Radio Resource Allocation in Emerging Broadband Wireless Access Networks: Some Analytical Models and Their Applications

by

Mohammad Mamunur Rashid

M.Sc. in Electrical & Computer Engineering, University of Manitoba, 2004

B.Sc. in Computer Science & Engineering, Bangladesh University of Engineering & Technology, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
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)

April 2010

© Mohammad Mamunur Rashid, 2010

Abstract

New generation wireless networks are designed not only to carry voice but also to support data-intensive and multimedia applications. Broadband wireless networks offer high bandwidth necessary to support these applications. However, without proper resource allocation schemes, increased bandwidth is not sufficient to meet diverse application quality of service (QoS) requirements. In designing or deploying a resource allocation scheme, it is crucial to understand the inter-relationship of the resource allocation scheme and important system parameters with resulting QoS performance. Analytical models provide an opportunity to derive these relationships in an accurate and readily verifiable way.

In this thesis, we develop novel analytical models for radio resource allocation schemes in emerging broadband wireless access networks. These models are then adopted for in-depth analysis of QoS performance of the modeled schemes and in devising new solutions based on the models to either improve upon or complement those schemes. Our work primarily deals with Medium Access Control layer; however, in most of our contributions, we also consider cross-layer issues. First, we develop a queueing model for a downlink packet scheduling policy in IEEE 802.16e mobile broadband systems and propose a resource allocation framework based on this model. Compared to existing schemes, proposed framework offers a simple yet more effective way to provide QoS to a heterogeneous mix of

applications. Second, we develop a cross-layer model for a prominent multiuser scheduling scheme in multi-antenna-based broadband cellular systems. It captures cross-layer effects of important parameters of the multi-antenna physical layer. The model output is shown to have important applications in QoS provisioning. Next, we perform queueing analysis of controlled channel access mechanism in IEEE 802.11e-based Wireless Local Area Networks. Using the insight gained, we propose a novel channel access scheduling mechanism that achieves very robust performance in meeting QoS guarantees. Finally, we focus on a promising new technology called Cognitive Radio (CR), which can greatly improve spectrum utilization in next generation broadband systems. We develop a queueing model to analyze the performance of an opportunistic spectrum access mechanism in CR networks. The model has important applications including cross-layer analysis and admission control in CR-based broadband networks.

Table of Contents

Abstract . . . . . **ii**

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

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

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

Acknowledgments . . . . . **xiv**

Dedication . . . . . **xv**

Statement of Co-Authorship . . . . . **xvi**

1 Introduction **1**

1.1 Scope, Motivation and Objective of the Thesis 4

1.2 Background and Literature Review 7

1.2.1 IEEE 802.16e Mobile WiMAX Systems 7

1.2.2 Multiuser MIMO for Cellular Broadband Access 13

1.2.3 IEEE 802.11e based WLANs 16

1.2.4 Cognitive Radio for Broadband Wireless Access 20

1.3 Thesis Outline 22
1.4 Bibliography 25

2 A Model-Based Downlink Resource Allocation Scheme for IEEE 802.16e Mobile WiMAX Systems 34
2.1 Introduction 34
2.2 IEEE 802.16e and Its QoS Architecture 38
2.2.1 Overview of IEEE 802.16e Air Interface 38
2.2.2 MAC Layer and QoS Framework 40
2.3 System Model and Related Assumptions 42
2.4 A Queue-Length-Based Packet Scheduling Policy and Its Queueing Analysis 44
2.4.1 Queueing Analytical Model for the Packet Scheduling Policy . . . 46
2.5 Model-Based and Channel-Aware Downlink Resource Allocation 55
2.6 Results 60
2.6.1 Numerical Results 60
2.6.2 Simulation Results 62
2.7 Related Work 74
2.8 Conclusion 77
2.9 Bibliography 79

3 Cross-Layer Analysis of Downlink V-BLAST MIMO Transmission Exploiting Multiuser Diversity 84
3.1 Introduction 84
3.2 System Model and Assumptions 87
3.2.1 MIMO Transmission Model and Receiver Architecture 88

3.2.2 Modeling the MIMO Channel 90
3.2.3 Multiuser Diversity Scheme 91
3.3 Queueing Model and Analysis 92
3.3.1 Formulation as a Quasi-Birth-and-Death Process 96
3.3.2 Matrix-Analytic Solution for Steady State Distributions 100
3.3.3 Derivation of Performance Measures 100
3.4 Numerical Results and Discussions 103
3.4.1 Effect of Number of Transmit and Receive Antennas 105
3.4.2 Effect of Transmit Antenna Correlation 107
3.4.3 Effect of Doppler Spread 111
3.4.4 Effect of Buffer Size, Traffic Intensity, and Number of Users 112
3.4.5 Validation of the Queueing Model 115
3.5 Application of the Queueing Model 116
3.6 Conclusions 118
3.7 Bibliography 119

4 Controlled Channel Access Scheduling for Guaranteed QoS in 802.11e-Based WLANs 123
4.1 Introduction 123
4.2 Background 125
4.2.1 Overview of HCCA 125
4.2.2 Reference HCCA Scheduler Design 126
4.3 Queueing Model and Analysis for the Reference HCCA Scheduler 127
4.3.1 Formulation of the Queueing Model 128
4.3.2 Markov Chain Analysis 130

4.3.3 Matrix-Analytic Solution for Steady State Distributions 134
4.3.4 Performance Measures 134
4.4 The Prediction and Optimization-Based HCCA (PRO-HCCA) Scheduler . . 139
4.4.1 Accounting for Different Delay Bounds 140
4.4.2 Predicting VBR Traffic Intensities 142
4.4.3 Optimizing TXOP Allocation 145
4.5 Performance Evaluation of the PRO-HCCA Scheduler 147
4.5.1 Simulation Setup and Parameters 147
4.5.2 Simulation Results and Discussions 149
4.6 Conclusions 153
4.7 Bibliography 154

5 Opportunistic Spectrum Scheduling for Multiuser Cognitive Radio: A Queueing Analysis 157
5.1 Introduction 157
5.2 The System Model and Assumptions 161
5.2.1 Network Model 161
5.2.2 Primary User Activity 163
5.2.3 Channel Model and Adaptive Modulation 164
5.2.4 Opportunistic Spectrum Scheduling 165
5.3 Formulation of the Queueing Model 166
5.3.1 Packet Arrival Process 166
5.3.2 Joint System State Considering Primary User Activity, Channel State, and Opportunistic Scheduling 167
5.3.3 Quasi-Birth-and-Death Process Formulation and Analysis 170

5.3.4 Matrix-Analytic Solution for Steady State Distributions 175
5.3.5 Derivation of Performance Measures 175
5.4 Numerical Results and Discussions 177
5.4.1 Primary User Activity Level 178
5.4.2 Number of Channels Available for CR Users 179
5.4.3 Channel Fading 180
5.4.4 Number of Admitted CR Users 182
5.4.5 Queueing Model Validation 184
5.5 Application of the Queueing Model 185
5.6 Conclusions 186
5.7 Bibliography 188

6 Conclusions and Future Research Directions 192
6.1 Conclusions 192
6.2 Future Research 195
6.2.1 Models for End-to-End QoS in Multihop Wireless Broadband Networks 196
6.2.2 Cross-Layer Models for Various Multiuser MIMO Systems 197
6.2.3 Analyzing Effects of CQI Feedback Issues 197
6.2.4 Enhanced Queueing and Game Theoretical Models for Spectrum Sharing in CR-based Wireless Broadband Systems 198
6.3 Bibliography 199

List of Tables

Table 2.1 System parameters 64

Table 2.2 Modulation and coding schemes (MCS) used in the simulation 64

Table 3.1 Throughput and packet loss performance for different number of transmit and receive antennas ($f_d = 15$ Hz, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) 105

Table 4.1 Packet overhead and MAC parameters 148

Table 4.2 Properties and QoS requirements of the scheduled streams 148

Table 4.3 Summary of the results 152

Table 5.1 Effect of channel fading on the packet loss rate of CR users. 181

List of Figures

Figure 2.1 IEEE 802.16e frame structure 39

Figure 2.2 A simplified view of the system model 42

Figure 2.3 CDF of packet delay for various values of model parameter x 61

Figure 2.4 Model parameter x vs. packet loss rate 61

Figure 2.5 System-wide packet loss rate for non-real-time services 66

Figure 2.6 Number of supported non-real-time service flows vs. their arrival intensities 67

Figure 2.7 System-wide packet delay violation probability for real-time services 69

Figure 2.8 System-wide packet loss rate for real-time services 69

Figure 2.9 Number of supported service flows (real-time video streams) vs. their bit rates 71

Figure 2.10 Number of supported non-real-time and real-time service flows competing simultaneously 72

Figure 3.1 Packet delay distribution for different number of transmit and receive antennas ($f_d = 15$ Hz, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) 105

Figure 3.2 Packet delay distribution for different transmit antenna correlation ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) 108

Figure 3.3 Effect of Tx antenna correlation on packet loss rate ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) . 108

Figure 3.4 Average throughput for different levels of antenna correlation ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) 109

Figure 3.5 Packet delay distribution for various Doppler spread ($t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) 110

Figure 3.6 Packet loss rate for various fading rates ($t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) 110

Figure 3.7 Average throughput for various fading rates ($t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$) 111

Figure 3.8 Delay distribution for different buffer size and traffic arrival profiles (AP) ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users) 113

Figure 3.9 Effect of the number of users on packet delay distribution ($f_d = 15$ Hz, $Q = 50$, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, $\rho = 0.1$, Arrival profile = AP1) 113

Figure 3.10 MAC layer throughput for different number of users ($f_d = 15$ Hz, $Q = 50$, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, $\rho = 0.1$, Arrival profile = AP1) . 114

Figure 3.11 Effect of the number of competing users on packet loss rate ($f_d = 15$ Hz, $Q = 50$, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, $\rho = 0.1$, Arrival profile = AP1) 114

Figure 4.1 Timing diagram for the reference HCCA scheduler showing different timing periods and their relationships within a beacon interval having two SI periods. 127
Figure 4.2 Cumulative distribution function of access delay (ms) derived from the analytical model. 137
Figure 4.3 Cumulative distribution function (derived from the analytical model) of queue length at the end of a TXOP. 137
Figure 4.4 Cumulative distribution function of delay (ms) experienced by packets generated from the interactive video stream. 149
Figure 4.5 Cumulative distribution function of delay (ms) experienced by packets generated from the non-interactive video stream. 150
Figure 4.6 Cumulative distribution function of delay (ms) experienced by packets generated from the VBR VoIP stream. 151

Figure 5.1 Infrastructure-based cognitive radio network architecture (CR = Cognitive radio; PU = Primary user; BS = Base station). 162
Figure 5.2 Effect of primary user activity level on packet delay distribution of CR users. 179
Figure 5.3 Effect of primary user activity level on packet loss rate of CR users. . 180
Figure 5.4 Effect of number of cognitive channels on packet delay distribution of CR users. 181
Figure 5.5 Effect of number of cognitive channels on packet loss rate of CR users. 182
Figure 5.6 Effect of channel fading on the packet delay distribution of CR users 183
Figure 5.7 Packet delay distribution for different number of CR users in the cognitive radio network. 184

Figure 5.8 Packet loss rate for different number of CR users in the cognitive radio network. 185

Acknowledgments

Many thanks go to all those who were instrumental in the development of this thesis. First, I must thank my supervisor, Professor Vijay K. Bhargava for his great guidance, support and encouragement throughout my academic studies and this research. Second, I would like to thank Professor Ekram Hossain for all his invaluable suggestions and helpful advice. I am grateful to my past and present colleagues in the Information Theory and Systems (ITS) laboratory for providing me with their genuine and friendly support during the course of my PhD research. I would also like to acknowledge with gratitude the support from the University of British Columbia in the form of Graduate Fellowship and a number of other scholarships.

Great appreciation goes to my family, specially to my parents for their unconditional love and for always being behind me. I am profoundly indebted to my loving wife, Zaki Rezoana, for her continued encouragement, support and sacrifice, that have been instrumental in the successful accomplishment of this endeavor.

To My Parents...

Statement of Co-Authorship

I am the primary researcher and author for all the research contributions made in this thesis. I identified the research problems, conducted literature review and performed research to address those problems. Mathematical analysis and formulation of the problems and development of novel schemes were totally carried out by me. I wrote the computer programs for analyzing the mathematical models and for simulating performances of proposed schemes. I also prepared the associated manuscripts for publication. Prof. Ekram Hossain is a co-author for contributions in Chapters 3, 4 and 5. He provided some important feedbacks during the identification and formulation of the research problems addressed in those chapters. In Chapter 3, he provided some input during the derivation of a small part of the mathematical formulation. He also checked the mathematical derivations for correctness and provided some editorial corrections for the associated manuscripts. Dr. Md. J. Hossain is a co-author for the contribution in Chapter 5. I consulted him during identification of the research problem. He also provided some technical feedbacks that helped in improving a part of the mathematical model developed in that chapter. My supervisor Prof. Vijay Bhargava is a co-author for contributions made in Chapters 2, 3, 4 and 5. I consulted him during the identification and formulation of the research problems. He also provided editorial feedbacks during my preparation of the manuscripts for publication.

Chapter 1

Introduction

With the explosive growth of Internet and widespread deployment of multimedia applications, the role of wireless networks has altered significantly. Current wireless communications networks are designed not only to carry voice but also to support data and multimedia applications. Broadband wireless access networks such as WiFi (Wireless Fidelity), WiMAX (Worldwide Interoperability for Microwave Access), beyond 3G (Third Generation) systems offer high bandwidth necessary to support these new applications and services. However, these multimedia and other data-intensive applications come with their own stringent Quality of Service (QoS) requirements for their smooth operation. For example, live video streams need to meet certain latency requirements to avoid choppy video sequences at the viewers' ends; VoIP communications need to fulfil certain latency and packet loss requirement for the involved parties to continue a meaningful conversation. Fulfilling these QoS requirements still remain challenging in broadband wireless systems despite the fact that they are able to offer much higher bandwidth and data rate than those in traditional wireless systems. Many research problems remain open–especially in the area

of radio resource allocation. Studying the relevant issues and challenges and developing solutions to these open problems are therefore very crucial for successful deployment of these emerging broadband wireless systems.

As broadband wireless communications is a recent phenomenon, systems offering broadband wireless access are still undergoing very active research and development stage. IEEE 802.11 standard [1] was finalized for Wireless Local Area Networks (WLAN), which eventually have become the most successful and widely deployed broadband wireless access networks across the globe. The technology for IEEE 802.11 WLANs, also referred as WiFi, is aimed at providing broadband access in the range of 100 meters–typically covering university campuses, hospitals, home and business complexes. Although WiFi technology supports a very limited mobility, it has been mainly targeted towards fixed broadbnad access within a small geographic area. For mobile users, the cellular technologies before the introduction of third generation (3G) cellular systems were designed mainly to carry voice. By offering the capability to simultaneously carry voice and data services, 3G cellular systems have significantly altered mobile wireless access. Personal mobile phones could now be used not only for regular phone conversations but also for accessing email, browsing internet and accessing other data services. However, the data transmission speed offered by this technology was still not enough to offer true broadband quality. 3G+ came into picture as a natural evolution of 3G networks to offer mobile broadband services with much faster speed than 3G using High Speed Packet Access (HSPA) technology. Recently, Multiple Input Multiple Output (MIMO) technology has dramatically increased the data rate and throughput of 3G+ systems by employing multiple transmit and/or receive antennas. Although 3G+ is a significant advance towards realizing wireless broadband access, the technology still remains relatively costly and suffers from a number of limitations in

terms of QoS support and spectral efficiency. WiMAX is another emerging technology that has revolutionalized broadband wireless access. Compared to 3G/3G+, it offers simpler, faster and more cost-effective broadband access. Using mostly cellular architecture, WiMAX can offer such access in the range of few kilometers, typically capable of covering large metropolitan, suburban, or rural areas. The technology is IP-based, comes with QoS support and high spectral efficiency. Although the initial IEEE standard 802.16 [2] was finalized by IEEE in 2001, WiMAX networks are at the early stage of deployment. In spite of all these advancements, underutilization of radio spectrum remains a critical issue because of the static allotment of spectrum among wireless service providers. Cognitive radio (CR) technology [3] has been introduced to implement a new dynamic spectrum sharing paradigm that can dramatically improve spectrum utilization across different broadband access networks. With this technology, valuable spectrums that might remain unused otherwise (such as TV-bands during off-peak hours), can also be harnessed to enhance the performance, capacity and coverage of broadband wireless networks. A new IEEE standard, named IEEE 802.22 [4], is now in development that aims to offer cognitive radio support for broadband wireless access.

The abundance of raw bandwidth that these networks offer by using new transmission technologies, however, itself is not sufficient to support the QoS requirements of the new applications. In these multiuser systems, applications and services associated with different user-ends compete for a common pool of radio resource to meet a diverse set of QoS requirements. The improvement in the raw data rate can be easily squandered if this shared resource is poorly managed and inefficiently allocated among competing users. Besides, the diversity in QoS requirements of different applications necessitates QoS-aware resource allocation schemes. The resource allocation schemes determine how the available

radio resource will be allocated to competing users or services over time. A common design goal for these schemes is to improve utilization of the broadband quality bandwidth by increasing the number of users that can be supported by the system while fulfilling QoS pledges.

1.1 Scope, Motivation and Objective of the Thesis

A common thread in this thesis is the development of novel analytical models for radio resource allocation techniques in a number of emerging broadband wireless access networks. These models are then adopted to achieve two main objectives: in-depth analysis of QoS performance of the modeled radio resource allocation techniques and devising new solutions based on these models to either improve upon existing schemes or to complement the operations of those schemes.

In designing or deploying a resource allocation scheme, it is crucial to understand the intricate inter–relationship among different system parameters, the dynamics of the resource allocation scheme and resulting system performance. Analytical models provide the opportunity to derive these relationships in an accurate and readily verifiable way. These models also allow the system engineers to analyze their choice of design parameters quickly and efficiently without first developing comprehensive simulation programs that might take considerably long time to develop and may be extremely costly as well. The models offer excellent opportunity for optimizing the parameter selection so that those system parameters can be selected that can optimize some performance measures given resource constraints and QoS requirements. For example, some service providers may want to ensure that audio/video over their networks are served within some delay constraints and at the same time wants to maximize the number of users to optimize their revenue poten-

tial. The analytical models can also greatly assist system engineers in capacity planning for their wireless networks by offering time-efficient means for gauging the QoS effects of the number of admitted users in the system. In summary, such analytical models offer more accurate, timely and efficient analysis compared to detailed simulation studies for resource allocation schemes and can greatly expedite the design and optimization of these systems. This remains the primary motivation for our work.

The analytical models, however, need to be designed carefully to consider a number of important factors. For example, the analysis of physical layer performance such as bit-error-rate (BER) itself may not provide meaningful QoS information at user ends because many MAC layer factors such as scheduling and admission control affect the performance experienced by users. On the other hand, performance model of a scheduling algorithm that runs in the MAC layer will be much more accurate and realistic if important physical layer parameters such as channel fading are considered. Although many cross-layer analysis have been done in literature, the researchers vary widely on what to consider as cross-layer. Many researchers have limited their cross-layer analysis to the rate and power adaptation in the wireless links in response to varying channel conditions. However, such analysis fails to cover important MAC layer dynamics and are not truly cross-layer in nature. Other important factors to consider in developing the analytical models include realistic traffic modeling and proper selection of performance measures to analyze. Most of the existing MAC layer-specific or cross-layer analytical models for the wireless systems that we consider in this thesis have assumed unrealistic traffic models and/or buffer dynamics (e.g. infinite buffer space), and have derived performance measures that are often inadequate as QoS measure for many multimedia services. Besides, application of these analytical models have not been well-addressed in these works. These limitations in the existing works

have also motivated us to better address these factors in this thesis.

The scope of this thesis encompasses resource allocation techniques in a number of emerging broadband wireless technologies: IEEE 802.16e WiMAX, IEEE 802.11e WLAN, 3G+ MIMO cellular and CR networks. We mainly focus on the Medium Access Control (MAC) layer because it is responsible for coordinating communication access over the shared wireless medium and hence, plays the pivotal role in resource allocation techniques. Therefore, our work primarily deals with MAC layer techniques such as channel access scheduling, packet scheduling and admission control. In most of our contributions, we also consider cross-layer effects of important physical layer parameters on these MAC layer techniques. However, we do not model or propose power and rate allocation techniques that are usually performed at the physical layer. Instead, in the cross-layer models or schemes proposed in this thesis, we consider certain rate allocation mechanisms implemented in the physical layer and aim to incorporate their effects in the modeling.

The broad objectives of this thesis are to investigate QoS performance of certain resource allocation techniques in emerging broadband wireless systems through analytical modeling and to derive improved or complementary solutions using outputs from these analytical models. In the following, we list very briefly the specific objectives of this thesis:

1. To develop an analytical model for a downlink packet scheduling policy in IEEE 802.16e mobile WiMAX systems, and to design a novel downlink resource allocation scheme based on this analytical model. In designing the proposed scheme, we try to achieve better QoS performance than existing downlink resource allocation schemes for mobile WiMAX systems.

2. To develop an analytical model of a prominent multiuser scheduling scheme for MIMO downlink transmissions in 3G+ cellular systems (e.g. HSDPA) employing

spatial multiplexing and then to show the application of the model in a number of system engineering purposes that include cross-layer analysis and QoS provisioning.

3. To analyze the performance of the resource allocation mechanism for guaranteed QoS provisioning in the IEEE 802.11e WLAN standard through a queueing analytical model and then to apply the insight gained from the model in designing a better channel access scheduling scheme.

4. To analyze opportunistic spectrum access scheduling in cognitive radio based broadband wireless systems through a queueing analytical model and to show how this model can be applied for cross-layer analysis and model-based admission control in CR-based broadband wireless access systems

1.2 Background and Literature Review

In this thesis, we address the resource allocation problem with special emphasis on QoS provisioning in a number of emerging broadband wireless networks. This sections overviews the underlying standards and/or technologies and provides a literature review of related works on resource allocation and QoS provisioning in these broadband systems.

1.2.1 IEEE 802.16e Mobile WiMAX Systems

The first IEEE standard for WiMAX systems was completed in October 2001 and publsihed in April 2002. This initial standard, also known as IEEE 802.16-2001 [2], was proposed for fixed wireless broadband access over a Wireless Metropolitan Area Network (WMAN). The core IEEE 802.16 standard defined the physical layer (PHY) for the 10-66GHz range of (licensed) spectrum. With single carrier quadrature amplitude modulation (QAM) in

the PHY layer, the core IEEE 802.16 system is referred to as WirelessMAN-SC. Non-line-of-sight (NLOS) operation is infeasible in this frequency range. IEEE 802.16a then extended the PHY layer specification to support the 2-11GHz band of licensed and unlicensed spectrum. The standard also removed the requirement for LOS operation. To take into account the effects multipath in a NLOS environment, orthogonal frequency division multiplexing (OFDM) was added to the specification as modulation scheme. The release of 802.16-2004 [5] superseded the earlier 802.16 documents for fixed wireless broadband access. The IEEE 802.16e standard was introduced in 2005 as an amendment to IEEE 802.16-2004 to address user mobility. As such, it is also known as the standard for mobile WiMAX systems. Similar to the core IEEE 802.16 standard, the IEEE 802.16e only defines the architecture, air interface, MAC protocol specifications and a QoS infrastructure.

The Architecture and Air Interface

The IEEE 802.16 standard specifies two modes for sharing the wireless medium: point-to-multipoint (PMP) and mesh (optional). With PMP, the base station (BS) serves a set of stations (SSs) within the same antenna sector in a broadcast manner. Transmissions from SSs are directed to and centrally coordinated by the BS. On the other hand, in mesh mode, traffic can be routed through other SSs and can occur directly among SSs. In this thesis, we focus mainly on the PMP mode. In IEEE 802.16e, the architectures remain similar; however, an SS in the network is now refereed as a mobile station (MS) to account for potential user mobility.

The IEEE 802.16e air interface uses Orthogonal Frequency-Division Multiple Access (OFDMA) as its core technology. The communication over the air interface is frame-based, where a frame is constituted from a number of OFDMA slots that serve as the basic resource

allocation units. Each OFDMA slot combines a certain number of OFDMA symbols allocated in the time domain and subcarriers grouped in subchannels in the frequency domain. How many symbols and subchannels will be used to form an OFDMA slot depends on the subchannelization mechanism, which can be either diversity-based or contiguous. With OFDMA-TDD, an IEEE 802.16e frame is further divided into a downlink (DL)-subframe and an uplink (UL)-subframe. Control information for all connections scheduled in the DL and UL subframes is carried by DL- and UL-submaps that are broadcast as part of a DL-submap. The DL-and UL-MAPs are composed of smaller elements known as information elements (IEs). Each IE describes an MS's allocation and provides information, such as the Modulation and Coding Scheme (MCS), resource size, location of the allocation in the frame.

MAC Layer Specification and QoS Infrastructure

MAC layer in an IEEE 802.16e systems performs several important functions including resource allocation, link adaptation, error recovery, network entry, mobility management as well as security and authentication. The MAC layer of WiMAX systems is connection-oriented, where a transport connection serves as a context for data communications over a service flow between the BS and an MS. For QoS management, the concept of service flow was introduced, which defines the QoS parameters for a unidirectional flow of packets. A service flow is created and its QoS parameters are defined through Dynamic Service Addition (DSA) messages that are exchanged between the BS and the MS associated with the service flow. The service flow and its QoS parameters can subsequently be updated through Dynamic Service Change (DSC) messages. Upon admission into the system, an active service flow is associated with a transport connection–identified by a unique connec-

tion ID (CID). For applications needing bidirectional communication, two service flows and correspondingly, two transport connections are created–one for each direction.

Another concept relevant to QoS support in WiMAX systems is *scheduling service*, which essentially refers to the data handling mechanism associated with a connection. The IEEE 802.16 restricted all service flows, both in downlink and uplink, to four scheduling services: Unsolicited Grant Services (UGS), Real-time Polling Services (rtPS), Non-real-time Polling Services (nrtPS) and Best Effort (BE). UGS is designed to support real-time applications (with strict delay requirements) that generate fixed-size data packets at periodic intervals. Uplink grants are granted by the BS regardless of the current estimation of backlog; hence, UGS connections use the unsolicited granting bandwidth-request mechanism (i.e., UGS connections never request bandwidth). rtPS is designed to support real-time applications (with less stringent delay requirements) that generate variable-size data packets at periodic intervals, such as MPEG video and VoIP with silence suppression. Unlike UGS and rtPS, nrtPS and BE are designed for applications that do not have any specific delay requirement. The main difference between the two is that nrtPS connections are reserved a minimum amount of bandwidth (by means of the minimum reserved traffic rate parameter), which can boost performance of bandwidth-intensive applications (e.g. FTP).

IEEE 802.16e introduced major changes to the definition and scope of scheduling services (section 6.3.5 of IEEE 802.16e-2005 [6]). Although the underlying concept remains similar, association of service flows in the downlink need not confine to a mandatory list of scheduling services. The standard suggests but does not mandate a number of QoS parameters that would provide scheduling service for different types of applications such as real-time and non real-time VBR video, VoIP. In essence, incorporating a flexible interpretation of scheduling services, IEEE 802.16e allows scheduling services of a service flow to

be defined by its stated QoS parameters. Besides the parameters already defined in the standard, these QoS parameters can also include vendor defined parameters. In the uplink, the scheduling services are renamed as *scheduling types* and every service flow in the uplink needs to define its scheduling type from a set of five predefined types: UGS, rtPS, nrtPS, Extended Real-time Polling Service (ertPS) and BE. The scheduling type determines the bandwidth request/grant mechanism for uplink connections.

Resource Allocation Schemes

Overall, from the resourc allocation perspective, the IEEE 802.16e defines the necessary infrastructure, signaling and messaging mechanism, but leaves open the designs of the resource allocation schemes. Previous work on resource allocation on WiMAX systems has evolved over time as the standardization activities have matured. Previous work on resource allocation on WiMAX systems has evolved over time as the standardization activities have matured. For the basic 802.16 standard, which was introduced with single-carrier PHY, the resource allocation problem, however, is very similar to that in TDMA-based cellular systems. This also applies to OFDM PHY, as with IEEE 802.16d, where all subcarriers are allocated as a whole to a particular MS during a frame either in the uplink or the downlink direction. Many of the earlier resource allocation solutions were proposed assuming time division multipexing without focusing on any specific PHY. They are often packet scheduling schemes based on traditional time-domain radio resource scheduling solutions and are customized to make compatible with the scheduling service classes defined in the 802.16 standard. Some of these schemes are designed for specific traffic types such as video, VoIP and targets the relevant service classes. For delay sensitive applications that use rtPS scheduling service, EDF and Modified Largest Weighted Delay First (MLWDF)

were evaluated in previous research. Classical wired network schedulers such as Weighted Round Robin (WRR), WFQ and their variants were also proposed and evaluated for both delay-sensitive rtPS and throughput-guaranteed nrtPS applications. Wongthavarawat and Ganz [7] proposed a hierarchical scheduling scheme where EDF and WFQ are used for intra-calss scheduling of rtPS and nrtPS connections respectively, and a simple priority based scheme is adopted for allocating resource among different service classes. Other notable solutions include a hierarchical scheme proposed by Wang and Dittman in [8], a cross-layer scheme proposed by Liu et al. [9], a joint queueing and optimization based scheme by Niyato and Hossain [10] and a fuzzy logic based approach proposed by Chen et al. [11]

Resource allocation schemes specifically proposed for OFDMA-based WiMAX systems, however, follow two main avenues of research. The first avenue of research considers subcarrier/subchannel assignment along with rate and power adaptation across users mainly to improve system level performance such as capacity and throughput. These solutions have basic similarity with those for OFDMA resource allocation problem in general. Examples of such solutions are [12, 13, 14, 15] and their main limitation is that the goal is to improve system performance but not necessarily to provide specific QoS to applications and service. The second category of solutions involve frame-by-frame scheduling of OFDMA subchannels or slots among admitted service flows or connections. Yahiya et al. [16] proposed a cross-layer scheduling mechanism that assigns priorities to UGS, rtPS and nrtPS connections for each available subchannel. In [17], Mehrjoo et al. proposed a queue- and channel-aware schedular, called opportunistic stable queue scheduling for WiMAX downlink. The scheduler, however, aimed to improve overall throughput and packet dropping probability experienced by MSs in the system. In [18], Ali et al. proposed

a cross-layer scheme where the scheduling priorities of different connections are calculated using channel quality, service class and QoS requirements of the connections. Wan et al. took a similar approach in [19], but resorted to priority-based scheduling to avoid complexity of an optimized solution.

1.2.2 Multiuser MIMO for Cellular Broadband Access

After MIMO technology was introduced, extensive research has been done by many researchers to devise schemes to reach its promised capacity. In traditional single-user point to point MIMO systems, the extra spatial degree of freedom offered by multiple antennas is utilized by advanced transmission schemes such as Space-Time Block Coding (STBC) and advanced receiver techniques to boost signal quality at the receiver and better combat channel errors. In such systems, the MIMO technology can be viewed primarily as a way to boost the physical layer performance. Although the MAC layer and upper layer protocols can benefit from this increased physical layer performance without being aware of the available MIMO capability, the true potential of MIMO cannot be be realized this way. More recent advances in MIMO, therefore, have been more focused on multiuser MIMO, which essentially takes a coordinated design approach between the MAC layer shceduling/resource allocation and the availability of a MIMO physical layer [20] . Information theoretic results in this field have shown that such coordinated design approach is necessary for optimal system performance [21]. In fact, an uncoordinated approach can cause severe performance degradation as can be seen when both layers attempt to extract diversity (e.g. use of channel-hardening [22] single-user space-time codes at the PHY combined with multiuser diversity scheduling at the MAC). Multiuser MIMO provides lot of advantages such as providing multiuser diversity in addition to antenna diversity and user multiplexing

in addition to stream multiplexing. Multiuser MIMO, however, introduces its own set of problems–especially in resource allocation.

Resource Allocation in Multiuser MIMO Systems

Resource allocation problems arise in multiuser MIMO systems in both downlink and uplink. In the downlink, which is often referred as MIMO broadcast channel (BC), the objective is often to reach as close as possible to the information-theoretic capacity limit of the MIMO BC. A concept known as dirty paper coding (DPC) has been shown to make reaching this limit possible–at least in theory [23]. Despite significant theoretical interest, this type of scheme is impractical because of its sheer extremely complex and need for codes that are incompatible with existing communication systems and protocols [24].

A simpler and popular scheme is called spatial multiplexing, which is a suboptimal scheme where BS equipped with multiple antennas sends parallel data streams across its transmit antennas. Each user's data is encoded independently of the others, allowing the use of traditional modulation and coding schemes. One of the most prominent spatial multiplexing systems is known as Vertical-Bell Labs Layered Space-Time (V-BLAST) system [25, 26], which was proposed by Bell Laboratories as an extremely efficient scheme for high capacity communications in wireless environments. Spatial multiplexing, however, requires the interference produced by the transmitter to be mitigated or balanced among all of the users in order to meet their QoS requirements [24]. To this end, a number of receiver architectures, such as the zero-forcing (ZF) receiver, the minimum mean square error (MMSE) receiver and the successive interference canceller (SIC), have been investigated to decode the spatially multiplexed signals over the multiuser MIMO downlink [27]. With a ZF receiver, the basic idea is to drive inter-user interference to zero and the design offers an

attractive low-complexity solution. Moreover, ZF receivers in conjunction with multiuser diversity have been shown to scale identically to optimal DPC strategy when number of users approaches infinity [28, 29]. However, the assumption of infinite user population for such optimal performance is not applicable in practical MIMO systems.

Developing scheduling strategies to best exploit multiuser diversity in multiuser MIMO systems with finite number of users and evaluating the capacity limit achieved by those strategies remain an active area of research. To this end, Shao et. al. proposed opportunistic beamforming where resources are allocated to only one user who has the best equivalent channel created by the multiple antennas and beamforming. In [30], a random beamforming technique was proposed, where the best single user is selected in communication based on the limited feedback information from the users. In [31] the authors studied multiuser diversity gain by selecting a single user or selecting multiple users simultaneously communicating in downlink.

In the multiuser MIMO uplink, the resource allocation is a multiple access problem, which is often referred as space-division multiple access (SDMA). A nearly optimal transmission strategy would be to group a number of users with SDMA within each group and to implement orthogonal multiple access between the groups [32]. Combination of these user subsets can vary over time and in a SDMA/TDMA system a challenge is to determine how users belonging to same subset have to be time-multiplexed as well. There have been numerous proposals [33, 34, 35] mainly on how to select suitably small user subset for simultaneous transmission (multi-user scheduling) in system potentially with a large number of users. Similar to downlink, multiuser diversity is also considered in user selection in uplink scheduling in multiuser MIMO systems [36].

Cross-layer Performance Analysis of Multiuser MIMO Systems

Performance analysis of resource allocation in multiuser MIMO systems has been an active area of research. Heath Jr. et al. in [31] have presented simulation analysis on various ways to exploit the multiuser diversity in multiuser MIMO systems and their effects on the spectral efficiency of the system. Airy et al., on the other hand, have presented analytical models in [37] to compare two opportunistic scheduling techniques in this context. Other notable works on this topic are [38, 39, 40, 41, 42, 43]. In most of the research works, however, the focus has been mainly on the capacity, throughput and/or outage probability achievable at the physical layer. The performance measures are the raw bit rate achievable from the system and the average delays experienced by fixed sized files served from different users. Developing a truly cross-layer analytical model that captures the layered aspect of the system design, derives more comprehensive performance measures such as packet delay statistics, packet throughput and loss rate, and relates these performance to important physical layer, link layer and traffic parameters has not been addresses in the literature. One of our contributions is on such cross-layer modeling.

1.2.3 IEEE 802.11e based WLANs

After its introduction in 1999, the IEEE 802.11 standard became the *de facto* standard for the deployment of WLANs. In subsequent years, it has gone through a number of amendments to address limitations and to incorporate new features and capabilities such as QoS support. In this thesis, we focus mainly on QoS aspects of these networks, which are addressed in the IEEE 802.11e [44] amendment. In this subsection, we therefore start with the basic IEEE 802.11 standard and then gravitate towards the IEEE 802.11e standard.

The Architecture and Air Interface

A basic service set (BSS) is the fundamental building block of the architecture. It is comprised of a group of wireless stations (STAs) encompassing a geographical area called Basic Service Area (BSA). BSA is a similar concept to a cell in a cellular wireless network. According to IEEE 802.11 standard, a WLAN can be configured in two modes, either as an *infrastructure* network or an *ad-hoc network*. In infrastructure networking, all STAs are connected directly to an Access Point (AP) which is connected to a wired network. Communication between devices inside a BSS goes through the AP. An AP can also work as an integration point between multiple BSSs, thus forming an extended service set (ESS). In this situation, an AP connects to a distribution system (DS) that integrates multiple BSSs. The DS can be thought of as a backbone network and it could even be wired. In an ad-hoc mode, there is no AP and STAs can form a network "on the fly". The STAs communicate directly with each other. The network topology changes dynamically as mobile nodes come in and out of the network.

For the physical layer, three different alternatives were proposed: infrared PHY (never implemented widely), frequency hopping spread spectrum (FHSS) PHY and direct sequence spread spectrum (DSSS) PHY. The PHY specification, however, was superseded by three revisions namely 802.11b, 802.11g and 802.11a. The 802.11b uses DSSS in the PHY and offers a maximum data rate of 11 Mbps in the unlicensed 2.4GHz radio band. On the other hand, 802.11g and 802.11a both use OFDM in the PHY and support upto 54 Mbps data rate in the unlicensed radio band of 2.4 GHz and 5 GHz respectively.

The IEEE 802.11e standard developed by 802.11 Task Group E proposes to enable the much needed QoS support , which is absent in the basic 802.11 standard. The BSS is now called QBSS with QoS access point (QAP) and QoS stations (QSTAs).

MAC Layer Specification and Functionalities

The IEEE 802.11 standard specifies two MAC protocols: Distributed Coordination Function (DCF) and the optional Point Coordination Function (PCF). An interval between two successive beacon frames consists of an optional Contention Free Period (CFP) followed by a Contention Period (CP). CFP occurs only if PCF is enabled. DCF is a distributed access control protocol based on the carrier sensing multiple access with collision avoidance (CSMA/CA). The centrally-coordinated access mechanism of the IEEE 802.11 MAC, called the PCF, adopts a poll-and-response protocol to control the access to the shared wireless medium and eliminate contention among wireless stations.

The IEEE 802.11e introduces QoS support through the introduction of Hybrid Coordination Function (HCF). HCF offers channel access through two MAC schemes: Enhanced Distributed Channel Access (EDCA) for prioritized QoS and HCF Controlled Channel Access (HCCA) for parameterized QoS provisioning. One main feature of HCF is to introduce four access category (AC) queues and eight traffic stream (TS) queues at MAC layer. Another main feature of the HCF is the concept of transmission opportunity (TXOP), the time interval permitted for a particular STA to transmit packets. During the TXOP, there can be a series of frames transmitted by an STA separated by short interframe space (SIFS). EDCA is aimed at service differentiation which is achieved by using different IFS sizes and allocating different CW sizes for different ACs. HCCA improves over the Point Coordination Function (PCF) of legacy 802.11 MAC. Similar to PCF, HCCA is a polling based mechanism where the access to the medium is arbitrated centrally. However, it resolves the limitation of PCF by eliminating the unpredictability of beacon delays arising from incompatible cooperation between Contention Period (CP) and Contention Free Period (CFP) [45].

Resource Allocation Schemes for IEEE 802.11e WLANs

A few research efforts have addressed the scheduler and admission controller design in the context of IEEE 802.11e WLANs. The scheduler design is relevant to HCCA and is not necessary for EDCA, which itself is designed to schedule medium access for competing traffic.

Ni et al. proposed a fair scheduling algorithm named fair HCF (FHCF) in [46]. A QAP scheduler estimates the queue length for each QSTA before the next SI. Based on these estimated queue lengths, the QAP adapts the computation of TXOPs accordingly. A node scheduler then redistribute the unused time among its different traffic streams. For EDCA, schemes proposed in [47, 48] are examples of measurement-based admission control schemes. In [47], a threshold-based admission controller for EDCA has been proposed. The QSTAs monitor the wireless link for traffic conditions and take individual admission/rejection decisions based on the measured relative occupied bandwidth or average collision rates. Although the scheme is simple to implement, setting the proper threshold values or optimizing them is very difficult. Xiao and Li [48] have proposed a mechanism, what they call global parameter control, to achieve performance protection for video and audio applications from data services. Measurement based schemes, however, suffer from performance oscillations. Among model based schemes for EDCA, Pong and Moors in [49] propose an admission controller for EDCA using a Markov Chain Analysis. In [50], Gao and Cai propose an admission control scheme called Physical-Rate-Based Admission Control (PRBAC) for HCCA. The idea is to consider the variance of physical rates of the wireless channels due to background noise, pathloss, multipath effects instead of a fixed physical rate when admitting new flows. The scheme works for CBR traffic and have the same limitations with VBR traffic as those with the reference admission con-

troller. In the admission control scheme proposed for VBR traffic in [51], the key idea is to introduce a new variable, effective TXOP (Te), to replace TXOP. Te is defined as the necessary TXOPs which can statistically guarantee that the packet loss probability is less than a threshold.

1.2.4 Cognitive Radio for Broadband Wireless Access

Cognitive radio is a recent development which aims to achieve high spectrum efficiency– very essential for broadband wireless access, given the scarcity and high cost of available radio spectrum. Massive underutilization of radio spectrum is well documented under traditional static spectrum allocation policy [52, 53] and is a major hurdle in achieving widespread deployment of broadband wireless access. Cognitive radio shifts the spectrum allocation paradigm from static to dynamic by allowing unlicensed secondary users (SUs) to opportunistically access frequency bands originally licensed to some primary users (PUs). A CR network is thus formed by a number of SUs that use sophisticated spectrum sensing and dynamic access techniques to utilize resource holes from PU networks. Research interest in CR technology has grown tremendously off late and a reasonable number of ensuing research efforts have dealt with resource allocation issues and performance analysis of CR systems.

Opportunistic Spectrum Access and Performance Modeling

In a CR network, the availability of radio spectrum for CR users depends mainly on the activity of primary users. On the other hand, since wireless channel quality of CR users is time-varying in nature, communication resource available to CR users is highly dynamic. How the secondary users can access the opportunistically available spectrum most effectively in such dynamic scenarios with minimal interference to the primary users is a chal-

lenging problem. The access schemes not only have implications for interference to primary users, but also can drastically affect the QoS experienced by applications running at the secondary user ends. The primary users have the preemptive right to reoccupy the spectrum opportunistically accessed by secondary users and therefore, can reoccupy the channels accessed by secondary user anytime; such preemptions would mean potential throughput degradation, increased packet loss and connection dropping. The opportunistic spectrum access in CR networks, therefore, demands careful design considerations and has received considerable attention from wireless communications researchers. For a detailed surevey on the spectrum access solutions for CR networks, we refer to [54]. Since MAC layer controls access to the available spectrum, designing MAC schemes adapted for opportunistic spectrum access has also been investigated in CR literature. A number of MAC protocols for CR networks have been proposed in literature. An in-depth survey of the MAC schemes for CR networks can be found in [55].

Besides spectrum access schemes design, performance modeling of CR networks is also a very active area of research. In most of the cases, however, the focus of the performance analysis has been on the capacity, throughput and/or outage probability achievable at the physical layer (PHY). In [56], Devroye et al. presented an information theoretic analysis to derive the capacity regions of a cognitive radio channel. Information theoretic limit of throughput in cognitive networks was also analyzed in [57] and [58] assuming different cooperative behaviors between primary and CR users. Game theoretic models have also been used to analyze CR networks. Ji and Liu provided a review of theses game theoretic models in [59]. These models are useful to analyze achievable payoffs and associated fairness of service among CR users resulting from different spectrum sharing policies in CR networks.

