_{l,j}^{l,reg} \) means that the subcarrier \( j \) is assigned to the regular group sector \( l \). Because a subcarrier can be assigned to only one geographical region within a cell, the algorithm sets these variables such that the conditions listed in Section 4.2.2, which are \( C_{sup} \cap C_{reg} = \{ \} \) and \( \bigcap_{l} C_{l,reg}^{l} = \{ \} \), are satisfied. Thus, the resulting output values have the following relations,

\[
\sum_{l} x_{l,j}^{l,reg} \in \{0, 1\} \quad \text{and} \quad x_{j}^{sup} = \bigvee_{l} (x_{l,j}^{l,reg}),
\]

where operator \( \bigvee \) is logical OR operator applied on the \( j \)th column of \( \mathbf{X}^{reg} \). From the above relation, it is clear that once \( \mathbf{X}^{reg} \) is known then \( \mathbf{X}^{sup} \) can be uniquely determined. Therefore, the RNC algorithm determines \( \mathbf{X}^{reg} \) during its main loop (steps 1 to 10), and later computes \( \mathbf{X}^{sup} \) in step 11. During initialization phase, algorithm assigns zeros to all entries of \( \mathbf{X}^{reg} \). This assignment signals that no subcarrier is assigned to any sector of the regular group and all subcarriers are allocated to the super group.

Initialization phase of the algorithm declares some variables and computes utility values of the subcarriers. The variable \( g_i \) is assigned rate which is aggregated data capacity of a sector. Its value is updated after assigning each subcarrier in the main loop (steps 5 and 7). The utility values, \( W_{j}^{sup} \) and \( W_{j}^{l,reg} \), measure the utility of a generic subcarrier \( j \) assigned to
the super group or $l^{th}$ sector of the regular group, respectively. Because RNC DSA does not know which user or group of users are going to use the subcarrier eventually during scheduling, it assumes that the subcarrier will be time multiplexed between all the available users in the respective geographical region. The time multiplexing of a subcarrier to several users results in an aggregate data rate which is average rate of all users for the given subcarrier in the region. Therefore, the fraction in the parenthesis computes this average by adding achievable rates of all users and dividing the sum with the number of users in the region. As stated earlier, the proposed system architecture allows frequency reuse factor of 1, hence, the subcarrier allocation in one cell is repeated in all the cells in the grid. Thus, the utility computation adds the average data rate for all cells.

The $U_{l}^{t,reg}$ variables hold fractional utility gain if subcarrier $j$ is removed from the super group and assigned to sector $l$ of the regular group. It is a fraction of the difference between the utility values and the utility value of the best sector. A negative value of $U_{l}^{t,reg}$ for all $l$ means that the allocation of the subcarrier to any sector of the regular group is expected to lower the system data rate. In this case, algorithm assigns it to the super group (step 5). Because of the overlap of users between super and regular groups, the assigned subcarrier is available to the users of all sectors in the regular group. Consequently, the assignment to the super group increases the assigned data rates of all regular group sectors. The algorithm incorporates this fact in step 5 when it updates $g'$ by a fraction of $W_{j}^{sup}$ assuming time multiplexing of the subcarrier by the respective users.

The main part of the RNC algorithm consists of a loop that iterates on the subcarriers. As stated earlier, during initialization phase, the algorithm assigns all subcarriers to the super group. In the main loop, it iterates on subcarriers, in the descending order of $U_{j}^{t,reg}$, and attempts to assign the subcarrier to a sector in the regular group which increases the most system data rate. If the algorithm finds such a sector, then it assigns the subcarrier to it and increases the assigned rate for the sector and updates the corresponding output variable $x_{j}^{l,reg}$ (steps 6 and 7). Otherwise, the algorithm leaves the subcarrier to the super group (step 5).
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

So far our description has only reported the greedy aspects of the proposed algorithm. The fairness measure is added in step 2 and 3 by a simple condition. Here, during every iteration, the algorithm first looks for the sectors which are lacking in radio resources such that their assigned data rates are less than their required load. If algorithm finds such sectors, then it considers only those sectors in the subsequent assignment steps during the iteration. If algorithm does not find any sectors, it means all sectors have sufficient radio resources to satisfy their required load, then it behaves as greedy algorithm and considers all sectors in the subsequent allocation steps. Consequently, when all sectors are considered then the one that increases the most system data rate will be assigned the subcarrier.

The time complexity of the RNC DSA algorithm is linear as it consists of only one main loop that iterates \( N \) times. The initialization stage has complexity of \( O(L \times N) \) where \( L \) being the number of sectors in a cell which is generally a very small value. The whole algorithm does not need any sorting operation. The implementation of the algorithm can be carried out in such a way that the main loop does not need to run \( N \) iterations in most of the situations. For example, by combining conditions of step 3 and 5 together, the loop can exit when \( \mathcal{L}' = \{} \) and \( U^{j^*, reg} < 0 \). In this case, there is no point in continuing the loop because all the unassigned subcarriers do not increase the system performance if assigned to any sector of the regular group. These unassigned subcarriers are more suitable for the super group assignment which by default is being the case. Therefore, algorithm can break from the main loop and avoid unnecessary iterations of the unassigned subcarriers. For ease of presentation and the choice of output variables in the discussion, we have not included these tricks in the pseudocode.

4.4.2 BS DSA Algorithm

The RNC algorithm finds the subcarrier assignment to geographical locations which are represented in values of \( \tilde{X}^{sup} \) and \( X^{reg} \). These values are passed to the BS DSA which is a scheduling algorithm and makes decisions at every time slot level. The other inputs to the scheduling al-
algorithm are instantaneous achievable rates for all the user and subcarrier pairs in the super and regular group settings and the minimum data rate requirements. The minimum data rate requirements, $C_i$ of (4.6e), are constant whereas the instantaneous rates dynamically change as they depend on radio channel state. Since identical and independent copies of the BS algorithm run on all $K$ BSs in the network; therefore, in the discussion we assume cell $k$. Furthermore, a user can be present in only one sector at a time. We consider sector $l$. This simplifies the discussion. The pseudocode of the algorithm is given in Fig. 4.5.