For CR networks, cross-layer performance modeling is especially important because of networking functionalities in these networks directly depend on the properties of the spectrum band in use and the activity of primary users in those bands. The MAC layer protocols in a CR network should be able to adapt to this very dynamic availability of the communication resources. An adaptive MAC layer, which efficiently uses communication resources given the QoS requirements, is an important component for an adaptive wireless access system based on CR technology. Among the notable research works done on such cross-layer modeling, Simeone et al. [60] studied MAC layer throughput for a CR user over a cognitive link given a certain throughput at the primary user end. The system model, however, is very simplistic with a single CR user opportunistically sharing the cognitive link with a single primary user. The queueing model assumes infinite transmission queues and does not study delay or packet loss performance measures. In [61], an M/G/1 queueing model was used to analyze the transmission performance of multiple CR users on an aggregate basis considering both PHY and MAC layer parameters. Analytical models were developed in [62] for cognitive MAC protocols with dedicated control channel and with embedded control channel. Call-level QoS performance (e.g., blocking performance) for CR users was analyzed in [63] based on a G/M/N queueing model. In a similar spirit, call-level performances for both primary users and CR users were analyzed in [64] using a two-dimensional Markov chain. A cross-layer design methodology for CR networks based on fuzzy logic was presented in [65].

1.3 Thesis Outline

In chapter 2, we address the resource allocation problem in the downlink of a mobile WiMAX system and propose a novel mode-based resource allocation framework to meet

QoS guarantees for both real-time and non-real-time applications. At first, we develop a queueing model to link important performance measures of downlink service flows to a set of tunable parameters. We show how these parameters could be set to appropriate values to meet the QoS performances sought by the admitted service flows. We then introduce a resource allocation scheme that uses these parameter values for scheduling packets among these service flows. Inside the proposed queue- and channel-aware scheme, the queue-length-based packet scheduler is complemented by a cross-layer OFDMA slot allocation mechanism that adapts to channel conditions at the destination mobile stations (MSs). Compared to existing schemes, our model-based approach offers a simple yet more effective way to provide QoS to a heterogenous mix of applications. It also offers greater flexibility to service providers by allowing probabilistic delay guarantees to time-critical multimedia applications. Besides, unlike many existing schemes, our proposed solution is designed to be compatible with the updated definition of some key resource allocation concepts in IEEE 802.16e. The cross-layer aspect of the scheme ensures efficient resource utilization in presence of link adaptations due to mobility and channel fading. We also present our simulation results to show performance benefits of the proposed scheme in providing QoS support for both real-time and non-real-time applications in mobile WiMAX systems.

In chapter 3, we develop a queuing analytic model for cross-layer analysis of a downlink resource allocation scheme that exploits multiuser diversity in a V-BLAST MIMO-based cellular broadband system. The queuing model is developed considering the system dynamics resulting from the prominent resource allocation scheme, fading in the MIMO channel, and the traffic arrival process. The transmit antenna correlation in the base station and the possibility of bursty traffic arrival are also factored into the model. The queuing

model is then shown to have important applications in the QoS provisioning framework for a V-BLAST MIMO system as the model output can be used in admission control, QoS negotiation and setting proper values for QoS relevant parameters.

In chapter 4, we formulate an analytical model of the controlled channel access mechanism, also called HCCA, of 802.11e based WLANs considering VBR traffic applications. The model enables us to derive important performance measures in terms of the parameters used in the reference scheduler and admission controller. It also shows the underlying cause of the critical weakness of the reference scheduler in providing QoS for VBR applications. Using the insights gained from the model, we then develop a novel channel access scheduler for IEEE 802.11e WLANs. By simulation analysis, we then show that the proposed scheme, referred as PRO-HCCA scheduler, achieves very robust and stable performance in meeting the QoS guarantees for different type of VBR traffic applications

In chapter 5, we develop a queueing analytical model to study the performance of opportunistic spectrum access by CR users in a dynamic spectrum access network. For resource allocation, a channel-quality-based opportunistic scheduling scheme is considered. In formulating the analytical model, we performed stochastic modeling of primary users' activity and CR users' channel quality variation over time. The proposed model is shown to have important applications including cross-layer analysis and design for system engineering, and model-based admission control in CR based broadband systems.

In chapter 6, we summarize our contributions, conclude the thesis and highlight potential future research directions.

1.4 Bibliography

[1] IEEE 802.11, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," Aug. 1999.

[2] IEEE 802.16-2001, "IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Apr. 2002.

[3] J. Mitola and G. Q. Maguire,"Cognitive radio: Making Software Radios More Personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13-18, Aug. 1999.

[4] http://www.ieee802.org/22/, retrieved Dec. 2009.

[5] IEEE 802.16-2004, "IEEE standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Oct. 2004.

[6] IEEE 802.16e-2005, "IEEE Standard for Local and Metropolitan Networks – Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1," Feb. 2006.

[7] K. Wongthavarawat and A. Ganz, "Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems," *International Journal of Communication Systems*, vol. 16, no. 1, pp. 81-96, Feb. 2003.

[8] H. Wang and L. Dittmann, "Adaptive Radio Resource Allocation in Hierarchical QoS Scheduling for IEEE 802.16 Systems," *Proc. IEEE Global Telecommunications Conference*, pp. 4769-4774, Washington, D.C., Nov. 2007.

[9] Q. Liu, X. Wang, and G. B. Giannakis, "A Cross-Layer Scheduling Algorithm with QoS Support in Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 839-847, May 2006.

[10] D. Niyato and E. Hossain, "A Queuing-Theoretic and Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks," *IEEE Transactions on Computers*, vol. 55, no. 11, pp. 1473-1488, Nov. 2006.

[11] C-L. Chen, J-W. Lee, C.-Y. Wu, and Y.-H. Kuo, "Fairness and QoS Guarantees of WiMAX OFDMA Scheduling with Fuzzy Controls," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Article ID 512507, 14 pages, Jan. 2009.

[12] A. Biagioni, R. Fantacci, D. Marabissi, and D. Tarchi, "Adaptive Subcarrier Allocation Schemes for Wireless OFDMA Systems in Wimax Networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 2, pp. 217-225, Feb. 2009.

[13] G. Miao and N. Himayat, "Low Complexity Utility Based Resource Allocation for 802.16 OFDMA Systems," *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1465-1470, Las Vegas, NV, Mar.-Apr. 2008.

[14] V. Singh and V. Sharma, "Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks," *Proc. IEEE Wireless Communication and Networking Conference*, vol. 2, pp. 984-990, Las Vegas, NV, Apr. 2006.

[15] C. Tarhini and T. Chahed, "AMC-Aware QoS Proposal for OFDMA-Based IEEE 802.16 WiMAX Systems," *Proc. IEEE Global Telecommunications Conference*, pp. 4780-4784, Washington, D.C., Nov. 2007.

[16] T.A. Yahiya, A.-L. Beylot, and G. Pujolle, "Cross-Layer Multiservice Scheduling for Mobile WiMAX Systems," *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1531-1535, Las Vegas, NV, Mar.-Apr. 2008.

[17] M. Mehrjoo, X. Shen, and K. S. Naik, "A Joint Channel and Queue-Aware Scheduling for IEEE 802.16 Wireless Metropolitan Area Networks," *Proc. IEEE Wireless Communications and Networking Conference*, Hong Kong, pp. 1549-1553, Mar. 2007.

[18] N.A. Ali, M. Hayajneh, and H. Hassanein, "Cross Layer Scheduling Algorithm for IEEE 802.16," *Proc. IEEE International Conference on Communications*, pp. 3858-3862, Beijing, China, May 2008.

[19] L. Wan, W. Ma and Z. Guo, "A Cross-Layer Packet Scheduling and Subchannel Allocation Scheme in 802.16e OFDMA System," *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1865-1870, Hong Kong, Mar. 2007.

[20] D. Gesbert, M. Kountouris, R. W. Heath, Jr., C.-B. Chae, and T. Salzer, "Shifting the MIMO Paradigm: From Single User to Multiuser Communications," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 36-46, Oct. 2007

[21] A. El Gamal and T.M. Cover, "Multiple user information theory," *Proc. IEEE*, vol. 68, no. 12, pp. 1466–1483, Dec. 1980.

[22] B. Hochwald, T. Marzetta, and V. Tarokh, "Multi-Antenna Channel Hardening and Its Implications for Rate Feedback Andscheduling," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 18931909, Sept. 2004.

[23] M. Costa, "Writing on Dirty Paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439441, 1983.

[24] Q. H. Spencer and A. L. Swindlehurst, "A Hybrid Approach to Spatial Multiplexing in Multiuser MIMO Downlinks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2004, no. 2, pp. 236247, Dec. 2004.

[25] G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection Algorithm and Initial Laboratory Results using the V-BLAST Space-Time Communication Architecture," *Electronics Letters*, vol. 35, no. 1, pp. 14–15, 1999.

[26] G. J. Foschini, G. D. Golden, P. W. Wolniansky, and R. A. Valenzuela, "Simplified Processing for Wireless Communication at High Spectral Efficiency Employing Multi-Element Arrays," *IEEE Journal on Seleced Areas in Communications*, vol. 17, no. 11,pp. 1841–1852, 1999.

[27] C.–J. Chen and L.–C. Wang, "Performance Analysis of Scheduling in Multiuser MIMO Systems with Zero-Forcing Receivers," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1435–1445, Sept. 2007.

[28] D. P. Palomar, M. A. Lagunas, and J. M. Cioffi, "Optimum Linear Joint Transmit-Receive Processing for MIMO Channels with QoS Constraints," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1179–1197, 2004.

[29] D. P. Palomar and M. A. Lagunas, "Joint Transmit-Receive Space-Time Equalization in Spatially Correlated MIMO Channels: A Beamforming Approach," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 730-743, 2003.

[30] J. Chung, C. S. Hwang, K. Kim, and Y. K. Kim, "A Random Beamforming Technique in MIMO Systems Exploiting Multiuser Diversity," *IEEE Journal on Selected Areas in Communications*, vol. 21, pp. 848–855, Jun. 2003.

[31] R. W. Heath, Jr., M. Airy, and A. J. Paulraj, "Multiuser diversity for MIMO wireless systems with linear receivers," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Oct. 2001.

[32] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, New York, 2005.

[33] R. Zhang, "Scheduling for Maximum Capacity in SDMA/TDMA Networks," *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, vol. 3, pp. 2141–2144, Orlando, FL, USA, May 2002.

[34] S. Serbetli and A. Yener, "Time Slotted Multiuser MIMO Systems: Beamforming and Scheduling Strategies," *EURASIP Journal on Wireless Communications and Networks, Special Issue on Multiuser MIMO Networks,* pp. 286-296, Dec. 2004.

[35] H. Yin and H. Liu, "Performance of Space-Division Multiple Access (SDMA) With Scheduling," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 611–618, Oct. 2003.

[36] W. Ajib and D. Haccoun, "An Overview Of Scheduling Algorithms in MIMO-Based Fourth-Generation Wireless Systems," *IEEE Network*, vol. 19, no. 5, pp. 43–48, Sept.–Oct. 2005

[37] M. Airy, S. Shakkottai, and R. W. Heath, Jr., "Limiting queuing models for scheduling in multi-user mimo wireless systems," in *Proc. IASTED Conference on Communications, Internet & Information Technology*, Scottsdale, USA, Nov. 2003.

[38] C.-J. Chen, and L.-C. Wang, "A Unified Capacity Analysis for Wireless Systems With Joint Multiuser Scheduling and Antenna Diversity in Nakagami Fading Channels," *IEEE Transactions on Communications*, vol. 54, no. 3, pp. 469–478, Mar. 2006

[39] C. K. Sung, S. H. Moon, J. Choe-i, and I. Lee, "Performance Analysis of Multiuser MIMO Systems with Zero Forcing Receivers," *Proc. IEEE Vehicular Technology Conference*, Dublin, Ireland, Apr. 2007

[40] C. Pietsch, S. Sand, W. G. Teich, J. Lindner, "Modeling and Performance Evaluation Of Multiuser MIMO Systems Using Real-Valued Matrices," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 744–753, Jun. 2003

[41] X. Zhang; Z. Lv, and W. Wang, "Performance Analysis of Multiuser Diversity in MIMO Systems with Antenna Selection," *IEEE Transactions on Wireless Communications*, vol. 7, no. 1, pp. 15–21, Jan. 2008

[42] M. Torabi, W. Ajib, and D. Haccoun, "Performance Analysis of Multiuser MIMO Systems with Scheduling and Antenna Selection," *Proc. IEEE Vehicular Technology Conference*, Singapore, May 2008.

[43] G. Dimic and N. D. Sidiropoulos, "On Downlink Beamforming With Greedy User Selection: Performance Analysis and A Simple New Algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, Oct. 2005.

[44] IEEE 802.11e-2005, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements," Nov. 2005.

[45] Q. Ni, L. Romdhani, and T. Turletti, "Survey of QoS enhancements for IEEE 802.11 wireless LAN," *Wiley Journal of Wireless Communications and Mobile Computing*, vol. 4, no. 5, pp. 547-566, 2004.

[46] P. Ansel, Q. Ni, and T. Turletti, "FHCF: A Simple and Efficient Scheduling Scheme for IEEE 802.11e," *Springer/Kluwer Journal on Mobile Networks and Applications*, vol. 11, no. 3, pp. 391–403, 2006.

[47] D. Gu and J. Zhang, "A New Measurement-Based Admission Control Method for IEEE 802.11 Wireless Local Area Networks," *Proc. IEEE Personal, Indoor and Mobile Radio Communications Symposium*, Beijing, China,Sept. 2003.

[48] Y. Xiao and H. Li, "Voice and Video Transmissions With Global Data Parameter Control for The IEEE 802.11e Enhance Distributed Channel Access,"*IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 11, pp. 1041–1053, Nov. 2004.

[49] D. Pong and T. Moors, "Call Admission Control for IEEE 802.11 Contention Access Mechanism," *Proc. IEEE Global Communications Conference*, San Francisco, CA, Dec. 2003.

[50] D. Gao and J. Cai, "Admission Control With Physical Rate Measurement for IEEE 802.11e Controlled Channel Access," *IEEE Communications Letters*, vol. 9, no. 8, pp. 694–696, Aug. 2005.

[51] W.F. Fan, D.Y. Gao, D.H.K. Tsang, and B. Bensaou, "Admission Control for Variable Bit Rate Traffic in IEEE 802.11e WLANs," *Proc. IEEE LANMAN'04*, Apr. 2004.

[52] Federal Communications Commission, "Spectrum Policy Task Force," Rep. ET Docket no. 02-135, Nov. 2002.

[53] R. W. Broderson, A. Wolisz, D. Cabric, S. M. Mishra, and D. Willkomm, "CORVUS: A cognitive radio approach for usage of virtual unlicensed spectrum," *white paper* submitted at the University of Berkeley, CA, Jul. 2004.

[54] I. F. Akyildiz, Won-Yeol Lee, M. C. Vuran, and S. Mohanty, "A Survey on Spectrum Management in Cognitive Radio Networks," *IEEE Communications Magazine*, Vol. 46, Issue 4, pp. 40–48, Apr. 2008.

[55] D. Niyato and E. Hossain, "Medium Access Control Protocols for Dynamic Spectrum Access in Cognitive Radio Networks: A Survey," invited chapter in *Cognitive Radio Networks*, (Eds. Y. Xiao and F. Hu), Auerbach Publications, CRC Press, 2008.

[56] N. Devroye, P. Mitran and V. Tarokh, "Achievable Rates in Cognitive Radio," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 1813–1827, May 2006.

[57] S. A. Jafar and S. Srinivasa, "Capacity Limits Of Cognitive Radio With Distributed and Dynamic Spectral Activity," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 529–537, Apr. 2007.

[58] S. Srinivasa and S. A. Jafar, "The Throughput Potential Of Cognitive Radio: A Theoretical Perspective," *IEEE Communications Magazine*, vol. 45, pp. 73–79, May 2007.

[59] Z. Ji and K. J. R. Liu, "Dynamic Spectrum Sharing: A Game Theoretical Overview," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 88–94, May 2007.

[60] O. Simeone, Y. Bar-Ness and U. Spagnolini, "Stable Throughput of Cognitive Radios with and Without Relaying Capability," *IEEE Transactions on Communications*, vol. 55, no. 12, pp. 2351–2360, Dec. 2007.

[61] S. Shankar, "Squeezing The Most out of Cognitive Radio: A Joint MAC/PHY Perspective," *Proc. IEEE ICASSP'07*, Honolulu, Hawaii, USA, Apr. 2007.

[62] L.-C. Wang, Y.-C. Lu, C.-W. Wang, and D. S. L. Wei, "Latency Analysis for Dynamic Spectrum Access in Cognitive Radio: Dedicated or Embedded Control Channel?," *Proc. Personal, Indoor and Mobile Radio Communications Symposium* , Athens, Greece, Sept. 2007.

[63] S. Keshavamurthy and K. Chandra, "Multiplexing Analysis for Dynamic Spectrum Access," *Proc. IEEE* , Washington, DC, USA, Oct. 2006.

[64] S. Tang and B. L. Mark, "Performance Analysis of a Wireless Network with Opportunistic Spectrum Sharing," *Proc. IEEE Global Communications Conference*, Washington, DC, USA, Nov. 2007.

[65] N. Baldo and M. Zorzi, "Fuzzy Logic for Cross-Layer Optimization in Cognitive Radio Networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 64-71, Apr. 2008.

Chapter 2

A Model-Based Downlink Resource Allocation Scheme for IEEE 802.16e Mobile WiMAX Systems[1]

2.1 Introduction

WiMAX is an emerging broadband wireless technology that holds the promise of dramatically expanding the reach of broadband Internet services while eliminating huge deployment costs associated with last-mile wired connections to client ends. The technology, first standardized in the IEEE 802.16 standard [1], can offer broadband quality data rate necessary to support wide range of multimedia applications and services that are currently generating perhaps the major portion of traffic carried over the Internet. After its introduction, the IEEE 802.16 standard has gone through several amendments and enhancements.

[1] A version of this chapter has been submitted for publication. Rashid, M.M. and Bhargava, V.K. (2009) A Model-Based Downlink Resource Allocation Scheme for IEEE 802.16e Mobile WiMAX Systems.

The initial IEEE 802.16 standard was designed for fixed broadband services. WiMAX then ventured into the mobile broadband service though the introduction of IEEE 802.16e [2], which has introduced many advanced features to better handle user mobility. Among them, the OFDMA-based air interface offers better robustness against multipath fading experienced by mobile users and enables efficient resource multiplexing in a multiuser system.

From its onset, the IEEE 802.16 standard came with a Quality of Service (QoS) support framework. This is especially important for WiMAX systems as they are expected to support wide range of multimedia applications such as Voice over IP (VoIP), streaming video, network gaming that require certain levels of QoS for their smooth operation. Besides, WiMAX service providers would also like to offer different levels of QoS support to their clients according to their subscriptions. With IEEE 802.16e, the basic QoS framework remains intact, but new resource allocation issues have emerged with OFDMA at the air interface and for the need to support QoS in presence of user mobility. The standard, however, only defines the QoS infrastructure and the basic signaling mechanisms to support QoS, but leaves open the designs of necessary resource management functionalities such as packet scheduling, OFDMA resource allocation and admission control. The IEEE 802.16e standard also adds a number of amendments to the specifications of QoS classes and their associated scheduling services, especially for downlink transmissions.

In this chapter, we focus on a resource allocation framework that comprises of packet scheduling and OFDMA resource allocation with the goal of QoS provisioning in IEEE 802.16e systems. We design the framework to deal with various types of QoS-constrained multimedia and data applications. Most of these applications generate variable-bit-rate (VBR) traffic and require their diverse QoS requirements to be met in presence of other competing VBR as well as constant-bit-rate (CBR) applications. Due to their dynamic traf-

fic characteristics, VBR applications are more challenging to provide QoS for than their CBR counterpart. Hence, we aim to provide a framework that is primarily designed for QoS support of VBR applications, but can also be configured with ease to handle CBR applications. The QoS framework in IEEE 802.16e provides QoS on a per-service-flow basis, where a service flow defines a unidirectional flow of packets usually generated by a single application or service. The signaling mechanism allows to specify traffic parameters as well as QoS requirements of individual service flows seeking admission into the system. Each admitted service flow is then assigned to a transport connection and is allocated resource inside WiMAX frames, which are comprised of a limited number of OFDMA slots. In our proposed resource allocation framework, a joint packet scheduling and OFDMA slot allocation scheme determines how many packets have to be scheduled for each backlogged connection during a frame time and how the available OFDMA slots need to be allocated to each connection for transmitting those scheduled packets. The objective is to meet QoS guarantees for the admitted service flows and the constraint is the limited number of OFDMA slots that are available in the WiMAX frames. How many slots should be allocated in a frame to different connections also depends on the link quality of the destination mobile station (MS) and the signaling overhead involved in transmitting the allocation information. Therefore, a cross-layer adaptive resource allocation scheme is desirable and pursued.

More specifically, we propose a model-based resource allocation framework that builds on the queueing model of a simple queue-length-based packet scheduling policy. First, we develop the queueing model that links important performance measures of a particular service flow to the value of a *model parameter* associated with its connection. Each downlink connection transporting a QoS-constrained service flow is associated with a model param-

eter. We then propose a method that utilizes this queueing model to set appropriate values for these model parameters, given the QoS requirements and arrival statistics of the admitted service flows. Finally, we propose the resource allocation scheme that adopts these model parameter values along with the signaling overhead and channel quality feedbacks in scheduling packets and allocating OFDMA slots to these QoS connections in the downlink.

Compared to existing WiMAX resource allocation schemes, which we overview in section 2.7, our model-based approach has several advantages. Instead of complexities involved with installing separate scheduling schemes within the same resource allocation framework for real-time and non-real-time traffic, the model-based approach offers an efficient unified way to handle both types of traffics and their diverse QoS requirements. Many of the existing schemes for IEEE 802.16e continue to assume that a number of predetermined service classes are mandatory and aims to provide intra-class or inter-class downlink resource allocation to provide QoS specific to those service classes. Some researchers have proposed hierarchical designs that combine both intra- and inter-class resource allocation; the inter-class component is usually priority-based that might cause applications belonging to a lower- priority service class to lack enough resource to fulfil QoS requirements. Besides, the hierarchical solutions can significantly increase implementation complexity of the resource allocation module. We note that the IEEE 802.16e standard no longer has provisions for mandatory service classes for the downlink and offers much more flexible specifications of traffic and QoS parameters. Therefore, the existing solutions might be incompatible and inefficient for IEEE 802.16e mobile WiMAX systems. We, in contrast, design our proposed scheme according to the flexible interpretation of service classes in IEEE 802.16e. The scheme also offers flexibility to service providers by allowing different probabilistic guarantees to be made for delay experienced by packets belonging to connections

serving different customer classes (e.g. platinum, gold, silver). Besides, the cross-layer design of the resource allocation scheme ensures robustness and efficient resource utilization in presence of user mobility and channel fading. In addition, our proposed scheme is designed to meet QoS of individual applications and is more comprehensive than some existing resource allocation schemes that focus mainly on improving overall system performance of OFDMA-based WiMAX systems.

The rest of the chapter is organized as follows. In section 2.2, we provide a brief overview of IEEE 802.16e standard focusing mainly on the air interface and QoS architecture. We discuss the system model and assumptions in section 2.3. We introduce a queue-length-based scheduling policy and its queueing analytical model in section 2.4, which forms the basis of our proposed resource allocation scheme. We describe our proposed resource allocation scheme in section 2.5. Next, we present and discuss performance evaluation results of our proposed scheme in section 4.5. A brief summary of related work is provided in Section 2.7. Section 6 concludes the chapter.

2.2 IEEE 802.16e and Its QoS Architecture

2.2.1 Overview of IEEE 802.16e Air Interface

The IEEE 802.16e air interface is OFDMA-based, where OFDMA symbols are allocated in the time domain and subcarriers grouped in subchannels are alocated in the frequency domain. Combination of certain number of symbols and subchannels constitutes an OFDMA slot. How many symbols and subchannels will be used to form an OFDMA slot depends on the subchannelization mechanism, which can be either diversity-based or contiguous. Among the subchannelization schemes, Partial Usage of Subchannels (PUSC) and Full

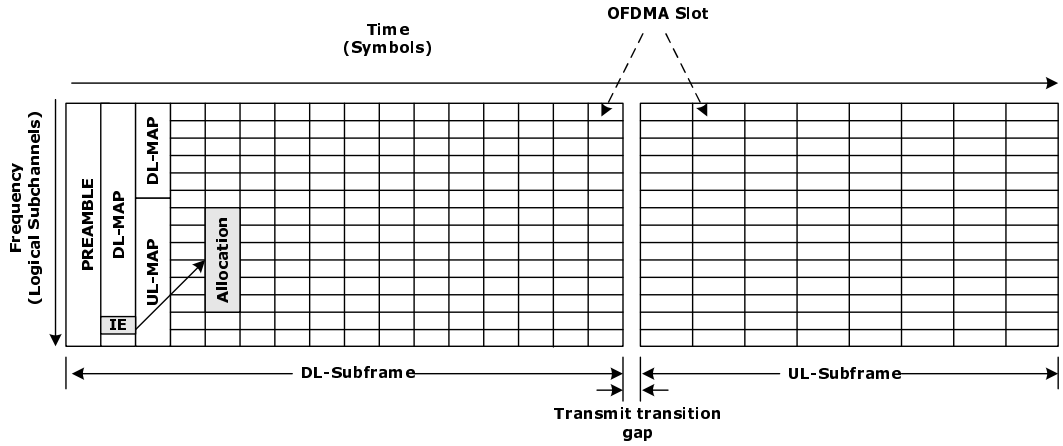


Figure 2.1: IEEE 802.16e frame structure

Usage of Subchannels (FUSC) are diversity-based and Adaptive Modulation and Coding (AMC) is contiguous subchanelization. In PUSC, which is mandatory for both downlink and uplink, logical subchannels are formed from randomly chosen subcarriers across the OFDM spectrum. A slot is constructed using two OFDMA symbols across one subchannel in PUSC and using one symbol across two subchannels in FUSC. On the other hand, AMC forms subchannels by grouping contiguous subcarriers and offers the opportunity for utilizing multiuser diversity. In general, AMC is desired for limited mobility or fixed WiMAX systems and PUSC/FUSC is recommended for mobile WiMAX.

Frame Structure

The IEEE 802.16e OFDMA air interface communication is frame-based. With OFDMA-TDD, an IEEE 802.16e frame is divided into a downlink (DL)-subframe and an uplink (UL)-subframe. The DL-subframe consists of a preamble followed by DL- and UL-MAPs respectively as shown in Fig. 2.1. The DL- and UL-MAPs are broadcast MAC management messages that specify the control information for all connections scheduled in the DL and UL subframes. The MAP is received and parsed by all MSs so that they can determine

the parameters associated with their allocations. The minimum unit of allocation is a slot and each MS's allocation is assigned an integral number of slots. The number of slots or resource size needed for an MS's allocation depends on its modulation and coding scheme (MCS) and the number of bits in the packets scheduled for transmission. The DL-and UL-MAPs are composed of smaller elements known as information elements (IEs). Each IE describes an MS's allocation and provides information such as the MCS, resource size, location of the allocation in the frame. A separate IE is usually required for each individual allocation in the frame.

2.2.2 MAC Layer and QoS Framework

The MAC layer of WiMAX systems is connection-oriented. For QoS management, the concept of service flow was introduced, which defines the QoS parameters for a unidirectional flow of packets. A service flow is created and its QoS parameters are defined through Dynamic Service Addition (DSA) messages that are exchanged between the BS and the MS associated with the service flow. The service flow and its QoS parameters can subsequently be updated through Dynamic Service Change (DSC) messages. Creation of a service flow may be initiated by an MS or the BS using a DSA-REQ message, which includes the parameters for the service flow. The MS-initiated service flow requests have to go through an admission control process. Upon admission into the system, an active service flow is associated with a transport connection–identified by a unique connection ID (CID). In essence, a transport connection serves as a context for data communications between the BS and an MS, whereas the associated service flow defines the connection's QoS requirements. For applications needing bidirectional communication, two service flows and correspondingly, two transport connections are created–one for each direction. In the rest of the chapter, the

terms “service flow” and “connection” will be used interchangeably.

Another concept relevant to QoS support in WiMAX systems is *scheduling service*, which essentially refers to the data handling mechanism associated with a connection. The IEEE 802.16 restricted all service flows, both in downlink and uplink, to four scheduling services: Unsolicited Grant Services (UGS), Real-time Polling Services (rtPS), Non-real-time Polling Services (nrtPS) and Best Effort (BE). Each of these scheduling services except BE has a mandatory set of QoS parameters; therefore, any service flow must declare the scheduling service that it wants to be handled with and as such must define the QoS parameters for that scheduling service.

IEEE 802.16e introduced major changes to the definition and scope of scheduling services (section 6.3.5 of IEEE 802.16e-2005 [2]). Although the underlying concept remains similar, association of service flows in the downlink need not confine to a mandatory list of scheduling services. The standard suggests but does not mandate a number of QoS parameters that would provide scheduling service for different types of applications such as real-time and non real-time VBR video, VoIP. In essence, incorporating a flexible interpretation of scheduling services, IEEE 802.16e allows scheduling services of a service flow to be defined by its stated QoS parameters. Besides the parameters already defined in the standard, these QoS parameters can also include vendor defined parameters. In the uplink, the scheduling services are renamed as *scheduling types* and every service flow in the uplink needs to define its scheduling type from a set of five predefined types: UGS, rtPS, nrtPS, Extended Real-time Polling Service (ertPS) and BE. The scheduling type determines the bandwidth request/grant mechanism for uplink connections.

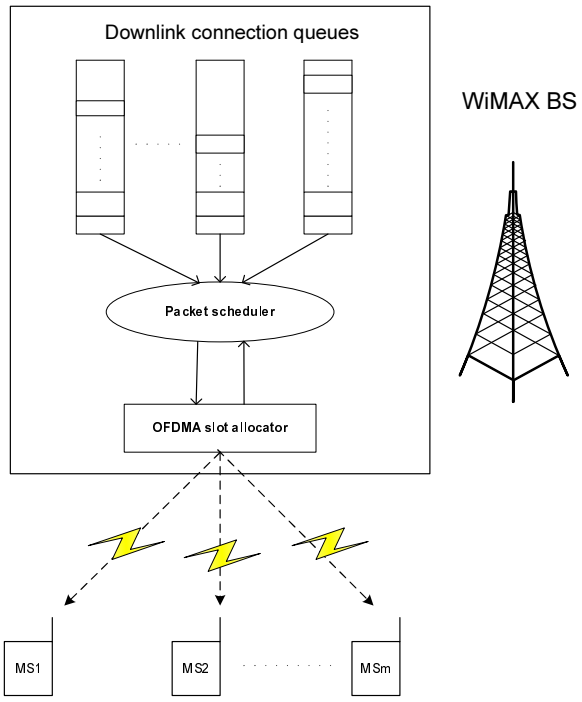


Figure 2.2: A simplified view of the system model

2.3 System Model and Related Assumptions

We consider a single cell in an IEEE 802.16e mobile WiMAX system operating in point-to-multipoint (PMP) mode, where a single BS communicates with a number of MSs located within a geographic area. We assume that TDD is used for multiplexing uplink and downlink connections. In this chapter, we focus on the resource allocation for admitted service flows in the system. Therefore, we do not propose any design for the admission control process, but assume that the service flows considered in the resource allocation process have successfully gone through such process. After being admitted, each of the service flows destined from the BS to an MS is assigned a downlink connection. Fig. 2.2 presents a simplified overview of the system model in consideration. We assume that a separate queue is maintained at the BS for each downlink connection and the packets belonging to

a downlink connection are served on an FCFS basis. The packets may arrive in variable intervals and may have variable packet size. Allocation decisions regarding the number of packets and the necessary number of slots to communicate these packets from each connection are made by the BS at the beginning of each frame. The packet scheduler is responsible for determining the number of packets to be scheduled from each downlink connection and can use the information from OFDMA slot allocator on available number of OFDMA slots in the physical layer (PHY). On the other hand, the OFDMA slot allocator, in coordination with the packet scheduler, determines how many slots will be allocated to the scheduled connections and in which order the connections will be allocated the slots. The DL-MAP is then constructed to indicate which connection is assigned which slots in the DL-subframe.

We also assume PUSC subchannelization in the downlink, therefore for a particular MS, all the subchannels offer similar fading dynamics. The MCS chosen by the BS to transmit to an MS depends on the channel quality information (CQI) feedback sent by the MS. A link adaptation process selects MCS levels for different MSs to maintain a target error performance in response to change in channel fading conditions. We assume that channel quality remains static over the length of a frame, so the MCS choice for a particular MS does not change within a frame time. The chosen MCS has the effect in the slot allocation scheme as it will be used in calculating how many slots need to be allocated for transmitting a certain number of packets from the BS to the MS.

Compared to IEEE 802.16 or its IEEE 802.16d amendment (later named 802.16-2004 [3]), the IEEE 802.16e standard has an updated scope for scheduling services in defining QoS requirement of service flows. The scheduling service of a service flow is no longer limited to four predefined classes namely UGS, rtPS, nrtPS, and BE. IEEE 802.16 and 802.16d also

specified a mandatory list of QoS parameters for each of these four scheduling services. IEEE 802.16e does not have such mandatory lists of QoS parameters; instead, the scheduling service for a service flow is now defined through its specified traffic characteristics and QoS parameters. Unlike many of the related works, we conform to these updates and do not assume four mandatory scheduling services. We also assume that three important QoS measures can be included in the scheduling service specification for a service flow: delay, packet loss rate and throughput. The *maximum latency* and *minimum reserved traffic rate* parameters defined in the standard can be used to specify delay and throughput requirements respectively. For delay performance, we allow tolerable delay limit to be specified probabilistically. Many service providers offer a number of QoS classes (e.g. platinum, gold, silver) and corresponding pricing policies for their customers; such providers can use probabilistic delay guarantees to differentiate the levels of service offered to customers belonging to different QoS classes. Therefore, we assume that the delay requirement of a service flow is specified by a maximum packet delay and a probability that the packets belonging to the service flow will be transmitted within this target delay. This probability measure and the packet loss rate are assumed to be specified through vendor specific QoS parameters, which the standard allows the vendors to define for QoS measures not covered by the predefined QoS parameters in the standard.

2.4 A Queue-Length-Based Packet Scheduling Policy and Its Queueing Analysis

A queue-length-based packet scheduling policy serves as the key to determining a set of tunable parameters in our proposed resource allocation scheme. In this section, we describe the scheduling policy and its queueing model that is used to calculate these parameters for

admitted service flows. Note that the packet scheduling policy described and analyzed in this section is not directly adopted in our proposed resource allocation scheme described later; rather, its queueing model is used to determine important parameters affecting the packet scheduler's operation in that scheme.

In this packet scheduling policy, the BS evaluates the number of packets in each backlogged downlink connection queue at the beginning of a frame and decides to schedule some fraction of the backlog at each queue in the current frame. More specifically, if queue-length of connection i at the beginning of frame n is denoted by $q_i^{(n)}$, then the number of packets, $w_i^{(n)}$ that will be scheduled for the connection in the current frame is given by

$$w_i^{(n)} = \begin{cases} \left\lceil x_i q_i^{(n)} \right\rceil, & \text{if} \quad 0 < q_i^{(n)} \leq L \\ 0, & \text{if} \quad q_i^{(n)} = 0, \end{cases} \tag{2.1}$$

where L is the maximum limit on the queue size in number of packets and x_i $(0 < x_i < 1)$ is a *model parameter* associated with connection i. The ceiling function in this expression is needed to convert any fractional values for the number of packets to the nearest integer and also to make sure that each non-empty queue gets to transmit at least one packet during each frame so that a queue does not remain without service for an extended period of time due to low traffic intensity.

The value of x_i for connection i is determined through a queueing analytical model when the associated service flow is admitted and its QoS parameters are known to the BS. The following subsection is devoted to formulating the queueing analytical model that relates the value of x_i with the QoS performance of connection i, given its traffic arrival statistics. If there are m admitted connections that have QoS requirements to fulfil, the analytical model will enable us to find the values for a set of m such model parameters.

Note that, if a service flow does not put forward any QoS requirement, it will be treated as BE traffic. Connections associated with BE traffic will be treated differently than those with QoS traffic and will not need association with a model parameter.

2.4.1 Queueing Analytical Model for the Packet Scheduling Policy

We develop the queueing analytical model from the perspective of the queue associated with a tagged downlink QoS connection. Without loss of generality, in the analysis, we can omit the CID of the tagged connection. Therefore, the model parameter for the connection will be denoted simply by x, and other parameters specific to the tagged connection will be used without any index referring to the CID. In developing the model, we consider the packet arrival process of the service flow associated with the connection and assume a given value for parameter x. The idea is to map the dynamics of this downlink queue to a discrete-time Markov chain (DTMC) and then solve the Markov chain for performance measures. The output of the queueing model will be the values for a set of performance measures given a particular value of x.

Packet Arrival Process

We model the packet arrival process to the downlink queue by a Markov Modulated Poisson process (MMPP), where the rate of arrivals varies depending on the current state of a finite-state Markov chain. MMPP has been extensively used by many researchers to model wide range of multimedia and IP traffic because of its ability to effectively model important properties of these traffic applications [4]. For example, it can closely match the observed autocovariance function and the marginal distribution associated with the packet arrival process. Besides, it offers excellent modeling for long-range dependence, burstiness and self-similar behavior observed in multimedia traffic [5]. Along with these desirable prop-

erties, it offers analytical tractability in modeling complex system dynamics.

In this chapter, we resort to a discrete-time analysis and therefore, we use a discrete-time equivalent of the MMPP arrival process, also known as d-MMPP. In general, a N-state d-MMPP is described by the transition probability matrix of a N-state Markov chain and a diagonal matrix containing the arrival rates associated with the various states of the Markov chain [6]:

$$S = \begin{bmatrix} 1 - \sum_{i \neq 1} s_{1i} & s_{12} & \cdots & s_{1N} \\ s_{21} & 1 - \sum_{i \neq 2} s_{2i} & \cdots & s_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ s_{N1} & s_{N2} & \cdots & 1 - \sum_{i \neq N} s_{Ni} \end{bmatrix}, \qquad \delta = \text{diag}(r_1, r_2, \cdots, r_N).$$

Now, assuming that the maximum number of packets that can arrive during a single time slot is M, we can form M transition probability matrices:

$$U^{(h)} = \beta^{(h)} S, \quad \text{where } 0 \leq h \leq M,\ \beta^{(h)} = \text{diag}(e^{-r_1} r_1^h/h!, e^{-r_2} r_2^h/h!, \cdots, e^{-r_N} r_N^h/h!).$$

Here, matrix $U^{(h)}$ corresponds to state transitions that happen after a batch of h packets arrives during a time slot. The value of M is chosen such that

$$\max_{1 \leq l \leq N} \sum_{i=M+1}^{\infty} e^{-r_l} r_l^i / i! < \epsilon, \tag{2.2}$$

where the value of ϵ is chosen to be very small to limit modeling error.

Assuming that matrix S is time-independent, the steady state probability vector $\boldsymbol{\psi}$ of

the arrival process can be found from

$$\boldsymbol{\psi} = \boldsymbol{\psi} S; \quad \boldsymbol{\psi}\mathbf{1} = 1, \tag{2.3}$$

where $\mathbf{1}$ is a column matrix of ones with appropriate dimension The mean packet arrival rate λ can be found from

$$\lambda = \boldsymbol{\psi}\delta\mathbf{1}. \tag{2.4}$$

Although the MMPP arrival process is typically used to model VBR multimedia applications, modeling CBR traffic with MMPP is also possible by appropriate selection of state space and transition probability matrix. For example, a CBR application that periodically generates a packet after every t time slots can be modeled by $U = \begin{bmatrix} \mathbf{0}_{(t-1)\times 1} & I_{t-1} \\ 1 & \mathbf{0}_{1\times(t-1)} \end{bmatrix}$, $\beta^{(0)} = \text{diag}(\mathbf{1}_{1\times(t-1)}, 0)$ and $\beta^{(1)} = \text{diag}(\mathbf{0}_{1\times(t-1)}, 1)$. Here, $\mathbf{0}_{(t-1)\times 1}$ and $\mathbf{0}_{1\times(t-1)}$ are column and row vectors of zeros respectively with dimension $t-1$. Similarly, $\mathbf{1}_{1\times(t-1)}$ is a row vector of ones with dimension $t-1$ and I_{t-1} is an identity matrix of dimension $t-1$. Also, VoIP applications with silence suppression are often modeled by an Interrupted Poisson Process (IPP), which is essentially a 2-state MMPP characterized by an arrival rate of 0 associated with one of the states.

Markov Chain Analysis

For this discrete-time analysis, we assume that the *time slot* is equal to the frame duration of the IEEE 802.16e WiMAX system. Note that a *time slot* in the context of the discrete-time queueing model in this analysis is different from an OFDMA slot in the WiMAX frame. To avoid further confusion, from hereon we will use the term *frame* to mean both the WiMAX frame and the *time slot* of the discrete-time queueing model. We proceed by

defining a process with state space $X^{(n)} \triangleq \{(q^{(n)}, a^{(n)}) | 0 \leq q^{(n)} \leq L; a^{(n)} = 1, 2, \cdots, N\}$, where $q^{(n)}$ is the number of packets in the queue and $a^{(n)}$ is the state of the arrival process at the beginning of frame n. Assuming the transition among the states occur at frame boundaries and as the state variables assume discrete values only, it can be proven that the process $X^{(n)}$ is a DTMC. We write the transition probability matrix P of the DTMC in the following form:

$$P = \begin{bmatrix} \mathcal{A}_{0\to 0} & \mathcal{A}_{0\to 1} & \cdots & \mathcal{A}_{0\to L} \\ \mathcal{A}_{1\to 0} & \mathcal{A}_{1\to 1} & \cdots & \mathcal{A}_{1\to L} \\ \vdots & \vdots & \cdots & \vdots \\ \mathcal{A}_{L\to 0} & \mathcal{A}_{L\to 1} & \cdots & \mathcal{A}_{L\to L} \end{bmatrix}.$$

Here, block submatrices $A_{k\to l}$ $(0 \leq k \leq L, 0 \leq l \leq L)$ denote transition probabilities associated with the events when queue length of the connection changes from k in frame n to l in frame $n+1$. These block submatrices can be derived as follows:

$$\mathcal{A}_{0\to l} = \begin{cases} U^{(l)}, & 0 \leq l \leq M \\ \mathbf{0}, & M < l \leq L \end{cases} \tag{2.5}$$

$$\mathcal{A}_{k\to l} = \begin{cases} U^{(l-k+\lceil kx \rceil)}, & \text{if} \quad l \leq M + k - \lceil kx \rceil \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{2.6}$$

$$\text{for} \quad 0 < k < l < L$$

$$\mathcal{A}_{k\to l} = \begin{cases} U^{(l-k+\lceil kx \rceil)}, & \text{if} \quad k - \lceil kx \rceil \leq l \leq min\left(M + k - \lceil kx \rceil, k - 1\right) \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

$$\text{for} \quad 0 < l < k < L \tag{2.7}$$

$$\mathcal{A}_{k\to L} = \begin{cases} \sum_{h=L-k+\lceil kx \rceil}^{M} U^{(h)}, & \text{if} \quad k \geq L - M + \lceil kx \rceil \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{2.8}$$

$$\text{for} \quad 1 < k < L$$

$$\mathcal{A}_{k\to k} = \begin{cases} U^{(\lceil kx \rceil)}, & \text{if} \quad \lceil kx \rceil \leq M \\ \mathbf{0}, & \text{if} \quad \lceil kx \rceil > M \end{cases} \tag{2.9}$$

$$\text{for} \quad 1 < k < L$$

$$\mathcal{A}_{L\to L} = \begin{cases} \sum_{h=\lceil Lx \rceil}^{M} U^{(h)}, & \text{if} \quad \lceil Lx \rceil \leq M \\ \mathbf{0}, & \text{if} \quad \lceil Lx \rceil > M, \end{cases} \tag{2.10}$$

where $\mathbf{0}$ denotes a zero matrix with appropriate dimensions.

Once the block submatrices are derived, we can find the steady state probability vector $\boldsymbol{\pi}$ for the Markov chain by solving the set of linear equations given by $\boldsymbol{\pi} = \boldsymbol{\pi}P;\ \boldsymbol{\pi}\mathbf{1} = 1$. By reblocking the entries in $\boldsymbol{\pi}$ according to the associated queue lengths, we can write the steady state probability vector as $\boldsymbol{\pi} \triangleq [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_L]$, where $\boldsymbol{\pi}_h$ stands for steady state vector associated with a queue length of h $(0 \leq h \leq L)$.