The output of the scheduling algorithm can be represented by a set of binary variables which describe the allocation of subcarriers to the users. For the dynamic FFR architecture presented in this chapter, a user can be allocated subcarriers from two sets. The first set is the set of subcarriers belonging to the super group, $C_{sup}$, and the second set is the set of subcarriers assigned to the regular group sector where user is present, $C_{reg}$. To represent the allocation of the available subcarriers for user $i$, we use variables $x_{i,j,t}^{sup}$ and $x_{i,j,t}^{reg}$ where $j \in C_{sup}$ and $j' \in C_{reg}$ and $t$ is the time slot.\(^8\)

The main feature of the BS DSA is the opportunistic scheduling of the subcarriers to the users. The objective of the opportunistic scheduling is to maximize the long-term expected data rate subject to complying with the minimum performance requirements. The other constraint is that the scheduler cannot schedule a subcarrier to two users in a cell. To satisfy these objectives, we use the minimum performance guarantee (MPG) algorithm of [20], which is also discussed in Chapter 3. There, MPG algorithm was opportunistically time-sharing single carrier among several users. Here, multiple subcarriers are to be opportunistically time shared between several users. To describe the multicarrier MPG, we need to declare some variables.

Let $R_i^T(x)$ be the average data rate of user $i$ up to time $T$ where $x$ represents the decisions made by the scheduler, that is, output of the BS DSA. For the system under consideration,\(^8\) For ease of notation, we have not used subscript $t$ in the pseudocode of Fig. 4.5. Furthermore, by abuse of notation, an iterator of a set is represented as a member of the set in this sub-section.
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

\( R^T_i(x) \) is estimated as:

\[
R^T_i(x) = \frac{1}{T} \sum_{t=1}^{T} \left( \sum_{j=1}^{\lvert C_{\text{sup}} \rvert} R^\text{sup}_{i,j,t} x^\text{sup}_{i,j,t} + \sum_{j'=1}^{\lvert C_{\text{reg}} \rvert} R^\text{reg}_{i,j',t} x^\text{reg}_{i,j',t} \right)
\]  \hspace{1cm} (4.7)

Additionally, \( R(x) = \sum_i^{\lvert Y_k \rvert} E(R_i(x)) \) is the average cell data rate where

\[
E(R_i(x)) \triangleq \limsup_{T \to \infty} R^T_i(x).
\]

The MPG problem can be written as

\[
\max_x R(x), \hspace{1cm} \text{s.t.} \hspace{0.5cm} E(R_i(x)) \geq C_i.
\]  \hspace{1cm} (4.8)

The solution to the above problem is found by multiplying the instantaneous achievable rates of all users for all carriers by user specific control variables and then for every subcarrier, finding a user which has the maximum product [17], [20]. This generic solution can be easily transformed for the BS DSA problem. For the assignment of the super group subcarriers \( j \in C_{\text{sup}} \), the algorithm finds the user \( i^* \) that satisfies the following expression at every scheduling slot \( t \).

\[
i^* = \arg \max_i \beta_i^* R^\text{sup}_{i,j,t}
\]  \hspace{1cm} (4.9)

Similarly, for the \( l \)th sector subcarrier assignments in the regular group, that is \( j' \in C_{\text{reg}}^l \), the BS DSA finds the user \( i^* \) that fulfills the following expression

\[
i^* = \arg \max_i \beta_i^* R^\text{reg}_{i,j',t}
\]  \hspace{1cm} (4.10)

The true controlling parameters \( \beta^*_i \) in the above solutions are chosen such that \( \beta^*_i \geq 1 \) for all \( i \), \( E(R_i(x)) \geq C_i \) for all \( i \), and if \( E(R_i(x)) > C_i \) then \( \beta^*_i = 1 \). Furthermore, employing stochastic approximation techniques [21], the true values of \( \beta_i \) can be estimated in real time as follows,

\[
\beta_i^{t+1} = \max \left[ \beta_i^t - \epsilon \left\{ E(R_i(x)) - C_i \right\}, 1 \right],
\]  \hspace{1cm} (4.11)
where $\epsilon$ is a small positive real number that acts as the step size for learning the true value of $\beta_i$.

A version of MPG for multicarriers is presented in [17] where an additional constraint that two subcarriers cannot be scheduled to a user makes the problem a bipartite graph matching problem. Because there is no need to have such a constraint in our system, our proposed solution remains simple with linear complexity. Even if there is a physical limit on the support of maximum number of subcarriers by the user equipment, our proposed algorithm can easily handle this constraint by removing the user from the set of users once its maximum subcarrier limit has reached.

### 4.5 Numerical Results

In this section, we report our simulation results for the proposed DSA algorithm and compare them with the full frequency reuse with full interference (FFFI), conventional sectored, and static FFR allocations. All the four schemes considered in this comparison have a frequency reuse factor of 1. The simulation parameters are given in Table 4.2. The algorithms are different in terms of the subcarrier allocation performed by the RNC. In case of the full frequency reuse with full interference (FFFI), all the subcarriers are available for allocation to all the BSs in the grid. Furthermore, cells are not sectored, and all the adjacent cells act as interferers for the subcarriers. Fig. 4.6(a) shows an example of frequency allocation for FFFI scheme. For the conventional sectored allocation, RNC randomly selects a subset of the subcarriers for a sector. This subset is repeated in the same sector of all cells. The size of the subset is approximately equal to $\left\lfloor N \frac{M_i}{M_k} \right\rfloor$ for $l = 1 \cdots L - 1$, and for the $L^\text{th}$ sector it is $N - \sum_{l=1}^{L-1} \left\lfloor N \frac{M_i}{M_k} \right\rfloor$. Fig. 4.6(b) shows an example of frequency allocation for the conventional sectored scheme. The static FFR scheme divides users according to a distance threshold from the serving BS. In the results, users within 70 percent of the cell radius are considered members of the super group. The remaining users are members of the regular group which is divided in 3 sectors. At the RNC level, the
subcarriers are randomly allocated to these regions whereas the number of subcarriers assigned to a region is proportional to the number of users in the region. Fig. 4.1 shows an example of partitioning of the cell surface and frequency allocation for the static FFR scheme.

The BS part of all the four allocation schemes is based on the minimum performance guarantee opportunistic scheduling [20]. Though, BS DSA algorithm is same for all the four schemes but the instances running for these schemes vary due to the subcarrier allocation provided by the RNC. The simulation runs for 64 sec, i.e., 100 super-frames. The channel undergoes fast fading according to the Jakes model. The number of users in each cell, the cell dimensions, and the BS locations remain the same for the 100 super-frames whereas user locations vary according to the random walk mobility model [22].

Fig. 4.7(a) and 4.8(a) compare the system throughput, which is the aggregate throughput received by all the users in all the cells, achieved by the proposed, conventional sector, FFFI and static FFR allocations for 7 cell and 19 cell grid sizes, respectively. This comparison is performed as a function of the cell radius where cell radius takes the values of \{1, 2, 3, 4, 5\} kms. From these two figures, it is clear that the proposed DSA scheme performs better than the other three schemes. The reason for this performance improvement is that the RNC DSA intelligently distributes subcarriers into geographical regions so that there is reduced interference and increased trunking gain that improves the system throughput. For smaller cell sizes, proposed algorithm significantly outperforms other three schemes because, for such cell structures, there is increased interference from the adjacent cells. Therefore, intelligently allocating subcarriers makes significant positive effect on the system performance. The proposed scheme reported gains of 49, 46 and 48 percent than the FFFI, conventional sector and static FFR allocations for the grid size of 7 cells where the cell radius is 1000 m.

The increase in cell radius reduces the interference from neighbors and the cell capacity. This reduction is due to the increased path loss. For such situations, all four schemes achieve comparable performance. For example as shown in Fig. 4.7(a), for grid size 7 and cell radius 3 km or higher, proposed DSA is only marginally better than the other three schemes. Because
there is little or no interference; therefore, RNC DSA cannot benefit significantly from the subcarrier allocation. The grid size 19 comparison also shows the same trends (see Fig. 4.8(a)), except that because of large number of interferers, the proposed scheme is able to achieve significantly higher performance for moderate size cells in comparison to the FFFI. For example, for cell size of 3000 m, proposed scheme showed 21 percent improvement than the FFFI scheme.

In general, FFFI experiences more interference than the conventional sector scheme; therefore, proposed scheme has larger gains against it. Likewise, for 19 cell grids, the FFFI has lower performance than the conventional sector allocation because of the increased interference in the grid whereas, for 7 cell grid, it has higher performance than the sectored allocation. The static FFR shows significant improvements over FFFI scheme for grid size 19. This result is in line with the results of [13]. However, in comparison to the proposed scheme and the traditional sectored allocation, static FFR performs badly due to the reduced trunking.

Fig. 4.7(b) and 4.8(b) show the fraction of the subcarrier allocated to the super group by the proposed scheme as a function of the cell radius. Here we observe that when cells are small there are no subcarriers assigned to the super group. In case of super group, all the adjacent cells act as interferers, and for small cell sizes, the interference is significant. This interference reduces the achievable data rates on subcarriers. Therefore, it is preferable to allocate them to sectors of the regular group because in that allocation there are fewer interferers, and subcarriers are capable of supporting a higher data rates. When cell sizes increase, the RNC DSA allocates a portion of the subcarriers to the super group. For larger cells, the super group faces less interference and increased trunking advantage results in a portion of subcarriers to have higher data rates in the super group than in the regular group.

Fig. 4.9(a) and (b) compare the cumulative distribution functions (CDF) of the achieved rates of users and their minimum data rates requirements for 7 and 19 cell grid sizes, respectively. These figures show that all four schemes have been able to comply with the minimum performance requirements. This compliance is due to the BS scheduling algorithm which is MPG opportunistic scheduling. Among the four DSA schemes considered, the proposed scheme is
able to provide increased data rates. A group of users in the proposed scheme is able to have very high data rates. This behavior is in line with the objectives of the proposed DSA as the RNC DSA is a greedy algorithm. As long as minimum requirements are satisfied, the algorithm allocates resources to the geographical locations (that is, set of users present in those locations) in greedy fashion. Therefore, a set of users have increased data rates.

Fig. 4.10 shows the network throughput comparison for four schemes as a function of time for 19 cell grid and a cell radius of 2000 m. The proposed scheme achieves higher network throughput in comparison to other three schemes. The increase is because of the way RNC DSA distributes subcarriers among geographical locations and the increased trunking advantage of the proposed BS DSA. In comparison to the FFFI allocation, the proposed scheme achieves an average gain of 64 percent. Whereas, in comparison to the conventional sectored allocation, the proposed scheme reports an average gain of 31 percent. The gain is higher against the FFFI allocation because FFFI experiences lower data rates due to higher interference from 18 interferers than conventional sectored allocation. The proposed scheme is also able to outperform the static FFR allocation.