Derivation of Performance Measures

In this subsection, we will derive a number of performance measures using the steady state probability vector achieved from the Markov chain analysis. These measures include delay, packet loss rate and throughput—three important performance measures that often constitute the QoS requirements of most multimedia and data intensive applications. In analyzing the delay performance, many researchers limit their analysis to computing average delay only. In contrast, we derive delay distribution because it offers a more meaningful perfor-

mance measure, especially for delay-sensitive multimedia applications. Individual packets belonging to these applications need to be delivered within a delay limit to maintain QoS; computing the average delay of these packets is, therefore, not very useful. The delay distribution of these packets, on the other hand, provides adequate information to determine how well the packets are served within the delay limit.

Delay Distribution We use an absorbing Markov chain to find the delay distribution for a packet in the tagged queue. In order to construct the state space of this Markov chain, we tag an arbitrary packet in the queue and consider the number of packets that are in front of it waiting to be served and the number of packets waiting behind it in the queue. The chain initializes from one of the possible states corresponding to the tagged packet's arrival at the corresponding queue and gets absorbed when the tagged packet reaches the top of the queue regardless of the number of packets waiting behind it. Let the state space Θ of this Markov chain be defined as follows:

$$\Theta \triangleq \{(u, v, w) | 0 \leq u \leq L-1; 0 \leq v \leq L-u-1; 1 \leq w \leq N\},$$

where u and v are the number of packets waiting in front and behind the tagged packet respectively and w is the MMPP state of the arrival process. Absorption of the chain occurs whenever any of the states in $\{(0, v, w) | 0 \leq v \leq L-1; 1 \leq w \leq N\}$ is reached. Let us denote the transition probability matrix of the absorbing Markov chain by Q and reblock it with block submatrices $\mathcal{B}_{(i_1,j_1)\rightarrow(i_2,j_2)}$ that contain transition probabilities from states representing i_1 $(0 \leq i_1 \leq L-1)$ packets in front and j_1 $(0 \leq j_1 \leq L-i_1-1)$ packets behind to states representing i_2 $(0 \leq i_2 \leq L-1)$ packets in front and j_2 $(0 \leq j_2 \leq L-i_2-1)$ packets behind the tagged packet. We now find these block submatrices of the absorbing

chain as follows:

$$\mathcal{B}_{(0,j_1)\to(i_2,j_2)} = \begin{cases} I, & \text{if } j_1 = j_2, i_2 = 0 \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{2.11}$$
$$\text{for } 0 \le j_1 \le L-1, 0 \le i_2 \le L-1, 0 \le j_2 \le L-i_2-1$$

$$\mathcal{B}_{(i_1,j_1)\to(i_2,j_2)} = \begin{cases} U^{(j_2-j_1)}, & \text{if } i_2 = i_1 - \lceil (i_1+j_1+1)x \rceil, j_2 \le j_1 + M \\ \mathbf{0}, & \text{otherwise} \end{cases}$$
$$\text{for } 1 \le i_1 \le L-1, 0 \le i_2 \le L-1, 0 \le j_1 \le L-i_1-1, 0 \le j_2 < L-i_2-1 \tag{2.12}$$

$$\mathcal{B}_{(i_1,j_1)\to(i_2,L-i_2-1)} = \begin{cases} \sum_{h=L-i_2-j_1-1}^{M} U^{(h)}, & \text{if } i_2 = i_1 - \lceil (i_1+j_1+1)x \rceil, j_1 \ge L-M-i_2-1 \\ \mathbf{0}, & \text{otherwise} \end{cases}$$
$$\text{for } 1 \le i_1 \le L-1, 0 \le i_2 \le L-1, 0 \le j_1 \le L-i_1-1 \tag{2.13}$$

where I and $\mathbf{0}$ denote identity matrix and zero matrix respectively of appropriate dimensions.

We then determine the initial probability vector for this absorbing Markov chain as the next step for finding the delay distribution. To this end, we first consider the probability vector associated with the number of packets in front and behind as seen by an arriving packet. Note that, the arriving packet might see packets behind because packet arrival may occur in bursts in a given frame and the arriving packet can be at any of the positions in that burst. Let us denote the probability vector by $\boldsymbol{\xi}^{(0)}$. By reblocking the vector in terms of packets seen in front by the arriving packet, we can write it in an expanded form as $\boldsymbol{\xi}^{(0)} \triangleq \left[\boldsymbol{\xi}_0^{(0)}, \boldsymbol{\xi}_1^{(0)}, \cdots, \boldsymbol{\xi}_{L+M-1}^{(0)}\right]$, where, $\boldsymbol{\xi}_i^{(0)}$ $(0 \le i \le L+M-1)$ is the probability vector associated with i packets in front as seen by the arriving packet. We can further reblock

the probability vector $\boldsymbol{\xi}_i^{(0)}$ as $\boldsymbol{\xi}_i^{(0)} \triangleq \left[\boldsymbol{\xi}_{(i,0)}^{(0)}, \cdots, \boldsymbol{\xi}_{(i,\min(M-1,L+M-i-1))}^{(0)}\right]$ on the basis of the number of packets behind the arriving packet that sees i packets in front of itself. These probability vectors can be derived as follows:

$$\boldsymbol{\xi}_{(i,j)}^{(0)} = \begin{cases} \frac{1}{(1-\alpha_0)(j+1)} \left(\boldsymbol{\pi}_0 + \sum_{\substack{y:1\le y\le L,\\ y=\lceil yx \rceil}} \boldsymbol{\pi}_y \right) U^{(j+1)}, i = 0, 0 \le j \le M-1 \\ \frac{1}{1-\alpha_0} \sum_{h=0}^{\min(M-j-1,i)} \frac{1}{j+h+1} \sum_{\substack{y:i-h<y\le L,\\ y-\lceil yx \rceil = i-h}} \boldsymbol{\pi}_y U^{(j+h+1)}, \ 1 \le i \le L, 0 \le j \le M-1 \\ \frac{1}{1-\alpha_0} \sum_{h=i-L+1}^{M-j-1} \frac{1}{j+h+1} \sum_{\substack{y:i-h<y\le L,\\ y-\lceil yx \rceil = i-h}} \boldsymbol{\pi}_y U^{(j+h+1)}, L+1 \le i \le L+M-1, 0 \le j \le L+M-i-1, \end{cases} \tag{2.14}$$

where α_0 denotes the probability associated with no packet arrivals during a frame and can be derived from $\alpha_0 = \boldsymbol{\psi}\beta^{(0)}\mathbf{1}$.

Note that the packets which find their destination queues full upon arrival are not admitted into the system. The probability p_o that a buffer overflow event will occur upon arrival of a packet can be found from:

$$p_o = \sum_{i=0}^{M-1} \boldsymbol{\xi}_{L+i}^{(0)} \mathbf{1} + \sum_{i=1}^{M-1} \sum_{j=i}^{M-1} \boldsymbol{\xi}_{(L-i,j)}^{(0)} \mathbf{1}. \tag{2.15}$$

Since we consider the delay experienced by the admitted packets only, the initial probability vector $\boldsymbol{\omega}^{(0)} \triangleq \left[\boldsymbol{\omega}_{(0,0)}^{(0)}, \cdots \boldsymbol{\omega}_{(0,L-1)}^{(0)}, \cdots, \boldsymbol{\omega}_{(i,j)}^{(0)} \cdots, \boldsymbol{\omega}_{(L-1,0)}^{(0)}\right]$ of the absorbing Markov

chain is derived from the above probability vector as follows:

$$\boldsymbol{\omega}_{(i,j)}^{(0)} = \begin{cases} \frac{\boldsymbol{\xi}_{(i,j)}^{(0)}}{1-p_o}, & \text{if } 0 \leq j \leq \min(M-1, L-i-1) \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{2.16}$$
$$\text{for } 0 \leq i \leq L-1, 0 \leq j \leq L-i-1,$$

which is due to the fact that packets that cannot enter the queue due to overflow are not considered in the state space of the absorbing Markov chain. After d frames, the state probability vector of the absorbing chain can be written as

$$\boldsymbol{\omega}^{(d)} = \boldsymbol{\omega}^{(0)} Q^d, \tag{2.17}$$

where Q^d is derived by multiplying the transition probability matrix Q of the absorbing Markov chain with itself d times. The cumulative distribution function (CDF) of packet delay $\mathcal{D}$ in frame times, not including the arrival frame, is given by:

$$F_{\mathcal{D}}(d) = \Pr(\mathcal{D} \leq d) = \sum_{j=0}^{L-1} \boldsymbol{\omega}_{(0,j)}^{(d)} \mathbf{1}. \tag{2.18}$$

Packet Loss Rate and Average Throughput A packet loss occurs at the downlink queue when a packet finds the queue full upon arrival. The packet loss rate, p_{loss} at the queue can be written as

$$p_{loss} = \sum_{i=0}^{M-1} \boldsymbol{\xi}_{L+i}^{(0)} \mathbf{1}. \tag{2.19}$$

Note that packets lost at the receiver side due to reception errors are not considered in the above expression. Assuming no error recovery protocol is used, the average throughput η

in packets per frame time can then be calculated as:

$$\eta = \lambda(1 - p_{loss})(1 - \overline{PER}), \tag{2.20}$$

where $\overline{PER}$ is the average packet error rate.

Queue Length Distribution and Average Delay The marginal probability distribution of queue length $\mathcal{L}$ can be written as:

$$f_{\mathcal{L}}(l) = \Pr(\mathcal{L} = l) = \boldsymbol{\pi}_l \mathbf{1}, 1 \leq l \leq L. \tag{2.21}$$

Subsequently, the average queue length $\bar{\mathcal{L}}$ and average delay $\bar{\mathcal{D}}$ can be found as

$$\bar{\mathcal{L}} = \sum_{l=1}^{L} l\boldsymbol{\pi}_l \mathbf{1}, \quad \bar{\mathcal{D}} = \frac{\bar{\mathcal{L}}}{\lambda}, \tag{2.22}$$

where the calculation of average delay follows from Little's law.

2.5 Model-Based and Channel-Aware Downlink Resource Allocation

In this section, we describe our proposed resource allocation scheme that utilizes the output from the analytical model developed in the preceding section. When a new downlink service flow, represented by connection i is admitted into the system, its QoS requirements are evaluated and the value of the model parameter x_i is computed. The outline of the computation procedure is listed in Algorithm 1 . The range of values of x_i is $0 < x_i < 1$, but searching over the entire range will be infeasible. Therefore, a step-size is selected and

Algorithm 1 Compute model parameter x_i for connection i

Input: T_i: Pkt. Arrival statistics for connection i; Δ : Step size; $\mathcal{Q}_{req}^{(i)}$: QoS requirement for connection i

Output: Model parameter x_i

1: $x_i \leftarrow 0$
2: **repeat**
3: $\quad x_i \leftarrow x_i + \Delta$
4: $\quad \mathcal{Q}_{out}^{(i)} \Leftarrow$ QueuingModel (x_i, T_i)
5: **until** $\mathcal{Q}_{out}^{(i)}$ meets QoS requirements $\mathcal{Q}_{req}^{(i)}$ of connection i

the value of x_i is incremented by this step size in iterations. In the simulation experiments, for example, we choose a step size of 0.05; however, vendors' choice for its value could be flexible depending on the available computational resource. During each iteration, the updated value of x_i is fed into the queueing model along with the queue-length information and the parameters of the packet arrival process. Increasing the value of model parameter x_i will allow more packets to be transmitted on average from the queue associated with connection i and hence, improve its QoS performance. The process continues until the set of performance measures, denoted by $\mathcal{Q}_{out}^{(i)}$ in Algorithm 1, calculated from the output of the queueing model satisfies the QoS requirement of connection i.

The value of the model parameter x_i is used in our resource allocation scheme to determine the *minimum* number of packets that need to be scheduled from connection i at each frame time. However, the actual number of packets that are scheduled is determined through a number of steps in the allocation procedure listed in Algorithm 2, which also shows how the available OFDMA slots are allocated once the number of packets to be scheduled for each admitted downlik service flows is determined

At the beginning of each frame, some important information become available to the resource allocation module at the BS: the backlog at the queues of the downlink connections

Algorithm 2 Schedule packets, allocate slots for downlink connections in the current frame

$\mathcal{N}_{rem} \leftarrow$ # of slots in the DL-subframe $-$ slots allocated to DL-subframe overhead;
$m \leftarrow$ # of downlink QoS connections; $[h_1, \cdots, h_m] \leftarrow$ CIDs of downlink QoS connections;
for $i \leftarrow 1 \cdots m$ **do**
  $\mathcal{K}_{h_i} \leftarrow 0$; // initialize slots allocated for each QoS connections to 0
  $\theta_{h_i} \leftarrow$ byte per slot with MCS selected for h_i;
  $\mathcal{M}_{h_i} \leftarrow \theta_{h_i} \times$ backlog (bytes) at connection h_i; //Metric $\mathcal{M}$ associated with h_i
end for
Sort QoS connections according to their $\mathcal{M}$ values in descending order. Let $[c_1, c_2, \cdots, c_m]$ be the sorted list and $[q_{c_1}, q_{c_2}, \cdots, q_{c_m}]$ be the corresponding queue-lengths
$\mathcal{G} \Leftarrow \emptyset$; // list of scheduled connections
repeat
  for $i \leftarrow 1 \cdots m$ **do**
    $p \leftarrow \lceil x_{c_i} \times q_{c_i} \rceil$; //# of packets selected based on queue-length & model parameter x_{c_i}
    $b \leftarrow$ number of bytes in top p packets in the queue of c_i;
    $a \leftarrow \lceil \frac{b}{\theta_{c_i}} \rceil$; // # of slots necessary to transmit p packets over connection c_i
    if $c_i \notin \mathcal{G}$ **then**
      $a \leftarrow a + IE$ overhead (slots); //IE added if not previously added for connection c_i
    end if
    if $\mathcal{N}_{rem} \geq a$ **then**
      $\mathcal{N}_{rem} \leftarrow \mathcal{N}_{rem} - a$;
      if $c_i \notin \mathcal{G}$ **then**
        $\mathcal{G} \Leftarrow \mathcal{G} \cup \{c_i\}$; // c_i added to list of scheduled connections
      end if
      $q_{c_i} \leftarrow q_{c_i} - p$; // queue length of c_i updated
      $\mathcal{K}_{c_i} \leftarrow \mathcal{K}_{c_i} + a$; // update slots allocated to c_i
    end if
  end for
until $(\mathcal{N}_{rem} = 0)$ or (no new allocation or allocation update happened in this iteration)
Allocate N_{rem} among BE connections

and the channel quality feedbacks for different subchannels at different MSs. At first, the list of active downlink QoS connections is sorted in descending order based on a metric $\mathcal{M}$. The value of $\mathcal{M}$ for a connection is calculated by multiplying the amount of backlog (in bytes) in its queue with the number of bytes that a single OFDMA slot will be able to transmit on the connection in the current frame. MCS selected for a connection is dependent on the CQI and will in turn determine the number of bytes an OFDMA slot can carry for the connection in the current frame. The metric affects the ordering of the connections while allocating available OFDMA slots. Note that such ordering becomes important when there are not enough OFDMA slots available to accommodate the minimum number of packets (as determined by model parameters) from all the backlogged downlink connections. If two connections have the same amount of backlog, then the connection that can transmit more bytes per slot will be given preference for the remaining slots. This cross-layer adaptive ordering is designed to improve resource utilization as the connection with the better channel quality will likely need less number of slots to clear the same amount of backlog. On the other hand, we also want connections with large backlogs to have better opportunity to transmit their packets. Therefore, among two connections with the same MCS level, the connection with larger backlog will be given preference in the ordering.

The scheduler then goes over the sorted list to determine how many packets to schedule in the current frame for each connection. Let us assume that connection with CID c_1 is the at the top of the sorted list. In the first iteration, it starts by selecting the top $\lceil x_{c_1} \times q_{c_1} \rceil$ packets for connection c_1, where q_{c_1} is the number of backlogged packets in the connection's queue and x_{c_1} is the associated model parameter. The number of slots necessary to transmit those packets is then calculated. If a new IE needs to be created in the DL-submap to carry the allocation information, the necessary overhead is added while calculating the

number of slots required. The other factor in this calculation is the transmission rate which corresponds to the selected MCS for the destination MS of connection c_1. If there is enough space left in the DL-subframe to accommodate the calculated number of slots, these packets are marked as selected for transmission in the current frame. The amount of available space is also updated by deducting the number of slots selected for the connection. The next connection in the sorted list is then considered and the above process is applied to choose the minimum number of packets and corresponding number of required OFDMA slots. The first iteration continues until all the connections in the list $[c_1, c_2, \cdots c_m]$ are considered for the above step or all the DL-subframe space is used up.

After the first iteration, if more space is left in the DL-subframe and some QoS connections in $[c_1, c_2, \cdots c_m]$ still have backlogged packets not selected yet for transmission in the current frame, the allocation module enters into more iterations. The aim of these subsequent iterations is to avoid wasting any empty space in the DL-subframe. In these iterations, the initial sorted order of connections is maintained, but the packets are selected by considering backlog containing those packets that have not been selected in previous rounds of iterations. The process is repeated until it is determined that further iterations will not allocate any more space to the QoS connections. The allocation module then considers BE connections if more space is left in the DL-subframe.

The selected packets for each connection are then scheduled to transmit in the current DL-subframe. A single IE with the allocation information is constructed in the DL-MAP for each scheduled connection, and slots allocated to the connection are chosen to be contiguous.

2.6 Results

In this section, we report numerical and simulation results for our proposed queueing model and the model-based resource allocation scheme. For the queueing model, we tag an arbitrary downlink service flow and analyze the sensitivity of the model outputs to the model parameter associated with its connection. From these results, we discuss how the appropriate value of the model parameter can be chosen for the connection so that the QoS requirement of the service flow can be met. The simulation results are then presented to show the performance of the proposed model-based resource allocation scheme in meeting QoS requirements of both real-time and non-real-time service flows.

2.6.1 Numerical Results

To derive the numerical results, we coded in Matlab the steps involved in the queueing model developed in section 2.4. For the packet arrival process, we use a 3-state MMPP model with the following parameters:

$$\mathrm{P} = \begin{bmatrix} 0.0005 & 0.9751 & 0.0244 \\ 0.2438 & 0.7245 & 0.0317 \\ 0.1463 & 0.0024 & 0.8513 \end{bmatrix}; \qquad \delta = \mathrm{diag}(0.62, 2.75, 4.95).$$

The maximum capacity of the downlink queue was set to 25 packets. Since numerical results are derived for a single connection, its model parameter is referred simply by x, without any subscript.

Fig. 2.3 shows the CDF of delay experienced by the packets transmitted over the tagged service flow for various values of model parameter x. As the value of x is increased, the delay CDF improves, albeit by different amount at different ranges of values of x. In the

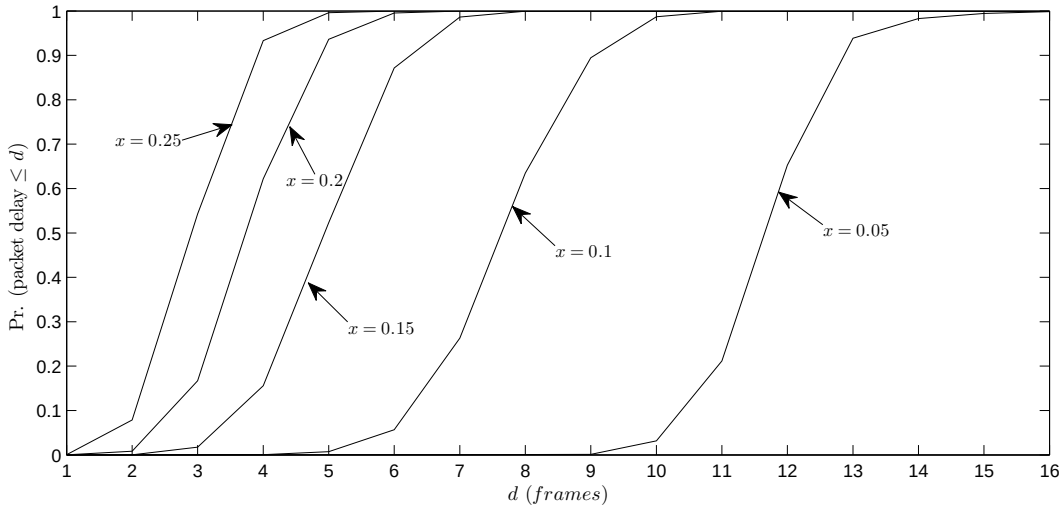


Figure 2.3: CDF of packet delay for various values of model parameter x

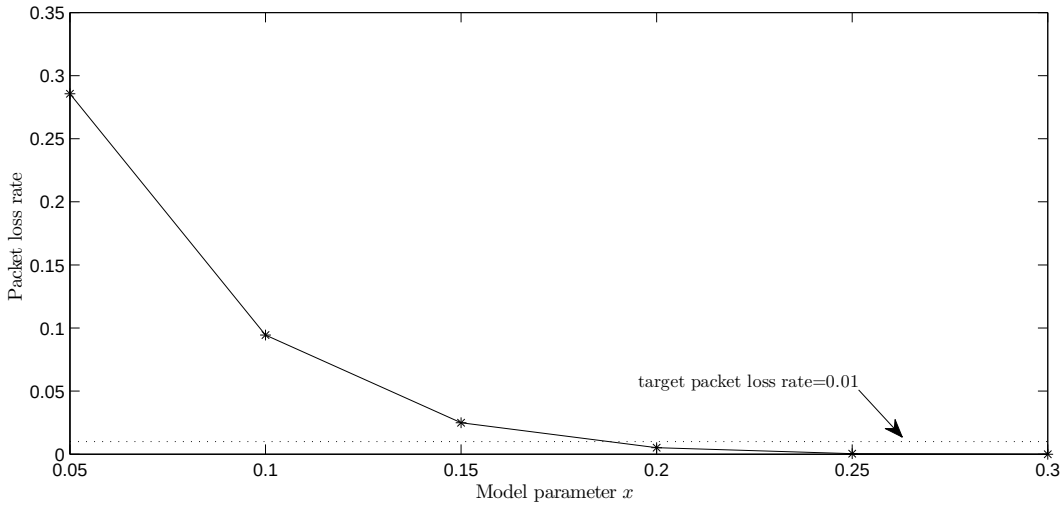


Figure 2.4: Model parameter x vs. packet loss rate

figure, the improvement is quite substantial as we increase x from 0.05 to 0.1 and then to 0.15. The gain then starts to level off when we increase x further. Although this trend should remain similar in general, the exact effects are likely to vary for different ranges of values for x, depending on the packet arrival statistics. Now let us assume that the tagged service flow has been provided a probabilistic guarantee that with 95% probability its packets will be served within 50 ms. Assuming a frame duration of 5 ms, we can see

from the figure that $x = 0.05$ will not be sufficient to meet this guarantee. If we only consider the values of x having an interval $\Delta = 0.05$, it is apparent that the delay guarantee could be met with $x = 0.1$. As we will see in the next paragraph, other QoS constraints, however, might demand the value of x to be raised further.

In Fig. 2.4, we show the effect of parameter x on the packet loss rate experienced by the tagged service flow. The interval Δ associated with the values of x is again set to 0.05. The effect of x on packet loss rate seems very pronounced when the value of x is changed inside the range 0.05 to 0.15. With x higher than 0.2, the probability of buffer overflow becomes negligible. Now if we assume that the service flow has been guaranteed a maximum packet loss rate of 1%, this obligation clearly cannot be met with $x = 0.1$ or $x = 0.15$, which incur a packet loss rate of 9.44% and 2.49% respectively. With $x = 0.2$, the packet loss rate comes down to 0.5%, which is below the allowed packet loss rate. Although $x = 0.1$ was sufficient to meet the delay guarantee mentioned in the previous paragraph, $x = 0.2$ is required instead if this packet loss target needs to be met at the same time.

2.6.2 Simulation Results

We have also performed extensive simulations to evaluate the performance of the proposed resource allocation framework. To this end, we have developed a MATLAB-based WiMAX System Level Simulator (SLS) that includes a resource allocation module implementing our proposed scheme. We have also implemented Earliest Deadline First (EDF), Weighted Deficit Round Robin (WDRR) and a hierarchical scheme in our SLS, which allows to selectively run these schemes to compare the performance of these schemes with that of our proposed scheme. The EDF is a well known scheme designed for providing QoS for delay-sensitive multimedia applications in many wired and wireless systems including WiMAX.

At each scheduling instant, priority is given to packets that have earlier deadlines. WDRR is a modified version of Deficit Round Robin (DRR) [7], which essentially is a O(1) approximation of Weighted Fair Queueing (WFQ) [8]. In [9], WFQ was proposed and evaluated for intra-class scheduling of throughput guaranteed nrtPS connections. However, due to the frame-based transmissions in mobile WiMAX systems, where multiple connections can be scheduled within a frame time, WFQ does not perform significantly better than round-robin based schemes [10]. The high complexity involved in implementing WFQ is thus no longer justified; rather DRR becomes a better alternative due its low-complexity. DRR was also selected by Cicconetti et al. in [11] for downlink scheduling . The WDRR scheme proposed by Lakkakorpi et al. in [12] further modifies the standard DRR to allow dynamic adjustment of the *quantum* parameter for each connection according to the connection's channel quality. The authors also found that, in a mobile WiMAX system, the channel-adaptive design of WDRR offers better performance over the standard DRR. We therefore compare our proposed scheme to WDRR instead of the standard DRR.

Simulation Setup and Parameters

We have simulated a single IEEE 802.16e WiMAX cell with PMP architecture, in which a single BS communicates with multiple MSs. The MSs are assumed to be distributed uniformly across the cell. To model the link adaptations due to mobility and channel variations, we have used the results reported in [13] from the Mobile WiMAX link-level simulations carried out at Intel's Wireless Standards and Technology Laboratory. The IEEE 802.16e link-level simulator assumed the following mix of mobility profiles [14] for the MSs served by the network: 60% ITU Pedestrian B 3 Km/h, 30% ITU Vehicular A 60 km/h and 10% ITU Vehicular A 120 km/h. It also assumed a 2×2 MIMO model in the DL with two trans-

Table 2.1: System parameters

Parameters	Value
PHY	OFDMA
Duplexing method	TDD
Frame duration	5 ms
Bandwidth	10 MHZ
DL:UL ratio	26:21
Number of DL subchannels	30
Number of DL symbols per slot	2
Number of Symbols in DL-subframe	26
DL subchannelization scheme	PUSC
Number of antennas (TX:RX)	2:2
FFT size	1024
Compressed MAP	enabled
MAP MCS	QPSK (1/2)
HARQ scheme	disabled
Queue size limit (in packets)	50
Fragmentation/packing	OFF
Mobility model	60% ITU PB3, 30% VA30, 10% VA120

Table 2.2: Modulation and coding schemes (MCS) used in the simulation

Level (without SMX/with SMX	MCS	Bytes per slot (without SMX/with SMX)
1/-	QPSK 1/12	1/-
2/-	QPSK 1/8	1.5/-
3/-	QPSK 1/4	3/-
4/12	QPSK 1/2	6/12
5/13	QPSK 3/4	9/18
6/14	16-QAM 1/2	12/24
7/15	16-QAM 3/4	18 /36
8/16	64-QAM 1/2	18/36
9/17	64-QAM 2/3	24/48
10/18	64-QAM 3/4	27/54
11/19	64-QAM 5/6	30/60

mit antennas at the BS and two receive antennas at each MS. The assumed MCS levels and their corresponding per-slot data capacities are listed in Table 2.2. MCS levels 1, 2 and 3 employ QPSK 1/2 with repetition 6, 4 and 2 respectively. MIMO mode for MCS levels 4 to 11 is space-time block coding (STBC), whereas MCS levels 12 to 19 adopt spatial multiplexing (SMX). We select the MCS level for each MS during each WiMAX frame according to the reported MCS distribtion in Fig. 3 (for 1.4 km BS-BS separation) of [13]. We list selected values for important simulation parameters involving the OFDMA-based air interface in Table 2.1, other physical layer specific parameters were chosen in accordance with [13]. We simulated 11000 WiMAX frames for each of the simulation experiments and took average from several runs to smooth out the results.

QoS Support for Non-Real-Time Service Flows

Non-real-time applications can tolerate delays, but demand a limited level of packet loss. Our first set of simulations is aimed at evaluating the effectiveness of our proposed scheme in providing QoS for such applications. We use a non-real-time application modeled by the following set of MMPP parameters:

$$S = \begin{bmatrix} 0.975 & 0.025 \\ 0.05 & 0.95 \end{bmatrix}, \qquad \delta = \text{diag}\,(0.5, 2.0).$$

These MMPP parameters can be derived considering a peak rate of 2 packets/frame, minimum rate of 0.5 packets/frame and mean rate of 1 packets/frame, which are equivalent to 747 Kbps, 187 Kbps and 373 Kbps respectively, considering the following packet size distribution: truncated pareto with shape paramter a =1.1, minimum packet size of 81 bytes and maximum packet size of 1500 bytes. The truncated pareto distribution is often

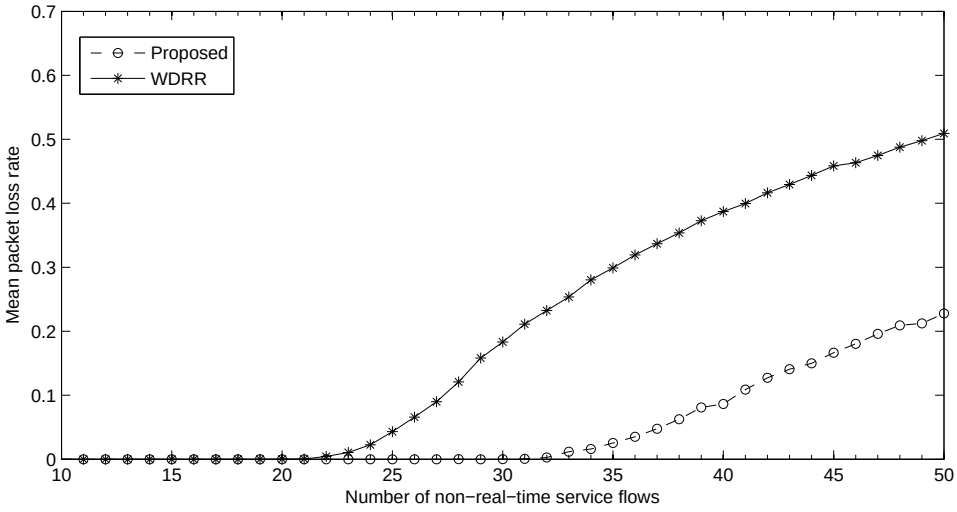


Figure 2.5: System-wide packet loss rate for non-real-time services

used to model size of the packets carried by Internet data traffic. For evaluating WiMAX MAC layer performance, same parameter values as above were also used in [10] to model packet size distribution of non-real-time data traffic. The packet loss performance results in Fig. 2.5 are obtained for a homogenous traffic scenario where each of the MSs is associated with a downlink service flow transporting traffic modeled by above listed parameters. The maximum tolerated packet loss rate for each of the service flows is set to 1%. QoS for non-real-time services are also often sought in terms of required throughput and our scheme is capable of handling such requirement too. However, we note that, for a given $\overline{PER}$ and the offered load on a service flow, throughput and packet loss rate measures of the service flow can be calculated directly from each other. We, therefore, consider packet loss rate only in our simulation analysis. The results for WDRR are also shown for comparison. We can clearly see that, compared to WDRR, our proposed scheme significantly improves the system-wide packet loss performance. The packet loss rates drawn in Fig. 2.5, however, is averaged over all connections. Nevertheless, the results show that the proposed scheme's pro-active management of backlogs at the connection queues is very effective in reducing

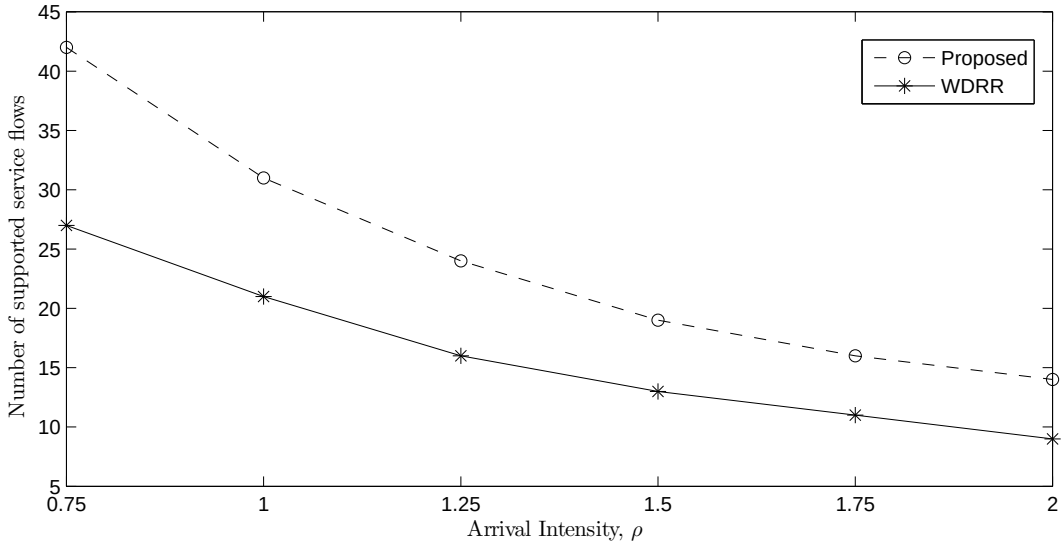


Figure 2.6: Number of supported non-real-time service flows vs. their arrival intensities

unwanted packet losses.

In Fig. 2.6, we show the number of service flows that can be supported without violating the target packet loss rate of 1% for any of the service flows. To show how different levels of traffic load affect the performance, we plot the capacity values against *intensity* of the packet arrival process. We use the same traffic parameters as described in the previous paragraph, except that the rate matrix is now specified as $\delta = \rho \times \text{diag}\ (0.5, 2.0)$, where ρ is the *intensity* parameter. We vary the value of ρ by 0.25 from 0.75 to 2.0, which translates to an average data rate of 280 Kbps to 747 Kbps. We can see that our proposed scheme significantly outperforms WDRR in a wide range of arrival intensities. Unlike WDRR, our proposed scheme is queue-aware and therefore, better adapts to backlog buildups at the connection queues. Thanks to the model-based approach, our scheme also has the advantage of having the precise information on the minimum amount of backlog that needs to be transmitted from each backlogged QoS connection during every frame. The proposed scheme can, therefore, make more-informed scheduling decisions and is more successful

in avoiding under-allocation of resources to the scheduled connections. These factors contribute positively to the performance benefit of our proposed scheme over WDRR in terms of the number of supported non-real-time service flows in the downlink.

QoS Support for Real-Time Service Flows

In the next set of simulations, we evaluate how effectively our model-based resource allocation scheme can meet QoS guarantees for real-time multimedia applications such as streaming video, audio/video conferencing, VoIP. These delay-sensitive applications need their packets to be transmitted within a certain delay limit and can tolerate only a very limited amount of packet losses. In our analysis, we assume a homogeneous traffic scenario, where each MS is associated with a real-time service flow characterized by identical traffic parameters. We use the following 3-state MMPP model mapped to the real traffic trace of a VBR MPEG-4 coded live video stream:

$$P = \begin{bmatrix} 0.9475 & 0.0525 & 0 \\ 0.1042 & 0.0750 & 0.8208 \\ 0 & 0.8333 & 0.1667 \end{bmatrix}; \qquad \delta = \text{diag}(0.58, 1.37, 4.65).$$

The parameters associated with the three states of this MMPP model were derived by matching the arrival and frame size statistics of I, P and B frames in the traffic trace [15]. To avoid confusion, we will refer to the video frames as packets in our analysis. The average bit rate of the VBR video stream is 256 Kbps. For each of these connections, packet delay limit is set to 50 ms with 95% probability level and the maximum allowable packet loss rate is set to 1%.

Fig. 2.7 compares our proposed scheme with EDF in terms of system-wide packet de-

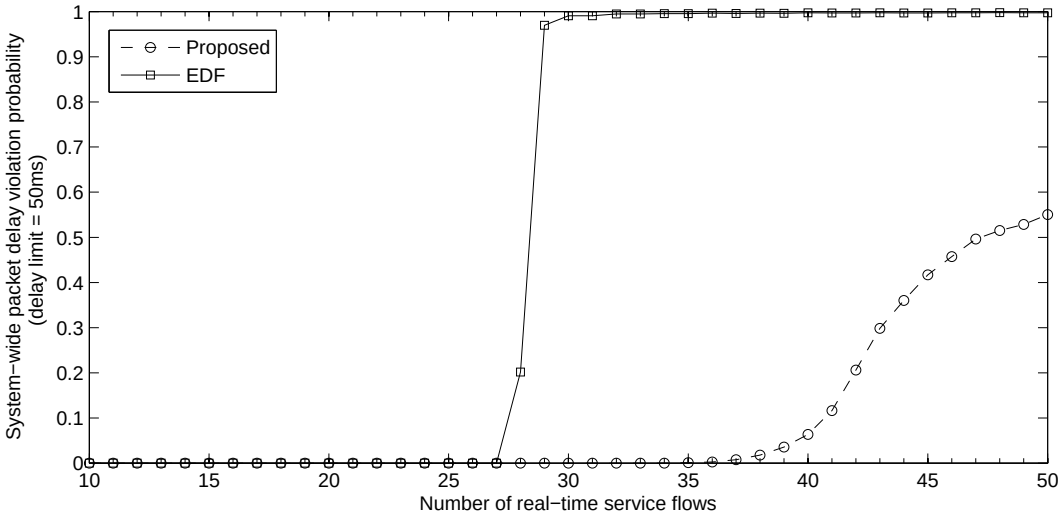


Figure 2.7: System-wide packet delay violation probability for real-time services

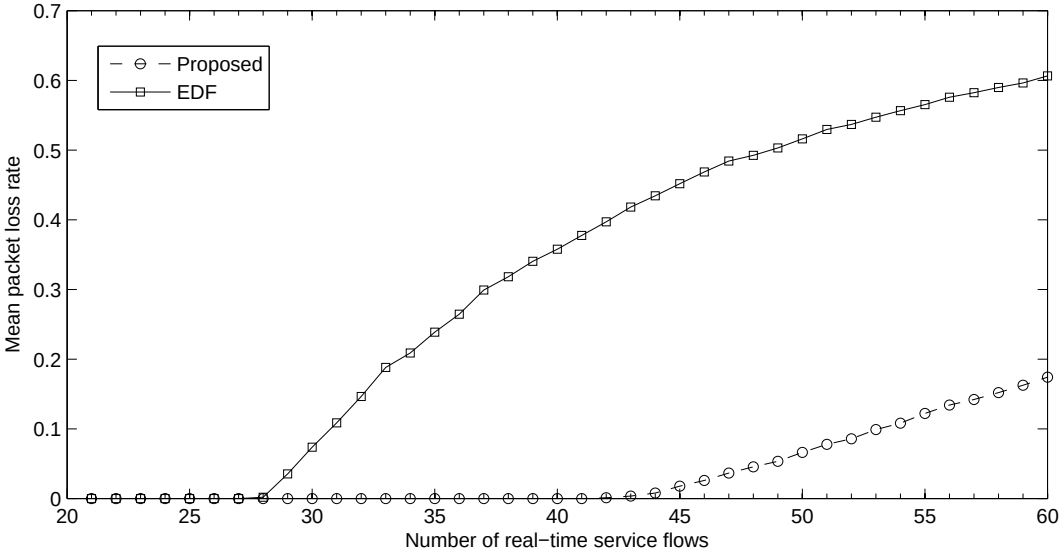


Figure 2.8: System-wide packet loss rate for real-time services

lay violation probabilities with various number of service flows in the system. To find the system-wide delay violation probability, we compute the delay experienced by each transmitted packet and then determine how many of these packets have failed to transmit within the guaranteed delay limit. We then divide the number of packets exceeding delay limit by the total number of transmitted packets. The figure clearly shows that the delay violation probability with EDF starts to rise much earlier than that with our proposed scheme. With increasing number of service flows, the limited number of slots in the DL-subframe gradually raises the earliest deadline that the EDF can serve during a frame time. After the number of service flows increases to a certain value, the earliest deadline the EDF can serve during a frame time crosses the maximum packet delay limit of all the backlogged connections. This can explain the sharp rise of packet delay violation probability with EDF when the number of service flows in Fig. 2.7 increases beyond 27. Our proposed scheme, on the other hand, pro-actively manages the backlogs at the connection queues and is targeted towards meeting the packet delay guarantees in contrast to EDF's goal of minimizing overall packet delay in the system. It can thereby achieve a lower system-wide packet delay violation probability for more number of service flows, as evident in Fig. 2.7. The delay violation probability starts to rise when the number of service flows is more than 36; the rate of this increase, however, declines considerably with more than 45 service flows. Note that we consider the delay experienced by admitted packets only. Therefore, this gradual decline means the packet loss rate is also on the rise with more than 45 service flows, as the scheme is working to keep delay violation probability of admitted packets in check.

In Fig. 2.8, we compare the packet loss performances achieved by EDF and our proposed scheme. Until the number of service flows in the system reaches 28, the system wide packet loss rates of both EDF and our proposed scheme remain insignificant. The packet

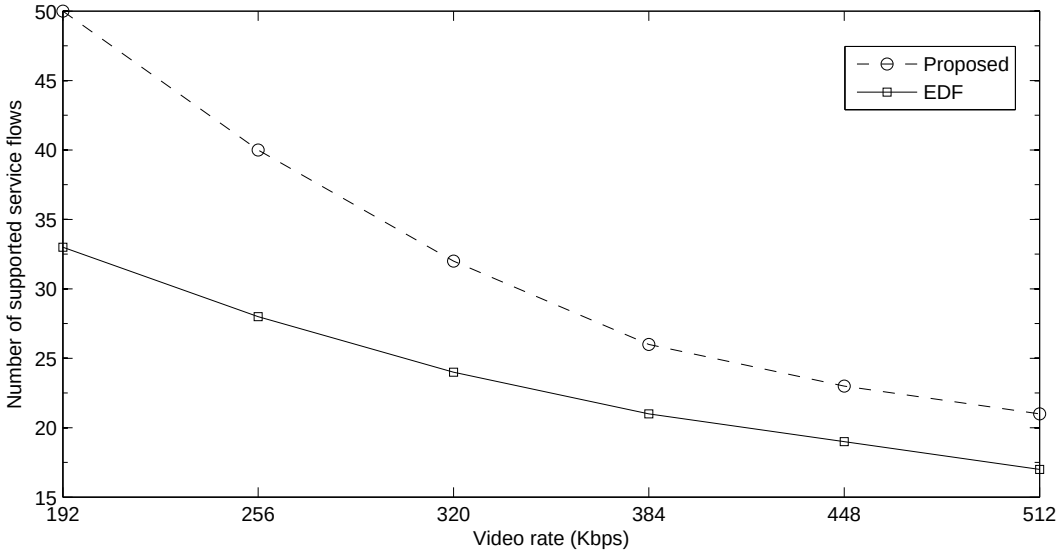


Figure 2.9: Number of supported service flows (real-time video streams) vs. their bit rates

loss rate with EDF then starts rising pretty rapidly. With our proposed scheme, on the other hand, we can increase the number of service flows to 45 without incurring any significant packet loss. Note that, the model parameters used in our resource allocation scheme were derived to keep packet loss rate at individual service flows within 1%. The packet loss rates drawn in Fig. 2.8, however, is averaged across all the real-time service flows in the downlink. Nevertheless, the results show that our scheme is more effective in preventing unwanted packet losses by better controlling the backlog buildups at the connection queues associated with the service flows.

Although the results presented in Figs. 2.7-2.8 clearly shows that our proposed scheme achieves significant improvement of both delay and packet loss performances across the system, we further investigate how the individual service flows are affected. Fig. 2.9 shows how many service flows can be served without violating the delay and packet loss guarantees for any of the service flows. We vary the bit rate of the video connection associated

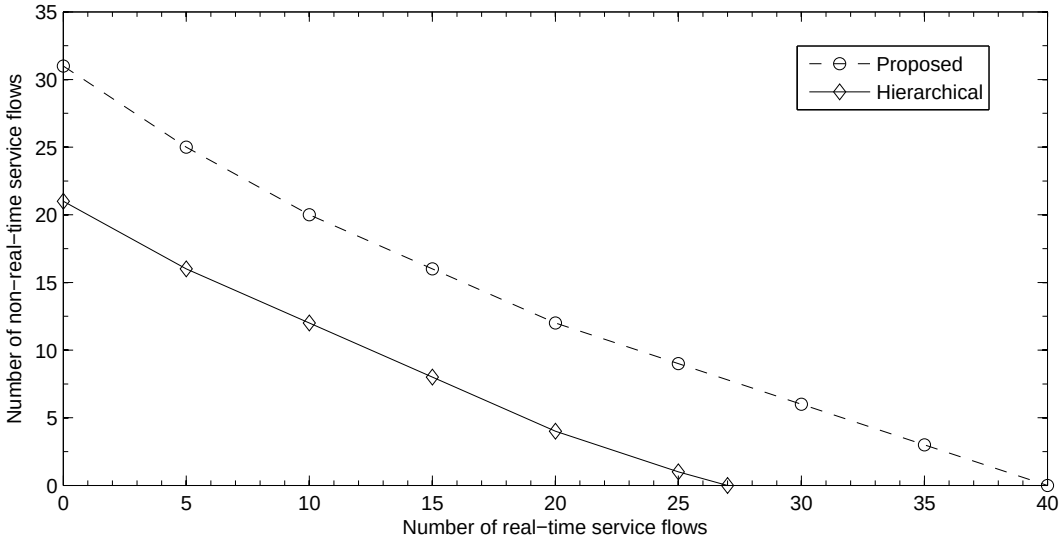


Figure 2.10: Number of supported non-real-time and real-time service flows competing simultaneously

with each downlink service flow. To vary the bit rate, we modify the video frame sizes by multiplying them by a common factor; I,P and B frames are equally affected. For each service flow, the maximum packet delay is guaranteed to be 50 ms with 95% probability level and maximum allowable packet loss rate is set to 1%. We compare the performances of the proposed scheme and the EDF. As can be seen, our proposed scheme can serve more service flows with the specified QoS support. The performance advantage of our scheme remains superior on a wide range of offered loads on the video connections. The results show that, with our proposed scheme, the WiMAX BS can admit significantly more number of MSs that are seeking service for real-time applications.

QoS Support for a Heterogenous Traffic Mix

The above results are indicative of the performance benefits of our proposed scheme in providing QoS for both real-time and non-real-time services. We further our analysis to heterogeneous traffic scenarios where both real-time and non-real-time applications are

supported simultaneously in the system. To this end, we attach real-time traffic service flows with downlink connections to a given number of MSs and configure the rest of the MSs in the system as the receiving ends of non-real-time service flows. We attach one service flow only to each MS. We vary the number of real-time service flows in the system and determine how many non-real-time service flows can be supported simultaneously without violating the QoS requirements of any of the real-time or non-real-time service flows. The traffic stream for the real-time service flows are generated from the same MMPP model as was used in the previous set of set of simulations. The bit rate of the real-time video stream was set to 256 Kbps. For the non-real-time traffic, we adopt the 2-MMPP model used before and set its intensity parameter $\rho = 1$. We compare the performance of our proposed scheme with a hierarchical resource allocation scheme. The design of the hierarchical scheme is similar to [9], which assigned higher priority to rtPS class over nrtPS class and proposed EDF and WFQ for the intra-class scheduling for these service classes respectively. Accordingly, the hierarchical scheme implemented in our simulation assigns higher priority to real-time service flows over non-real-time service flows. Given the benefits of WDRR over WFQ as discussed earlier, the scheme implements WDRR instead of WFQ for allocating resource among non-real-time service flows, whereas EDF is retained for resource allocation among real-time service flows.

Fig. 2.10 presents the number of non-real-time service flows that can be supported simultaneously with different number of real-time service flows. The QoS requirement for the non-real-time service flows is set to a maximum packet loss rate of 1%, whereas the real-time service flows are guaranteed a maximum packet delay of 50 ms with 95% probability and a maximum packet loss rate of 1%. We can see that, compared to the hierarchical scheme, our proposed scheme can provide QoS for more number of service

flows when both real-time and non-real-time applications are competing for service. With the hierarchical scheme, the queues attached to non-real-time traffic are serviced only after the queues of real-time traffic are exhausted. As a result, some non-real-time traffic queues might not receive adequate resource to meet their QoS or might even starve while the real-time traffic queues are allocated more than enough resource. Our proposed scheme, on the other hand, does not need such priority-based mechanism to protect the QoS of real-time service flows; individual service flows are allocated resource according to their QoS needs regardless of their service classes. The benefit of such isolation of service compared to a priority-based hierarchical scheme is evident in the results plotted in Fig. 2.10. These findings confirm that our proposed scheme not only excels in providing QoS for individual application types but also better protects their QoS from each other in a heterogeneous traffic scenario.

2.7 Related Work

Previous work on resource allocation on WiMAX systems has evolved over time as the standardization activities have matured. In the uplink, the resource allocation problem is often jointly addressed with bandwidth request/grant mechanisms as a WiMAX BS does not have direct access to queues located at individual MSs. In this work, we focus on downlink resource allocation problem, and for a survey on uplink resource allocation schemes, we refer readers to [16].

For the basic 802.16 standard, which was introduced with single-carrier PHY, the resource allocation problem, however, is very similar to that in TDMA-based cellular systems. This also applies to OFDM PHY, as with IEEE 802.16d, where all subcarriers are allocated as a whole to a particular MS during a frame either in the uplink or the downlink

direction. Many of the earlier resource allocation solutions were proposed assuming time division multipexing without focusing on any specific PHY. They are often packet scheduling schemes based on traditional time-domain radio resource scheduling solutions and are customized to make compatible with the scheduling service classes defined in the 802.16 standard. Some of these schemes are designed for specific traffic types such as video, VoIP and targets the relevant service classes. For delay sensitive applications that use rtPS scheduling service, EDF and Modified Largest Weighted Delay First (MLWDF) [17] were evaluated in previous research. Classical wired network schedulers such as Weighted Round Robin (WRR), WFQ and their variants were also proposed and evaluated for both delay-sensitive rtPS and throughput-guaranteed nrtPS applications. Wongthavarawat and Ganz [9] proposed a hierarchical scheduling scheme where EDF and WFQ are used for intra-calss scheduling of rtPS and nrtPS connections respectively, and a simple priority based scheme is adopted for allocating resource among different service classes. The solution proposed by Wang and Dittman in [18] is another example of hierarchical scheduling scheme. Among other solutions, Niyato and Hossain [20] proposed a scheduling mechanism that distributes the available bandwidth of the WiMAX channel to the connections belonging to different service classes by using a joint queueing and optimization model; the aim was to maximize the utility which is dependent on the average throughput and delay performance achieved by connections belonging to different QoS classes. In [19], Liu et al. proposed a cross-layer scheduling scheme that assigns a fixed amount of bandwidth to UGS connections and then uses a priority function to determine the connection to be scheduled for the remaining bandwidth in a frame. The priority function for each connection is calculated using the channel quality, service status and QoS requirments. Only one non-UGS connection is scheduled in a frame, which might result in inefficient resource usage. Lera

et al. proposed a schemer that tweaked the WF^2Q+ algorithm in [21] to work for per-class queues rather than per-flow queues that WF^2Q is originally designed for. Chen et al. in [22] introduced a fuzzy logic based scheduling algorithm to provide intra-class QoS guarantees while maintaining intra- and inter-class fairness. The online computation requirement with the scheme, however, could be prohibitively high.

Resource allocation schemes specifically proposed for OFDMA-based WiMAX systems, however, follow two main avenues of research. The first avenue of research considers subcarrier/subchannel assignment along with rate and power adaptation across users mainly to improve system level performance such as capacity and throughput. These solutions have basic similarity with those for OFDMA resource allocation problem in general. Examples of such solutions are [23, 24, 25, 26] and their main limitation is that the goal is to improve system performance but not necessarily to provide specific QoS to applications and service. The second category of solutions involve frame-by-frame scheduling of OFDMA subchannels or slots among admitted service flows or connections. Yahiya et al. [27] proposed a cross-layer scheduling mechanism that assigns priorities to UGS, rtPS and nrtPS connections for each available subchannel. The scheme, however, has the limitation of assuming an AWGN channel model and of possible starvation of applications that have low traffic intensity. Performance results only showed average system throughput and packet loss probability achieved for different number of users. In [28], Mehrjoo et al. proposed a queue- and channel-aware schedular, called opportunistic stable queue scheduling for WiMAX downlink. The scheduler, however, aimed to improve overall throughput and packet dropping probability experienced by MSs in the system and did not provide mechanism to meet QoS guarantees for individual connections. In [29], Ali et al. proposed a cross-layer scheme where the scheduling priorities of different connections are calcu-

lated using channel quality, service class and QoS requirements of the connections. The optimization-based solution is likely to suffer from very high implementation complexity. The average delay and throughput performances achieved by different service classes over time were presented, but how well QoS guarantees for individual connections could be met was not shown in the simulation results. Wan et al. took a similar approach in [30], but resorted to priority-based scheduling to avoid complexity of an optimized solution. Average system throughput, average rtPS packet loss rate and average nrtPS packet delay were presented against different channel SNR. Effectiveness in meeting QoS guarantees for individual connections was not presented.

2.8 Conclusion

In this chapter, we have proposed a novel resource allocation framework for downlink service flows in IEEE 802.16e mobile WiMAX systems. At first, we have developed a queueing analytical model that relates important performance measures of an admitted service flow to its packet arrival statistics and a parameter in the packet scheduling scheme. Referring to this parameter as the model parameter for the service flow, we have demonstrated how the output from the queueing model can be utilized to set the appropriate value for this model parameter. This value of the model parameter essentially allows to compute the minimum number of packets that need to be scheduled from the queue of the service flow in order to meet its QoS requirement. Each admitted service flow is associated with a model parameter. Next, we have proposed a cross-layer adaptive and queue-aware resource allocation scheme that utilizes these parameter values along with the channel quality and backlog information to schedule packets and allocate OFDMA slots among downlink QoS connections. The proposed model-based approach offers simplicity of design for the on-

line component, flexibility to the service providers in defining QoS parameters and better compatibility to the IEEE 802.16e standard. Instead of complexities involved with installing separate scheduling schemes within the same resource allocation framework for real-time and non-real-time traffic, our model-based approach offers an efficient unified way to handle both types of traffics and their diverse QoS requirements. We have run simulation experiments to evaluate the performance of our proposed scheme in providing QoS for VBR real-time and non-real-time multimedia and data services. Compared to EDF scheduling, which is specialized for allocating resource to delay-sensitive applications, our proposed scheme significantly reduces delay violation probability and packet loss rate for real-time services. As a result, the number of real-time service flows that can be supported by the system is also improved considerably. For non-real-time services, our proposed scheme achieves much better packet loss rate performance and can support significantly more service flows compared to throughput conscious WDRR scheduling. We have also shown that, even without priority-based or complex hierarchical approach, our proposed scheme can successfully protect QoS of individual service flows in a heterogenous traffic scenario involving both real-time and non-real-time services.

2.9 Bibliography

[1] IEEE Std. 802.16-2001, "IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Apr. 2002.

[2] IEEE 802.16e-2005, "IEEE Standard for Local and Metropolitan Networks – Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1," Feb. 2006.

[3] IEEE 802.16-2004, "IEEE standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Oct. 2004.

[4] H. Kobayashi and B. L. Mark, *System Modeling and Analysis: Foundations of System Performance Evaluation*, Prentice Hall, 2008.

[5] P. Salvador, A. Pacheco, and R. Valadas, "Multiscale Fitting Procedure using Markov Modulated Poisson processes," *Telecommunications Systems*, vol. 23, no. 1-2, pp. 123-148, Jun. 2003.

[6] D.P. Heyman and D. Lucantoni,"Modeling Multiple IP Traffic Streams with Rate Limits," *IEEE/ACM Transactions on Networking*, vol. 11, no. 6, pp. 948-958, Dec. 2003.

[7] M. Shreedhar and G. Varghese, "Efficient Fair Queuing Using Deficit Round Robin," *IEEE/ACM Transactions on Networking*, Vol.4, No.3, pp. 375-385, June 1996.

[8] A. Demers, S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm," *Proc. ACM SIGCOMM*, Austin, TX, pp. 1-12, Sep. 1989.

[9] K. Wongthavarawat and A. Ganz, "Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems," *International Journal of Communication Systems*, vol. 16, no. 1, pp. 81-96, Feb. 2003.

[10] O. Gusak, N. Oliver, and K. Sohraby, "Performance Evaluation of the 802.16 Medium Access Control Layer," *Lecture Notes on Computer Science*, vol. 3280/2004, pp. 228-237, Oct. 2004.

[11] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support," *IEEE Transactions on Mobile Computing*, vol. 6, no. 1, pp. 26-38, Jan. 2007.

[12] J. Lakkakorpi, A. Sayenko, and J. Moilanen, "Comparison of Different Scheduling Algorithms for WiMAX Base Station: Deficit Round-Robin vs. Proportional Fair vs. Weighted Deficit Round-Robin," *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1991-1996, Las Vegas, NV, Mar. 2008.

[13] R. Srinivasan, S. Timiri, A. Davydov, and A. Papathanassiou, "Downlink spectral efficiency of Mobile WiMax," *Proc. IEEE Vehicular Technology Conference*, Dublin, Ireland, Apr. 2007.

[14] ITU-R Task Group 8/1 "Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000," Recommendation ITUR M.1225, 1999.

[15] G. Dan, V. Fodor, and G. Karlsson, "On the Effects of The Packet Size Distribution on FEC Performance," *Computer Networks*, vol. 50, no. 8, pp. 1104-1129, Jun. 2006.

[16] N. A. Ali, P. Dhrona, and H. Hassanein, "A Performance Study of Uplink scheduling Algorithms in Point-to-Multipoint WiMAX Networks," *Computer Communications*, vol. 32, no. 3, pp. 511-521, Feb. 2009.

[17] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing Quality of Service over a Shared Wireless Link," *IEEE Communication Magazine*, vol. 39, no. 2, pp. 150-154, Feb. 2001.

[18] H. Wang and L. Dittmann, "Adaptive Radio Resource Allocation in Hierarchical QoS Scheduling for IEEE 802.16 Systems," *Proc. IEEE Global Telecommunications Conference*, pp. 4769-4774, Washington, D.C., Nov. 2007.

[19] Q. Liu, X. Wang, and G. B. Giannakis, "A Cross-Layer Scheduling Algorithm with QoS Support in Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 839-847, May 2006.

[20] D. Niyato and E. Hossain, "A Queuing-Theoretic and Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks," *IEEE Transactions on Computers*, vol. 55, no. 11, pp. 1473-1488, Nov. 2006.

[21] A. Lera, A. Molinaro, and S. Pizzi, "Channel-Aware Scheduling for QoS and Fairness Provisioning in IEEE 802.16/WiMAX Broadband Wireless Access Systems," *IEEE Network*, vol. 21, no. 5, pp. 34-41, Sept.-Oct. 2007.

[22] C-L. Chen, J-W. Lee, C.-Y. Wu, and Y.-H. Kuo, "Fairness and QoS Guarantees of WiMAX OFDMA Scheduling with Fuzzy Controls," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Article ID 512507, 14 pages, Jan. 2009.

[23] A. Biagioni, R. Fantacci, D. Marabissi, and D. Tarchi, "Adaptive Subcarrier Allocation Schemes for Wireless OFDMA Systems in Wimax Networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 2, pp. 217-225, Feb. 2009.

[24] G. Miao and N. Himayat, "Low Complexity Utility Based Resource Allocation for 802.16 OFDMA Systems," *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1465-1470, Las Vegas, NV, Mar.-Apr. 2008.

[25] V. Singh and V. Sharma, "Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks," *Proc. IEEE Wireless Communication and Networking Conference*, vol. 2, pp. 984-990, Las Vegas, NV, Apr. 2006.

[26] C. Tarhini and T. Chahed, "AMC-Aware QoS Proposal for OFDMA-Based IEEE 802.16 WiMAX Systems," *Proc. IEEE Global Telecommunications Conference*, pp. 4780-4784, Washington, D.C., Nov. 2007.

[27] T.A. Yahiya, A.-L. Beylot, and G. Pujolle, "Cross-Layer Multiservice Scheduling for Mobile WiMAX Systems," *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1531-1535, Las Vegas, NV, Mar.-Apr. 2008.

[28] M. Mehrjoo, X. Shen, and K. S. Naik, "A Joint Channel and Queue-Aware Scheduling for IEEE 802.16 Wireless Metropolitan Area Networks," *Proc. IEEE Wireless Communications and Networking Conference*, Hong Kong, pp. 1549-1553, Mar. 2007.

[29] N.A. Ali, M. Hayajneh, and H. Hassanein, "Cross Layer Scheduling Algorithm for IEEE 802.16," *Proc. IEEE International Conference on Communications*, pp. 3858-3862, Beijing, China, May 2008.

[30] L. Wan, W. Ma and Z. Guo, “A Cross-Layer Packet Scheduling and Subchannel Allocation Scheme in 802.16e OFDMA System,” *Proc. IEEE Wireless Communications and Networking Conference*, pp. 1865-1870, Hong Kong, Mar. 2007.

Chapter 3

Cross-Layer Analysis of Downlink V-BLAST MIMO Transmission Exploiting Multiuser Diversity[1]

3.1 Introduction

Ever increasing demand for high data rate transmission over wireless systems has opened many new avenues of research in wireless communication. Multiple Input Multiple Output (MIMO) wireless technology is receiving profound attention in this respect because it can enhance spectral efficiency significantly in rich scattering wireless environment [1]. Besides increased data rate, a major goal of future wireless systems design is to provide proper Quality of Service (QoS) in terms of performance measures such as delay, packet loss, packet throughput to both data and time intensive applications that are increasingly

[1]A version of this chapter has been published. Rashid, M.M., Hossain, E. and Bhargava V.K. (2009) Cross-Layer Analysis of Downlink V-BLAST MIMO Transmission Exploiting Multiuser Diversity. IEEE Trans. on Wireless Commun. 8(9):4568–4579.

being deployed on wireless systems. Although improving physical layer data rate is a major step toward this objective, this will often not be enough. This is especially true for multiuser systems where many users have to compete for the limited available radio resource. A cross-layer design approach is a more viable approach that would take into consideration both physical and Medium Access Control (MAC) layer aspects of the communication as well as the traffic characteristics and QoS requirement of the applications. Studying the cross-layer effects on the QoS experienced by the users in such systems is an important step towards this objective.

Among many signal processing techniques available for MIMO wireless systems, V-BLAST [2] is well known for its ability to offer excellent performance-complexity tradeoff. The V-BLAST aims to achieve MIMO potential by employing spatial multiplexing, which can theoretically offer a nearly linear increase in system capacity with increasing number of transmit antennas–provided thatthe number of receive antennas is greater than or equal to the number of transmit antennas. In a multiuser scenario, how well this improved capacity can be utilized also depends on how transmissions are scheduled among the users. Exploiting multiuser diversity in the scheduling process has been shown to achieve optimal system throughput for Single Input Single Output (SISO) systems [3]. Multiuser diversity can also be exploited to the efficient utilization of the increased system capacity in MIMO systems [4]. In fact, simple zero-forcing (ZF) receivers in conjunction with multiuser diversity have been shown to achieve a large portion of MIMO downlink capacity [5, 6, 7]. By considering the linear ZF receiver with the V-BLAST transmission model, the MIMO channel can be decoupled into parallel independent subchannels. The transmit antennas have a one-to-one association with these subchannels, which can now transmit independent streams or substreams [4]. In a multiuser scenario, multiuser diversity can be exploited

further by utilizing this extra dimension to opportunistically assign different transmit antennas to different users over time. The *independent stream scheduler* proposed by Heath Jr. *et al.* in [4] is designed based on this idea. As a generalization of the throughput optimal scheduling for multiuser diversity in SISO systems, it is expected to offer excellent throughput performance for the multiuser MIMO systems considered in this chapter.

In this chapter, we consider the *independent stream scheduler* in the downlink to exploit multiuser diversity in a V-BLAST MIMO wireless system and develop a queuing analytic model to determine the cross-layer effects on important QoS measures. We model the independent subchannels associated with each of the transmit antennas by a Finite State Markov chain (FSMC) and consider Adaptive Modulation and Coding (AMC) employed in the physical layer to achieve a predefined packet error rate. From the queuing model, we derive important performance measures at the MAC layer. From the model output, these performance measures can be related to the physical layer parameters and traffic arrival statistics. Example applications of the queuing model are also outlined which include a model based admission control scheme.

Although many related research efforts have dealt with performance analysis (e.g. [8, 9, 5, 10, 11]) and performance enhancement (e.g. [12, 13, 14, 15]) of multiuser MIMO systems, their focus in most cases has been on the capacity, throughput and/or outage probability achievable at the physical layer. We, in contrast, propose a very elaborate cross-layer analytical model to relate the physical layer parameters with the MAC layer performance measures in presence of multiuser diversity exploited in the resource scheduling process. As in multiuser SISO systems, multiuser diversity has been shown to have a great potential in improving system performance in multiuser MIMO systems as well. Heath Jr. *et al.* in [4] have presented simulation analysis on various ways to exploit the multiuser diver-

sity in multiuser MIMO systems and their effects on the spectral efficiency of the system. Airy *et al.*, on the other hand, have presented analytical models in [16] to compare two opportunistic scheduling techniques in this context. The performance measures are the raw bit rate achievable from the system and the average delays experienced by fixed sized files served from different users. The results from these two research works have motivated us to choose the particular scheduling scheme assumed in our chapter. However, unlike the authors in these works, we develop a cross-layer analytical model that captures the layered aspect of the system design, derives more comprehensive performance measures such as packet delay statistics, packet throughput and loss rate, and relates these performance to important physical layer and traffic parameters for better system design. Our work also considers more realistic traffic model that captures the potential burstiness in the arrival process.

The rest of the chapter is organized as follows. In section 5.2, we describe the MIMO system model and multiuser diversity scheme that we are considering in this chapter, and discuss relevant assumptions. We formulate the queuing model and derive important performance measures from the model in section 3.3. In section 5.4, we present some selected numerical results derived from the queuing model. Some example applications of the model are presented in section 5.5. Section 6 concludes the chapter with some future research directions.

3.2 System Model and Assumptions

We consider a MIMO wireless system comprising of a number of users covered by a single base station. The model presented in this chapter focuses on downlink scheduling. Therefore, the transmission buffers for the users are assumed to be located at the MAC layer in

the base station. We consider limited space in these transmission buffers and assume that downlink traffic arrive at these transmission buffers following some stochastic traffic arrival process. The MIMO system consists of V-BLAST transmission model with a ZF detector at the receiver side.

Similar to the MAC layer model used in [17], packets arrive from the higher layers to the MAC layer where they are grouped into MAC layer frames. The system is assumed to be time slotted with the slot size equal to the duration of a frame transmission. This duration, T_f is fixed although the size (in bits) of a frame might vary from frame to frame due to the AMC used to adaptively adjust the transmission rate at different transmit antennas. Considering a slowly varying fading environment, we assume that the channel remains static over a frame duration.Assuming a fixed packet size, the number of packets transmitted in a frame duration will vary accordingly. The MIMO transmission model allows independent streams to be transmitted through different transmit antennas, which are located at the base station for downlink transmission. The V-BLAST MIMO transmission model and the receiver architecture are briefly described in the next subsection.

3.2.1 MIMO Transmission Model and Receiver Architecture

We consider V-BLAST transmission model with t transmit antennas and r receive antennas ($r \geq t$). If $\mathbf{s} = [s_1, s_2, \cdots, s_t]^T$ is the vector of symbols transmitted from the t transmit antennas during a symbol time and $\mathbf{H}$ is the $t \times r$ channel matrix at that time instant, then the received signal vector $\mathbf{v} = [v_1, v_2, \cdots, v_r]^T$ can be expressed as

$$\mathbf{v} = \mathbf{sH} + \mathbf{w}, \tag{3.1}$$

where $\mathbf{w} = [w_1, w_2, \cdots, w_r]^T$ is the noise vector. We assume that the elements of this noise vector are zero mean i.i.d Gaussian random variables with variance equal to N_0.

For the detection of the symbol vector from the V-BLAST transmitter, several detection algorithms including minimum mean square error (MMSE), successive interference cancellation (SIC) and ZF have been proposed for the MIMO receiver [2]. Choice of a receiver algorithm, in general, is a tradeoff between performance and implementation cost/-complexity. ZF is simple and offers low cost implementation, but is known to suffer from the effect of noise enhancement. MMSE is more complex than ZF, but offers better noise suppression. In contrast, with much higher complexity, SIC-based receiver could potentially achieve the full capacity gain of the MIMO system [5, 18]. However, in a multiuser scheduling system, multiuser diversity can offer a natural way to mitigate the drawback of poor noise performance by ZF; in fact, it has been shown that the combination of multiuser diversity with zero forcing can lead to a low-cost and effective way for achieving a large portion of MIMO downlink capacity [5, 6, 7]. Besides these, the fact that the post-detection SNR distribution for ZF receiver is known allows for nice analytical tractability for our cross-layer queueing analysis.

In ZF detection, the receiver, which is assumed to have the perfect knowledge of the channel, multiplies the signal vector received during a symbol time by the pseudo-inverse $\mathbf{H}^{\dagger}$ of the channel matrix at that time instant. Assuming a transmission power of E_s per transmit antenna, the post-detection SNR associated with the i^{th} transmit antenna i.e. the post-detection SNR at the i^{th} subchannel can be expressed as [19]

$$\gamma_i = \frac{E_s}{N_0[\mathbf{H}^*\mathbf{H}]_{ii}^{-1}} = \frac{\gamma_0}{[\mathbf{H}^*\mathbf{H}]_{ii}^{-1}}, i = 1, 2, \cdots, t, \tag{3.2}$$

where $\mathbf{H}^*$ is the Hermitian transpose of $\mathbf{H}$ and $\gamma_0 = E_s/N_0$. Due to this MIMO system model, t independent streams can be fed into t transmit antennas. However, a single data stream can be divided into t independent substreams for transmission as well.

3.2.2 Modeling the MIMO Channel

We assume a point-to-point Rayleigh flat fading channel with transmit antenna correlation, but without any receive antenna correlation. In cellular mobile systems, in general there are several local scatterers around the receiver or hand set. Due to the effect of these local scatterers, fading of the received signals are more likely to be uncorrelated across the receiver antennas. However, at the base stations trasmit antennas are often mounted on top of high buildings or towers. Local scatterers are not usually present around these transmit antennas and therefore, the signal fading across these antennas tends to be correlated. With this assumption, the probability density function (pdf) of the post-detection SNR associated with the i^{th} subchannel can be shown to be Gamma distributed as follows due to [19]:

$$f(\gamma_i) = \frac{\gamma_i^{r-t}}{(\gamma_0/\sigma_i^2)^{r-t+1}\Gamma(r-t+1)} \exp\left(-\frac{\gamma_i}{\gamma_0/\sigma_i^2}\right), \tag{3.3}$$

where σ_i^2 is the i^{th} diagonal entry of the inverse of the transmit correlation matrix.

We assume that the post detection SNR of each subchannel is quantized into a number of discrete values which are fed back to the transmitter to adjust the modulation level. These discrete values essentially represent the subchannel states. If $Z+2$ thresholds $\{\phi_0, \phi_1, \cdots, \phi_{Z+1}\}$ are used to quantize a subchannel, the subchannel will have $Z+1$ states $\mathcal{C} = \{0, 1, \cdots, Z\}$ and it will be in state j when the post-detection SNR is within the range $(\phi_j, \phi_{j+1}]$. Here it is assumed that the SNR thresholds satisfy the following inequality: $0 = \phi_0 \leq \ldots \leq \phi_j \leq \ldots \leq \phi_{Z+1} = \infty$.

Assuming correlated block fading, we model the variation in each subchannel over time using an FSMC, which is a well accepted channel model for slowly varying environment [20]. For a given doppler spread, the transition probabilities between the states of the FSMC can be found from the post-detection SNR distribution following few steps as described, for example, in [17]. In the transmitter, for each of the transmit antennas, AMC is applied to map a unique transmission rate to each of the states of the associated subchannel. When the subchannel is in state j, ζ_j packets can be transmitted per frame time through this subchannel using transmission mode j. These transmission modes are selected such that a prescribed average packet error rate $\overline{PER}$ can be maintained at each subchannel state.

3.2.3 Multiuser Diversity Scheme

The objective of downlink multiuser diversity systems is to maximize long run average system throughput. The opportunistic scheduling scheme proposed in [3] for SISO systems has been shown to achieve this objective by assigning the time slot in consideration to the user that has the best channel quality at that time instant. Heath Jr. *et al.* in [4] have proposed generalization of this scheme for MIMO systems where the MIMO channel can be modeled as a series of parallel SISO equivalent channels. The scheduler is called *independent stream scheduler* which assigns each transmit antenna at each time slot independently to the user who has the highest state in the subchannel associated with the transmit antenna. If more than one users exhibit the highest state for that subchannel, then one of those users is picked up randomly to be assigned to that transmit antenna. As a result, a single user may be assigned zero, one or more than one transmit antennas at any time instant.

Note that optimality results similar to [3] is not possible for MIMO multiuser diversity systems due to the dependence of throughput not only on the received SNR but also on

the subspace structure of the channel matrix and the receiver [4]. The *independent stream scheduler* considered in this chapter is thus not a provably optimal MIMO multiuser diversity scheme. However, as a generalization of the throughput optimal SISO multiuser diversity scheme, it is expected to offer excellent throughput performance for the multiuser MIMO systems considered in this chapter.

3.3 Queueing Model and Analysis

We consider a system where all the users have the same number of receive antennas and have similar channel fading characteristics. These assumptions of a homogeneous system allow us to model the system from the perspective of a tagged user. Without loss of generality, the following analysis is done considering user 1 as the tagged user and therefore, the user number index for user 1 is omitted. Also note that the time slot duration in the discrete time queuing model is equal to a MAC layer frame transmission time, T_f.

We assume that packets arrive in the buffer following a Batch Bernoulli arrival process. We also assume that packets arriving during time interval $n-1$ cannot be served until time interval n at the earliest. This arrival process can be described by a vector $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \cdots, \alpha_N]$, where N is the maximum batch size, the maximum number of packets that can arrive during a time slot. In our model, α_0 is the probability that no packets arrives and $\alpha_i (1 \leq i \leq N)$ is the probability that i packets arrive in a time slot. The average arrival rate λ can be calculated as follows:

$$\lambda = \sum_{i=1}^{N} i\alpha_i. \tag{3.4}$$

The assumed Batch Bernoulli arrival model is very general which can capture different

level of burstiness in the traffic arrival process. The waiting packets in the queue are served in a first-come first-served manner. Besides, packet transmissions in a time slot are assumed to finish before arriving packets enter the queue.

As we have seen in previous section, due to the MIMO transmission model and the detection scheme we consider, the allocation of a transmit antenna to a user is equivalent to allocating the associated subchannel to that user. In the following, the term subcannel, when used, should be interpreted as such. Let the random variable $m_n^{(i)} \in \{1, 0\}$ represent whether i^{th} subchannel is assigned to the tagged user or not at time slot n due to the multiuser diversity scheme. Let $m_n^{(i)} = 1$ denote that the i^{th} subchannel is assigned to the tagged user at time slot n. Similarly $c_n^{(i)} \in \mathcal{C}$ denotes the state of the i^{th} subchannel at time slot n at tagged user's end. Therefore, for the i^{th} subchannel, the joint state indicating whether it is assigned to the tagged user and its state at time slot n at tagged user's end, is a vector denoted as $(m_n^{(i)}, c_n^{(i)})$. We can write the state transition matrix for vector states $(m_n^{(i)}, c_n^{(i)})$ as follows:

$$\mathbf{M}^{(i)} = \begin{bmatrix} \mathbf{M}_{0\to 0}^{(i)} & \mathbf{M}_{0\to 1}^{(i)} \\ \mathbf{M}_{1\to 0}^{(i)} & \mathbf{M}_{1\to 1}^{(i)} \end{bmatrix}, \tag{3.5}$$

where $\mathbf{M}_{k\to l}^{(i)}$ is the $Z \times Z$ matrix whose elements are defined as follows:

$$\mathbf{M}_{k\to l}^{(i)}(f, g) = \Pr\left(m_{n+1}^{(i)} = l, c_{n+1}^{(i)} = g | m_n^{(i)} = k, c_n^{(i)} = f\right). \tag{3.6}$$

The probability in (3.6) can be expressed as

$$\Pr(m_{n+1}^{(i)}, c_{n+1}^{(i)} | m_n^{(i)}, c_n^{(i)}) = \Pr(m_{n+1}^{(i)} | c_{n+1}^{(i)}, m_n^{(i)}, c_n^i) \times \Pr(c_{n+1}^{(i)} | c_n^{(i)}), \tag{3.7}$$

where we have used the fact that $\Pr(c_{n+1}^{(i)} | m_n^{(i)}, c_n^{(i)}) = \Pr(c_{n+1}^{(i)} | c_n^{(i)})$. Also, $\Pr(c_{n+1}^{(i)} | c_n^{(i)})$

can be determined from the transition probability matrix of the FSMC associated with the subchannel as $\Pr(c_{n+1}^{(i)} = g | c_n^{(i)} = f)$ is the probability that the i^{th} subchannel transits from state f to g. In addition, $\Pr(m_{n+1}^{(i)} | c_{n+1}^{(i)}, m_n^{(i)}, c_n^{(i)})$ can be written as :

$$\Pr(m_{n+1}^{(i)} | c_{n+1}^{(i)}, m_n^{(i)}, c_n^{(i)}) = \frac{\Pr(m_{n+1}^{(i)}, m_n^{(i)} | c_{n+1}^{(i)}, c_n^{(i)})}{\Pr(m_n^{(i)} | c_n^{(i)})}, \tag{3.8}$$

where we used the fact that $\Pr(m_n^{(i)} | c_{n+1}^{(i)}, c_n^{(i)}) = \Pr(m_n^{(i)} | c_n^{(i)})$. We show the computation process for these probabilities in the following. Since $m_{n+1}^{(i)}$ can be either 0 or 1, we can write $\Pr(m_{n+1}^{(i)} = 0 | c_{n+1}^{(i)}, m_n^{(i)}, c_n^{(i)}) = 1 - \Pr(m_{n+1}^{(i)} = 1 | c_{n+1}^{(i)}, m_n^{(i)}, c_n^{(i)})$. Therefore, showing the derivations for $m_{n+1}^{(i)} = 1$ will be sufficient.

Let us express the state of subchannel i at user $h (1 \leq h \leq S)$ at time slot n as $c_n^{(i,h)}$, where S is the total number of users in the system. Now, to find the denominator of the right side expression in (3.8), we can write

$$\Pr(m_n^{(i)} = 1 | c_n^{(i)} = f) = \sum_{c_n^{(i,2)}=0}^{Z} \cdots \sum_{c_n^{(i,S)}=0}^{Z} \Pr(m_n^{(i)} = 1, c_n^{(i,2)}, \cdots, c_n^{(i,S)} | c_n^{(i)} = f), \tag{3.9}$$

in which the inner probability expression of the right side can be derived as follows:

$$\Pr(m_n^{(i)} = 1, c_n^{(i,2)}, \cdots, c_n^{(i,S)} | c_n^{(i)} = f) = \begin{cases} 0, & \text{if } \exists_h (2 \leq h \leq S) \text{ s.t. } c_n^{(i,h)} > f \\ \frac{1}{u} \prod_{h=2}^{S} \Pr(c_n^{(i,h)}), & \text{otherwise} \end{cases}, \tag{3.10}$$

where u is the number of users that have the same state as that of the tagged user (user 1) on subchanne i at time slot n. Probabilities $\Pr(c_n^{(i)})$ can be found from steady state probabilities of FSMC for subchannel i. We can also derive $\Pr(m_n^{(i)} = 0 | c_n^{(i)} = f)$ as

$\Pr(m_n^{(i)}=0|c_n^{(i)}=f)=1-\Pr(m_n^{(i)}=1|c_n^{(i)}=f)$.

Now, let us derive the numerator of the right side expression of (3.8) for $m_{n+1}^{(i)}=1$. We can rewrite the expression as

$$\begin{aligned}&\Pr(m_{n+1}^{(i)}=1,m_n^{(i)}|c_{n+1}^{(i)},c_n^{(i)})\\&=\sum_{c_{n+1}^{(i,2)}=0}^{Z}\cdots\sum_{c_{n+1}^{(i,S)}=0}^{Z}\sum_{c_n^{(i,2)}=0}^{Z}\cdots\sum_{c_n^{(i,S)}=0}^{Z}\Pr(m_{n+1}^{(i)}=1,m_n^{(i)},c_{n+1}^{(i,2)},\cdots,c_{n+1}^{(i,S)},c_n^{(i,2)},\cdots,c_n^{(i,S)}|c_{n+1}^{(i)},c_n^{(i)}),\end{aligned}\tag{3.11}$$

in which the inner probability expression in the right side can be derived as follows for $\Pr(m_{n+1}^{(i)}=1,m_n^{(i)}=1|c_{n+1}^{(i)}=g,c_n^{(i)}=f)$ at the left side:

$$\begin{aligned}&\Pr(m_{n+1}^{(i)}=1,m_n^{(i)}=1,c_{n+1}^{(i,2)},\cdots,c_{n+1}^{(i,S)},c_n^{(i,2)},\cdots,c_n^{(i,S)}|c_{n+1}^{(i)}=g,c_n^{(i)}=f)\\&=\begin{cases}0, & \text{if } \exists_h(2\le h\le S)\text{ s.t. } (c_n^{(i,h)}>f \text{ or } c_{n+1}^{(i,h)}>g)\\ \frac{1}{uv}\prod_{h=2}^{S}\Pr(c_{n+1}^{(i)},c_n^{(i)}), & \text{otherwise}\end{cases},\end{aligned}\tag{3.12}$$

where v is the number of users that have same state for subchannel i as that of the tagged user (user 1) at time slot $n+1$. The probabilities $\Pr(c_{n+1}^{(i)},c_n^{(i)})$ can be derived from the FSMC of subchannel i as $\Pr(c_{n+1}^{(i)},c_n^{(i)})=\Pr(c_{n+1}^{(i)}|c_n^{(i)})\times\Pr(c_n^{(i)})$. Similarly, for $\Pr(m_{n+1}^{(i)}=1,m_n^{(i)}=0|c_{n+1}^{(i)}=g,c_n^{(i)}=f)$ at the left side of (3.11), we can derive the inner probability expression in the right side as

$$\begin{aligned}&\Pr(m_{n+1}^{(i)}=1,m_n^{(i)}=0,c_{n+1}^{(i,2)},\cdots,c_{n+1}^{(i,S)},c_n^{(i,2)},\cdots,c_n^{(i,S)}|c_{n+1}^{(i)}=g,c_n^{(i)}=f)\\&=\begin{cases}0, & \text{if } (\exists_h(2\le h\le S)\text{ s.t. } c_{n+1}^{(i,h)}>g)\text{ or }(\forall_h(2\le h\le S)\ c_n^{(i,h)}<f)\\ \frac{1}{v}\prod_{h=2}^{S}\Pr(c_{n+1}^{(i)},c_n^{(i)}), & \text{if } \exists_h(2\le h\le S)\text{ s.t. } c_n^{(i,h)}>f\\ \frac{u-1}{uv}\prod_{h=2}^{S}\Pr(c_{n+1}^{(i)},c_n^{(i)}), & \text{otherwise}\end{cases}\end{aligned}\tag{3.13}$$

Now considering all the transmit antennas or the associated subchannels, the process defined by state space $\mathcal{W}_n \triangleq \{(m_n^{(1)}, c_n^{(1)}, \cdots, m_n^{(t)}, c_n^{(t)}) | \forall_{i \in \{1,2,\cdots t\}} m_n^{(i)} \in \{0,1\}, c_n^{(i)} \in \mathcal{C}\}$ is a discrete-time Markov Chain (DTMC), whose transition matrix $\mathcal{W}$ can be written as $\mathbf{W} = \mathbf{M}^{(1)} \otimes \mathbf{M}^{(2)} \otimes \cdots \otimes \mathbf{M}^{(t)}$. Its steady state probability distribution $\boldsymbol{\varphi}$ can be found from solving the set of linear equations: $\boldsymbol{\varphi}\mathbf{W} = \boldsymbol{\varphi}$ and $\boldsymbol{\varphi}\mathbf{1} = 1$, where $\mathbf{1}$ stands for a column matrix containing 1 only in each entry.

3.3.1 Formulation as a Quasi-Birth-and-Death Process

Assuming a finite buffer size of Q, let us define a process $\mathcal{P}_n \triangleq \{(x_n, a_n) | 0 \leq x_n \leq Q, 0 \leq a_n \leq L\}$, where state variable x_n is the number of packets at the tagged user's queue at the beginning of time slot n, state variable a_n is the maximum number of packets that can be transmitted from that queue at time slot n and $L = t\zeta_Z$. Assuming that state transition occurs only at the slot boundaries and noting the fact that the state variables are designed to assume discrete values only, we can model the system process $\mathcal{P}_n$ using a DTMC. To derive the transition probability matrix of the DTMC, we consider the system dynamics as it evolves over one frame to the next frame. With an infinite buffer space $(Q = \infty)$, the state transition probability matrix for the DTMC describing the system can be written as

$$\mathbf{P} = \left[\begin{array}{c|c|c|c}
A_0^{(0)} & A_{1^+}^{(0)} \cdots \cdots A_{N^+}^{0} & & \\
\hline
\begin{array}{c} A_{1^-}^{(1)} \\ \vdots \\ A_{(L-N+1)^-}^{(L-N+1)} \\ \vdots \\ A_{L^-}^{(L)} \end{array} & \begin{array}{c} A_0^{(1)} \; A_{1^+}^{(1)} \cdots \cdots A_{N^+}^{(1)} \\ \vdots \qquad\qquad \ddots \\ A_{(L-N)^-}^{(L-N+1)} \cdots \cdots \cdots \cdots A_{1^-}^{(L-N+1)} \; A_0^{(L-N+1)} \; A_{1^+}^{(L-N+1)} \cdots A_{(N-1)^+}^{(L-N+1)} \\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ddots \qquad \vdots \\ A_{(L-1)^-}^{(L)} \cdots \cdots \cdots \cdots \cdots \cdots \cdots A_{1^-}^{(L)} \; A_0^{(L)} \end{array} & \begin{array}{c} A_{N^+} \\ \vdots \qquad \ddots \\ A_{1^+} \cdots \cdots A_{N^+} \end{array} & \\
\hline
\begin{array}{c} A_{L^-} \\ \qquad \ddots \\ \qquad\qquad A_{L^-} \\ \qquad\qquad\qquad\qquad \ddots \\ \qquad\qquad\qquad\qquad\qquad\qquad A_{L^-} \end{array} & \begin{array}{c} \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots A_{1^-} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots A_{(L-N+1)^-} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots A_{L^-} \end{array} & \begin{array}{c} A_0 \; A_{1^+} \cdots \cdots A_{N^+} \\ \vdots \qquad\qquad\qquad \ddots \\ A_{(L-N)^-} \cdots \cdots \cdots \cdots A_{1^-} \; A_0 \; A_{1^+} \cdots A_{(N-1)^+} \\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ddots \qquad \vdots \\ A_{(L-1)^-} \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots A_{1^-} \; A_0 \end{array} & \begin{array}{c} A_{N^+} \\ \vdots \qquad \ddots \\ A_{1^+} \cdots A_{N^+} \end{array} \\
\hline
 & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ddots & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ddots & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ddots
\end{array}\right]. \tag{3.14}$$

By constructing blocks of submatrices as shown in (3.14), we can obtain a *quasi-birth-and-death* (QBD) process with a transition matrix as follows:

$$\mathbf{P_{inf}} = \begin{bmatrix} \mathbf{B} & \mathbf{E} & & & & \\ \mathbf{C} & \mathbf{D} & \mathbf{G}_0 & & & \\ & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G}_0 & & \\ & & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G_0} & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{3.15}$$

However, we are interested in finding performance measures for systems with finite buffer space, which is more realistic. Setting $K = \lfloor Q/L \rfloor$, we can write the transition matrix in the following form:

$$\mathbf{P} = \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ K-1 \\ K \end{matrix} \begin{bmatrix} \mathbf{B} & \mathbf{E} & & & & & \\ \mathbf{C} & \mathbf{D} & \mathbf{G}_0 & & & & \\ & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G}_0 & & & \\ & & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G_0} & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G'}_0 \\ & & & & & \mathbf{G'}_2 & \mathbf{G'}_1 \end{bmatrix}. \tag{3.16}$$

The next step in the model involves deriving the submatrices inside $\mathbf{P}$. We proceed by considering a process defined by the state space $\mathcal{V}_n \triangleq \{a_n | 0 \leq a_n \leq L\}$. Now let b_k denote the set of states from state space $\mathcal{W}_n$ resulting in $a_n = k$, where $0 \leq k \leq L$. With $\mathbf{W}_{h \to h'}$ denoting the transition probability from state h to h', where both h and h' are elements of state space $\mathcal{W}_n$, we can construct the transition probability matrix $\mathbf{V}$ for state

space $\mathcal{V}_n$ as $\mathbf{V}_{i\to j} = \frac{\sum\limits_{h\in b_i}\sum\limits_{h'\in b_j}\varphi_h \mathbf{W}_{h\to h'}}{\sum\limits_{h\in b_i}\varphi_h}$, where $\mathbf{V}_{i\to j}$ denotes the transition probability from state i to state j.