For the above experiment, we compare the performance of the cell edge users in Fig. 4.11(a). In this comparison, the cell edge users are those users which are at a distance of 1400 m or higher from the serving BS. The static FFR is able to report the highest data rates for the cell edge users followed by the FFFI and the sectored allocation. These three schemes are able to report better results than the proposed scheme. The static FFR is designed for the improvements of the cell edge users. The sectored allocation is able to show this performance because every sector gets subcarriers proportional to the number of users in that sector. In comparison to the other three schemes, the proposed scheme is a greedy allocation strategy which attempts to satisfy the lower bounds. Once those lower bounds are satisfied, the scheme mostly allocates its resources to the group of users that can provide the maximum gain to the performance. Because cell edge users generally have weaker channels, the proposed solution allocates any additional subcarriers, those left over after satisfying the minimum performance
requirements, to the users with stronger channels. Hence, the proposed scheme reports reduced throughput for the edge users. However, it is important to mention that the proposed scheme is able to satisfy the minimum requirements of all cell edge users as shown in Fig. 4.11(b) where CDF of achieved data rates of cell edge users and their requirements are shown. Therefore, the proposed scheme is not in violation of any design objectives. The problems of low data rates for edge users in comparison to those in the center of the cell can be solved by increasing the minimum requirements of such users or by placing upper bounds on the maximum data rate that inner cell users can achieve by employing algorithms like TCOS of Chapter 3.

4.6 Related Work

DSA for OFDMA systems improve system performance. Furthermore, employing adaptive modulation and coding (AMC), the achievable data rate is a function of the power allocation, that is, the data rate increases with transmit power and vice versa. Therefore, it is expected that adaptive power allocation (APA) can further improve the system data rate. Due to these reasons, downlink DSA and APA for single cell have attracted considerable research interests (see [7], [23], [24], and references therein).

In [7], authors adopt optimization-theoretic approach to solve the DSA and APA problems. It assumes continuous data rates and an infinite number of subcarriers in the OFDMA system. These ideal conditions render the DSA and APA problems tractable and enable the system data rate to be maximized through a greedy algorithm. It is shown in [23] that the data rate of multiuser OFDM system is maximized if subcarrier is assigned to the best user and the transmit power is distributed by the water-filling policy. According to the solution, more power is allocated when the channel gain is high and vice versa. In [24], authors have proposed a solution which minimizes the total transmission power while satisfying lower bounds on individual data rates. It employs Lagrangian relaxation technique to provide significant performance gain in comparison to the fixed assignment strategies. The algorithm requires large number of iterations
to converge. The practical OFDM systems have finite number of subcarriers, thus, the DSA problem becomes discrete optimization problem. As long as there is no fairness requirements, the greedy algorithm approach of ideal case is applicable for such a system [25].

In general, the joint DSA and APA problem can be formulated as the minimization of the overall transmit power under fairness constraints on users data rates or the maximization of a particular fairness criterion. In the case that the problem is convex, the solution is straightforward. However, in most cases, since the actual problem is combinatorial in nature, it is attractive to use efficient heuristic solution techniques.

Several previous studies, for example [7] and [23], have shown that the performance improvements of the APA solution are marginal over a wide range of signal-to-noise ratios (SNRs) due to the statistical effects. Therefore, a simpler solution involving DSA with equal power per subcarrier is preferred over more complex joint DSA and APA solution. The DSA with equal power per subcarrier can be solved by Hungarian method [17].

4.7 Conclusion

In this chapter, we have presented a novel dynamic channel allocation and opportunistic scheduling scheme for the multicell OFDMA networks. The scheme proposes a new dynamic fractional frequency reused system architecture where a cell surface is virtually partitioned into two regions. Both of these regions cover the whole cell surface. The first region is called super group, and the user-subcarrier pairs in this group experience interference from all the neighboring cells in the grid. The second region is called the regular group which is physically partitioned into sectors. The user-subcarrier pairs for this group experience reduced interference. The system architecture supports full frequency reuse. Thus, the subcarrier allocated to a group in a cell are repeated in the corresponding groups of all cells of the grid. By covering whole cell surface, both groups include all the cell users which results in increased trunking gain. The proposed DSA scheme makes best use of the FFR architecture and consists of two algorithms. The
first algorithm runs at the RNC and allocates subcarriers to the groups so that the system performance is increased while satisfying the sum of the minimum performance requirements of those groups. The second algorithm runs at every BS where scheduling decisions are made and subcarriers are assigned to the users. The scheduling algorithm is a popular opportunistic scheduling algorithm that satisfies the minimum performance requirements. The dynamic system architecture, the intelligent distribution of subcarriers to groups by the RNC, and the opportunistic scheduling at the BS level helps the proposed scheme to perform significantly better than the conventional allocation schemes. Simulation results compare the performance of the proposed scheme against traditional allocation schemes. For small to medium cell and large grid sizes, the proposed scheme outperforms the traditional schemes. The performance improvements are a result of the reduced interference and increased trunking gain of the proposed solution.
Table 4.1: Estimated Distance between BS and a User in Central Cell

<table>
<thead>
<tr>
<th>BS number (Fig. 4.3)</th>
<th>Estimated Distance (Y is inner cell radius &amp; y is the user distance from the serving BS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>y</td>
</tr>
<tr>
<td>2, 6</td>
<td>$\sqrt{(Y-y)^2 + 3Y^2}$</td>
</tr>
<tr>
<td>3, 5</td>
<td>$\sqrt{(Y+y)^2 + 3Y^2}$</td>
</tr>
<tr>
<td>4</td>
<td>$2Y + y$</td>
</tr>
<tr>
<td>7</td>
<td>$2Y - y$</td>
</tr>
<tr>
<td>8, 10, 12, 14, 16, 18</td>
<td>$4Y$</td>
</tr>
<tr>
<td>9, 11, 13, 15, 17, 19</td>
<td>$2\sqrt{3}Y$</td>
</tr>
</tbody>
</table>