Now, let us define a number of $(L+1)\times(L+1)$ matrices as follows

$$\mathbf{U}^{(g)} = \boldsymbol{\varepsilon}_{g+1}\mathbf{V}, \quad 0 \le g \le L,$$

where $\boldsymbol{\varepsilon}_{g+1}$ is a row vector of size $L+1$, in which all the elements are 0 except a 1 at position $g+1$. With these matrices defined, we can derive the inner submatrices of $\mathbf{P}$ in (3.14) as follows:

$$\mathbf{A}_{0}^{(0)} = \alpha_0\mathbf{V}; \quad \mathbf{A}_{d^+}^{(0)} = \alpha_d\mathbf{V}, 1 \le d \le N \tag{3.17}$$

$$\mathbf{A}_{d^-}^{(e)} = \begin{cases} \alpha_0 \sum\limits_{j=e}^{L} \mathbf{U}^{(j)}, & d = e \\ \sum\limits_{j=0}^{e-d-1} \alpha_j \mathbf{U}^{(j+d)} + \alpha_{e-d} \sum\limits_{j=e}^{L} \mathbf{U}^{(j)}, & e-N \le d \le e-1 \\ \sum\limits_{j=0}^{N} \alpha_j \mathbf{U}^{(j+d)}, & 1 \le d < e-N, \end{cases} \tag{3.18}$$

for $\quad 1 \le e \le L-1$

$$\mathbf{A}_{d^+}^{(e)} = \sum_{j=d}^{d+e-1} \alpha_j \mathbf{U}^{(j-d)} + \alpha_{d+e} \sum_{j=e}^{L} \mathbf{U}^{(j)}, 1 \le d \le N, 1 \le e \le L-1 \tag{3.19}$$

$$\mathbf{A}_{0}^{(e)} = \begin{cases} \sum\limits_{j=0}^{e-1} \alpha_j \mathbf{U}^{(j)} + \alpha_e \sum\limits_{j=e}^{L} \mathbf{U}^{(j)}, & 1 \le e \le N \\ \sum\limits_{j=0}^{N} \alpha_j \mathbf{U}^{(j)}, & N < e \le L \end{cases} \tag{3.20}$$

$$\mathbf{A}_{d^-} = \sum_{j=0}^{N} \alpha_j \mathbf{U}^{(j+d)}, 1 \leq d \leq L \tag{3.21}$$

$$\mathbf{A}_{d^+} = \sum_{j=d}^{N} \alpha_j \mathbf{U}^{(j-d)}, 1 \leq d \leq N \tag{3.22}$$

$$\mathbf{A}_0 = \sum_{j=0}^{N} \alpha_j \mathbf{U}^{(j)} \tag{3.23}$$

$$\mathbf{G'}_0 = \begin{bmatrix} \mathbf{G}_0 & \mathbf{0}_{L^2 \times K'L} \end{bmatrix} \tag{3.24}$$

$$\mathbf{G'}_1 = \begin{bmatrix} \mathbf{A}_0 & \mathbf{A}_{1^+} & \cdots & \mathbf{A}_{N^+} & & \\ \mathbf{A}_{1^-} & \mathbf{A}_0 & \mathbf{A}_{1^+} & \cdots & \mathbf{A}_{N^+} & \\ \vdots & & \ddots & & \ddots & \\ \mathbf{A}_{(L-N+K')^-} & \cdots & \cdots & \cdots & & \mathbf{A}'_{(N-1)^+} \\ \vdots & & \ddots & & & \vdots \\ \mathbf{A}_{L^-} & \cdots & \cdots & \cdots & \cdots & \mathbf{A}'_{(K'-1)^+} \\ \ddots & & \ddots & & & \vdots \\ & \mathbf{A}_{L^-} & \cdots & \cdots & \cdots & \mathbf{A}'_{1^+} \\ & & \mathbf{A}_{L^-} & \cdots & \cdots & \mathbf{A}'_0 \end{bmatrix} \tag{3.25}$$

$$\mathbf{A}'_{k^+} = \mathbf{U}^{(0)}\alpha_k + \sum_{j=k+1}^{N} \alpha_j \left[\mathbf{U}^{(j-k)} + \mathbf{U}^{(0)}\right], 1 \leq k \leq N-1 \tag{3.26}$$

$$\mathbf{G'}_2 = \begin{bmatrix} \mathbf{G}_2 \\ \mathbf{0}_{K'L \times L^2} \end{bmatrix}, \tag{3.27}$$

where $K' = Q - \lfloor Q/L \rfloor$.

3.3.2 Matrix-Analytic Solution for Steady State Distributions

With the system described by a QBD type Markov chain whose block matrices are derived in the previous subsection, we can now apply matrix-analytic method [21] to find its steady state distributions. From these steady state probability distributions, we can derive the steady state probability vectors $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_Q]$ for queue lengths in the buffer. For the detailed procedure to calculate $\boldsymbol{\pi}$, we refer readers to [21].

3.3.3 Derivation of Performance Measures

After obtaining the state vector, we can determine the performance measures of interest. We are mainly interested in the following four important performance measures: queue length distribution, packet loss rate, average throughput, and the distribution of access delay for packets belonging to the target queue. We denote by access delay the time between the moment a packet arrives at the queue and the moment the packet gets access into the medium for transmission. All these performance measures are assumed to be obtainable at the MAC layer.

Queue Length Distribution

The marginal probability $p(\ell)$ of finding ℓ packets in the queue at an arbitrary time instant can be written as:

$$p(\ell) = \boldsymbol{\pi}_\ell \mathbf{1}, 1 \leq \ell \leq Q. \tag{3.28}$$

Delay Distribution

The delay distribution for a packet belonging to the tagged user queue is derived by considering an absorbing Markov chain. The chain initializes from the state when a packet arrives the corresponding queue and gets absorbed in the states associated with the transmission of the packet. Note that, in this case, there is no arrival in that particular queue while the process moves towards the absorbing state. Therefore, the transition matrix $\mathbf{P}_{abs}$ of the chain can be constructed by setting $\alpha_0 = 1$ and $\alpha_i = 0 \quad (1 \leq i \leq N)$. The next step for finding the delay distribution is to determine the initial probability vector for this absorbing Markov chain. To this end, we first consider the probability vector associated with the number of packets including the arriving packet and the packets waiting ahead of it in the queue for service as seen by the arriving packet. This probability vector $\boldsymbol{\xi}^{(0)} = \left[\boldsymbol{\xi}_0^{(0)}, \boldsymbol{\xi}_1^{(0)}, \cdots, \boldsymbol{\xi}_{Q+N}^{(0)}\right]$ can be derived as:

$$\boldsymbol{\xi}_h^{(0)} = \begin{cases} \mathbf{0}, & h = 0 \\ \frac{1}{1-\alpha_0}\left[\sum_{i=1}^{h}\sum_{j=1}^{i}\sum_{k=0}^{L}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+k}\mathbf{U}^{(k)} + \sum_{i=h+1}^{N}\sum_{j=1}^{h}\sum_{k=0}^{L}\frac{\alpha_i}{i}\pi_{h-j+k}\mathbf{U}^{(k)}\right], & 1 \le h \le N \\ \frac{1}{1-\alpha_0}\left[\sum_{i=1}^{N}\sum_{j=1}^{i}\sum_{k=0}^{L}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+k}\mathbf{U}^{(k)}\right], & N < h \le Q-L+1 \\ \frac{1}{1-\alpha_0}\left[\sum_{i=1}^{N}\sum_{j=1}^{i}\sum_{k=0}^{Q-h+j}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+k}\mathbf{U}^{(k)}\right], & Q-L+1 < h \le Q \\ \frac{1}{1-\alpha_0}\left[\sum_{i=h-Q}^{N}\sum_{j=h-Q}^{i}\sum_{k=0}^{Q-h+j}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+k}\mathbf{U}^{(k)}\right], & Q < h \le Q+N \end{cases} \tag{3.29}$$

Note that the packets which find their destination queues full upon arrival are not admitted into the system. The probability that an arriving packet finds the corresponding queue full can be written as

$$p_{full} = \sum_{i=1}^{N} \boldsymbol{\xi}_{Q+i}^{(0)}. \tag{3.30}$$

Since we consider the delay experienced by the admitted packets only and since an arriving packet seeing a full queue will be dropped, the initial probability vector $\boldsymbol{\omega}^{(0)} = \left[\boldsymbol{\omega}_0^{(0)}, \boldsymbol{\omega}_1^{(0)}, \cdots, \boldsymbol{\omega}_Q^{(0)}\right]$ of the absorbing Markov chain is derived from the above probability vectors as follows:

$$\boldsymbol{\omega}_i^{(0)} = \frac{\boldsymbol{\xi}_i^{(0)}}{1 - p_{full}}, 0 \le i \le Q. \tag{3.31}$$

After d time slots the state vector of the absorbing chain is

$$\boldsymbol{\omega}^{(d)} = \boldsymbol{\omega}^{(0)} \mathbf{P}_{abs}^{<d>} \tag{3.32}$$

where $\mathbf{P}_{abs}^{<d>}$ is the transition matrix derived by multiplying $\mathbf{P}_{abs}$ with itself d times.

We can then write the *cumulative distribution function* (cdf) of the packet delay D as

$$F_D(d) = \boldsymbol{\omega}_0^{(d)} \mathbf{1} \tag{3.33}$$

Packet Loss Rate and Average Throughput

A packet loss occurs when a packet finds the queue full upon arrival or when a packet is received in error in the receiver. Assuming that no error recovery protocol (e.g., ARQ) is used, the packet loss rate, p_{loss} can be found as follows:

$$p_{loss} = 1 - (1 - p_{full})(1 - \overline{PER}). \tag{3.34}$$

Once the packet loss rate is derived as above, the average throughput η in packets per time slot can be calculated as follows:

$$\eta = \lambda(1 - p_{loss}). \tag{3.35}$$

3.4 Numerical Results and Discussions

The main objective of the analytical model developed in this chapter is to facilitate a better cross-layer system design. Two types of system parameters can be involved in the design

process. The operating environment such as fading conditions may determine the values for a number of system parameters. On the other hand, some system parameters such as number of admitted users can be tunable to obtain different levels of performance from the system. To this end, we run the model to obtain numerical results relating the obtainable QoS measures to different values of system parameters, both pre-determined and tunable, for the MIMO system in consideration. Since our primary interest lies in the cross-layer aspects of the system design, the performance measures of interest were mainly relevant to the MAC layer whereas the system parameters were mainly chosen from the physical layer. For the sake of completeness, we also consider traffic model parameters as well as the number of users in the system in our analysis.

Since we assume a homogeneous system, we tag one of the users in the system and evaluate the performance measures from its perspective. The frame duration, packet size and target packet error rate are assumed to be $T_f = 2$ ms, $L_p = 1080$ bits, and $\overline{PER} = 1\%$ respectively. We assume a four-state FSMC model for the subchannels associated with the transmit antennas. In addition, the following parameter values are assumed in general, but are changed when their individual effects are evaluated. Number of transmit and receive antennas are assumed to be $t = 2$ and $r = 2$ respectively. The average transmit SNR at each transmit antenna, transmit antenna correlation coefficient and Doppler spread are set to $\gamma_0 = 10$ dB, $\rho = 0.1$ and $f_d = 15$ Hz respectively. A mix of five homogeneous users in the system with a packet arrival probability given by $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$ and a buffer size of $Q = 50$ packets are assumed. The values of other important parameters are listed in the relevant subsections below. We used MATLAB to code and execute the different steps in the development of the queuing model. The numerical results are discussed below.

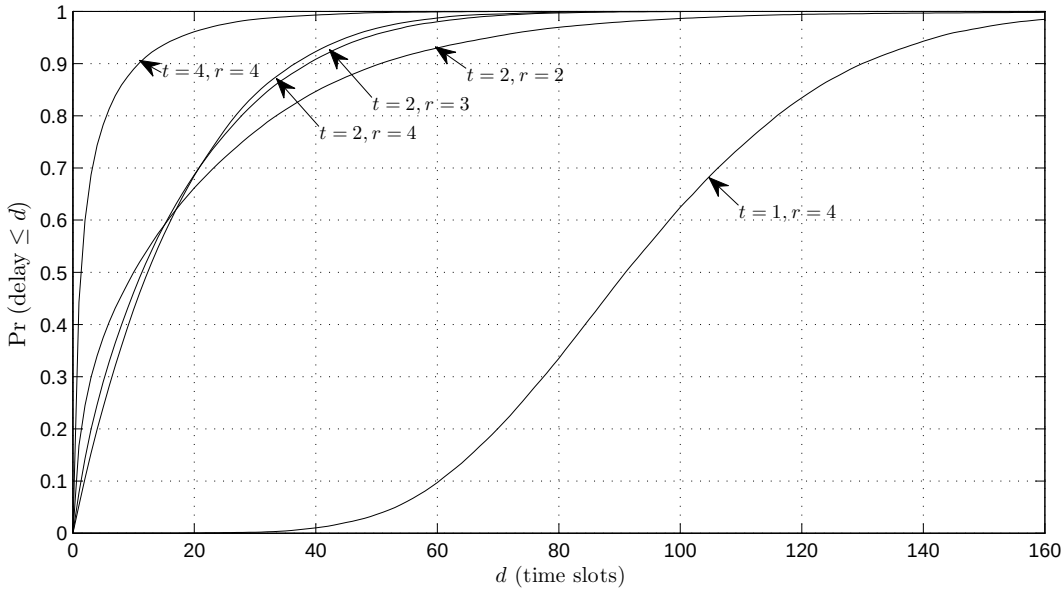


Figure 3.1: Packet delay distribution for different number of transmit and receive antennas ($f_d = 15$ Hz, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

Table 3.1: Throughput and packet loss performance for different number of transmit and receive antennas ($f_d = 15$ Hz, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

Rx Antennas	Tx Antennas	Avg. Throughput (pkts/slot)	Packet Loss (%)
4	1	0.5	43.93
4	2	0.879	0.95
4	4	0.89	0.11
3	2	0.883	1.45
2	2	0.79	5.93

3.4.1 Effect of Number of Transmit and Receive Antennas

The number of transmit antennas directly affects the multiplexing gain from a V-BLAST MIMO system. On the other hand, the number of transmit and receive antennas both affect the achievable diversity in a V-BLAST system. Both the multiplexing and receiver diversity gain are likely to effect how fast a user device can transmit or receive data. We vary the

number of transmit antennas in the base station and the number of received antennas at the user end to see the effect of the number of transmit and receive antennas on MAC layer performance measures.

Fig. 3.1 shows the effect on the delay distribution when the number of transmit and receive antennas are varied. The effect on packet loss rate and throughput is shown in Table 3.1. We observe that the delay performance improves significantly when we increase the number of transmit antennas from one to two keeping the number of receive antennas fixed at four. With only a single transmit antenna at the base station, the downlink queue of a particular user shows poor packet loss rate and throughput performance as well. This observation suggests that the probability of the transmit antenna being assigned to the tagged user is much less than sufficient to clear the backlogs created in the queue. Two transmit antennas, at the base station means the user will have access to data transfer when any of these antennas is assigned to it. This greatly improves the downlink queue's chance of getting access to transmission at a given time slot. The packet loss rate drops sharply to only 0.95% with two transmit antennas, which indicates that the access probability is almost sufficient to clear the backlog at the downlink queue. The improvement in the throughput and delay performance can be similarly explained. Although increasing the number of transmit antennas from two to four shows better performance, the improvement is less pronounced. Similar observations can be made for throughput and packet loss rate experienced by the user. This simply indicates that the QoS improvement in the MAC is not linear with the increasing number of transmit antennas.

We also vary the number of receive antennas keeping the number of transmit antennas fixed at two. We see a moderate level of performance improvement in terms of delay, packet loss rate and throughput when the number of receive antennas are increased from

two to three. The performance gain can be seen to diminish significantly when the number of receive antennas is increased from three to four. We can conclude that increase in performance by the addition of further receive antennas is markedly less for each additional antenna.

For a given system setup and for specific system parameter values, the above observations imply that going beyond a certain number of transmit or receive antennas may not be beneficial from the QoS perspective. Moreover, antennas are expensive to deploy and increasing their numbers can come with a steep price tag. The queuing model thus offers an objective way to find a tradeoff between the cost associated with the increased number antennas and the improvement achievable in the MAC layer performance for a given system setup.

3.4.2 Effect of Transmit Antenna Correlation

From (3.3), we see that the transmit antenna correlation coefficients are factored into the received SNR from the independent subchannels associated with the transmit antennas. Transmit antenna correlation reduces available multiplexing gain; in equation (3.3) the effect is reflected in the reduction of received SNR at the subchannel associated with each transmit antenna. On the other hand, the *independent stream scheduler* is designed to allocate the antennas based on the received SNR levels. Therefore, it should be interesting to see how the performance measures are affected by the transmit antenna correlation when paired with the scheduling scheme. In Figs. 3.2, 3.3, and 3.4, we consider two transmit antennas at the base station and vary their correlation to see the effect on packet delay distribution, loss rate, and throughput respectively. We can see that increasing transmit antenna correlation negatively affects all three performance measures. The performance

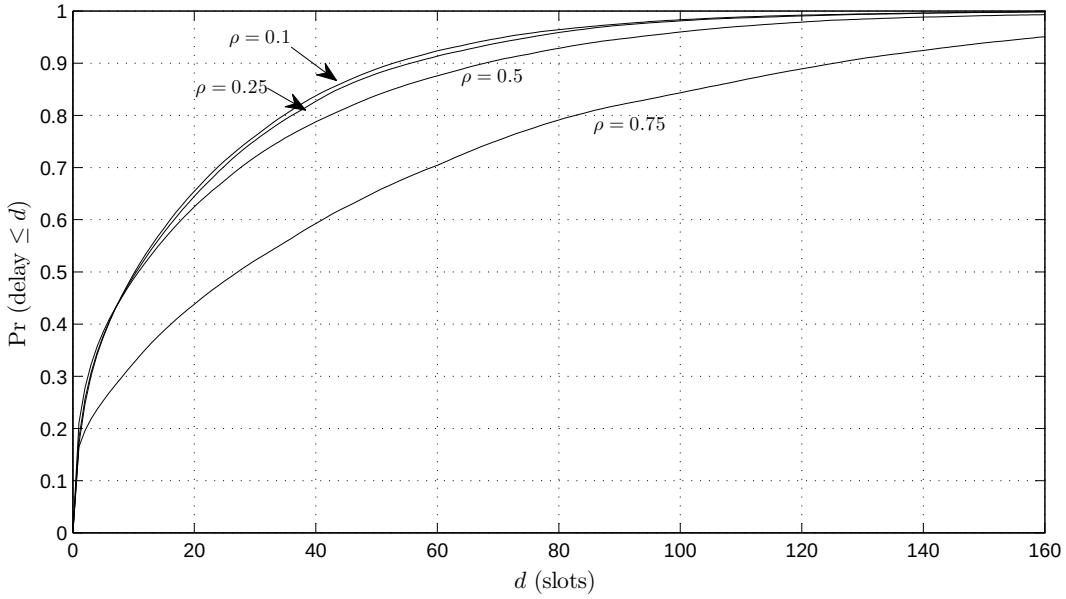


Figure 3.2: Packet delay distribution for different transmit antenna correlation ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

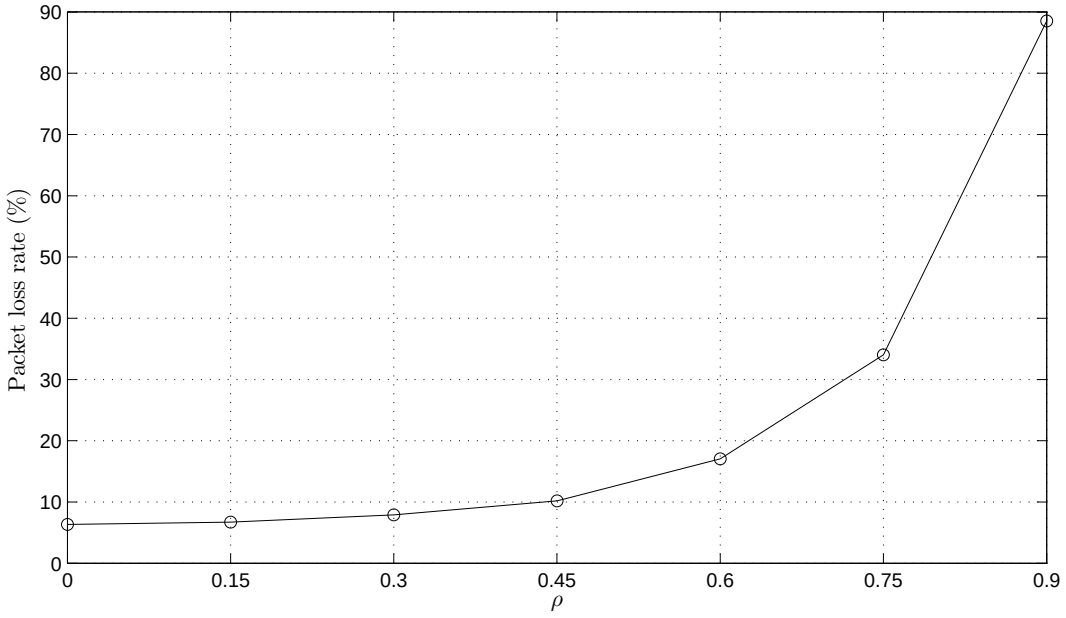


Figure 3.3: Effect of Tx antenna correlation on packet loss rate ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

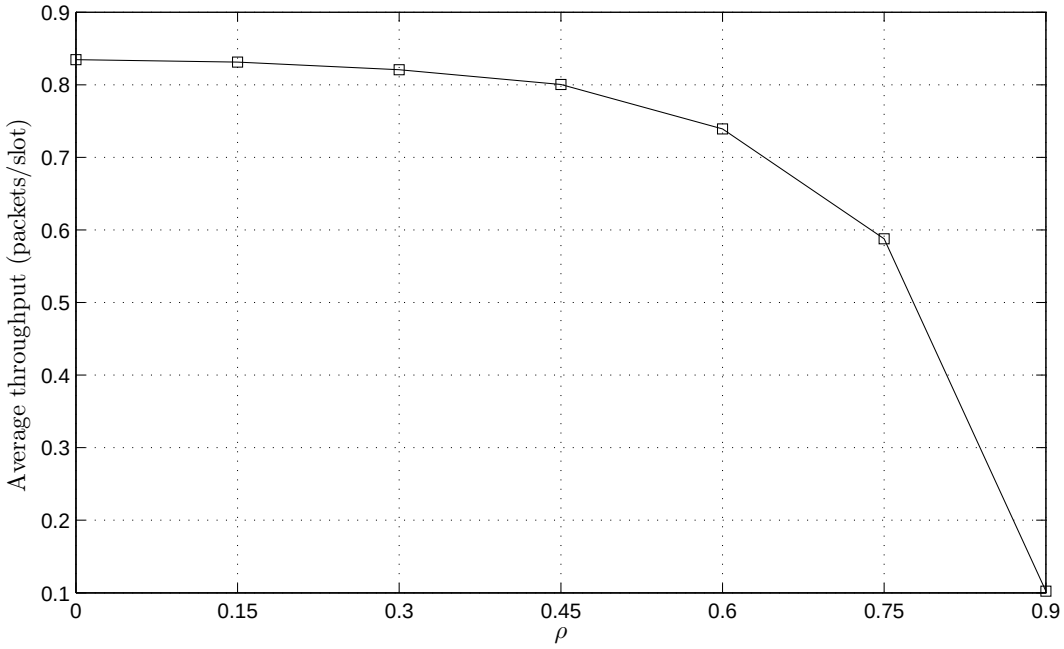


Figure 3.4: Average throughput for different levels of antenna correlation ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

degradation, however, is not linear with the antenna correlation. We see in the figures that the rate at which the performances degrade also increases with increased correlation. In fact, the effect remains rather subdued until certain level of transmit correlation is present and then picks up very fast. This is consistent with how the average received SNR at receiver changes with varying transmit antenna correlation. When the correlation co-efficient reaches near 1.0, the high correlation causes extreme loss in received SNR resulting in a very poor transmission rates at all the subchannels. Consequently, packets take long time to transmit and create large backlogs in the queues. As a result, very high delay and packets loss rates are observed along with a severe throughput degradation.

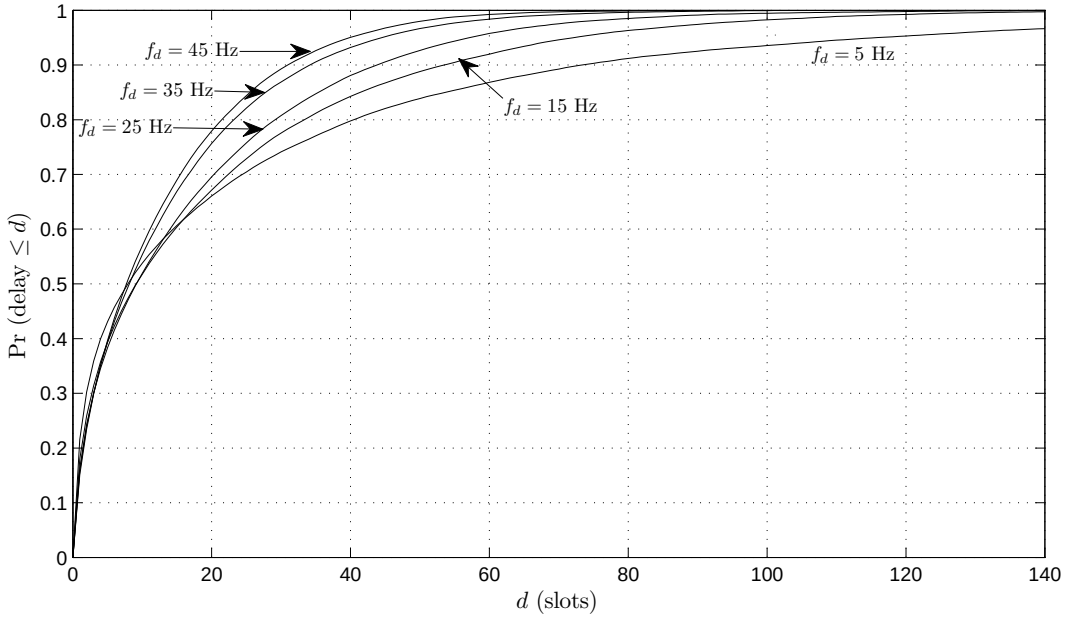


Figure 3.5: Packet delay distribution for various Doppler spread ($t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

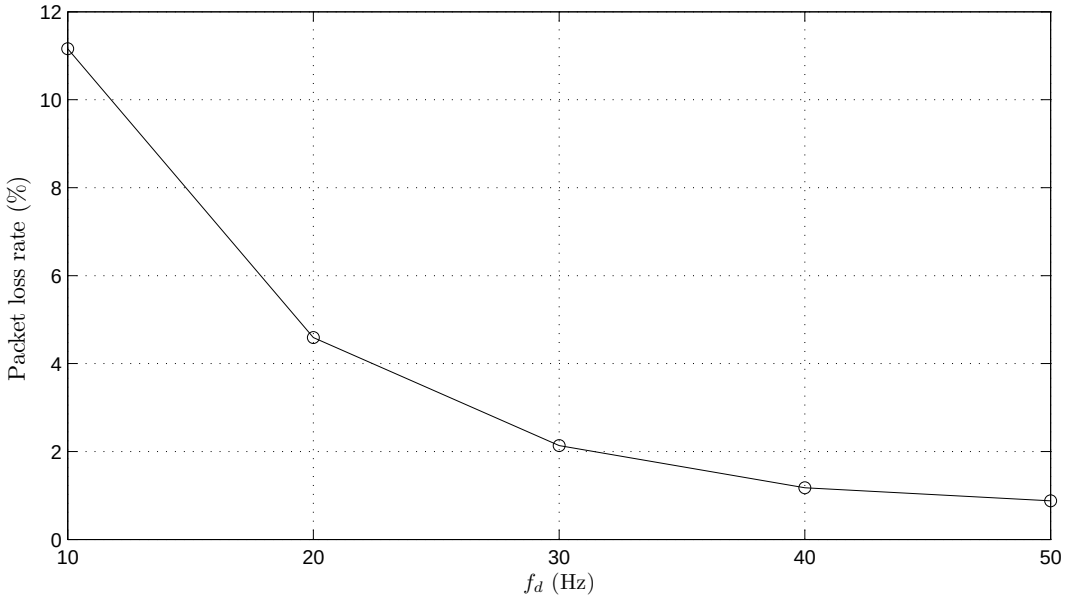


Figure 3.6: Packet loss rate for various fading rates ($t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

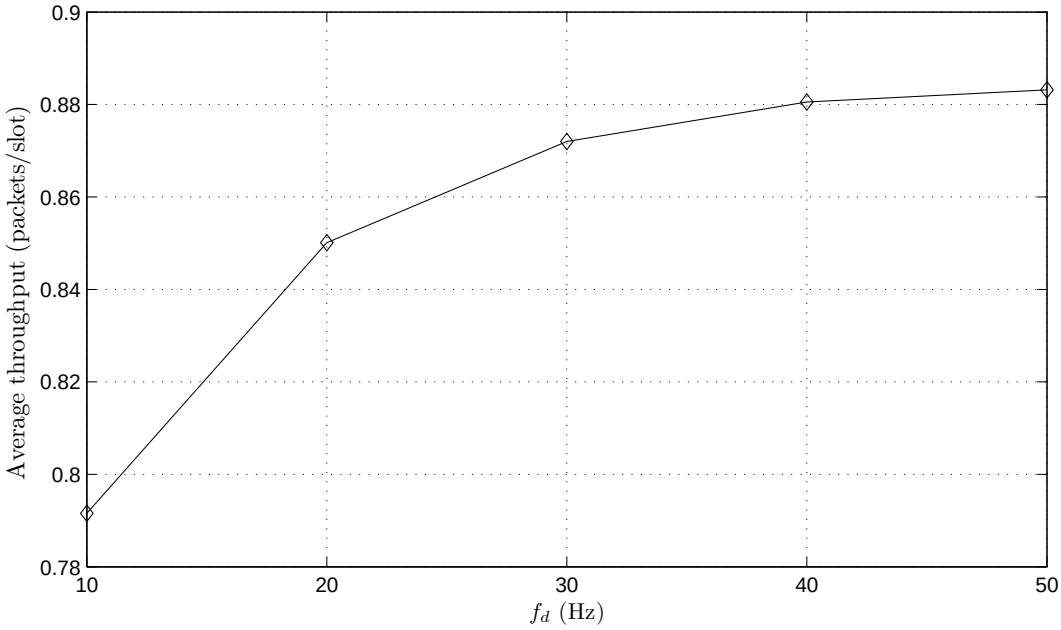


Figure 3.7: Average throughput for various fading rates ($t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users, $Q = 50$, $\boldsymbol{\alpha} = [0.4, 0.4, 0.1, 0.1]$)

3.4.3 Effect of Doppler Spread

Doppler spread has been shown to have significant performance implications in single antenna wireless systems exploiting multiuser diversity. Therefore, it is imperative to examine this relationship in the MIMO system in our consideration. The Doppler spread is primarily governed by the relative motion of the base station and the user mobile device. Fig. 3.5 shows that the delay distribution improves when the Doppler spread increases. A higher Doppler spread means less correlation over time in the subchannel associated with a particular transmit antenna. As a result, the state of the subchannel at any particular user will change more rapidly over time. This phenomenon in conjunction with the scheduling scheme will make it less likely that a particular user can monopolize the subchannel for a long time. Our tagged user will thus have more probability of access to the subchannel. The same conclusion can be drawn for other subchannels as well. Consequently, the tagged

user will have more opportunities to reduce the backlog in its queue. A reduced backlog in the queue implies less delay for the incoming packets. Moreover, with the less probability of queue size reaching the buffer limit, the packet loss rate drops and subsequently, the average throughput is also improved (see Figs. 3.6-3.7)

3.4.4 Effect of Buffer Size, Traffic Intensity, and Number of Users

The queuing model also allows us to investigate quantitatively the effects of buffer size, traffic characteristics as well as the number of users on the system performance. Three different traffic arrival patterns AP1, AP2, and AP3 were characterized by the packet arrival probability vectors as below:

$$\boldsymbol{\alpha}_{\mathrm{AP1}} = [0.7, 0.3]$$
$$\boldsymbol{\alpha}_{\mathrm{AP2}} = [0.3, 0.7]$$
$$\boldsymbol{\alpha}_{\mathrm{AP3}} = [0.3, 0.5, 0.2]$$

AP1 represents an arrival pattern that has very low packet arrival probability and is non-bursty in nature. AP2 is also non-bursty, but has higher packet arrival probability in a given time. On the other hand, although AP3 has the same packet arrival probability as that of AP2, it introduces burstiness into the traffic.

Fig. 3.8 shows the delay distribution curves for different values of buffer size and these traffic arrival profiles. For a given traffic profile, we should see increased packet delay when a larger buffer size is chosen because the backlogs in the queues can grow larger. For AP2 and AP3, we can easily observe this behavior. For AP1, however, the effect of buffer size is observed to be very insignificant. The packet arrival probability in AP1 is very low, and therefore, the backlog tends to remain short even with a large buffer space. On the

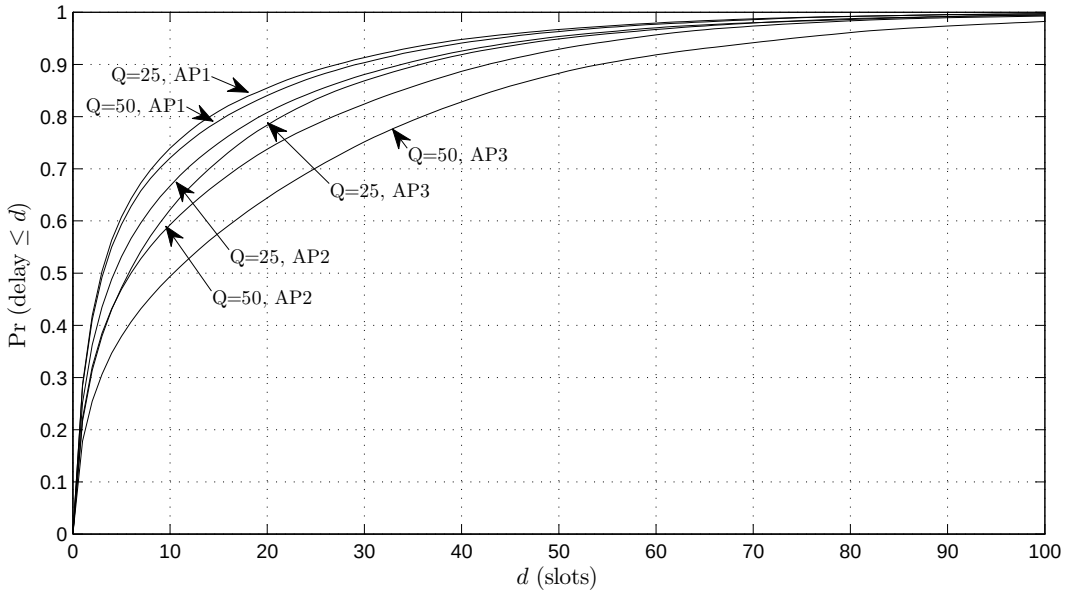


Figure 3.8: Delay distribution for different buffer size and traffic arrival profiles (AP) ($f_d = 15$ Hz, $t = 2$, $r = 2$, $\rho = 0.1$, $\gamma_0 = 10$ dB, 5 users)

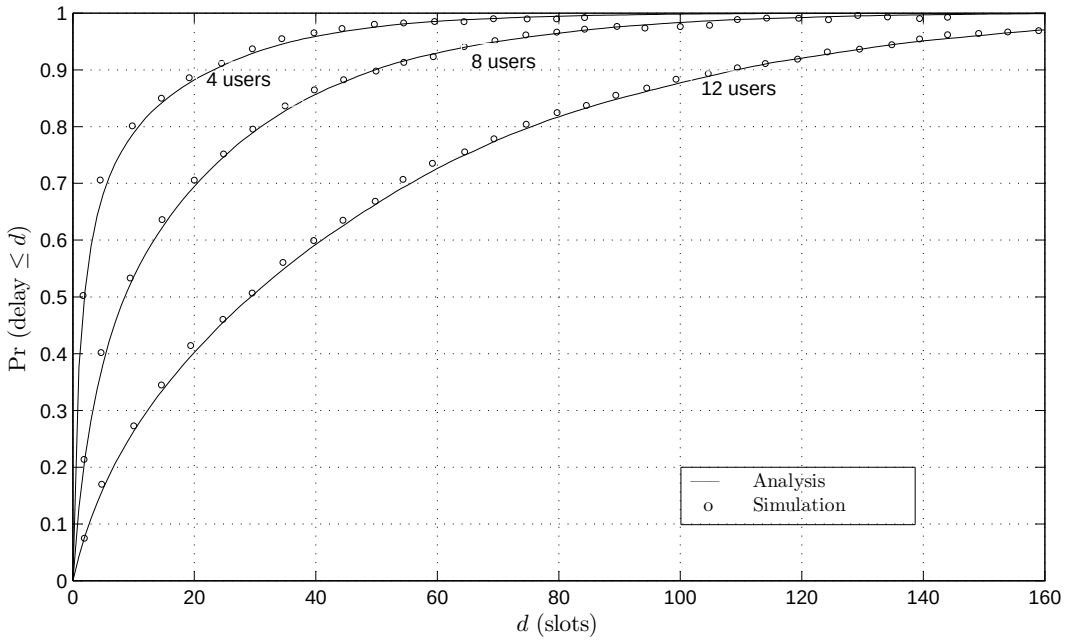


Figure 3.9: Effect of the number of users on packet delay distribution ($f_d = 15$ Hz, $Q = 50$, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, $\rho = 0.1$, Arrival profile = AP1)

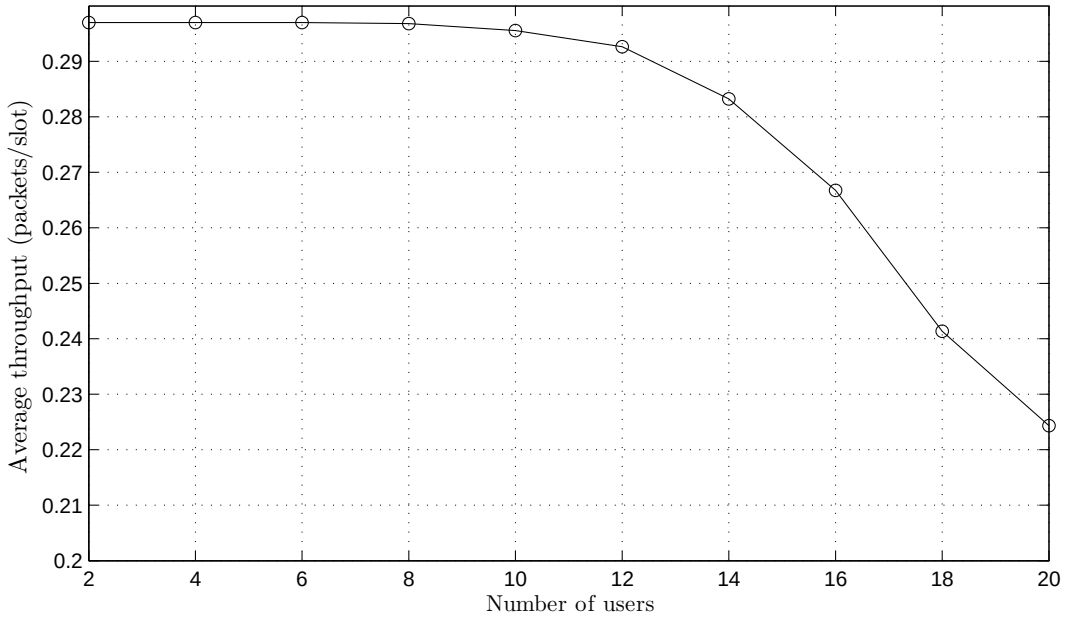


Figure 3.10: MAC layer throughput for different number of users ($f_d = 15$ Hz, $Q = 50$, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, $\rho = 0.1$, Arrival profile = AP1)

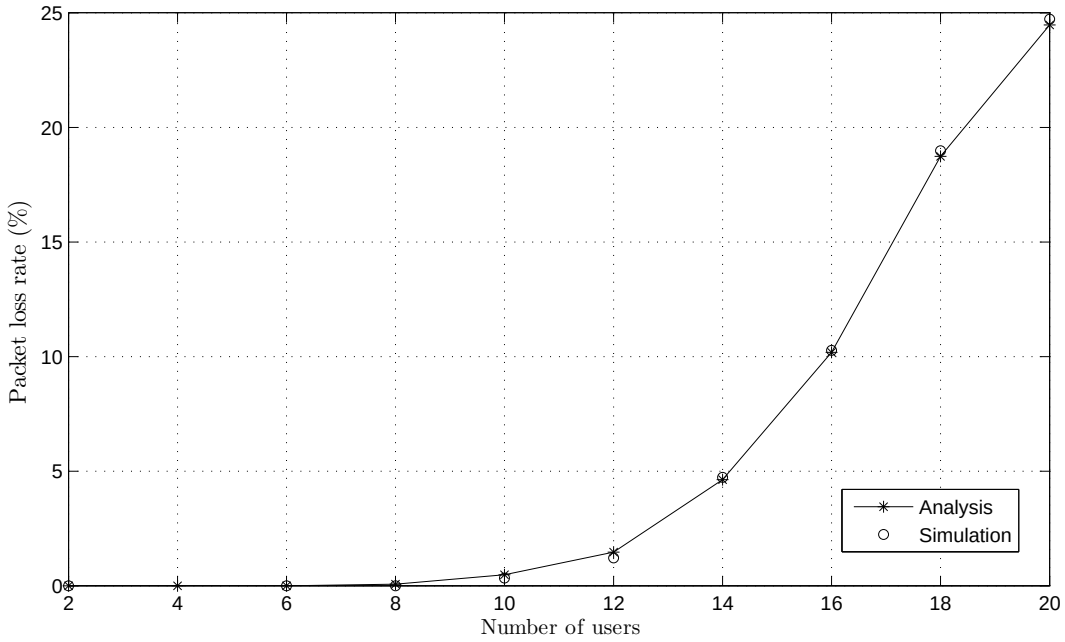


Figure 3.11: Effect of the number of competing users on packet loss rate ($f_d = 15$ Hz, $Q = 50$, $t = 2$, $r = 2$, $\gamma_0 = 10$ dB, $\rho = 0.1$, Arrival profile = AP1)

other hand, with a fixed buffer size, traffic profile AP1 shows the best delay performance among the three traffic profiles. As expected, the delay increases with higher packet arrival probabilities associated with AP2 and AP3. However, if we compare the delay distribution curves for AP2 and AP3, it also becomes apparent how burstiness in case of AP3 contributes to the increased packet delay. Bursty packet arrivals can give rise to sudden rise in queue length, which can take time to clear. Consequently, delay performance is more likely to suffer.

We also investigate the effect of the number of users in the V-BLAST system in consideration. We consider four receive antennas and four transmit antennas with no antenna correlation. Fig. 3.9 shows how the number of users affects the delay distribution. As expected, more users in the system create more competition, and therefore, longer packet delays are experienced. Similar observations can be made for the throughput and packet loss rate performance (see Figs. 3.10-3.11).

3.4.5 Validation of the Queueing Model

We also developed a discrete-event simulator to validate the correctness of the proposed queueing model. At each of the downlink queue that is simulated for each admitted user in the system, packets are generated following a Batch Bernouli distribution. The V-BLAST MIMO downlink to each of these users is simulated by t (t is the number of transmit antennas at the base station) parallel subchannels, each of which is modeled by an FSMC parameterized by a transition probability matrix. All the parameter values related to the traffic arrival process and FSMC channel model are chosen to be identical to the values used for the corresponding numerical analysis. In each time slot, the state for each subchannel is randomly generated according to the FSMC transition probability matrix. The

simulated *independent stream scheduler* then considers these subchannel states for all users to determine which transmit antenna will be assigned to which user in the current time slot. Once the user selection and antenna assignment are done, the total number of packets that will be transmitted from the downlink queue associated with each user is determined through the same AMC scheme as assumed in the queueing model. We keep the statistics on the number of packets that arrived, dropped and transmitted at each of the downlink queue over the entire simulation time. We also record the arrival and departure times of each of the transmitted packets at all the downlink queues. We simulate 100000 time slots for each simulation run. We present two representative sets of simulation results along with analytical results in Fig. 3.9 and Fig. 3.11, where the delay statistics and packet loss rate performances achieved from simulation and analytical model are compared for a tagged user (user 1). As can be seen, the simulation and numerical results match very closely, thereby validating the correctness of our proposed queueing model.

3.5 Application of the Queueing Model

The queuing model developed in this chapter can be used for system engineering through cross-layer design in a multiuser V-BLAST MIMO system. Some of the example applications are as follows:

- The model essentially draws a relationship between different system specific parameters mainly at the physical layer and the obtainable QoS at the MAC layer. It thus can give the operating points for these system parameters to achieve a specific QoS level. Resource limitations or operating environments can put constraints on the values of these parameters. For example, with the given parameter values used to find the relationship between the delay distribution and the number of transmit/receive

antennas in Fig. 3.1, two transmit and two receive antennas are sufficient to keep the packet delay within a target delay bound of 80 ms with a 90% probability. Therefore, the model gives a design guideline on the number of antennas that should be deployed under given resource constraints.

- The model allows to gauge beforehand the level of QoS the system can entertain for a new incoming request given the resource constraints and operating environments. Therefore, the model output can also be used by base stations in QoS negotiation for new connections or service flows. The QoS negotiation usually involves a counter offer on the level of service that the base station can offer when a new service flow puts forward its desired QoS level when seeking entry into the system. The base station can make use of the output from the queuing model to furnish counter offers.

- The queuing model can be utilized in a model based admission controller design as well. When a user wants to establish a new request for an application or service, the admission controller can use the queuing model to determine whether the required QoS of the new and existing applications or services can be met given the number of users with active connection already in the system. If the QoS of the requested and existing applications or services cannot be met, the connection request may be refused or can be accepted otherwise. For example, in Fig. 3.11, with the given parameter values used to find the relationship between packet loss rate and the number of users in the system, only 14 users can be served with a target packet loss rate of 5%. If the system is already serving 14 active users at some point in time, the admission controller could refuse admission of a connection from a new user on the basis of the achievable packet loss rate measure determined by the queuing model output.

3.6 Conclusions

We have developed a queuing analytic model for cross-layer analysis of downlink resource allocation in a V-BLAST MIMO system exploiting multiuser diversity. The MIMO channel in the V-BLAST system is decomposed into subchannels associated with the transmit antennas and the users are then allocated the transmit antennas opportunistically based on the received SNR on the associated subchannels. The queuing model is developed considering the system dynamics resulting from the scheduling mechanism, fading in the MIMO channel, and the traffic arrival process. The transmit antenna correlation in the base station and the possibility of bursty traffic arrival are also factored into the model. After developing the queuing model, we elaborate the steps to compute important QoS performance measures including the delay distribution, throughput, and packet loss performance at the MAC layer. We also present selected numerical results that show cross-layer effects of the physical layer parameters as well as the effects of the traffic arrival statistics and the number of users in the system on the MAC layer performance. The queuing model has important applications in the QoS provisioning framework for a V-BLAST MIMO system as the model output can be used in admission control, QoS negotiation and setting proper values for QoS relevant parameters.

3.7 Bibliography

[1] G. Foschini and M. J. Gans, "On Limits of Wireless Communication in a Fading Environment when using Multiple Antenna Arrays," *Proc. Wireless Personal Communication*, vol. 6, pp. 311-335, Mar. 1998

[2] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Vlenzuela, "V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel," *Proc. URSI International Symposium on Signals, Systems, and Electronics*, Pisa, Italy, Sept. 1998.

[3] D. N. C. Tse, "Optimal Power Allocation over Parallel Gaussian Broadcast Channels," *Proc. IEEE International Symposium on Information Theory*, Ulm, Germany, Jul. 1997.

[4] R. W. Heath, Jr., M. Airy, and A. J. Paulraj, "Multiuser Diversity for MIMO Wireless Systems with Linear Receivers," *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Oct. 2001.

[5] C.-J. Chen and L.-C. Wang, "Performance Analysis of Scheduling in Multiuser MIMO Systems with Zero-Forcing Receivers," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1435-1445, Sept. 2007.

[6] T. Yoo and A. Goldsmith, "On The Optimality Of Multiantenna Broadcast Scheduling Using Zero-Forcing Beamforming," IEEE Journals on Selected Areas in Communications, vol. 24, no. 3, pp. 528-541, Mar. 2006.

[7] M. Airy, R. W. Heath and S. Shakkottai, 'Multi-User Diversity For The Multiple Antenna Broadcast Channel With Linear Receivers: Asymptotic Analysis," Proc. Asilomar Conference on Signals, Systems, amd Computers, pp. 886-890, Nov. 2004.

[8] W. Rhee and J. M. Cioffi, "On the Capacity of Multiuser Wireless Channels with Multiple Antennas," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2580-2595, Oct. 2003.

[9] C.-J. Chen and L.-C. Wang, "A Unified Capacity Analysis for Wireless Systems with Joint Antenna and Multiuser Diversity in Nakagami Fading Channels," *Proc. IEEE International Conference on Communications*, Paris, France, Jun. 2004.

[10] L. B. Le and E. Hossain, "On the Performance of Spatial Multiplexing MIMO Cellular Systems with Adaptive Modulation and Scheduling," *Proc. IEEE Wireless Communications and Networking Conference*, Atlanta, USA, Mar. 2004.

[11] E. S. Lo, P. W. C. Chan, V. K. N. Lau, R. S. Cheng, K. B. Letaief, R. D. Murch, and W. H. Mow, "Adaptive Resource Allocation and Capacity Comparison of Downlink Multiuser MIMO-MC-CDMA and MIMO-OFDMA," *IEEE Transactions on Wireless Communications*, vol. 6, no. 3, pp. 1083-1093, Mar. 2007.

[12] K. K. Wong, R. D. Murch, and K. B. Letaief, "Performance Enhancement of Multiuser MIMO Wireless Communication Systems," *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 1960-1970, Dec. 2002.

[13] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek, "Subchannel Allocation in Multiuser Multiple-Input-Multiple-Output Systems," *IEEE Transactions on Information Theory*, vol. 52, no. 10, Oct. 2006.

[14] O.-S. Shin and K. B. Lee, "Antenna-Assisted Round Robin Scheduling for MIMO Cellular Systems," *IEEE Communications Letters*, vol. 7, no. 3, pp. 109-111, Mar. 2003.

[15] G. Aniba and S. Aissa, "Adaptive Scheduling for MIMO Wireless Networks: Cross-Layer Approach and Application to HSDPA," *IEEE Transactions on Wireless Communications*, vol. 6, no. 1, pp. 259-268, Jan. 2007.

[16] M. Airy, S. Shakkottai, and R. W. Heath, Jr., "Limiting Queuing Models for Scheduling in Multi-User MIMO Wireless Systems," *Proc. IASTED Conference on Communications, Internet & Information Technology*, Scottsdale, USA, Nov. 2003.

[17] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with Adaptive Modulation and Coding Over Wireless Link: Cross-Layer Analysis and Design," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1142- 1153, May 2005.

[18] G. J. Foschini, D. Chizhik, M. Gans, C. Papadias and R. A. Valenzuela, "Analysis and performance of some basic spacetime architectures," IEEE Journals on Selected Areas oin Communications., vol. 21, pp. 303-320, Apr. 2003.

[19] D. Gore, R. W. Heath, Jr., and A. J. Paulraj, "On Performance of the Zero Forcing Receiver in Presence of Transmit Correlation," *Proc. IEEE International Symposium on Information Theory*, Lausanne, Switzerland, Jul. 2002.

[20] H. S. Wang and N. Moayeri, "Finite-State Markov Channel-A Useful Model for Radio Communication Channels," *IEEE Transactions on Vehicular Technology*, vol. 44, pp. 163-171, Feb. 1995.

[21] M. F. Neuts, *Matrix geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, MD: John Hopkins Univ. Press, 1981.

Chapter 4

Controlled Channel Access Scheduling for Guaranteed QoS in 802.11e-Based WLANs[1]

4.1 Introduction

Wireless Local Area Network (WLAN) technology offers a simple, cost effective, and flexible solution to ubiquitous wireless access in homes, offices, campuses, hospitals etc. With the surging popularity of WLANs, multimedia applications such as voice, streaming audio/video, network gaming, teleconferencing are going to be widely deployed and supported in WLANs. These applications require certain Quality of Service (QoS) support for their smooth operation. However, the current WLAN technology standardized in the IEEE 802.11 standard lacks this QoS support. The IEEE 802.11e standard [1] aims to offer this

[1] A version of this chapter has been published. Rashid, M.M., Hossain, E. and Bhargava V.K. (2008) Controlled Channel Access Scheduling for Guaranteed QoS in 802.11e-Based WLANs. IEEE Trans. Wireless Commun., 7(4):1287–1297

QoS support to the 802.11-based WLANs. The 802.11e introduces the Hybrid Coordination Function (HCF) that offers two medium access control (MAC) mechanisms, namely, Enhanced Distributed Channel Access (EDCA) for prioritized QoS and HCF Controlled Channel Access (HCCA) for parameterized/guaranteed QoS.

Most of the multimedia applications require hard QoS requirements that could only be provided through the HCCA mechanism instead of EDCA which is able to only provide a "soft" or "relative" QoS. HCCA is a polling based mechanism in which the Hybrid Controller (HC) collocated with Access Point (AP) grants the transmission opportunities to the contending traffic streams generated from these applications. However, the HC has to properly schedule the granting of the transmission opportunities in order to provide these applications with their pledged QoS. Moreover, the HC should not grant admission to those traffic streams for which it cannot provide the required QoS. The standard provides a reference scheduler design along with the design of an admission controller unit to complement the HCCA scheme. Many multimedia applications such as real-time multimedia streaming, videoconferencing, and network gaming produce Variable Bit Rate (VBR) traffic. The packets are generated in variable intervals with fluctuating packet sizes. This type of traffic poses significant challenge to the HCCA scheduler due to their time-varying nature. Although the transmission opportunities could be granted according to the expected bit rate and required service intervals, the traffic might be generated at much higher or lower rates at times.

In this chapter, first, we develop a queueing model to analyze the performance of the reference scheduler in HCCA that captures the traffic dynamics from a VBR multimedia source. We model the system as a vacation queue with general arrival and service processes. Then *matrix-analytic method* [2] is applied to solve the queueing model and derive

the performance measures of interest including the access delay statistics. Although EDCA has been the focus of a number of analytical studies (e.g. [3, 4, 5]), there has been no analytical framework proposed in the literature so far for HCCA. The analytical model in this chapter reveals some interesting performance behaviors that point out the problems with the reference HCCA scheduler. Then, based on the insights obtained from the performance analysis, we design the prediction and optimization-based HCCA (PRO-HCCA) scheduler which can solve these performance problems and enable HCCA to meet the desired QoS guarantees.

The rest of the chapter is organized as follows. Section II provides background on the IEEE 802.11e standard and the HCCA mechanism along with the reference HCCA scheduler design. We develop and analyze a queueing model for the HCCA reference scheduler in Section III. Section IV presents the proposed PRO-HCCA scheduling algorithm and its implementation aspects in a IEEE 802.11e network. The performance evaluation results for the PRO-HCCA scheduler in a simulated IEEE 802.11e WLAN are presented in Section V. Section VI states the conclusions.

4.2 Background

4.2.1 Overview of HCCA

The HCCA mechanism, which improves over the Point Coordination Function (PCF) of legacy 802.11 MAC, is a polling based mechanism where the access to the medium is arbitrated centrally. However, it resolves the limitation of PCF by eliminating the unpredictability of beacon delays arising from incompatible cooperation between Contention Period (CP) and Contention Free Period (CFP) [6]. Note that, since PCF is unable to guar-

antee the transmission times of the polled stations, it is not a viable solution for QoS support in 802.11-based WLANs.

IEEE 802.11e introduces the concept of Traffic Stream (TS) which can be thought of as a set of data units (MSDU) that has to be delivered conforming to a corresponding Traffic Specification (TSPEC). A TSPEC characterizes the traffic streams and its QoS requirements. A TSPEC negotiation takes place between a QoS Station (QSTA) and the HC collocated with the QoS Access Point (QAP) before a TS can be served through the HCCA. The TSPEC parameters which are considered during the negotiation include nominal MSDU size in octets, mean data rate in bps, Maximum Service Interval (MaxSI) which is the maximum interval in micro seconds (μs) between two successive polls for the stream, minimum PHY rate in bps, and delay bound in μs. Once the TSPEC negotiation is successful, the TS is admitted and offered transmission opportunities (TXOPs) by the HC in each polling cycle. The length of the TXOPs offered to the stream is decided through a scheduling scheme. The periods in which the HC has the exclusive control of the channel for data transmission are called Controlled Access Periods (CAPs). The HC polls the QSTAs during these CAPs and offers TXOPs to the admitted streams.

4.2.2 Reference HCCA Scheduler Design

The scheduler first determines a Scheduled Service Interval (SI), which is the time interval used by the QAP to periodically poll each non-AP QSTA that has one or more streams admitted by the admission controller. The SI is calculated as a number which is smaller than all the MaxSIs of the admitted streams and a submultiple of the beacon interval. Any admitted stream will be able to get a TXOP at the end of each SI. The idea is to provide even the most QoS constrained stream with at least one TXOP within its MaxSI time limit.

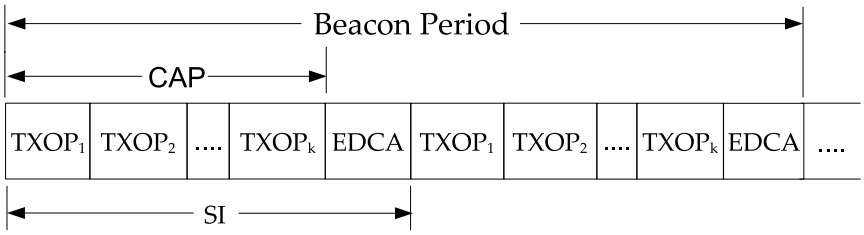


Figure 4.1: Timing diagram for the reference HCCA scheduler showing different timing periods and their relationships within a beacon interval having two SI periods.

The scheduler then determines the TXOP duration needed for each stream by considering the number of packets that may arrive within an SI. The TXOP for an admitted TS i in a QSTA is calculated as follows:

$$TXOP_i = max(\frac{G_i \times S_i}{R} + X, \frac{Z}{R} + X) \tag{4.1}$$

where G_i is the number of MSDUs arrived in the QSTA at mean data rate λ_i during the SI and can be calculated from $G_i = \lceil \frac{SI \times \lambda_i}{S_i} \rceil$, S_i is the nominal MSDU size, R is the physical transmission rate, Z is the maximum allowable size of MSDU, and X denotes the overhead (which includes inter-frame spaces, acknowledgements and polling frames) in time unit. Fig. 4.1 shows the channel timing diagram encompassing a beacon period in an HCCA enabled 802.11e WLAN with the reference HCCA scheduler in operation, where k in the figure is number of scheduled stream.

4.3 Queueing Model and Analysis for the Reference HCCA Scheduler

In this section, we develop and analyze a queueing model for the IEEE 802.11e HCCA when coupled with the reference HCCA scheduler. We analyze the QoS performance of

the admitted traffic streams which would enable us to see the effectiveness of the HCCA reference scheduler in providing guaranteed QoS in a IEEE 802.11e WLAN. We assume that each stream is queued in a buffer that can accommodate a limited number of packets.

4.3.1 Formulation of the Queueing Model

The QAP maintains separate queues for the downlink traffic streams, whereas each of the QSTAs maintains separate queues for its uplink traffic streams. Through the scheduling operation, each of these TS queues are assigned a TXOP in each SI. The HC polls the queues during these TXOPs when queued packets are transmitted to their destinations. From the perspective of any of these queues, the HC acts like a server that serves the queue for a preassigned TXOP time with the service rate proportional to the channel data rate and then goes for a vacation before serving it again. The duration of the vacation is the SI period calculated during the scheduling process. In the analysis, without loss of generality, we focus on a particular VBR traffic stream and consider its queue as the target queue.

Pleas note that we develop a discrete-time queueing model and therefore, assume that time is slotted with a pre-defined slot duration. In rest of the analysis, the time-related parameters such as TXOP, SI should thus be interpreted in terms of this slot duration.

Arrival and Service Processes

In order to model packet arrivals into the target queue, we consider the VBR nature of the traffic source. Unlike a CBR source, packets are generated from a VBR source at random intervals. However, the distribution of these packet inter-arrival times depends on the characteristics and parameters associated with the traffic source (video, audio etc.). To generalize the arrival process, we use a discrete phase-type distribution [2] to model the inter-arrival times. The phase-type distribution for inter-arrival time is represented by

$(\boldsymbol{\alpha}, \mathbf{U}, M_1)$, where M_1 is the number of transient phases of the absorbing Markov chain describing the distribution, $\boldsymbol{\alpha}$ is the initial probability vector of order M_1, and $\mathbf{U}$ is the transition matrix among the transient phases. The mean inter-arrival time is $\frac{1}{\lambda}$=$\boldsymbol{\alpha}(\mathbf{I}_{M_1} - \mathbf{U})^{-1}\mathbf{e}_{M_1}$, where λ is the rate of packet arrival, $\mathbf{I}_j$ represents an identity matrix of order j, and $\mathbf{e}_j$ is a column vector of order j containing only 1s. The column vector for the transitions to absorption phase is $\widetilde{\mathbf{U}} = \mathbf{e}_{M_1} - \mathbf{U}\mathbf{e}_{M_1}$.

To model the service process, we follow a similar approach. The variable-length packets generated from a VBR source when added with fixed length headers form variable length MAC layer frames. Once the opportunity to transmit arrives, a MAC frame is transmitted at the channel using a fixed data rate. We assume that rate adaptation algorithms such as auto rate fall back [7] are not implemented. The 802.11e protocol requires that there be a SIFS time gap between the reception of a poll or acknowledgement (ACK) and the start of the next packet transmission. Therefore, a SIFS period and an ACK duration of T_{ACK} have to be added to the packet transmission time. Let us denote the length of a packet as L_P. For a typical video transmission over UDP, the MAC layer header consists of RTP, UDP, IP, MAC protocol headers, whose lengths are denoted by L_{RTP}, L_{UDP}, L_{IP}, L_{MAC}, respectively. Moreover, a PLCP header transmission time T_{PLCP} is added by the physical layer. Therefore, the total transmission time of the packet can be written as $T_{total} = \frac{L_P+L_{layer}}{R} + T_{PLCP} + SIFS + T_{ACK}$, where $L_{layer} = L_{RTP} + L_{UDP} + L_{IP} + L_{MAC}$ and R is the channel data rate. Assuming that the header size, channel data rate, the SIFS, PLCP, and ACK durations are fixed, the transmission time is a random variable which varies with the length of the packet.

The distribution of the transmission time of a packet (i.e., service time) is modeled as a phase-type distribution. It is represented by $(\boldsymbol{\beta}, \mathbf{S}, M_2)$, where M_2 is the number of

transient phases of the absorbing Markov chain describing the distribution, $\boldsymbol{\beta}$ is the initial probability vector of order M_2, and $\mathbf{S}$ is the transition matrix among the transient phases. The column vector for the transitions to absorption phase is $\widetilde{\mathbf{S}} = \mathbf{e}_{M_2} - \mathbf{S}\mathbf{e}_{M_2}$.

Mapping into a Vacation Queue Model

From the perspective of the target queue, the system can be modeled as a time-limited Ph/Ph/1 vacation queueing system, where "Ph" stands for phase-type distribution, the time limit is the TXOP duration and the vacation period is the SI period. T Note that in the rest of our analysis, we use the term "server" as an abstraction of HC with a service rate proportional to the channel data rate and the term "service completion" of a packet as the end of transmission of the packet in the channel.

Although the server has the maximum limit of TXOP allowed to serve the queue in one round, it can leave the queue before TXOP if the queue becomes empty. In such scenarios, the HC will poll the next TS in its polling list. The duration of the vacation is the SI period and the system is of multiple vacation types. We represent the deterministic vacation period SI by a phase-type distribution $(\boldsymbol{\delta}, \mathbf{V}, SI)$, where $\boldsymbol{\delta} = [\underbrace{1, 0, 0, \cdots, 0}_{SI}]$ and $\mathbf{V} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{SI-1} \\ 0 & 0 \end{bmatrix}$. The phase-type distributions for the inter-arrival and the service processes can be fitted from their known distributions or the traces of inter-arrival times and the lengths of the incoming packets.

4.3.2 Markov Chain Analysis

We analyze the Markov chain describing the queueing model above by adopting a matrix-analytic method [8]. Let $\boldsymbol{\theta}_j(i)$ denote the ith column of $\mathbf{I}_j$ and $\boldsymbol{\theta}'_j(i)$ denote the transpose

of $\boldsymbol{\theta}_j(i)$. Let $\mathbf{e}$ (when used without a subscript) represent a column vector of appropriate order (equal to the number of columns of the matrix that it is multiplied with) consisting of only 1s, $\mathbf{H}_k$ represent $\begin{bmatrix} \mathbf{0} & \mathbf{I}_{k-1} \\ 0 & 0 \end{bmatrix}$, and $\otimes$ denote the Kronecker product.

The discrete-time Markov chain (DTMC) describing the model has state space $\Psi = \Psi_V \cup \Psi_S$, where Ψ_V is the state space of the system when the server is in vacation and Ψ_S is the state space when the server is serving the target queue. The state space Ψ_V and Ψ_S can further be described by considering the number of packets in the queue, the time elapsed in the current visit of the server, the phase of the arrival process, and the phase of the service when the server is serving or the phase of the vacation when the server is in vacation i.e. serving other queues. We can write

$$\begin{aligned} \Psi_V &= \{(x_n, a_n, v_n) | 0 \leq x_n \leq q, 1 \leq a_n \leq M_1, 1 \leq v_n \leq SI\} \\ \Psi_S &= \{(x_n, y_n, a_n, s_n) | 0 \leq x_n \leq q, 1 \leq y_n \leq TXOP, 1 \leq a_n \leq M_1, 1 \leq s_n \leq M_2\} \end{aligned},$$

where x_n is the number of packets in the queue, q is the buffer size in number of packets, y_n is the time elapsed during the current visit of the server and s_n is the phase of the service if the server is serving the queue, a_n is the phase of the arrival process, v_n is the phase of the vacation if the server is in vacation at time n.

Let $\mathbf{P}$ denote the transition matrix of the DTMC. Since more than one arrival or service completion cannot take place in a single time slot, it is straightforward to see that the transition matrix $\mathbf{P}$ will have the following structure of a quasi-birth-and-death (QBD)

process

$$\mathbf{P} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{B}_1 & & & & \\ \mathbf{B}_2 & \mathbf{A}_1 & \mathbf{A}_0 & & & \\ & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 \\ & & & & \mathbf{A}_2 & \mathbf{A}_{1q} \end{bmatrix}. \tag{4.2}$$

By considering the possible state transitions when the queue is empty, the block matrices $\mathbf{B}_0$ and $\mathbf{B}_1$ can be written as

$$\mathbf{B}_0 = \mathbf{U} \otimes (\mathbf{V} + \widetilde{\mathbf{V}}\boldsymbol{\delta}) \tag{4.3}$$

$$\mathbf{B}_1 = \begin{bmatrix} (\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes \mathbf{V} & \boldsymbol{\theta}'_{txop}(1) \otimes (\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes (\widetilde{\mathbf{V}}\boldsymbol{\delta}) \end{bmatrix} \tag{4.4}$$

Block matrix $\mathbf{B}_2$ represents transitions in which number of packets in the queue changes from 1 to 0. The server goes for vacation if the queue becomes empty regardless of the time spent in the current visit. We, therefore, can write

$$\mathbf{B}_2 = \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_{txop} \otimes \mathbf{U} \otimes \boldsymbol{\delta} \otimes \widetilde{\mathbf{S}} \end{bmatrix}. \tag{4.5}$$

On the other hand, block matrix $\mathbf{A}_2$ contains the transitions corresponding to the reduction in the number of packets in the queue by 1. If the time spent in the current visit to the queue has reached the TXOP allocated for the queue, the server leaves the queue for other

queues i.e. the server enters another vacation period. Consequently,

$$\mathbf{A}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{\theta}_{txop}(txop) \otimes \mathbf{U} \otimes \boldsymbol{\delta} \otimes \widetilde{\mathbf{S}} & \mathbf{H}_{txop} \otimes \mathbf{U} \otimes \left(\widetilde{\mathbf{S}}\boldsymbol{\beta}\right) \end{bmatrix}. \tag{4.6}$$

When no service completion or arrival occurs in a time slot, the number of packets in the queue will not change. However, the phase of service may change for the packet currently in service (if the server is currently serving the queue) or the phase of server in vacation may change (if the server is in vacation). In addition, the server will go to vacation if TXOP time has already elapsed in the current visit. Block matrix $\mathbf{A}_1$ contains these transitions and thus can be written as

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{U} \otimes \mathbf{V} & \boldsymbol{\theta}'_{txop}(1) \otimes \mathbf{U} \otimes (\widetilde{\mathbf{V}}\boldsymbol{\beta}) \\ A_1^{sv} & \mathbf{A}_1^{ss} \end{bmatrix} \tag{4.7}$$

where $\mathbf{A}_1^{sv} = \boldsymbol{\theta}_{txop}(txop) \otimes (\mathbf{U} \otimes \boldsymbol{\delta} \otimes \mathbf{e}_{M_2} + (\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes \boldsymbol{\delta} \otimes \widetilde{\mathbf{S}})$ and $\mathbf{A}_1^{ss} = \mathbf{H}_{txop} \otimes (\mathbf{U} \otimes \mathbf{S} + (\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes (\widetilde{\mathbf{S}}\boldsymbol{\beta}))$.

The boundary block matrix $\mathbf{A}_{1q}$ can be found out by replacing $\mathbf{U}$ by $\mathbf{U} + \widetilde{\mathbf{U}}\boldsymbol{\alpha}$ in $\mathbf{A}_1$. This is because when the buffer is full, a new packet arrival will not increase the buffer length (i.e. the new packet will be dropped).

Since there cannot be more than one arrival or service completion within a time slot, the number of packets in the queue can only increase if there is a new arrival and no service completion in the time slot. Considering the possible transitions, we can write $\mathbf{A}_0$ as

$$\mathbf{A}_0 = \begin{bmatrix} \left(\widetilde{\mathbf{U}}\boldsymbol{\alpha}\right) \otimes \mathbf{V} & \boldsymbol{\theta}'_{txop}(1) \otimes (\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes \left(\widetilde{\mathbf{V}}\boldsymbol{\beta}\right) \\ A_0^{sv} & \mathbf{H}_{txop} \otimes \left(\widetilde{\mathbf{U}}\alpha\right) \otimes \mathbf{S} \end{bmatrix} \tag{4.8}$$

where $\mathbf{A}_0^{sv} = \boldsymbol{\theta}_{txop}(txop) \otimes (\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes \boldsymbol{\delta} \otimes \mathbf{e}_{M_2}$.

4.3.3 Matrix-Analytic Solution for Steady State Distributions

With the system described by a QBD type chain whose block matrices are derived in the previous subsection, we can now apply matrix-analytic method [2] to find its steady state distributions. The distributions are represented by a steady state vector $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_q]$, which satisfies $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\sum_{i=0}^{q} \boldsymbol{\pi}_i \mathbf{e} = 1$. We can use the boundary and the normalization conditions to find the vectors $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_q$. For the detailed procedure to calculate $\boldsymbol{\pi}$, please see [2].

4.3.4 Performance Measures

The performance measures including queue length distribution at the end of a TXOP, packet loss rate, and distribution of access delay for the packets from the VBR traffic source can be determined based on the state vector. We assume that the steady state vector $\boldsymbol{\pi}_i, 0 \leq i \leq q$ is partitioned as $[\boldsymbol{\pi}_i^v, \boldsymbol{\pi}_{i,1}^s, \boldsymbol{\pi}_{i,2}^s, \cdots, \boldsymbol{\pi}_{i,txop}^s]$.

Queue Length Distribution

The objective of the reference scheduler for HCCA is to leave a queue empty at the end of every visit to the queue. Therefore, it is interesting to find out the distribution of the queue length at the end of a TXOP to see how successful the scheduler is to meet this objective. The marginal probability of finding l packets in the queue at the end of a TXOP can be written as $f_{L_{txop}}(l_{txop}) = \boldsymbol{\pi}_{l,txop}\mathbf{e}, \quad 0 \leq l_{txop} \leq q$, where $f_{L_{txop}}(l_{txop})$ is the pmf of queue length at the end of a TXOP.

Packet Loss Rate

When a packet arrives at the queue which has reached its limit, the newly arrived packet will be lost in two scenarios. In the first scenario, the server is in vacation. In the second scenario, the server is serving the queue but no service completion occurs with the new arrival. Therefore, packet loss rate P_{loss} can be written as follows:

$$P_{loss} = \frac{1}{\boldsymbol{\lambda}} \left[\left(\boldsymbol{\pi}_q^v ((\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes \mathbf{V}) \right) \mathbf{e} + \left(\boldsymbol{\pi}_{q,txop}^s ((\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes \boldsymbol{\delta} \otimes \mathbf{e}_{M_2}) \right) \mathbf{e} + \left(\sum_{i=1}^{txop-1} \boldsymbol{\pi}_{q,i}^s ((\widetilde{\mathbf{U}}\boldsymbol{\alpha}) \otimes \mathbf{S}) \right) \mathbf{e} \right]. \tag{4.9}$$

Access Delay Distribution

Access delay is the time between the moment a packet arrives at a station and the moment the packet gets access into the medium for transmission. The access delay distribution for a packet from the target stream is derived by considering an absorbing Markov chain. The chain initializes from the state when the packet arrives at the corresponding queue and gets absorbed when the packet reaches the head of the queue. Note that, in this case, there is no arrival in that particular queue while the process moves towards the absorbing state. Therefore, the transition matrix $\mathbf{P}_{abs}$ of the chain can be constructed by setting $\mathbf{U} = 1$ and $\widetilde{\mathbf{U}}\boldsymbol{\alpha} = 0$. The next step for finding the delay distribution is to determine the initial probability vector for this absorbing Markov chain. This vector $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_q]$ can be found by considering the steady state vector of the system as seen by an arriving packet and then summing up the probabilities due to the phases of any new arrival. We can partition vector $\boldsymbol{\gamma}_i, 0 \leq i \leq q$ as $\boldsymbol{\gamma}_i = [\boldsymbol{\gamma}_i^v, \boldsymbol{\gamma}_{i,1}^s, \boldsymbol{\gamma}_{i,2}^s, \cdots, \boldsymbol{\gamma}_{i,TXOP}^s]$, which can be computed as follows:

$$\boldsymbol{\gamma}_0^v = \tfrac{1}{\lambda}\left[\boldsymbol{\pi}_0^v((\widetilde{\mathbf{U}}\boldsymbol{\alpha})\otimes(\mathbf{V}+\widetilde{\mathbf{V}}\boldsymbol{\delta}))+\sum_{y=1}^{TXOP}\boldsymbol{\pi}_{1,y}^s((\widetilde{\mathbf{U}}\boldsymbol{\alpha})\otimes(\widetilde{\mathbf{S}}\boldsymbol{\beta}))\right](\mathbf{e}_{M_1}\otimes\mathbf{I}_{SI}) \tag{4.10}$$

$$\boldsymbol{\gamma}_i^v = \tfrac{1}{\lambda}\left[\boldsymbol{\pi}_i^v((\widetilde{\mathbf{U}}\boldsymbol{\alpha})\otimes\mathbf{V})+\boldsymbol{\pi}_{i,TXOP}^s((\widetilde{\mathbf{U}}\boldsymbol{\alpha})\otimes\boldsymbol{\delta}\otimes\mathbf{e}_{M_2})+\boldsymbol{\pi}_{i+1,TXOP}^s((\tilde{\mathbf{U}}\boldsymbol{\alpha})\otimes\boldsymbol{\delta}\otimes\mathbf{S})\right](\mathbf{e}_{M_1}\otimes\mathbf{I}_{SI}) \tag{4.11}$$

$$\boldsymbol{\gamma}_{i,1}^s = \tfrac{1}{\lambda}\left[\boldsymbol{\pi}_i^v((\widetilde{\mathbf{U}}\boldsymbol{\alpha})\otimes(\widetilde{\mathbf{V}}\boldsymbol{\beta}))\right](\mathbf{e}_{M_1}\otimes\mathbf{I}_{M_2}) \tag{4.12}$$

$$\boldsymbol{\gamma}_{i,y}^s = \tfrac{1}{\lambda}\left[\boldsymbol{\pi}_{i,y}^s((\widetilde{\mathbf{U}}\alpha)\otimes\mathbf{S})+\boldsymbol{\pi}_{i+1,y-1}^s((\widetilde{\mathbf{U}}\boldsymbol{\alpha})\otimes(\widetilde{\mathbf{S}}\boldsymbol{\beta}))\right](\mathbf{e}_{M_2}\otimes\mathbf{I}_{SI}). \tag{4.13}$$

After d time slots the state vector of the absorbing chain is $\boldsymbol{\gamma}^{(d)} = \boldsymbol{\gamma}\mathbf{P}_{abs}^d$. Assuming that $\boldsymbol{\gamma}^{(d)}$ can be partitioned as $\gamma^{(d)} = [\boldsymbol{\gamma}_0^{(d)}, \boldsymbol{\gamma}_1^{(d)}, \boldsymbol{\gamma}_2^{(d)}, \cdots]$, we can write the Cumulative Distribution Function (CDF) of the access delay D as

$$F_D(d) = \Pr\{D \le d\} = \boldsymbol{\gamma}_0^{(d)}\mathbf{e}. \tag{4.14}$$

Figs. 4.2 and 4.3 show typical numerical results obtained from the analytical model using real VBR traffic traces. We consider a video traffic stream admitted in a QSTA of a 802.11e WLAN. The real traffic trace was collected using VIC videoconferencing tool [9] and H.261 coding in QCIF format. The mean sending rate, average packet size, and mean

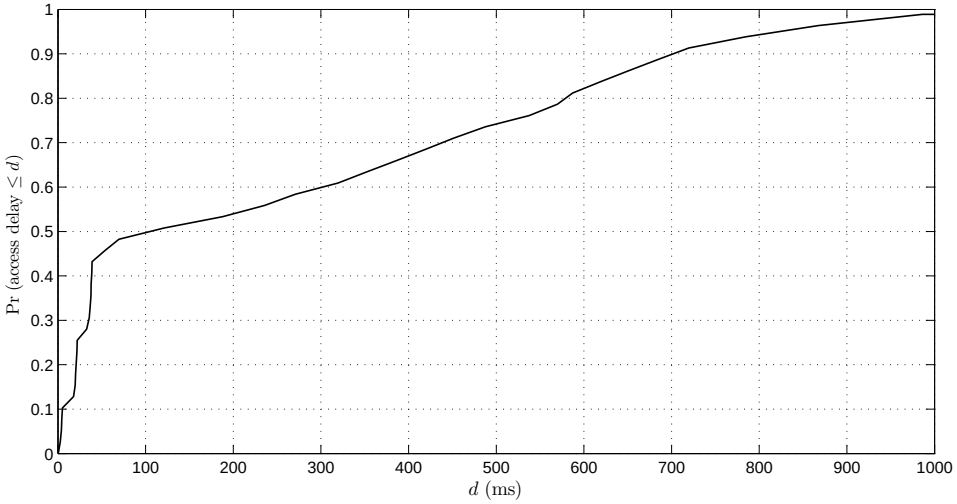


Figure 4.2: Cumulative distribution function of access delay (ms) derived from the analytical model.

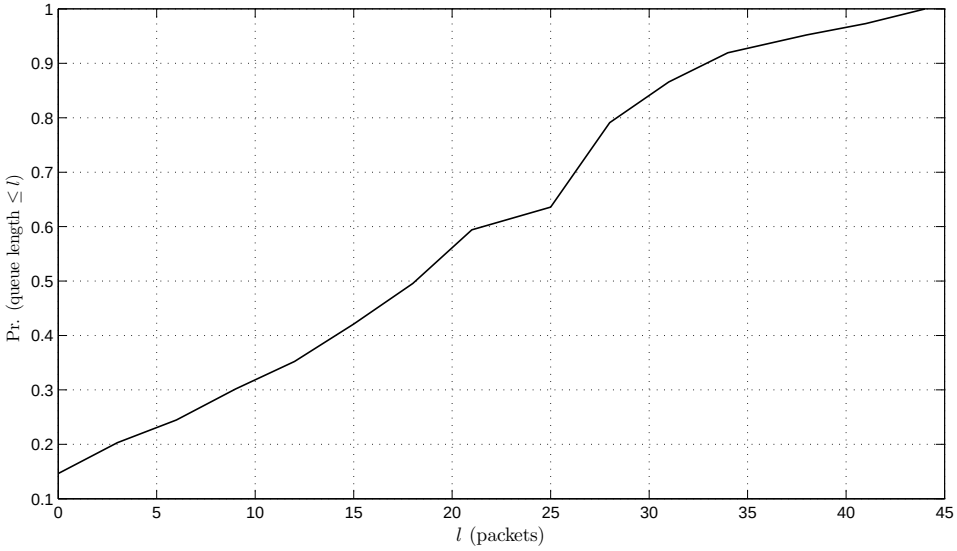


Figure 4.3: Cumulative distribution function (derived from the analytical model) of queue length at the end of a TXOP.

packet inter-arrival time computed from the trace are 200 kbps, 687 bytes, and 26 ms, respectively. The maximum service interval, delay bound, and mean data rate in the QoS specification are set to be 50 ms, 100 ms, and 20 kbps, respectively. The physical layer conforms to the 802.11a standard. We selected 36 Mbps as the PHY data rate (802.11a offers data rates: 6, 9, 12, 18, 24, 36, 48, or 54 Mbps). For the analysis, we assume that the SI period after considering the maximum service intervals needed by the admitted traffic streams is 50 ms. The TXOP duration is then calculated by using (4.1). We collect the inter-arrival and the packet size information from the VBR traffic trace. The individual packet sizes are then converted to corresponding transmission times (i.e. service times) considering the different header sizes and the physical layer transmission rate. The inter-arrival and service times are then fitted to discrete phase-type distributions using the PhFit [10] program. We then run our analysis using Matlab.

The CDF of access delay for the VBR video traffic is plotted in Fig. 4.2. It shows that the access delay fails to remain under the delay requirement (100 ms) for 50% of the packets. Moreover, there remains a significant chance of some packets' suffering excessive delays (over 500 ms).

The observations suggest that a large number of packets remain in the queue for a duration much longer than the MaxSI time interval. In an ideal scenario, the queue of the VBR stream should become empty at the end of TXOPs, because the scheduler is designed to allocate enough TXOP to transmit all the packets within the scheduled SI. However, Fig. 4.3 shows for the VBR stream the probability that the queue is empty at the end of the TXOPs is rather very low (about 15% only). The packets remaining in the queue after a TXOP must have to wait another one or more scheduled SIs. This accumulated backlog might further add to the delay of the packets arriving during the next SIs in addition to the

delays caused by the insufficient TXOPs. The backlog eases up when packets arrive at a rate lower than the expected rate for some period of time. However, a significant number of packets eventually end up experiencing high access delays due to the queue buildup.

Now let us assume that a similar stream is generated from the same traffic trace and is simultaneously admitted through another QSTA. The stream, however, demands a less stricter access delay bound of 300 ms and a maximum service interval of 150 ms. Due to the presence of lower MaxSI values of other admitted streams the SI period remains 50 ms. Obviously, the plot of the CDF of access delay will be the same as in Fig. 4.2. However, it also means that percentage of packets failing to meet the access delay bound has reduced from 50% to 40%; the former is the percentage of packets failing to meet the deadline for the same stream with a stricter delay bound of 100 ms. This observation can be explained if we consider that the reference scheduler treats the admitted streams similarly by serving them with similar urgencies in equal intervals. Consequently, streams with stricter delay bounds are penalized more when backlogs buildup at the end of TXOPs and higher percentage of their packets fail to meet the delay requirement.

4.4 The Prediction and Optimization-Based HCCA (PRO-HCCA) Scheduler

The PRO-HCCA scheduler avoids the reference scheduler's problem of unduely penalizing the more delay constrained packets. This is due to the fact that the reference scheduler treats the streams having different delay requirements with equal priority. Therefore, in the PRO-HCCA scheduler we introduce an accounting mechanism to treat packets from different streams according to their urgencies for service. Another idea here is to minimize the backlog that contributes to unwanted delay at the end of TXOPs. This is achieved in

PRO-HCCA scheduler by using a proper prediction mechanism to account for the dynamic intensity of VBR traffic and then finding an optimal allocation of available transmission time among the competing traffic streams. The optimal allocation strives to strike a balance between the urgencies of services for different traffic streams and the available medium time. To further reduce unnecessary backlogs in the queues, we take into consideration the fact that after certain queuing delay a packet would become useless and should be dropped.

4.4.1 Accounting for Different Delay Bounds

In the design of PRO-HCCA scheduler, we keep the computation of SI similar to that in the reference scheduler. This interval is fixed for a given traffic mix until a new stream is admitted or one or more of the existing streams finish transmission. Having a dynamic service interval is likely to result in significant complexity in the scheduling process without offering benefits when compared to an optimization-based fixed service interval scheduling.

For each admitted TS i, the QAP will maintain a partition list PL_i with entry j containing the amount of traffic backlogged for time period between $(j-1)*SI$ and $j*SI$, where $0 < j \leq \lceil \frac{D_i}{SI} \rceil$ and D_i is the delay bound of the stream. We refer to this backlogged traffic recorded in the jth entry of the list as the jth partition of the corresponding TS queue and denote it by PL_{ij}. The index j literally represents the degree of urgency for service. This list will be updated at each scheduling instant. A pseudo code for the updating procedure is provided in Algorithm 1, where N is the number of admitted streams, QL_i is the queue length (in bits) of TS i, X is overhead in time units, and 0 in the superscript stands for the previous value of the variable.

Let us consider the traffic amount in the $(j-1)$th partition at the previous scheduling instant. When the current scheduling instant arrives, this traffic amount has been either

Algorithm 1: Updating the partition list PL

1. for $i := 1$ to N
2. for $j := \lceil \frac{D_i}{SI} \rceil$ downto 3
3. $PL_{ij} := PL^0_{i(j-1)} - (t^0_{i(j-1)} - X) * R$
4. endfor
5. $PL_{i2} := AR^0_i * SI - (t^0_{i1} - X) * R$
6. if i is a downlink stream
7. $PL_{i1} := QL_i - \sum_{j=2}^{\lceil \frac{D_i}{SI} \rceil} PL_{ij}$
8. else $PL_{i1} := PR_i * SI$
9. endif
10. endfor

fully or partially transmitted. The transmitted traffic amount can be determined from the transmission time $t_{i(j-1)}$ that was computed for the partition in the optimization process (described in Subsection 4.4.3) during the previous scheduling instant. The rest of the traffic in the partition has been backlogged for another SI period. Therefore, this remaining amount of traffic should be set as the value of PL_{ij} in the current scheduling instant. This update method is applied to all the partitions except partition 1 and 2. Therefore, we choose 3 as the end value of the loop in line 2 of the pseudo code listed in Algorithm 1. The updating method for PL_{i1} is different because this is the amount of traffic that has arrived only after the last scheduling instant. On the other hand, since PL_{i1} may contain the predicted amount of traffic instead of the actual amount, it cannot be used to calculate the new value of PL_{i2}.