Table 4.2: Simulation parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel bandwidth</td>
<td>5 MHz</td>
</tr>
<tr>
<td>Carrier frequency</td>
<td>2 Ghz</td>
</tr>
<tr>
<td>Number of subcarriers</td>
<td>300</td>
</tr>
<tr>
<td>Subcarrier spacing ($\Delta f$)</td>
<td>15 KHz</td>
</tr>
<tr>
<td>$N_0$</td>
<td>-174 dBm/Hz</td>
</tr>
<tr>
<td>Grid layout</td>
<td>3-sectored hexagonal 7 and 19 cells</td>
</tr>
<tr>
<td>Cell radius</td>
<td>1000 to 5000 m</td>
</tr>
<tr>
<td>Fast fading</td>
<td>Jakes Model</td>
</tr>
<tr>
<td>BS Transmit power</td>
<td>43 dBm</td>
</tr>
<tr>
<td>Channel model</td>
<td>ITU Vehicular, 30 km/h</td>
</tr>
<tr>
<td>Slot duration</td>
<td>2 ms</td>
</tr>
<tr>
<td>Frame length</td>
<td>5 slots</td>
</tr>
<tr>
<td>Super-frame length</td>
<td>64 frames</td>
</tr>
<tr>
<td>Number of users</td>
<td>Uniformly distributed between 15 to 20 per cell</td>
</tr>
<tr>
<td>Location of users</td>
<td>Uniformly distributed on cell surface</td>
</tr>
<tr>
<td></td>
<td>magnitude: 50 m to radius, and angle: 0 to $2\pi$</td>
</tr>
<tr>
<td>Minimum rate requirement</td>
<td>Uniformly distributed between 0 and 512 Kbps</td>
</tr>
<tr>
<td>Target BER</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>$\epsilon$ (4.11)</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

Figure 4.1: Frequency band allocation in a static fractional frequency reused architecture.

Figure 4.2: Frequency band allocation in the proposed fractional frequency reused architecture.
Figure 4.3: Fractional frequency reused based cell models (a) 7-cell grid (b) 19-cell grid
Algorithm RNC DSA

Inputs: $R_{i,j}^{\text{sup}}, R_{i,j}^{\text{reg}}, C_i, i = 1 \cdots M_c, j = 1 \cdots N$

Outputs: $x_j^{\text{sup}}, x_j^{l,\text{reg}}, j = 1 \cdots N, l = 1 \cdots L$

Initialization: $x_j^{l,\text{reg}} = 0, g_l = 0, j = 1 \cdots N, l = 1 \cdots L$

1. FOR $j=1 : N$ do //channel iteration
   2. Find set of sectors, $L'$, that satisfy $C_l - g_l > 0, l \in L$
   3. IF $L' == \{\}$ THEN $L' = L$ ENDIF
   4. Find sector subcarrier pair ($l^*, j^*$) that satisfies $\max_{l,j} \left( U_{j,l}^{l,\text{reg}} \right), l' \in L', j \in Z$
   5. IF $U_{j^*,l'^*}^{l'^*,\text{reg}} < 0$ THEN $g_l = g_l + W_{j^*,l'^*}^{\text{sup}} M_{l'^*}, l \in L$
   6. ELSE $x_j^{l'^*,\text{reg}} = 1$ //assign $j^*$ subcarrier to $l^*$ sector
   7. $g_l = g_l + W_{j^*,l'^*}^{l'^*,\text{reg}}$ //update assigned rate for $l^*$ sector
   8. ENDIF
   9. Remove $j^*$ from $Z$
10. End FOR
11. $x_j^{\text{sup}} = \bigvee \left( x_j^{l,\text{reg}} \right), j = 1 \cdots N$

Figure 4.4: Pseudocode for RNC DSA algorithm
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

Algorithm BS DSA (Cell K, Time slot t)

Inputs: Instantaneous $R_{i,j}^{sup}, R_{i,j}^{l,reg}, C_{i}, C_{sup}, C_{reg}^{l}$ $i = 1 \ldots M_k, j = 1 \ldots N, l = 1 \ldots L$

Outputs: $x_{i,j}^{sup}, x_{i,j}^{l,reg}, j = 1 \ldots N, l = 1 \ldots L$

Stored Local Variables: $\beta_i^t, R_i^T(x)$ $\forall i$

Initialization: $x_{i,j}^{sup} = 0, x_{i,j'}^{l,reg} = 0, j \in C_{sup}, j' \in C_{reg}^{l}$

//Algorithm segment that runs on super group
1. FOR $j = 1 : |C_{sup}|$ do  //channel iteration for super group assignments
2. Find user $i^*$ that satisfies $\arg \max_{i \in Y_k} \beta_i^t R_{i,j}^{sup}$
3. $x_{i^*,j}^{sup} = 1$
4. End FOR

//Algorithm segment that runs on regular group
5. FOR $l = 1 : L$ do  //sector iteration for regular group sectors
6. FOR $j = 1 : |C_{reg}^{l}|$ do  //channel iteration for regular group assignments
7. Find user $i^*$ that satisfies $\arg \max_{i \in Y_k^l} \beta_i^t R_{i,j}^{l,reg}$
8. $x_{i^*,j}^{l,reg} = 1$
9. End FOR
10. End FOR

//Algorithm segment updates and stores local variables $\forall i$
11. Update $R_i^T(x)$ with the data assignments of current iteration according to (4.7).
12. Update $\beta_i^{t+1}$ according to (4.11).