Regarding the calculation of PL_{i1}, the uplink streams have to be considered differently than the downlink streams. The traffic queue of an uplink stream is located at the QSTA

where the stream is attached. Since the queue is not at the QAP, the scheduler does not know the amount of traffic that has arrived to that traffic queue during the last SI period. The scheduler, therefore, has to predict this amount for an uplink TS i by multiplying a predicted traffic intensity PR_i with the value of SI. On the other hand, the instantaneous queue lengths of downlink streams are known to the scheduler since the downlink queues are collocated with the scheduler at the QAP. The value of PL_{i1} of a downlink stream is, therefore, easily obtainable by considering its queue length and the rest of the partitions of its queue.

To update PL_{i2}, the actual amount of TS i traffic that arrived during the SI period preceding the previous scheduling instant has to be considered. This amount can be calculated from traffic arrival intensity AR_i of TS i during that SI period. For downlink streams, the calculation is straightforward. In subsection 4.4.2 we will discuss how AR_i can be calculated for uplink streams. To get the updated value of PL_{i2}, the amount of traffic that has been transmitted by the transmission time allocated for partition 1 of TS i during the last SI period is deducted from the amount of traffic that arrived during the SI period preceding the previous scheduling instant.

4.4.2 Predicting VBR Traffic Intensities

From the previous subsection, we can see that the first entry in the PL list of an uplink stream cannot be determined like the other entries in the list. To determine this entry the scheduler has to know the amount of traffic that has arrived within the SI period since the start of the last service to the corresponding TS. The downlink traffic volumes can be determined from the downlink queues, which are collocated with the QAP. However, since the arrivals occur at the QSTAs, the actual information on the newly arrived uplink traffic

volume is not available to the QAP when it makes the scheduling decision. Due to the potentially VBR nature of the traffic, the QAP needs to make predictions. The predicted traffic intensities will be used to estimate the amount of newly arrived traffic for each TS.

The prediction mechanism should meet some requirements. Due to the limited memory and computational power available in the WLAN routers, the scheme has to be simple. On the other hand, it should be able to deal with the short term and long term correlation between traffic intensities over time. It is well known that VBR video and audio traffic show this property. The least mean square (LMS) predictor is a simple prediction mechanism and has been shown to predict VBR traffic reasonably well [11].

However, the performance of the LMS filter largely depends on the eigenvalue spread of the input data. A large spread in the eigenvalue badly affects the performance of the LMS predictor. As a result, the LMS predictor can perform poorly if the time series shows strong correlation property. This limitation can be largely overcome by using transform domain LMS predictors [12]. Wavelet domain LMS prediction could be an ideal choice due to its simplicity and its ability to smooth out the effect of large correlation among the input data.

Wavelet transform is a powerful signal processing technique. The idea behind wavelet transform-based predictor is to decompose the original data series into a different series that will offer better stability and convergence. As shown in [12], the eigenvalue spread of the input data is reduced substantially when they are decomposed using wavelet transform. The discrete transform gives a wavelet co-efficient vector of the same length as that of the input vector. The wavelet transform-based LMS predictor can be described by the following set

of equations [12]:

$$\begin{aligned}
\mathbf{z}(n) &= DWT(x(n-1), x(n-2), \cdots, x(n-K)) \\
\hat{x}(n) &= \mathbf{w}^w(n) \cdot \mathbf{z}(\mathbf{n}) \\
e(n) &= x(n) - \hat{x}(n) \\
\mathbf{w}^w(n+1) &= \mathbf{w}^w(n) + 2\mu e(n) \Gamma_z^{-1}(n) \mathbf{z}(n)
\end{aligned} \tag{4.15}$$

where DWT stands for Discrete Wavelet Transform, $\mathbf{w}^w(n)$ is the filter co-efficient vector, $\hat{x}(n)$ is the predicted and $x(n)$ is the real signal value at the nth time step, K is the order of the prediction, Γ_z is the input correlation matrix in the DWT domain. The condition of convergence is $0 < \mu < 1$. Considering a self-orthogonalizing WLMS, the Γ_z matrix can be approximated following the methods proposed in [13] as $\hat{\Gamma}_z(n) = diag(\hat{\zeta}_{n,1}, \hat{\zeta}_{n,2}, \cdots, \hat{\zeta}_{n,K})$, in which $\hat{\zeta}_{n,m} = \varphi \hat{\zeta}_{n-1,m} + (1-\varphi) z_m^2(n), \quad 1 \leq m \leq K$ where φ is an adjustable parameter ($0 \leq \varphi \leq 1$) that can be adjusted for better convergence.

To predict the traffic intensity over the next SI period, the PRO-HCCA scheduler entity in the QAP keeps record of the K observations in the recent past over the last K SIs for each admitted stream i. Here, K is the order of the WLMS predictor. By substituting these K past observations into the variables $x(n-1)$ to $x(n-K)$ of the WLMS predictor equations, the predicted traffic intensity PR_i can be calculated as $\hat{x}(n)$.

For the adaptation step of the predictor, the actual intensity of traffic that have arrived during the last SI has to be computed. The IEEE 802.11e specification allows that after a QSTA is done with its transmission opportunity in a scheduling interval, the QSTA informs the QAP about the queue length information of the TSs attached to itself. This information usually is piggybacked with the last packet transmitted within that transmission opportunity. Let the queue length of a TS i attached to a particular QSTA at this time instant be

denoted as QL_{txop_i} (in bits). If $QL^0_{txop_i}$ denotes this value at the end of TXOP during the previous SI then the QAP can calculate the actual intensity AR_i of traffic arrival for TS i as follows:

$$AR_i = \frac{QL_{txop_i} - QL^0_{txop_i} + (TXOP_i - X) * R}{SI}. \tag{4.16}$$

This value of AR_i is then used as $x(n)$ in the above mentioned WLMS equations.

4.4.3 Optimizing TXOP Allocation

At each scheduling instant that will occur after each SI interval, the QAP has to distribute (in the form of TXOPs) a constrained amount of medium time T_{avail} $(0 \leq T_{avail} \leq SI)$ among the admitted traffic streams. We formulate this allocation as an optimization problem. Thanks to the accounting and prediction mechanism described in Subsections 4.4.1 and 4.4.2, respectively, the scheduler entity inside the QAP has a picture on the urgencies of service for different portions of the TS queues. We propose to use a utility function to take into account these urgencies and to formulate an optimization model accordingly. The utility received from transmitting data belonging to partition PL_{ij} for a single time unit (in slot time) is $U_{ij} = \frac{1}{\lceil \frac{D_i}{SI} \rceil - j + 1}$.

The constraint on the available medium time for HCCA within each scheduling interval is described by two parameters. The first parameter ρ_{hcca} is the fraction of medium time allocated for HCCA. If T and T_{CP} denote the beacon interval and the time used by EDCA traffic within a beacon interval, respectively, then $\rho_{hcca} = \frac{T - T_{CP}}{T}$. The second parameter is the 802.11e *dot11CAPlimit*, which is set during the WLAN setup. Therefore, $T_{avail} = max(\rho_{hcca} * SI, dot11CAPlimit)$.

The output from the optimization procedure will be the transmission times allocated for different partitions of the TS queues. For TS queue i, let us denote by t_{ij} the transmission

time allocated for the jth partition. The optimization model can be described mathematically as follows:

$$\begin{aligned}
\underset{t_{ij}}{\textbf{maximize}}\ \Omega \quad &= \quad \sum_{i=1}^{N} \sum_{j=1}^{\left\lceil \frac{D_i}{SI} \right\rceil} U_{ij}.t_{ij} \\
\text{s.t.} \quad &: \quad \sum_{i=1}^{N} \sum_{j=1}^{\left\lceil \frac{D_i}{SI} \right\rceil} t_{ij} \leq T_{avail} \\
& \qquad 0 \leq t_{ij} \leq \left(\frac{PL_{ij}}{R} + X \right), \\
& \qquad i = 1, 2, \cdots, N; j = 1, 2, \cdots, \left\lceil \tfrac{D_i}{SI} \right\rceil
\end{aligned} \tag{4.17}$$

A closer inspection into the formulation of the optimization reveals that it can be mapped into the classical fractional Knapsack problem [14] which can be solved efficiently with a greedy algorithm in an online fashion. From different TSs, we can combine those transmission time units which have the same utility. Since we only have discrete values for utility, there will be a discrete set of transmission time units. Let us denote the set as $G = [g_1, g_2, \cdots, g_p]$, where p is the number of distinct values of per transmission time unit utilities at the scheduling instant. The scheduler at the QAP maintains the information about which TSs are contributing how much to the value of each element in the set. For each element, an entire amount or a fraction will be selected by the solution [14] of the fractional Knapsack problem.

If g_h $(1 \leq h \leq p)$ is selected as a whole, then from each TS the entire amount contributing to g_h is selected for service. When a fraction is selected, the amounts for different TSs are selected in proportion to their contributions to g_h. The outcome of the procedure will

be the t_{ij} values which are the variables in the optimization problem. Note that, for a given TS, the transmission times will be allocated such that partition $j+1$ will be able to transmit completely before partition j gets the opportunity to transmit. This is simply due to the fact that transmitting a data unit from partition $j+1$ will yield higher utility than transmitting a data unit from partition j and that the optimization is aimed at maximizing the total utility. With the t_{ij} values determined as above, the TXOPs for the TSs are calculated as follows:

$$TXOP_i = \sum_{j}^{\left\lceil \frac{D_i}{SI} \right\rceil} t_{ij}, i = 1, 2, \cdots, N \tag{4.18}$$

Once the TXOP values for the admitted streams have been calculated as above during the scheduling instant, the QAP broadcasts these allocations to the QSTAs.

4.5 Performance Evaluation of the PRO-HCCA Scheduler

4.5.1 Simulation Setup and Parameters

We simulate IEEE 802.11e HCCA using a discrete-event simulator written in Parsec [15]. We code the reference scheduler and the proposed PRO-HCCA scheduler in the simulator. Simulations can be run with either of the scheduler enabled. In the WLMS predictor, we use Daubechie-4 as the mother wavelet to perform wavelet transforms. The physical layer of the WLAN conforms to 802.11a specification with a physical rate of 36 Mbps. The PHY and MAC parameters are listed in Table 4.1.

The WLAN topology consists of 3 QSTAs each generating a traffic stream. The mix of the streams consists of VBR VoIP and video streams. Two types of video streams are

Table 4.1: Packet overhead and MAC parameters

Parameter	Value
RTP header size	12 bytes
UDP header size	8 bytes
IP header size	20 bytes
MAC header size	36 bytes
PLCP header size	3 bytes
SIFS interval	16 μsec
ACK size	14 bytes
PHY data rate	36 Mbps
Minimum data rate	6 Mbps
Slot time	9 μsec
Beacon period	500 ms
IFQ Length	50

Table 4.2: Properties and QoS requirements of the scheduled streams

TS No.	Type	TSPEC Parameters				Attached QSTA
		Nominal MSDU Size	Mean Data Rate	MaxSI	Delay Bound	
1	Interactive video	1000 bytes	210 kbps	50 ms	100 ms	1
2	Non-Interactive video	920 bytes	180 kbps	100 ms	200 ms	2
3	VoIP	160 bytes	22.4 kbps	25 ms	50 ms	3

used: interactive and non-interactive. QSTA 1 generates the interactive video whose traffic trace is collected from the lecture room cam trace (high quality MPEG-4) available at [16]. A non-interactive or streaming video stream is generated from real traffic trace and is attached to QSTA 2. The stream is generated from the medium quality MPEG-4 trace of the movie *Mr. Bean*. This trace is also collected from [16]. A VBR VoIP stream with silence suppression is simulated and attached to QSTA 3. The stream has exponentially distributed ON and OFF state times with the average state duration of 352 ms and 650 ms, respectively. Table 4.2 summarizes the properties and QoS requirements of the streams used in the simulation. The amount of time simulated is 300 seconds.

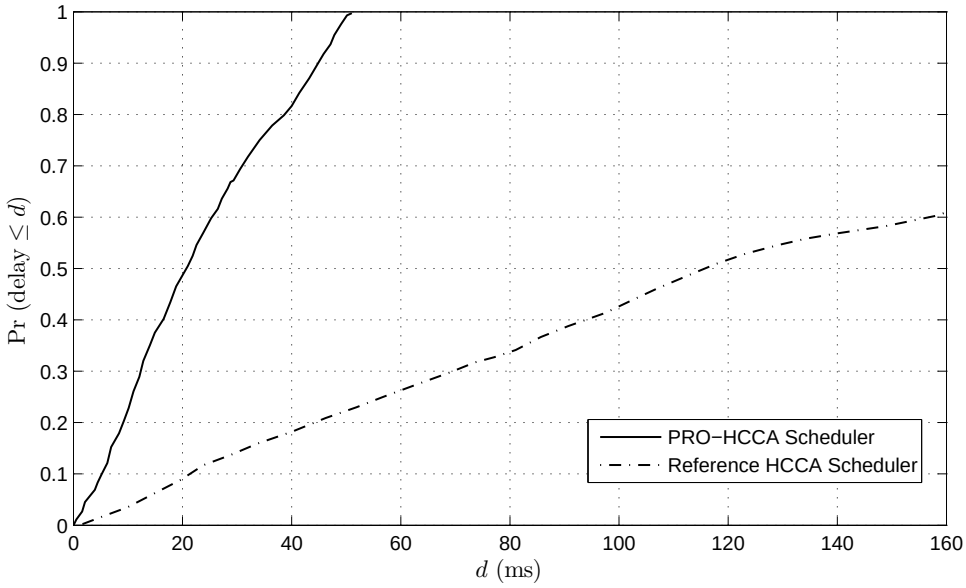


Figure 4.4: Cumulative distribution function of delay (ms) experienced by packets generated from the interactive video stream.

4.5.2 Simulation Results and Discussions

Fig. 4.4 shows the CDF of the delays experienced by packets belonging to the interactive video stream. The packet delay includes the access delay and the transmission time of the packet itself. We can observe that the reference scheduler performs very badly in terms of meeting the delay bound. Only 43% of the packets could be served within the pledged delay bound. On the other hand, the PRO-HCCA scheduler clearly shows its ability to meet the delay guarantee. All the packets generated by the stream are served within the guaranteed delay although the interactive video requires a very stringent delay bound of 100 ms. Table 4.3 also shows that this delay performance is achieved with no packet loss at all. Note that the PRO-HCCA is designed to drop any packet that fails to meet the delay bound. The reference scheduler does not have such provisioning. Even with this extra passage for packet dropping, the PRO-HCCA scheduler achieves 0% packet loss, which is

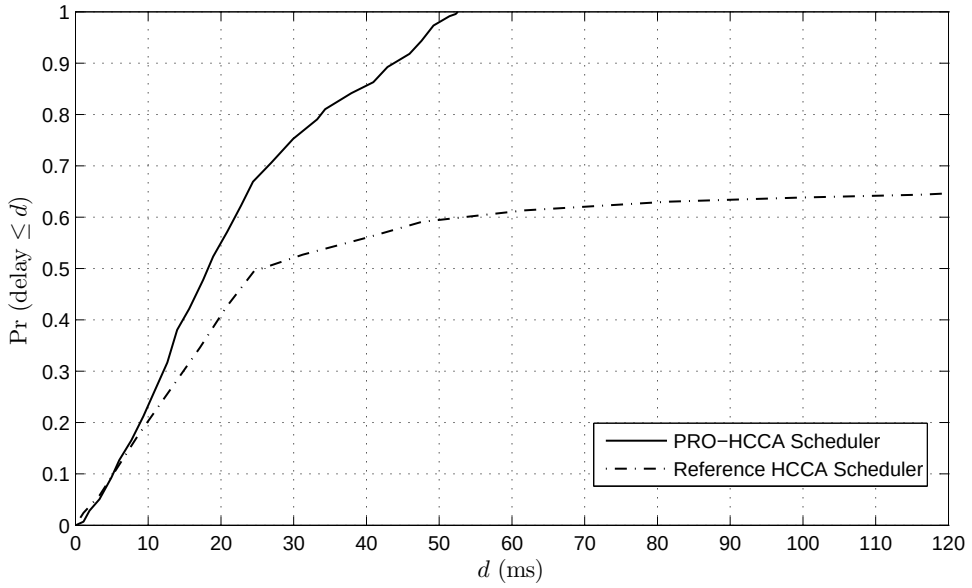


Figure 4.5: Cumulative distribution function of delay (ms) experienced by packets generated from the non-interactive video stream.

quite remarkable. On the other hand, the reference scheduler drops a significant percentage (almost 19%) of packets.

Fig. 4.5 shows the CDF of delays experienced by packets belonging to the non-interactive or streaming video traffic. Compared to the interactive video stream, this stream has better delay tolerance with a delay bound of 200 ms. The figure shows that the reference scheduler still performs very badly. Although the percentage of packets meeting the delay bound improves to nearly 65% because of the less strict delay bound, the performance is still unacceptable. In contrast, the proposed PRO-HCCA scheduler excels in performance in terms of meeting the delay bound. The figure clearly shows that 100% of the packets belonging to the stream are served with latencies comfortably within the delay guarantee. For the reference scheduler, the packet drop rate (2.5%) for the non-interactive stream is not as bad as it is with the interactive video stream. This improvement can be attributed to the lower intensity of the traffic arrival (since the video is of a medium quality) and also

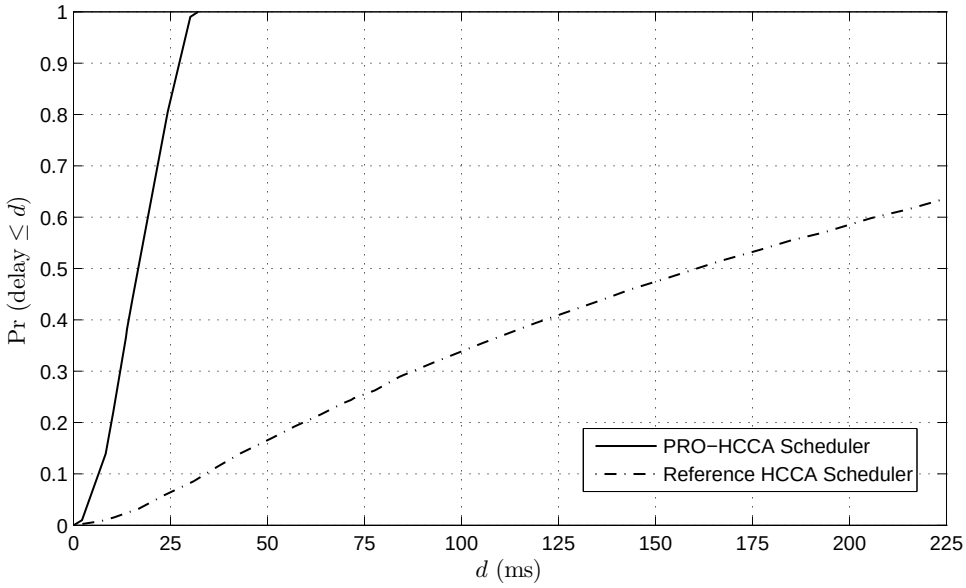


Figure 4.6: Cumulative distribution function of delay (ms) experienced by packets generated from the VBR VoIP stream.

to the less strict delay bound. On the other hand, the PRO-HCCA scheduler maintains 0% packet loss with the stream as well.

Finally, the CDF of packet delays for the VoIP stream is shown in Fig. 4.6. Although this stream has the most stringent delay bound of 50 ms, with the PRO-HCCA scheduler, all the packets of the stream are served within this delay bound. As is evident from the figure, the reference scheduler performs the worst in terms of meeting delay guarantees for the stream out of the three streams served. This also confirms that the reference scheduler penalizes more the streams with stricter delay bounds. On the other hand, the PRO-HCCA scheduler treats the streams fairly and can provide delay guarantees irrespective of the strictness of the delay bound. As shown in Table 4.3, the reference scheduler does not drop packet for the considered VBR VoIP stream because of the very low traffic arrival intensity of the stream. The proposed PRO-HCCA scheduler also exhibits no packet loss for the stream. The packet loss performance of the reference scheduler is highly sensitive

Table 4.3: Summary of the results

Stream Type	Packets served within delay bound		Packet loss	
	Reference HCCA Scheduler	PRO-HCCA Scheduler	Reference HCCA Scheduler Scheduler	PRO-HCCA Scheduler Scheduler
Interactive video	43%	100%	19%	0%
Non-Interactive video	65%	100%	2.5%	0%
VBR VoIP	17%	100%	0%	0%

to the traffic intensity of the relevant stream, whereas the PRO-HCCA scheduler shows stable performance irrespective of the traffic intensity. A summary of the numerical results is listed in Table 4.3.

Overall, the simulation results show the effectiveness of the PRO-HCCA scheduler in adapting to various multimedia traffic scenarios while delivering necessary QoS requirements. Although the results are shown for a WLAN with three QSTAs, the performance of PRO-HCCA is expected to show its effectiveness in much larger sized WLAN deployment. Due to the low computational complexity of the prediction mechanism used for the scheduler, the QAP will be able to handle the VBR traffic prediction with acceptable accuracy for a reasonable number of traffic streams. The optimization module is designed to achieve the maximum utilization of the available medium time and bandwidth. If the number of scheduled streams grows, the delays experienced by the streams are expected to grow towards their delay bounds. However, since the packets needing the most urgent service are serviced with the most priority, the delay guarantees will be met as much as possible with the available resource constraint. Moreover, the PRO-HCCA scheduler is designed to not schedule the packets that have already missed their deadlines. Therefore, meeting delay guarantees will enjoy higher priority over packet loss if the available medium time cannot accommodate the instantaneousness demand. Multimedia applications typically can tolerate some level of packet loss. By continuing to meet the delay guarantees at the expense of

some packet loss could be the worst case scenario if the number of streams and their aggregate QoS demand exceeds the available resource in the system. However, we assume that the HCCA admission controller will complement the operation of the scheduler to restrict the number of admitted streams so that aggregate QoS demand remains within the available resource in the WLAN system. We conjecture that the PRO-HCCA scheduler will be able to meet the QoS guarantees of increased number of multimedia streams from increased number of WLAN stations so long as the admission controller can effectively regulate the admission of traffic streams based on the QoS demands and the system capacity.

4.6 Conclusions

We have formulated an analytical model for the controlled channel access mechanism (also called HCCA) in IEEE 802.11e based WLANs considering VBR traffic applications. The model enables us to derive important performance measures in terms of the parameters used in the reference scheduler. The results obtained from the analytical model have shown that the QoS guarantees for VBR traffic cannot be met without dynamically adapting the transmission opportunities to variable traffic intensities and to urgencies for service of competing connections. Using the insights gained into the cause of the problems in the reference scheduler, we have developed a novel HCCA scheduler for IEEE 802.11e WLANs. The proposed scheduler adapts dynamically with dynamic traffic intensities that may arise from multimedia applications. The proposed PRO-HCCA scheduler utilizes simple but smart prediction mechanism along with an efficient online optimization process. The urgency of service for different parts of the backlogged queues is considered in the optimization step. Simulation experiments have shown that the PRO-HCCA scheduler achieves very robust and stable performance in meeting QoS guarantees for various VBR traffic applications.

4.7 Bibliography

[1] IEEE Std. 802.11e-2005, *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, Nov. 2005.

[2] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models - An Algorithmic Approach*, John Hopkins Univ. Press, Baltimore, MD, 1981.

[3] Z.-N. Kong, D. H. K. Tsang, B. Bensaou, and D. Gao, "Performance Analysis of IEEE 802.11e Contention-based Channel Access," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 10, pp. 2095- 2106, Dec. 2004.

[4] Z. Tao and S. Panwar, "An Analytical Model for the IEEE 802.11e Enhanced Distributed Coordination Function," *Proc. IEEE International Conference on Communications*, pp. 4111 - 4117, Paris, France, Jun. 2004.

[5] O. Tickoo and B. Sikdar, "Queueing Analysis and Delay Mitigation in IEEE 802.11 Random Access MAC Based Wireless Networks," *Proc. IEEE Infocom 2004*, Anchorage, AK, USA, Mar. 2004.

[6] S. Mangold, S. Choi, G. Hiertz, O. Klein, B. Walke, "Analysis of IEEE 802.11e for QoS Support in Wireless LANs," *IEEE Wireless Communications*, vol. 10, no. 6, pp. 40-50, Dec. 2003.

[7] G. Holland, N. Vaidya, and P. Bahl, "A Rate-Adaptive MAC Protocol for Multi-Hop Wireless Networks," *Proc. ACM/IEEE MOBICOM'01*, Rome, Italy, pp. 236–250, Jul. 2001.

[8] A. S. Alfa, "Vacation Models in Discrete Time," *Queueing Systems*, vol. 44, no. 1, pp. 5-30, May 2003.

[9] S. McCanne and V. Jacobson, "VIC: A Flexible Framework for Packet Video," *Proc. ACM Multimedia'95*, S. Francisco, California, Nov. 1995.

[10] A. Horváth and M. Telek, "Phfit: A General Phase-Type Fitting Tool," *Proc. 12th Performance TOOLS*, Lecture Notes in Computer Science, vol. 2324, London, UK, Apr. 2002.

[11] T. S. Randhawa and R. H. S. Ardy, "Estimation and Prediction of VBR Traffic in High-Speed Networks using LMS Filters," *Proc. IEEE International Conference on Communications*, Atlanta, GA, USA, June 1998.

[12] B. Farhang-Boroujeny, *Adaptive Filters: Theory and Applications*, New York: Wiley, 1998.

[13] J. C. Lee and C. K. Un, "Performance of Transform-Domain LMS Adaptive Digital Filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 499-510, 1986.

[14] T. Cormen, C. Leisorson, R. Rivest, and C. Stein, *Introduction to Algorithms, Second edition*, Cambridge: MIT Press, 2001.

[15] R. Bagrodia, R. Meyer, M. Takai, Y. Chen, X. Zeng, J. Martin, and H. Y. Song, "PARSEC: A Parallel Simulation Environment for Complex Systems," *IEEE Computer*, vol. 31, no. 10, pp. 77-85, Oct. 1998.

[16] MPEG-4 and H.263 Video Traces for Network Performance Evaluation, http://www.tkn.tu-berlin.de/research/trace/trace.html, Oct. 2006.

Chapter 5

Opportunistic Spectrum Scheduling for Multiuser Cognitive Radio: A Queueing Analysis[1]

5.1 Introduction

Widespread deployment of broadband wireless access systems have dramatically increased the demand for radio spectrum. This is expected because such systems require certain amount of bandwidth to offer broadband quality data rates. However, radio spectrum is one of the most scarce and valuable resources for wireless communications. As the available radio spectrum is limited, improving its utilization has gained renewed interest with the proliferation of broadband wireless technologies. New insights into the use of radio spectrum have challenged the traditional static spectrum allocation policy. Actual measurements

[1] A version of this chapter has been published. Rashid, M.M., Hossain, M.J., Hossain, E. and Bhargava V.K. (2009) Opportunistic Spectrum Scheduling for Multiuser Cognitive Radio: A Queueing Analysis. IEEE Trans. on Wireless Commun. 8(10):5259–5269.

have shown that most of the allocated spectrum is largely underutilized [1]. Similar views on the underutilization of allocated spectrum were reported by the Spectrum-Policy Task Force appointed by Federal Communications Commissions (FCC) [2]. Spectral efficiency can be increased significantly by giving opportunistic access of the frequency bands to unlicensed users. Cognitive radio (CR) has been proposed as a way to improve spectrum efficiency by exploiting the unused spectrum in dynamically changing environments [3]. The CR technology is based on an innovative radio design philosophy which involves smartly sensing the swaths of radio spectrum and then determining the transmission characteristics (e.g., symbol rate, power, bandwidth) for cognitive radio users (CR users) based on the perceived behavior of the primary users (i.e., users who are licensed for the spectrum in consideration).

Although opportunistic spectrum access would allow CR users to identify and access available spectrum resources, one of the main concerns is to utilize the available spectrum resources efficiently. As such CR users' quality-of-service (QoS) requirements are met without violating the priority right of primary users. In a CR network, the availability of radio spectrum for CR users depends mainly on the activity of primary users. On the other hand, since wireless channel quality of CR users is time-varying in nature, communication resource available to CR users is highly dynamic. The medium access control (MAC) layer protocols in a CR network should be able to adapt to the availability of the communication resources. An adaptive MAC layer, which efficiently uses communication resources given the QoS requirements, is an important component for an adaptive wireless access system based on CR technology. In order to design smart radio resource management schemes for this adaptive system, the system engineers have to determine how important system and operating parameters affect important performance measures relevant to the design. These

parameters are most likely to encompass different layers of the protocol stack. Therefore, a cross-layer analytical model will be very useful for efficient design, analysis, and optimization of resource management schemes such as scheduling and admission control.

Research interest on CR technology has grown tremendously off late and a reasonable number of ensuing research efforts have dealt with performance analysis of CR systems. In most of the cases, the focus of the performance analysis has been on the capacity, throughput and/or outage probability achievable at the physical layer (PHY). In [4], Devroye *et al.* presented an information theoretic analysis to derive the capacity regions of a cognitive radio channel. Information theoretic limit of throughput in cognitive networks was also analyzed in [5] and [6] assuming different cooperative behaviors between primary and CR users. Game theoretic models have also been used to analyze CR networks. Ji and Liu provided a review of theses game theoretic models in [7]. These models are useful to analyze achievable payoffs and associated fairness of service among CR users resulting from different spectrum sharing policies in CR networks. The problem of efficient spectrum sensing and spectrum access in CR networks was addressed in [8]-[10]. While [11] and [12] studied the spectrum access problem in a single-user scenario, [13] and [10] considered a multi-user scenario. A framework for medium access control (MAC) protocol adaptation in a dynamic spectrum access scenario was proposed in [14]. Simeone et al. [15] studied MAC layer throughput for a CR user over a cognitive link given a certain throughput at the primary user end. The system model, however, is very simplistic with a single CR user opportunistically sharing the cognitive link with a single primary user. The queueing model assumes infinite transmission queues and does not study delay or packet loss performance measures. In [16], an M/G/1 queueing model was used to analyze the transmission performance of multiple CR users on an aggregate basis considering both PHY and MAC

layer parameters. Analytical models were developed in [17] for cognitive MAC protocols with dedicated control channel and with embedded control channel. Call-level QoS performance (e.g., blocking performance) for CR users was analyzed in [18] based on a G/M/N queueing model. In a similar spirit, call-level performances for both primary users and CR users were analyzed in [19] using a two-dimensional Markov chain. A survey of the MAC schemes for CR networks can be found in [20]. A cross-layer design methodology for CR networks based on fuzzy logic was presented in [21].

Different from the above works, we develop a cross-layer (MAC layer and PHY) queueing model that considers multiple CR users competing for spectrum opportunities, and exploitation of multiuser diversity for allocating spectrum opportunities among the CR users. We consider a spectrum overlay scenario in an infrastructure-based dynamic spectrum access environment where the CR controller or the cognitive radio base station (BS) allocates the available spectrum opportunities among CR users. The cognitive radio BS uses an opportunistic scheduling scheme based on channel quality to allocate the spectrum resources among CR users, because such a scheduling scheme offers throughput gain with increased number of users in a wireless system [22]. For the queueing analysis, we model the activity of primary users in a channel as a two state Markov chain and the channel fading in the link between the cognitive radio BS and a CR user as a finite-state Markov chain (FSMC). The model also considers realistic traffic model that captures the potential burstiness in the arrival process. The queueing analysis enables use to relate the packet-level QoS performance measures at the MAC layer with a number of important system and PHY parameters. Therefore, this will facilitate cross-layer (MAC layer and PHY) design of CR networks. Also, the analysis facilitates the design of a model-based admission controller in order to provide the required QoS for the CR users. Selected numerical results are presented to show

the effect of important system and traffic parameters on the performance measures of interest including the delay distribution, packet loss rate, and throughput. The matrix-analytic method used in this chapter has a nice algorithmic approach, which is much more elegant than traditional queueing analysis methods that usually involve complex Laplace-Stieltjes Transforms (LST). Being free from LST, our method allows to numerically compute the delay and queue length distributions with much greater ease.

The rest of the chapter is organized as follows. In Section 5.2, we describe the system model which includes details of primary users' activity level modeling, wireless channel modeling, and the spectrum scheduling scheme. We develop the queueing analytic model in Section 5.3 and present some selected numerical results in Section 5.4. Example applications of the model are presented in Section 5.5. Section 6 concludes the chapter.

5.2 The System Model and Assumptions

5.2.1 Network Model

We consider an infrastructure-based CR system consisting a cognitive radio BS and multiple CR users which opportunistically utilize the spectrum licensed to some primary users (Fig. 5.1). The cognitive radio BS controls transmissions to/from the CR users in its coverage area. For downlink transmission, the packet queues for the CR users will be located at the cognitive radio BS. In case of uplink transmission, the location of the packet queues will be at the CR users. The analytical model presented later in the chapter will be valid for both the cases.

Following the cross-layer model in [23], we assume that packets arriving from the higher layers are buffered and scheduled at the MAC layer and then transmitted on air

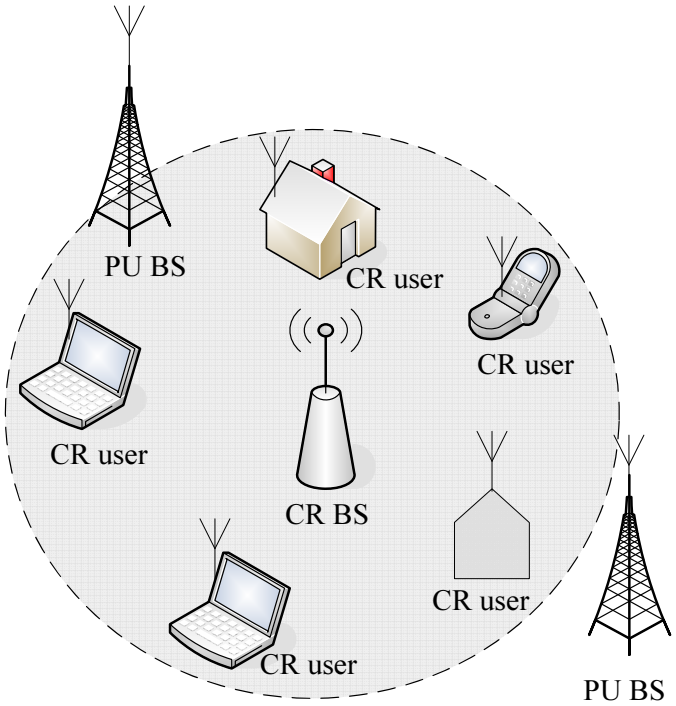


Figure 5.1: Infrastructure-based cognitive radio network architecture (CR = Cognitive radio; PU = Primary user; BS = Base station).

one or more at a time inside physical layer *frames*. The system is assumed to be time slotted with the slot size fixed to the duration (T_f) of a frame transmission. Adaptive modulation and coding (AMC) is used to adjust the PHY transmission mode for each radio channel between a CR user and the cognitive radio BS. Therefore, the size (in bits) of a frame might vary from frame to frame due to AMC and the number of channels assigned to the user. Assuming a fixed packet size, the number of packets transmitted in a frame duration will vary accordingly. Note that in the rest of the chapter, the term *time slot* will be used synonymously for the frame duration T_f.

Let the total bandwidth available for communication be W_{T}[Hz] which is licensed to a group of primary users. We assume that the total bandwidth is divided into M channels. Portions of the spectrum i.e., channels licensed to the primary users may remain unused

and thus be available to the CR users. The spectrum sensing technique implemented in the cognitive radio BS will enable it to discover these spectrum opportunities at each time slot n[2]. We assume that if a channel is not used by primary users at the beginning of a time slot n, it remains unoccupied by primary users during that slot and the cognitive radio BS can assign this channel to a CR user under its coverage. The availability of spectrum opportunities for CR users will vary over time depending on the activity level and pattern of the usage by the primary users. The wireless channel quality of CR users is also time-varying in nature. Since in a given time slot, a number of CR users may demand for the spectrum resource, a scheduling method will be required to allocate the spectrum opportunities among CR users.

5.2.2 Primary User Activity

We assume that the channel occupancy by primary users for i^{th} $(1 \leq i \leq M)$ channel is a two-state time-homogenous first-order Markov Process, and the process is independent and identical from one channel to another. Similar channel occupancy model was used in [11]. Let O denote the Markov process representing the channel occupancy by primary users and $\mathbf{P}_O$ denote its transition probability matrix. In this chapter, we refer $\mathbf{P}_O$ as the *occupancy matrix* and define it as follows:

$$\mathbf{P}_O = \begin{bmatrix} 1 - p_{0\to1} & p_{0\to1} \\ p_{1\to0} & 1 - p_{1\to0} \end{bmatrix} \tag{5.1}$$

where $p_{0\to1}$ denotes the probability that the channel is occupied in time n and will be released in the next time slot $n+1$. Similarly, $p_{1\to0}$ denotes the probability that an unoc-

[2]Spectrum sensing can be performed through separate sensing devices installed by the cognitive radio service provider [24].

cupied channel will be occupied by a primary user in the next time slot. The activity factor of primary users in a given channel can be defined as the probability that the channel will be occupied in a given time slot. This probability can be obtained by finding the steady state probabilities of the state transition matrix in (5.1). We assume that theses transition probabilities are known at the cognitive radio BS. We also assume a perfect channel sensing scheme in the CR network in consideration, and therefore, ignore any channel sensing error when estimating primary user activity.

5.2.3 Channel Model and Adaptive Modulation

The power gain for channel fading of CR user j at channel i, $\gamma^{i,j}$ is assumed to be slowly varying with Gamma distribution which corresponds to the Nakagami-m distribution [25] for fading amplitude variation. The channel fading process can be modeled as an FSMC [26], for which, $\gamma^{i,j}$ maps to a channel state in set $\mathcal{C} = \{0, 1, \ldots, Z\}$ with a total of $Z + 1$ states. In a time slot, let the state of channel i for CR user j be k if $\sigma_k \leq \gamma^{i,j} < \sigma_{k+1}$. The power gain thresholds satisfy the following inequality: $0 = \sigma_0 \leq \ldots \leq \sigma_k \leq \ldots \leq \sigma_{Z+1} = \infty$. A set of AMC schemes are chosen as transmission modes according to a one-to-one relationship with these channel states. The following relationship between packet transmission rates and transmission modes is also assumed: $\xi_k = bk, 1 \leq k \leq Z$, where ξ_k denotes the number of packets that can be transmitted during a time slot when transmission mode k (which corresponds to channel state k) is selected, b is an integer which depends on time slot duration and modulation parameters. Channel state 0 simply represents that there will be no transmission in the channel when the fading SNR remains below σ_1. The thresholds σ_k $(1 \leq k \leq Z)$ can be obtained to meet the average packet error rate (PER) target for for each of these transmission modes [23]. For a given Doppler shift,

the transition probabilities between the channel states of the FSMC can be obtained as in [26]. Note that, in our model, a channel can change from one frame to the next, and the dynamics of such changes is captured through the FSMC channel model. A session, on the other hand, usually contains many frame times and therefore, signal strength of a channel can change many times during a session. The assumption made in our model only restricts a channel state to remain unchanged over a frame time, which is usually in the range of few milliseconds. However, we allow the possibility that a channel can change many times during a session, which is usually in the order of many seconds.

5.2.4 Opportunistic Spectrum Scheduling

For wireless networks, opportunistic scheduling, which exploits multiuser diversity, has been proved to maximize system throughput [27]. Due to the inherent gain from multiuser diversity, we also assume an opportunistic spectrum scheduling scheme. In this scheme, an empty channel is assigned to a CR user for whom the highest transmission rate can be supported in that channel. The spectrum scheduling scheme can be summarized as follows:

- At time slot n, the cognitive radio BS identifies the set of unoccupied channels, $\mathcal{L}(n)$.
- From the list $\mathcal{L}(n)$, a channel is picked up and the channel state for each CR user is checked for this channel. It is assigned to the CR user with the highest channel state. If there are more than one CR users with the highest state, a user is selected randomly. The channel is assigned to this user and removed from the list.
- This procedure is repeated until all the channels in $\mathcal{L}(n)$ are assigned to the CR users

According to the above spectrum scheduling scheme, a CR user may be allocated more than one channel in a given time slot. Also, multiple CR users could be allocated spectrum

opportunities in the same time slot. For this, we assume that a hybrid FDMA/TDMA mechanism is employed at the cognitive radio BS.

Note that the channel fading information for the CR users need to be made available at the scheduler through some Channel Quality Information (CQI) feedback mechanism implemented at the BS. The main cost of such feedback is the extra uplink radio resource necessary to carry the CQI. How to minimize the amount of radio resources necessary for the CQI feedback method is itself a very active research area and in our analysis, we do not assume any particular scheme to obtain channel fading information. For the system model analyzed in our chapter, the possible CQI (i.e. SNR) values are divided into a predefined number of channel states; therefore, for a particular channel, a user can report the channel state instead of reporting the exact SNR. As a result, a CR user can forgo sending CQI to the cognitive radio BS for a channel when channel fading in the current frame does not differ from that in the previous frame to warrant a transition in channel states. Therefore, we conjecture that the cost of getting the channel fading information would be reasonably low. Since there may not be any common control channel available for periodic feedback of channel state information from the CR users to the cognitive radio BS, a predictive scheme can be used [28].

5.3 Formulation of the Queueing Model

5.3.1 Packet Arrival Process

Assuming a homogenous system, we develop the queueing model from the perspective of a tagged CR user. We assume that packet arrivals for a CR user follow a Batch Bernoulli process which can be described by a vector $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \cdots, \alpha_N]$, where N is the maxi-

mum number of packets that can arrive during a time slot. Here, α_0 is the probability of no packet arrival and $\alpha_i \quad (1 \leq i \leq N)$ is the probability that i packets arrive in a time slot. The average arrival rate λ (in packets/slot) can be calculated as $\lambda = \sum_{i=1}^{N} i\alpha_i$.

The assumed Batch Bernoulli arrival model is very general which can capture different level of burstiness in the traffic arrival process. We also assume that packets arriving during time slot $n-1$ cannot be served until time slot n at the earliest. The packets in the queue are served in a first-come first-served manner and packet transmissions in a time slot are assumed to finish before arriving packets enter the queue.

5.3.2 Joint System State Considering Primary User Activity, Channel State, and Opportunistic Scheduling

Let random variable $m_n^{(i)} \in \{1, 0\}$ represent whether ith channel is assigned to the tagged user (CR user 1) or not at time slot n. We assume that $m_n^{(i)} = 1$ denotes that CR user 1 is assigned to ith channel at time slot n. Similarly $c_n^{(i)} \in \mathcal{C}$ denotes the channel state of ith channel at time slot n. Therefore, the joint state for the ith channel at time slot n can be denoted as $(m_n^{(i)}, c_n^{(i)})$. The corresponding state transition probability matrix $\boldsymbol{\mathcal{M}}^{(i)}$ can be determined as follows:

$$\boldsymbol{\mathcal{M}}^{(i)} = \begin{bmatrix} \boldsymbol{\mathcal{M}}_{0\to0}^{(i)} & \boldsymbol{\mathcal{M}}_{0\to1}^{(i)} \\ \boldsymbol{\mathcal{M}}_{1\to0}^{(i)} & \boldsymbol{\mathcal{M}}_{1\to1}^{(i)} \end{bmatrix} \tag{5.2}$$

where $\boldsymbol{\mathcal{M}}_{k\to l}^{(i)}$ is a $(Z+1) \times (Z+1)$ matrix whose element at row y and column z denoted here by $\boldsymbol{\mathcal{M}}_{k\to l}^{(i)}(y, z)$ is defined as $\boldsymbol{\mathcal{M}}_{k\to l}^{(i)}(y, z) = \Pr\,(m_{n+1}^{(i)} = l, c_{n+1}^{(i)} = z | m_n^{(i)} = k, c_n^{(i)} = y)$. These probabilities can be obtained using similar equations as (3.6)-(3.13) of Chapter 3. For brevity, we do not include the details of the procedure here. Also note that these transition probabilities are calculated for a specific number of users inside the network.

When any CR user moves in and out of the CR network, these transition probabilities can be recalculated with the updated number of CR users in the system.