//End of algorithm

Figure 4.5: Pseudocode for one scheduling slot of BS DSA for cell $k$
Figure 4.6: Traditional frequency allocation schemes (a) full frequency reuse with full interference (b) conventional sectored allocation.
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

Figure 4.7: Grid size 7 results as a function of cell radius (a) comparison of proposed DSA with conventional schemes, (b) fraction of subcarriers in super group of proposed scheme
Figure 4.8: Grid size 19 results as a function of cell radius (a) comparison of proposed DSA with conventional schemes, (b) fraction of subcarriers in super group of proposed scheme
Figure 4.9: CDF of achieved data rates and lower bounds by the proposed and conventional schemes for cell radius of 1000 m: (a) grid size 7, (b) grid size 19.
Chapter 4. Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks

Performance comparison for grid size=19, cell radius = 2000m

![Graph showing system throughput comparison](image)

Figure 4.10: Network throughput comparison
Figure 4.11: Cell edge throughput comparison for grid size=19, cell radius = 2000 m (a) data rates (b) CDF of data rates.
Bibliography


Chapter 5

Conclusions and Future Work

This thesis considered three closely related cross layer scheduling and radio resource allocation problems. The three problems deal with the resource allocation of multiple users in the downlink of time-slotted and frequency-slotted cellular data networks. The problems maximize the system capacity, which is the sum of the long-term expected data rates of the users, under certain fairness constraints. For each of these problems, we have proposed algorithms that dynamically adapt to the changing environment and intelligently use the available information in their decision rules. The proposed algorithms are based on the cross layer technique of channel aware resource allocation. The opportunistic scheduling falls within the broader channel aware resource allocation framework because it improves the system capacity by taking advantage of the fast time scale fluctuations of the radio channel [1].

In the first problem, presented in Chapter 2, we extend the basic opportunistic scheduling by combining the user’s mobility and channel state information in resource scheduling. The use of the mobility information in the opportunistic scheduling helps in exploiting not only the fading-related fast variations, but also the path-loss-related slow variations in the radio channel. To combine the mobility information with the opportunistic scheduling, we present a discrete state mobility model that is suitable for the user mobility among hot-spot locations within a cell. Next, we present novel optimal and approximate mobility assisted opportunistic scheduling (MAOS) algorithms. Both of the proposed scheduling algorithms dynamically adjust a user’s fairness constraint according to its mobility information. The adaptation of the fairness constraints implicitly results in increasing the priority to those users who have better channel
Chapter 5. Conclusions and Future Work

states and vice versa. The optimal algorithm precomputes constraint values according to a fully known mobility model, whereas the approximate algorithm depends on the future prediction of the user mobility information. Through numerical results, we show that the use of the mobility information in opportunistic scheduling increases the system data rate and the channel capacity. We also provide an analytical lower bound on the performance of the approximate algorithm by generalizing a linear programming inequality of Dyer et al. [2],[3]. Extensive simulation results confirm that the combined use of the mobility and channel state information significantly benefits the system performance.

The second problem, discussed in Chapter 3, maximizes the aggregate user performance subject to the lower and upper bounds on user data rates. These bounds are user specific and provide the system operator with a better control of radio resource allocation and service differentiation among different classes of users. The proposed algorithm, called as throughput constrained opportunistic scheduling (TCOS), is capable of adapting to the radio channel conditions. The novelty of the proposed solution stems from the fact that it is capable of trading feasibility with throughput. By relaxing the feasible region marginally, the algorithm achieves significant performance gains in comparison to the existing solution under practical channel conditions. Finally in Chapter 4, we have solved a dynamic fractional frequency reuse (FFR) problem in the multicell OFDMA systems. For these systems, dynamic subcarrier allocation is an important issue to avoid inter-cell interference, increase frequency reuse, and exploit multi-user diversity. We propose a dynamic fractional frequency reuse cell architecture. Next, we formulate the subcarrier allocation problem on the lines of set partitioning problem and propose a simple and efficient hierarchical solution which divides the available subcarriers into groups and opportunistically schedules them to the users. The resulting solution outperforms the traditional allocation strategies.

The three proposed solutions are designed for existing and future wireless systems and technologies. For example, MaOS and TCOS are designed for the time-shared wireless systems like HDR (High Data Rate) [4] of cdma2000 1xEV-DO and HSDPA (High Speed Downlink
Packet Access) [5] of 3GPP specification. The dynamic FFR solution presented in Chapter 4 is designed for the time and frequency shared wireless systems like fixed and mobile wireless metropolitan area networks (IEEE 802.16 and 802.16e) [6], [7] and downlink of 3GPP’s Evolved Universal Terrestrial Radio Access (E-UTRA) [8]. However, the underlying principles of the proposed solutions are general and can be extended to other systems and technologies. For example, all the three solutions can be applied to any resource sharing current or future wireless system that employs adaptive modulation and coding (AMC) techniques to adapt data rates according to the radio channel states.

The proposed solutions have a common theme of making the best use of the additional information already present in or can be easily collected from the system. For example, the current cellular networks have the capability to estimate any user’s location. Furthermore, frameworks for location services are in place and new services are being developed [9]. The next generation of location information will also include user’s mobility information. Therefore, the presence of the mobility information in the cellular network provides new opportunities for intelligent resource allocation schemes. Moreover, scheduling algorithms can benefit by enforcing upper bounds on the user performance to provide better service differentiation and control. The upper bounds are already been exchanged between the current versions of the core networks and the user equipments. Similarly, the channel state information is being collected and used by the recent wireless systems. The use of the average channel information for dynamic allocation of subcarriers in a multicell environment provides an opportunity to control inter-cell interference and to benefit from the increased trunking and frequency reuse factors.

The proposed solutions can be implemented independently or can be combined together. For example, the MAOS algorithm of Chapter 2 can also benefit from the TCOS solution presented in Chapter 3 by enforcing upper bounds on the individual resource allocation. The presence of the upper bounds will provide additional benefits of an increased service differentiation and a better radio access control along with the increased capacity achieved by the mobility assisted scheme. Similarly, the dynamic FFR solution presented in Chapter 4 can benefit from the
upper bounds on the individual data rates applied through the base station scheduler. The upper bounds will assist the cell edge users with increased data rates because the subcarriers and time slots left unused by the strong users due to the upper bounds will now be available for the cell edge users.

5.1 Future Work

In this section, we identify open problems and provide recommendations for future research.

- *Use of Mobility Information in Multicell Networks*: The proposed mobility assisted opportunistic scheduling algorithm assumes a single cell environment. This work can be extended to the multicell environments where user’s mobility information is collectively used in scheduling, handoff and admission control procedures. For example, network can preallocate resources in cells, which are on the trajectory of the mobile users, to assist in the handoff operations. It can deny admission of new sessions in order to support the load of the mobile users. Similarly, knowing the user mobility, network can foresee sudden increase or decrease of load in a cell. Thus, it can dynamically adjust its admission control procedures before hand. Furthermore, mobility information can also be used with the dynamic subcarrier allocation (DSA) in multicell networks.

- *Joint Adaptive Power and Dynamic Subcarrier Allocation in Multicell Environments*: The proposed dynamic fractional frequency reuse scheme, discussed in Chapter 4, solves the DSA problem and assumes that the power is constant for all the subcarriers. For multicell environments with inter-cell interference, the adaptive power allocation (APA) can benefit the system performance. The APA can be used to lower the interference to the neighborhood. Thus, the proposed solution can be extended to include APA in its design.

- *Frequency Reuse DSA in OFDMA Mesh Networks*: Mesh topology offers very attractive solutions in providing the high performance and flexible deployment for fixed and mobile
broadband wireless access [10]. In mesh topology, a subscriber station (SS) can be directly connected to another SS and/or to the base station (BS), and it can route traffic on behalf of another SS by acting as a relay station. Furthermore, multihop routing (MHR) over a mesh topology allows SSs to exploit more reliable channels at lower transmission costs. This is because, even if the channel condition between an SS and the BS is bad, the channel condition between the SS and one of its neighboring SS’s may be good, and the SS may successfully route its traffic via the neighbor SS at a relatively lower cost than the BS [11]. The opportunities for lower transmission costs by MHR increase with the node density in the network. It is expected that the mesh mode operation can significantly benefit by designing efficient frequency reused DSA solutions.
Bibliography


Appendix A

Proof of Theorem 1

Proof: LP (2.14) and (2.17) differ with each other with respect to constraints (2.14c) and (2.17c). Thus it is sufficient to show that (2.17c) results in a larger set than (2.14c).

Let $\phi^* (\vec{h}, .)$ maximize the objective function of (2.14). Then multiplying both sides of (2.14c) by $\tilde{R} (\vec{m}, i)$ and summing $\forall i$ and $\forall \vec{m} \in S$, we get:

$$q^* := \sum_{\vec{m} \in S} \sum_{i=1}^{2} \tilde{R} (\vec{m}, i) \sum_{\tilde{h} \in \mathbb{H} (\vec{m})} \phi^* (\tilde{h}, i) = \sum_{\vec{m} \in S} \sum_{i=1}^{2} \tilde{R} (\vec{m}, i) r_i.$$

Similarly, let $\omega^* (\vec{h}, .)$ maximize the objective function of (2.17). Repeating the above procedure on (2.17c), we get:

$$o^* := \sum_{\vec{m} \in S} \sum_{i=1}^{2} \tilde{R} (\vec{m}, i) \sum_{\tilde{h} \in \mathbb{H} (\vec{m})} \omega^* (\tilde{h}, i) = \sum_{\vec{m} \in S} \sum_{i=1}^{2} \tilde{R} (\vec{m}, i) \tilde{r} (\vec{m}, i).$$

From (2.7) and given $r^* (., )$, there is a linear mapping between $o^*$ and the maximum objective function value of (2.6a). Thus, $o^* \geq q^*$. 

\[\blacksquare\]
Appendix B

Proof of Theorem 2

Proof: Without loss of generality, assume that the constraint-coefficient matrix, $A = (a_{ij})$, is of full rank $K$, and (2.21) is non-degenerate. Further assume that $P$ be the polyhedral feasible region to the LP (2.21), and it has $W$ feasible bases with $B_1, B_2, ..., B_W$ be the corresponding index set. Let $A_{B_e}, 1 \leq e \leq W$, be a feasible bases consisting of columns of $A$, $a_j = (a_{1j}, a_{2j}, ..., a_{Kj})'$, $j \in B_e$, where $(.)'$ represents vector transpose operation.

By the optimality criterion of the maximization problem, $B_e$ is optimal if and only if [1]:

$$c_j \leq c_{B_e} A_{B_e}^{-1} a_j, \quad \text{for all } j \notin B_e, \quad (B.1)$$

where $c_{B_e}$ is the row vector consisting of $c_j, j \in B_e$.

Let $V_e$ represent an event when optimality condition (B.1) is satisfied. The $\bigcup_{e=1}^{W} V_e$ has probability equal to one because problem (2.21) has at least one basic optimal solution. Applying optimality condition (B.1) and the conditional expectation hypothesis (2.23), we get an identity on expected values for $c_j$, for $j \notin B_e$:

$$E(c_j \mid V_e) = E(c_j \mid c_j \leq c_{B_e} A_{B_e}^{-1} a_j \text{ and } B_e), \quad (B.2a)$$

$$\leq \beta c_{B_e} A_{B_e}^{-1} a_j \quad (B.2b)$$

Multiplying (B.2a) with feasible solution $\hat{x}_j$, and summing over all $j$, we get,

$$E \left( \sum_{j=1}^{J} c_j \hat{x}_j \mid V_e \text{ and } B_e \right) = \sum_{j \in B_e} c_j \hat{x}_j + \sum_{j \notin B_e} E \left( c_j \mid V_e \text{ and } B_e \right) \hat{x}_j \quad (B.3a)$$

$$\leq \sum_{j \in B_e} c_j \hat{x}_j + \beta \sum_{j \notin B_e} c_{B_e} A_{B_e}^{-1} a_j \hat{x}_j \quad (B.3b)$$

146
Appendix B. Proof of Theorem 2

For optimal $B_e$, the optimal reward $u^*$ satisfies the following equation:

$$u^* = c_{B_e}A_{B_e}^{-1}b,$$

(B.4)

where $b$ is the column vector consisting of right side of (2.21b) constraints. Furthermore, the constraints of (2.21b) are satisfied, i.e.,

$$\sum_{j=1}^{J} a_j\hat{x}_j = b.$$  

(B.5)

Thus, applying (B.5) and (B.4) on (B.3b), we get:

$$E\left(\sum_{j=1}^{J} c_j\hat{x}_j \mid V_e \text{ and } c_{B_e}\right) \leq \sum_{j \in B_e} c_j\hat{x}_j + \beta c_{B_e}A_{B_e}^{-1}\left( b - \sum_{j \in B_e} a_j\hat{x}_j \right),$$

(B.6a)

$$\leq \sum_{j \in B_e} (c_j - \beta c_{B_e}A_{B_e}^{-1}a_j)\hat{x}_j + \beta E(u^* \mid V_e, c_{B_e})$$

(B.6b)

The $a_j$ and $A_{B_e}$ definitions result in:

$$c_j = c_{B_e}A_{B_e}^{-1}a_j \quad j \in B_e.$$

The above identity means that the coefficients of $\hat{x}_j$ are $(1 - \beta)c_j$ in (B.6b). Thus, we get:

$$E\left(\sum_{j=1}^{J} c_j\hat{x}_j \mid V_e \text{ and } c_{B_e}\right) + (\beta - 1) \sum_{j \in B_e} c_j\hat{x}_j \leq \beta E(u^* \mid V_e, c_{B_e})$$

Taking expectation on values of $c_{B_e}$ conditioned on $B_e$, we get,

$$\sum_e P(V_e)\left(\sum_{j=1}^{J} c_j\hat{x}_j \mid V_e\right) + (\beta - 1) \sum_{j \in B_e} E(c_j)\hat{x}_j \leq \beta \sum_e P(V_e)E(u^* \mid V_e)$$

$$\sum_{j=1}^{J} E(c_j)\hat{x}_j + (\beta - 1) \sum_{j \in B_e} E(c_j)\hat{x}_j \leq \beta E(u^*)$$

$$\sum_{j \notin B_e} E(c_j)\hat{x}_j + \beta \sum_{j \in B_e} E(c_j)\hat{x}_j \leq \beta E(u^*)$$

$\blacksquare$
Appendix C

Proof of Proposition 1

This proof is based on the optimality proof of the minimum-performance guarantee algorithm in [2].

Let $Q$ be a policy satisfying $C_i^u \geq E\left(R_i\mathbf{1}_{(Q(\bar{R})=i)}\right) \geq C_i^l$ for all $i$. Since $\alpha_i^* \geq 1$ and $\gamma_i^* \geq 0$ by definition,

$$E\left(R_{Q(\bar{R})}\right) \leq E\left(R_Q(\bar{R})\right)$$

$$\quad + \sum_{i=1}^{N} (\alpha_i^* - 1) \left( E\left(R_i\mathbf{1}_{(Q(\bar{R})=i)}\right) - C_i^l \right)$$

$$\quad + \sum_{i=1}^{N} \gamma_i^* \left( C_i^u - E\left(R_i\mathbf{1}_{(Q(\bar{R})=i)}\right) \right)$$

$$= \sum_{i=1}^{N} E\left(R_{Q(\bar{R})}\right)$$

$$\quad + \sum_{i=1}^{N} (\alpha_i^* - 1) E\left(R_i\mathbf{1}_{(Q(\bar{R})=i)}\right) + \sum_{i=1}^{N} \gamma_i^* C_i^u$$

$$\quad - \sum_{i=1}^{N} (\alpha_i^* - 1) C_i^l - \sum_{i=1}^{N} \gamma_i^* E\left(R_i\mathbf{1}_{(Q(\bar{R})=i)}\right)$$

$$= \sum_{i=1}^{N} (\alpha_i^* - \gamma_i^*) E\left(R_i\mathbf{1}_{(Q(\bar{R})=i)}\right) + \sum_{i=1}^{N} \gamma_i^* C_i^u$$

$$\quad - \sum_{i=1}^{N} (\alpha_i^* - 1) C_i^l.$$
Appendix C. Proof of Proposition 1

\[ \sum_{i=1}^{N} (\alpha_i^* - \gamma_i^*) R_i 1_{\{Q_i(\vec{R}) = i\}} \leq \sum_{i=1}^{N} (\alpha_i^* - \gamma_i^*) R_i 1_{\{Q^*(\vec{R}) = i\}}. \]

Therefore,

\[ \mathbb{E}(R_{Q_i(\vec{R})}) \leq \sum_{i=1}^{N} \mathbb{E}(\alpha_i^* - \gamma_i^*) \left( R_i 1_{\{Q^*(\vec{R}) = i\}} \right) \]

\[ + \sum_{i=1}^{N} \gamma_i^* C_i^u - \sum_{i=1}^{N} (\alpha_i^* - 1) C_i^l \]

\[ = \mathbb{E}(R_{Q^*(\vec{R})}) \]

\[ + \sum_{i=1}^{N} (\alpha_i^* - 1) \left( \mathbb{E}(R_i 1_{\{Q^*(\vec{R}) = i\}}) - C_i^l \right) \]

\[ + \sum_{i=1}^{N} \gamma_i^* \left( C_i^u - \mathbb{E}(R_i 1_{\{Q^*(\vec{R}) = i\}}) \right) \]

\[ = \mathbb{E}(R_{Q^*(\vec{R})}) \]

which implies there is no better policy than \( Q^* \).
Bibliography