However, assignment of a given channel to a CR user does not guarantee that the user will be able to access the channel since it can be occupied by a primary user. Therefore, we introduce another random variable $o_n^{(i)} \in \{1, 0\}$ to denote the occupancy state of ith channel by primary users at time slot n. Specifically, $o_n^{(i)} = 0$ means ith channel is occupied by primary users and $o_n^{(i)} = 1$ represents that the channel is available to the CR user at time slot n. In the previous section, we have also introduced the occupancy matrix $\mathbf{P}_O$ to describe the Markov process assumed for channel occupancy by primary users. Therefore, for ith channel, the joint state considering the spectrum scheduling scheme and primary user activity can be denoted as

$$\mathbf{\Phi}^{(i)} \triangleq \left\{ \left(o_n^{(i)}, m_n^{(i)}, c_n^{(i)} \right) | 0 \leq o_n^{(i)} \leq 1, 0 \leq m_n^{(i)} \leq 1, 0 \leq c_n^{(i)} \leq Z \right\}.$$

The transition probabilities among the states in $\mathbf{\Phi}^{(i)}$ can be expressed using matrix $\mathbf{H}^{(i)}$ as: $\mathbf{H}^{(i)} = \mathbf{P}_O \bigotimes \boldsymbol{\mathcal{M}}^{(i)}$, where $\bigotimes$ represents the Kronecker product.

Let us now define a new state variable $s_n(M)$ calculated as: $s_n(M) = \sum_{1 \leq i \leq M} o_n^{(i)} m_n^{(i)} c_n^{(i)}$ and construct a state space $\mathbf{\Theta}(M) \triangleq \{s_n(M) | 0 \leq s_n(M) \leq MZ\}$. Next, we describe a procedure to calculate the transition probability matrix $\mathbf{S}(M)$ for state space $\mathbf{\Theta}(M)$.

Let us consider channels 1 and 2 first and start by constructing a joint state space

$$\mathbf{\Omega}^{(2)} \triangleq \left\{ \left(o_n^{(1)}, m_n^{(1)}, c_n^{(1)}, o_n^{(2)}, m_n^{(2)}, c_n^{(2)} \right) | 0 \leq o_n^{(1)}, o_n^{(2)} \leq 1; 0 \leq m_n^{(1)}, m_n^{(2)} \leq 1; 0 \leq c_n^{(1)}, c_n^{(2)} \leq Z \right\}.$$

The transition probabilities among these states can be represented by matrix $\mathfrak{R}^{(2)}$ as: $\mathfrak{R}^{(2)} = \mathbf{H}^{(1)} \bigotimes \mathbf{H}^{(2)}$.

The steady state probability vector $\boldsymbol{\tau}^{(2)}$ for the state space $\boldsymbol{\Omega}^{(2)}$ can be derived by solving the following set of linear equations: $\boldsymbol{\tau}^{(2)}\mathfrak{R}^{(2)} = \boldsymbol{\tau}^{(2)}$ and $\boldsymbol{\tau}^{(2)}\mathbf{1} = 1$, where $\mathbf{1}$ stands for a column matrix containing 1 only in each entry.

With $s_n(2) = o_n^{(1)}m_n^{(1)}c_n^{(1)} + o_n^{(2)}m_n^{(2)}c_n^{(2)}$, let y_t ($y_t \subset \boldsymbol{\Omega}^{(2)}$) denote the set of states that result in $s_n(2) = t$, where $0 \leq t \leq 2Z$. With $\mathfrak{R}^{(2)}_{h\to h'}$ denoting the transition probability from state h to h', where both h and h' are elements of state space $\boldsymbol{\Omega}^{(2)}$, we can construct the transition probability matrix $\mathbf{S}(2)$ for state space $\boldsymbol{\phi}(2)$ as follows:

$$\mathbf{S}_{k\to k'}(2) = \frac{\sum\limits_{h\in y_k}\sum\limits_{h'\in y_{k'}} \boldsymbol{\tau}_h^{(2)}\mathfrak{R}^{(2)}_{h\to h'}}{\sum\limits_{h\in y_k}\boldsymbol{\tau}_h^{(2)}} \tag{5.3}$$

where $\mathbf{S}_{k\to k'}(2)$ denotes the transition probability from state k to state k', where $0 \leq k, k' \leq 2Z$.

We now describe a generalized procedure to recursively calculate transition probability matrix $\mathbf{S}(M)$ for $M \geq 3$. We combine state variables $o_n^{(M)}, m_n^{(M)}$, and $c_n^{(M)}$ for channel M with state variable $s_n(M-1)$ to construct the joint state space

$$\boldsymbol{\Omega}^{(M)} \triangleq \left\{\left(s_n(M-1), o_n^{(M)}, m_n^{(M)}, c_n^{(M)}\right) | 0 \leq o_n^{(i)} \leq 1; 0 \leq m_n^{(i)} \leq 1; 0 \leq c_n^{(i)} \leq Z\right\},$$

whose transition probability matrix $\mathfrak{R}^{(M)}$ can be derived, assuming $\mathbf{S}(M-1)$ has been calculated in the previous iteration of the recursion, as $\mathfrak{R}^{(M)} = \mathbf{S}(M-1) \bigotimes \mathbf{H}^{(M)}$.

From definition of $s_n(M)$, we can also write $s_n(M) = s_n(M-1) + o_n^{(M)}m_n^{(M)}c_n^{(M)}$. Therefore, the transition probability matrix $\mathbf{S}(M)$ for M channels can be derived similarly as was done for the two channel case by grouping states having same $s_n(M)$ values from state space $\boldsymbol{\Omega}^{(M)}$.

Now, let $r_n(M)$ denote the sum rate (in packets/time slot) available to the tagged CR user due to M channels at time slot n. Since at any time slot n, the number of packets that can be transmitted over a single channel is bc_n when the channel's state is c_n, it is straightforward to show that $r_n(M) = bs_n(M)$. For the sake of clarity in mathematical expressions, we will omit the index M from $s_n(M)$ and $\mathbf{S}(M)$ in the rest of the analysis.

5.3.3 Quasi-Birth-and-Death Process Formulation and Analysis

The system can be viewed as a time-slotted system where transitions among the states occur at the slot boundaries. In addition, the state variables assume discrete values only. Therefore, we can model the system using a discrete-time Markov chain (DTMC). To derive the transition probability matrix of the DTMC, we consider the system dynamics as it evolves from one time slot to the next time slot. With a finite buffer size of Q packets, the state space of the system can be written as follows:

$$\Psi \triangleq \{(x_n, s_n) \,|0 \leq x_n \leq Q; 0 \leq s_n \leq MZ\},$$

where x_n is the queue length in number of packets at time slot n.

Once the transition probabilities in (5.2) are calculated, assuming an infinite buffer space ($Q = \infty$) and setting $L = bMZ$, the state transition matrix for the DTMC can be

written as

$$\mathbf{P} = \left[\begin{array}{c|c|ccccc|ccccccc|ccccccccc|cccc}
A_0^{(0)} & A_{1^+}^{(0)} & \cdots & \cdots & A_{N^+}^{0} & & & & & & & & & & & & & & & & & & & & \\
\hline
A_{1^-}^{(1)} & A_0^{(1)} & A_{1^+}^{(1)} & \cdots & \cdots & A_{N^+}^{(1)} & & & & & & & & & & & & & & & & & & & \\
\vdots & \vdots & & & & \ddots & & & & & & & & & & & & & & & & & & & \\
A_{(L-N+1)^-}^{(L-N+1)} & A_{(L-N)^-}^{(L-N+1)} & \cdots & \cdots & \cdots & \cdots & A_{1^-}^{(L-N+1)} & A_0 & A_{1^+}^{(L-N+1)} & \cdots & A_{(N-1)^+}^{(L-N+1)} & A_{N^+} & & & & & & & & & & & & & \\
\vdots & \vdots & & & & & & & \ddots & & \vdots & \vdots & \ddots & & & & & & & & & & & & \\
A_{L^-}^{(L)} & A_{(L-1)^-}^{(L)} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & A_{1^-}^{(L)} & A_0 & A_{1^+} & \cdots & \cdots & A_{N^+} & & & & & & & & & & \\
\hline
 & A_{L^-} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & A_{1^-} & A_0 & A_{1^+} & \cdots & \cdots & A_{N^+} & & & & & & & & & \\
 & & \ddots & & & & & & & & \vdots & \vdots & & & & \ddots & & & & & & & & & \\
 & & & A_{L^-} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & A_{(L-N+1)^-} & A_{(L-N)^-} & \cdots & \cdots & \cdots & \cdots & A_{1^-} & A_0 & A_{1^+} & \cdots & A_{(N-1)^+} & A_{N^+} & & & \\
 & & & & & & & & & \ddots & \vdots & \vdots & & & & & & & \ddots & & \vdots & \vdots & \ddots & \\
 & & & & & & & & & & A_{L^-} & A_{(L-1)^-} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & A_{1^-} & A_0 & A_{1^+} & \cdots & A_{N^+} \\
\hline
 & & & & & & & & & \ddots & & & & & & & & & & & & & & \ddots
\end{array}\right]. \tag{5.4}$$

By constructing blocks of submatrices as shown in (3.14), we can obtain a quasi-birth-and-death (QBD) process with a transition matrix of following form:

$$\mathbf{P_{inf}} = \begin{bmatrix}
\mathbf{B} & \mathbf{E} & & & & \\
\mathbf{C} & \mathbf{D} & \mathbf{G}_0 & & & \\
 & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G}_0 & & \\
 & & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G_0} & \\
 & & & \ddots & \ddots & \ddots
\end{bmatrix}. \tag{5.5}$$

However, we are interested in finding performance measures for systems with finite buffer space, which is more realistic. Assuming a finite buffer size of Q and setting $K = \lfloor Q/L \rfloor$, we can write the transition matrix in the following form:

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ K-1 \\ K \end{array} \begin{bmatrix} \mathbf{B} & \mathbf{E} & & & & & & \\ \mathbf{C} & \mathbf{D} & \mathbf{G}_0 & & & & & \\ & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G}_0 & & & & \\ & & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G_0} & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \mathbf{G}_2 & \mathbf{G}_1 & \mathbf{G'}_0 & \\ & & & & & \mathbf{G'}_2 & \mathbf{G'}_1 & \end{bmatrix}. \tag{5.6}$$

To proceed further, let us define a set of matrices $\mathbf{U}^{(k)}(0 \leq k \leq MZ)$ as follows:

$$\mathbf{U}_{i,j}^{(k)} = \begin{cases} \mathbf{S}_{i,j} & \text{if} \quad i = k \\ 0, & \text{if} \quad i \neq k \end{cases}$$

where $0 \leq i, j \leq MZ$.

With these matrices defined, we can derive the inner submatrices of $\mathbf{P}$ in (3.14) as follows:

$$\mathbf{A}_0^{(0)} = \alpha_0 \mathbf{S}; \quad \mathbf{A}_{d^+}^{(0)} = \alpha_d \mathbf{S}, 1 \leq d \leq N \tag{5.7}$$

$$\mathbf{A}_{d^-}^{(e)} = \begin{cases} \alpha_0 \sum\limits_{\lceil e/b \rceil \leq j \leq MZ} \mathbf{U}^{(j)}, & \text{if} \quad d = e \\ \sum\limits_{\substack{0 \leq j \leq e-d-1; \\ (j+d) \bmod b = 0}} \alpha_j \mathbf{U}^{\left(\frac{j+d}{b}\right)} + \alpha_{e-d} \sum\limits_{\lceil e/b \rceil \leq j \leq MZ} \mathbf{U}^{(j)}, & \text{if} \quad e - N \leq d \leq e-1 \\ \sum\limits_{\substack{0 \leq j \leq N; \\ (j+d) \bmod b = 0}} \alpha_j \mathbf{U}^{\left(\frac{j+d}{b}\right)}, & \text{if} \quad 1 \leq d < e - N, \end{cases}$$
$$\text{for} \quad 1 \leq e \leq L-1. \tag{5.8}$$

$$\mathbf{A}_{d^+}^{(e)} = \sum_{\substack{d\le j\le d+e-1;\\ (j-d) \bmod b=0}} \alpha_j \mathbf{U}^{(\frac{j-d}{b})} + \alpha_{d+e} \sum_{\lceil e/b\rceil \le j\le MZ} \mathbf{U}^{(j)}, 1 \le d \le N, 1 \le e \le L-1 \quad (5.9)$$

$$\mathbf{A}_0^{(e)} = \begin{cases} \sum\limits_{\substack{0\le j\le e-1;\\ j \bmod b=0}} \alpha_j \mathbf{U}^{\left(\frac{j}{b}\right)} + \alpha_e \sum\limits_{\lceil e/b\rceil \le j\le MZ} \mathbf{U}^{(j)}, & \text{if} \quad 1 \le e \le N \\ \sum\limits_{\substack{0\le j\le N;\\ j \bmod b=0}} \alpha_j \mathbf{U}^{\left(\frac{j}{b}\right)}, & \text{if} \quad N < e \le L \end{cases} \quad (5.10)$$

$$\mathbf{A}_{d^-} = \begin{cases} \sum\limits_{\substack{0\le j\le N;\\ (j+d) \bmod b=0}} \alpha_j \mathbf{U}^{\left(\frac{j+d}{b}\right)}, & \text{if} \quad 1 \le d < L-N \\ \\ \sum\limits_{\substack{0\le j\le L-d;\\ (j+d) \bmod b=0}} \alpha_j \mathbf{U}^{\left(\frac{j+d}{b}\right)}, & \text{if} \quad L-N \le d < L \end{cases} \quad (5.11)$$

$$\mathbf{A}_{d^-} = \sum_{\substack{0\le j\le N;\\ (j+d) \bmod b=0}} \alpha_j \mathbf{U}^{\left(\frac{j+d}{b}\right)}, 1 \le d \le L \quad (5.12)$$

$$\mathbf{A}_{d^+} = \sum_{\substack{d\le j\le N;\\ (j-d) \bmod b=0}} \alpha_j \mathbf{U}^{\left(\frac{j-d}{b}\right)}, 1 \le d \le N \quad (5.13)$$

$$\mathbf{A}_0 = \sum_{\substack{0 \le j \le N; \\ j \bmod b = 0}} \alpha_j \mathbf{U}^{\left(\frac{j}{b}\right)}; \quad \mathbf{G'}_0 = \begin{bmatrix} \mathbf{G}_0 & \mathbf{0}_{LMZ \times K'MZ} \end{bmatrix} \tag{5.14}$$

$$\mathbf{G'}_1 = \begin{bmatrix} \mathbf{A}_0 & \mathbf{A}_{1^+} & \cdots & \mathbf{A}_{N^+} & & \\ \mathbf{A}_{1^-} & \mathbf{A}_0 & \mathbf{A}_{1^+} & \cdots & \mathbf{A}_{N^+} & \\ \vdots & & \ddots & & \ddots & \\ \mathbf{A}_{(L-N+K')^-} & \cdots & \cdots & \cdots & & \mathbf{A}'_{(N-1)^+} \\ \vdots & & \ddots & & & \vdots \\ \mathbf{A}_{L^-} & \cdots & \cdots & \cdots & \cdots & \mathbf{A}'_{(K'-1)^+} \\ \ddots & & \ddots & & & \vdots \\ & \mathbf{A}_{L^-} & \cdots & \cdots & \cdots & \mathbf{A}'_{1^+} \\ & & \mathbf{A}_{L^-} & \cdots & \cdots & \mathbf{A}'_0 \end{bmatrix} \tag{5.15}$$

$$\mathbf{A}'_0 = \sum_{0 \le j \le N} \sum_{\substack{0 \le h \le j; \\ h \bmod b = 0}} \alpha_j \mathbf{U}^{\left(\frac{h}{b}\right)} \tag{5.16}$$

$$\mathbf{A}'_{k^+} = \sum_{k \le j \le N} \sum_{\substack{0 \le h \le j-k; \\ h \bmod b = 0}} \alpha_j \mathbf{U}^{\left(\frac{h}{b}\right)}, 1 \le k \le N-1 \tag{5.17}$$

$$\mathbf{G'}_2 = \begin{bmatrix} \mathbf{G}_2 \\ \mathbf{0}_{K'MZ \times LMZ} \end{bmatrix}, \tag{5.18}$$

where $K' = Q - \lfloor Q/L \rfloor$.

5.3.4 Matrix-Analytic Solution for Steady State Distributions

With the system described by a QBD type Markov chain whose block matrices are derived in the previous subsection, we can now apply matrix-analytic method [29] to find its steady state probability vector $\boldsymbol{\pi}$. For the detailed procedure to calculate $\boldsymbol{\pi}$, we refer the readers to [29]. By blocking the entries in $\boldsymbol{\pi}$ according to the associated queue lengths, we can write the steady state probability vector as $\boldsymbol{\pi} \triangleq [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_Q]$, where $\boldsymbol{\pi}_i$ stands for steady state probability vector associated with a queue length of i $(0 \leq i \leq Q)$.

5.3.5 Derivation of Performance Measures

After obtaining the steady state probability vector for queue length, we can determine a number of performance measures of interest at the MAC layer. We are mainly interested in the following three important performance measures: packet loss rate, average throughput, and the distribution of access delay for packets in the tagged queue. We denote by access delay the time between the moment a packet arrives at the queue and the moment the packet gets access into the medium for transmission.

Delay Distribution

The delay distribution for a packet in the tagged user queue is derived by considering an absorbing Markov chain. The chain initializes from the state when a packet arrives at the

corresponding queue and gets absorbed in the states associated with the transmission of the packet. Note that, in this case, there is no arrival in that particular queue while the process moves towards the absorbing state. Therefore, the transition matrix $\mathbf{P}_{abs}$ of the chain can be constructed by setting $\alpha_0 = 1$ and $\alpha_i = 0 \quad (1 \leq i \leq N)$. The next step for finding the delay distribution is to determine the initial probability vector for this absorbing Markov chain. To this end, we first consider the probability vector associated with the number of packets including the arriving packet, and the packets ahead of it as seen by the arriving packet. This probability vector $\boldsymbol{\xi}^{(0)} = \left[\boldsymbol{\xi}_0^{(0)}, \boldsymbol{\xi}_1^{(0)}, \cdots, \boldsymbol{\xi}_{Q+N}^{(0)}\right]$ can be derived as:

$$\boldsymbol{\xi}_h^{(0)} = \begin{cases} \mathbf{0}, \quad \text{if} \quad h = 0 \\ \frac{1}{1-\alpha_0}\left[\sum_{i=1}^{h}\sum_{j=1}^{i}\sum_{k=0}^{MZ}\frac{\alpha_i}{i}\pi_{h-j+bk}\mathbf{U}^{(k)} + \sum_{i=h+1}^{N}\sum_{j=1}^{h}\sum_{k=0}^{MZ}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+bk}\mathbf{U}^{(k)}\right], \quad \text{if} \quad 1 \leq h \leq N \\ \frac{1}{1-\alpha_0}\left[\sum_{i=1}^{N}\sum_{j=1}^{i}\sum_{k=0}^{MZ}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+bk}\mathbf{U}^{(k)}\right], \quad \text{if} \quad N < h \leq Q-L+1 \\ \frac{1}{1-\alpha_0}\left[\sum_{i=1}^{N}\sum_{j=1}^{i}\sum_{\substack{0\leq k\leq Q-h+j;\\ k \bmod b=0}}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+k}\mathbf{U}^{\left(\frac{k}{b}\right)}\right], \quad \text{if} \quad Q-L+1 < h \leq Q \\ \frac{1}{1-\alpha_0}\left[\sum_{i=h-Q}^{N}\sum_{j=h-Q}^{i}\sum_{\substack{0\leq k\leq Q-h+j;\\ k \bmod b=0}}\frac{\alpha_i}{i}\boldsymbol{\pi}_{h-j+k}\mathbf{U}^{\left(\frac{k}{b}\right)}\right], \quad \text{if} \quad Q < h \leq Q+N \end{cases} \tag{5.19}$$

Note that the packets which find their destination queues full upon arrival are not admitted into the system. The probability that an arriving packet finds the corresponding queue full can be written as: $p_{full} = \sum_{i=1}^{N} \boldsymbol{\xi}_{Q+i}^{(0)}$.

Since we consider the delay experienced by the admitted packets only, the initial prob-

ability vector $\boldsymbol{\omega}^{(0)} \triangleq \left[\boldsymbol{\omega}_0^{(0)}, \boldsymbol{\omega}_1^{(0)}, \cdots, \boldsymbol{\omega}_Q^{(0)}\right]$ of the absorbing Markov chain is derived from the above probability vectors as follows:

$$\boldsymbol{\omega}_i^{(0)} = \frac{\boldsymbol{\xi}_i^{(0)}}{1 - p_{full}}, 0 \leq i \leq Q \tag{5.20}$$

which is due to the fact that an arriving packet seeing a full queue will be dropped and consequently, will not be considered in the delay calculation. After d time slots the state vector of the absorbing chain is $\boldsymbol{\omega}^{(d)} = \boldsymbol{\omega}^{(0)}\mathbf{P}_{abs}^d$, where $\mathbf{P}_{abs}^d$ is the transition matrix derived by multiplying $\mathbf{P}_{abs}$ with itself d times. The cumulative distribution function (cdf) of packet delay D is given by: $F_D(d) = \boldsymbol{\omega}_0^{(d)}\mathbf{1}$.

Packet Loss Rate and Average Throughput

A packet loss occurs when a packet finds the queue full upon arrival or when a packet is received in error in the receiver. Assuming that no error recovery protocol is used, the packet loss rate, p_{loss} can be found as: $p_{loss} = 1 - (1 - p_{full})(1 - \overline{PER})$. The average throughput η in packets per time slot can then be calculated as: $\eta = \lambda(1 - p_{loss})$.

5.4 Numerical Results and Discussions

Using the queueing analytical model, the MAC layer QoS performance measures for a tagged CR user are obtained considering PHY as well as some other system parameters (e.g., primary users' activity level, number of radio channels, number of CR users). Unless otherwise specified in the following subsections, we also assume the parameter values listed below:

- Duration of a time slot, $T_f = 2ms$, packet size, $L_p = 1080$ bits

- Doppler spread, $f_d = 15$ Hz, average SNR = 15 dB, Nakagami fading parameter, $m = 1.1$
- Number of channels, M = 12, number of CR users = 10, buffer size, $Q = 30$ packets
- Packet arrival probability vector, $\alpha = [0.0, 0.3, 0.4, 0.3]$, primary user activity factor, θ = 0.5

Each of the channels are modeled by an FSMC derived for target $\overline{PER} = 1\%$; all the channels are assumed to be homogeneous in terms of their fading characteristics. We use MATLAB to code and execute the different steps in the development of the queueing model.

5.4.1 Primary User Activity Level

We capture the activity level of primary users by occupancy matrix derived from two parameters: θ and the average occupancy duration t of primary users on a given channel. For example, with an activity factor $\theta = 0.6$ and average occupancy duration of $t = 5$ time slots, the occupancy matrix $\mathbf{P}_o$ will be $\begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}$.

Fig. 5.2 shows the effect of primary user activity level on packet delay distribution of the CR user. As expected, the delay performance deteriorates when the primary users' activity factor or the average occupancy duration increases. The effect of the average occupancy duration, however, is not very prominent in the figure when $\theta = 0.4$ and becomes much more significant with $\theta = 0.8$ than with $\theta = 0.6$. Therefore, the average occupancy duration tends to affect the delay performance more when the activity factor is higher. Fig. 5.3 shows the relationship between packet loss rate at the CR user end and the activity level of the primary users. The packet loss rate increases rapidly with increasing value of

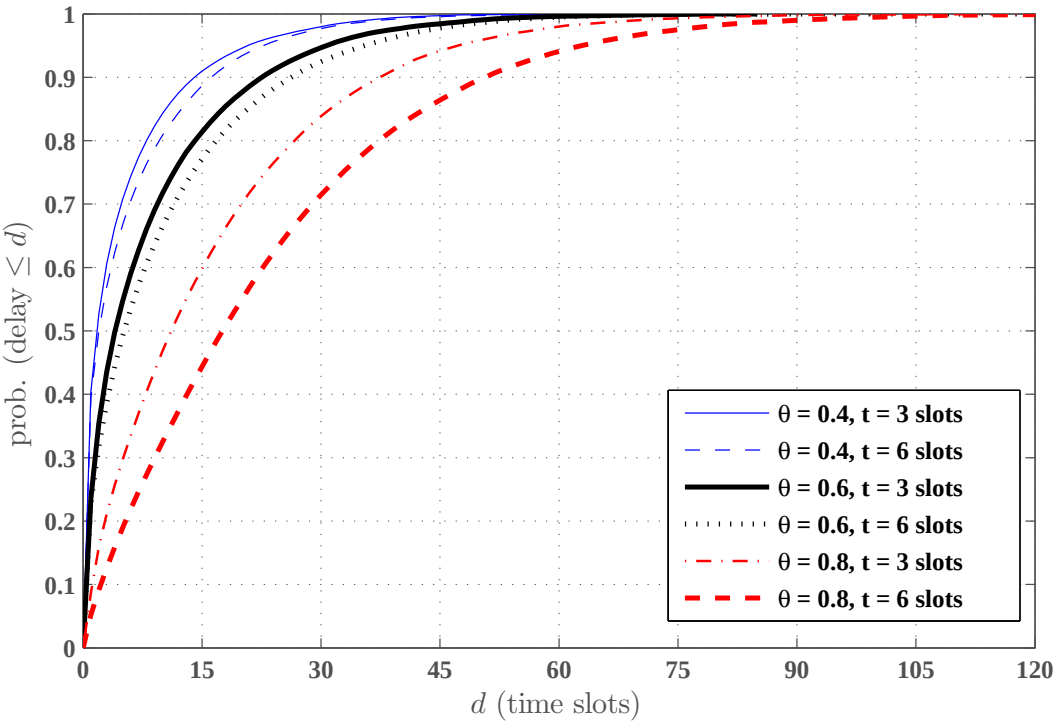


Figure 5.2: Effect of primary user activity level on packet delay distribution of CR users.

θ. There is also a positive correlation between packet loss rate and the average occupancy time. However, the effect is much less significant than that of the activity factor.

5.4.2 Number of Channels Available for CR Users

Figs. 5.4-5.5 show the delay and packet loss performances, respectively, with varying number of available channels. As expected, given the primary users' activity level and other system parameters remain unchanged, both delay and packet loss performance improve with increasing number of channels. The effect on packet loss rate, however, diminishes gradually once the number of channels becomes high enough to provide adequate radio resource to serve the waiting packets without causing any buffer overflow.

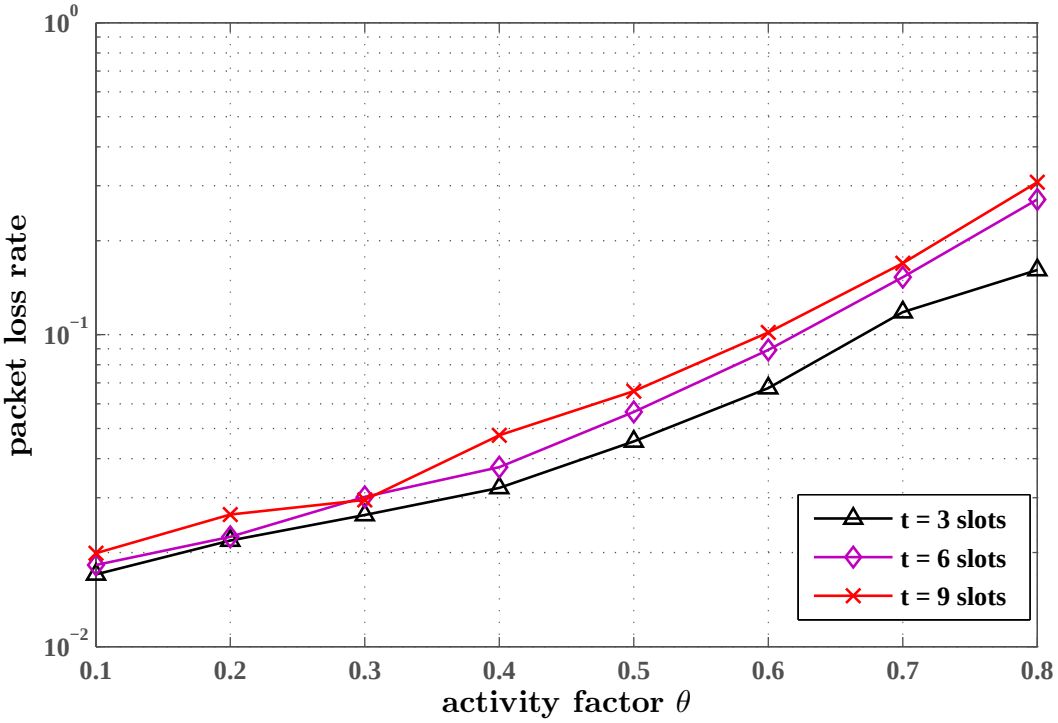


Figure 5.3: Effect of primary user activity level on packet loss rate of CR users.

5.4.3 Channel Fading

The analytical model enables use to demonstrate the effect of channel fading on MAC layer performance. Fig. 5.6 and Table 5.1 show how the delay distribution and packet loss rate vary with Doppler spread and average SNR. At higher Doppler spread, the channel state for a CR user will change more rapidly over time. This phenomenon in conjunction with the scheduling scheme will make it less likely that a particular user can monopolize the channel for a long time. The tagged user will thus have more probability of access to the channel. Consequently, it will have more opportunities to reduce its queue backlog. A reduced backlog implies smaller delay for the incoming packets. Moreover, with the less probability of queue size reaching the buffer limit, the packet loss rate drops, and subsequently, the average throughput improves.

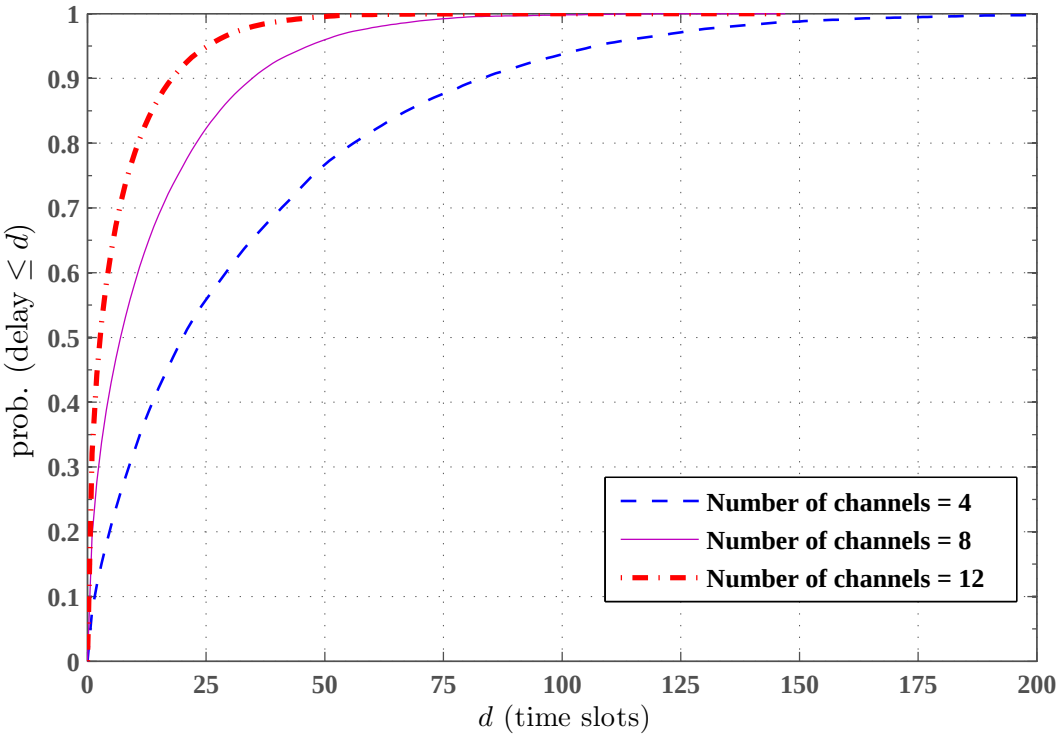


Figure 5.4: Effect of number of cognitive channels on packet delay distribution of CR users.

Table 5.1: Effect of channel fading on the packet loss rate of CR users.

Doppler spread	Average SNR	Packet loss rate
15 Hz	10 dB	12.3%
15 Hz	15 dB	4.7%
30 Hz	10 dB	9%
30 Hz	15 dB	1.9%

The effect of the Doppler spread in the CR network, however, is less profound than its expected effect in a traditional (i.e. not cognitive) cellular network that uses multiuser diversity for opportunistic scheduling. This observation can be explained as follows. The primary user occupancy in a channel that is currently assigned to a CR user makes the channel unavailable to the CR user for some time slots. As a result, the correlation in

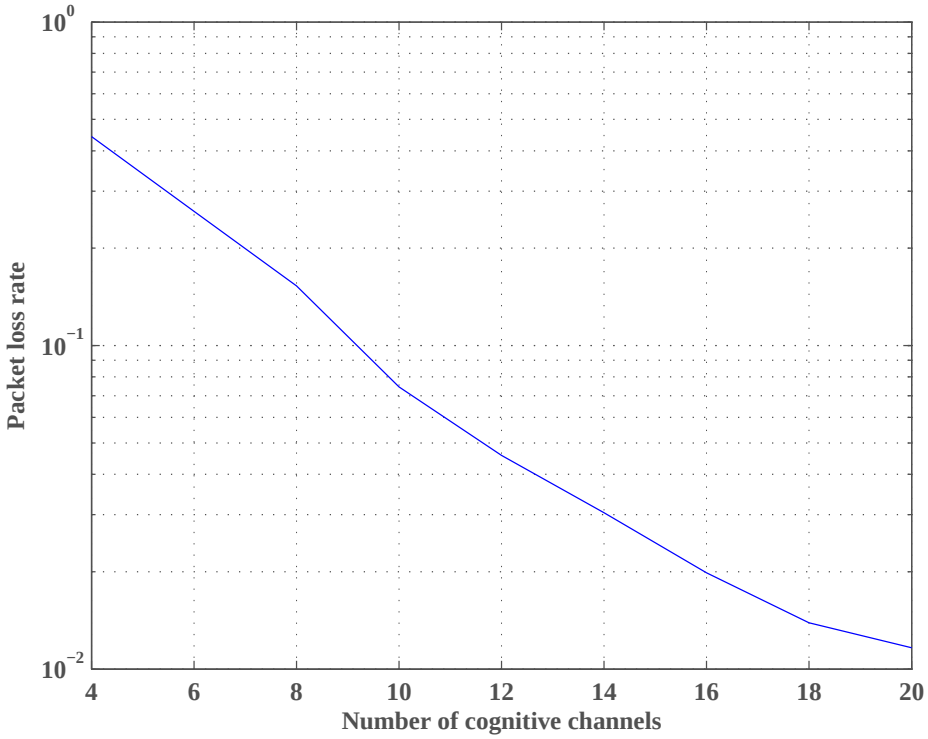


Figure 5.5: Effect of number of cognitive channels on packet loss rate of CR users.

the channel fading diminishes by the time the channel is assigned back to the same CR user. The fading correlation remains effective for the CR user as long as the channel is not reoccupied by some primary user. The average SNR of the homogenous channels has more direct effect on the delay and packet loss performances, both of which improve with increasing SNR.

5.4.4 Number of Admitted CR Users

The queueing model also allows us to investigate quantitatively the effects of the number of admitted users in the CR network on the QoS experienced by our tagged user. Fig. 5.7 shows how the delay distribution varies with different number of users admitted to the sys-

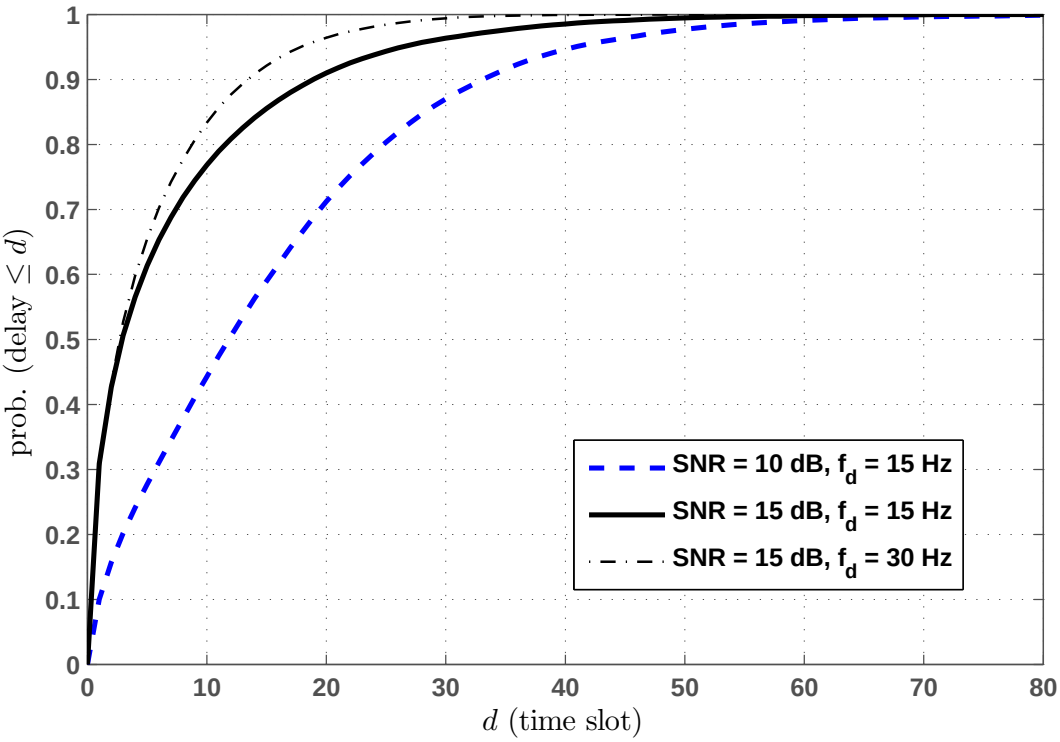


Figure 5.6: Effect of channel fading on the packet delay distribution of CR users

tem, given the system and traffic parameters stated earlier in the section. As expected, more users in the system create more competition, and therefore, longer packet delays are experienced. Similar observations can be made for the packet loss rate as shown in Fig. 5.8. Such relationships, however, can be drawn using our model for different values of system parameters such as primary user occupancy, number of available CR channels, channel fading, the characteristics of the packet arrival process of the applications requesting service. This can be very useful in deciding how many users could be supported for a given system condition and traffic parameters.

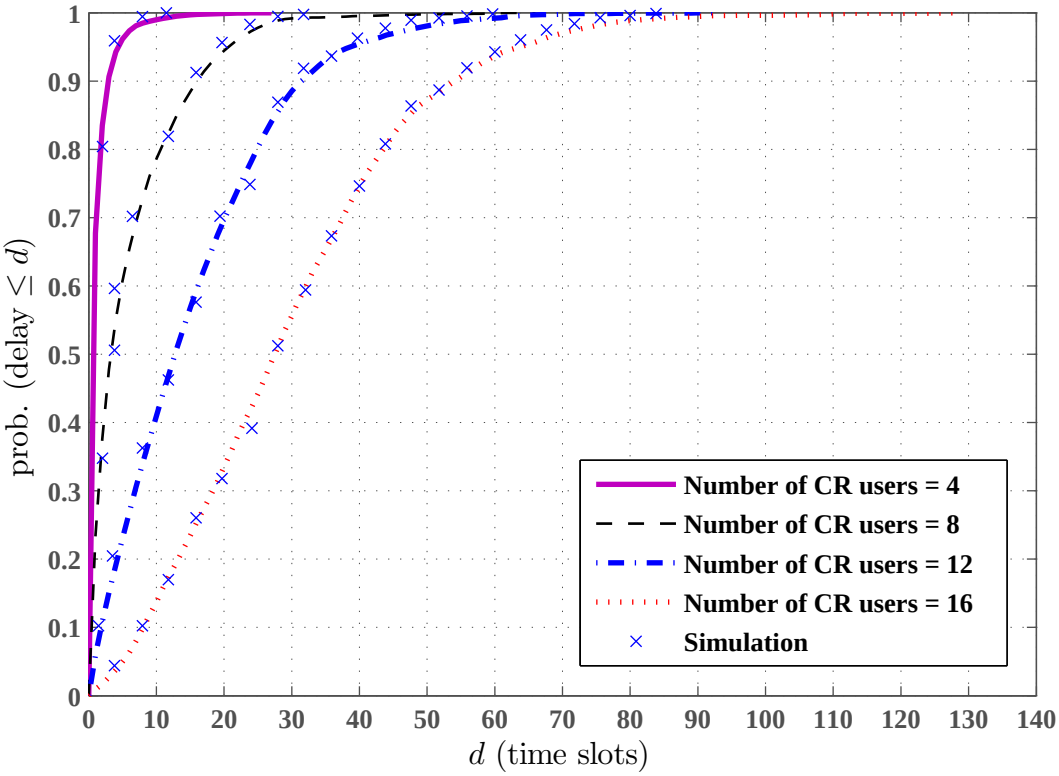


Figure 5.7: Packet delay distribution for different number of CR users in the cognitive radio network.

5.4.5 Queueing Model Validation

To validate our analytical model, we developed a discrete-event simulator using Matlab. At each time slot, the availability and channel state for each cognitive channel are generated randomly according to a channel occupancy matrix and FSMC transition probability matrix respectively. These matrices and the parameters for packet arrival process at each user queue are selected to be same as those used in obtaining the numerical results. The AMC scheme used to link channel states with the maximum number of packet transmission over a channel during a time slot is also selected to be identical to the one adopted in the analytical model. For the purpose of comparison, we have included two representative sets of

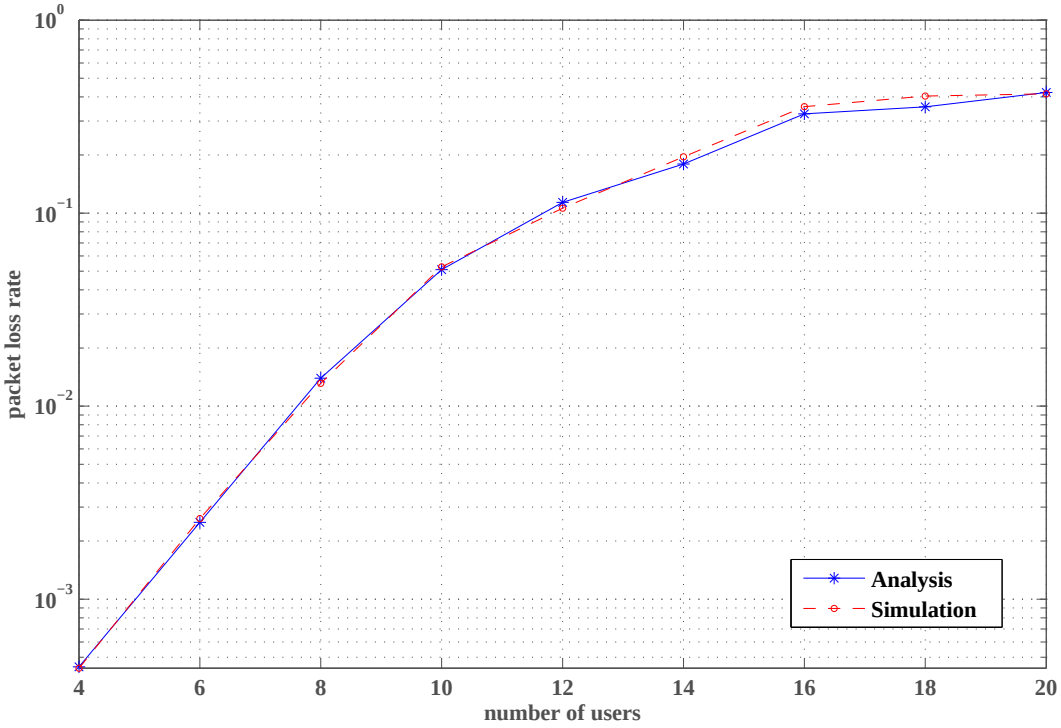


Figure 5.8: Packet loss rate for different number of CR users in the cognitive radio network.

simulation results in Fig. 5.7 and 5.8 beside the analytical results for a tagged user. Clearly, the presented simulation results match very closely with the corresponding analytical findings. We are, therefore, very confident about the validity of the analytical model presented in this chapter.

5.5 Application of the Queueing Model

The queueing model has important applications for system engineering in CR networks. Two example applications of the model are as follows:

- The model allows to gauge beforehand the level of QoS the system can entertain for a new incoming connection given the resource constraints and operating environments.

Therefore, the model output can also be used by the cognitive radio BS in QoS negotiation for new connections arriving from the CR users. The QoS negotiation usually involves a counter offer by the BS in response to the QoS level demanded by an incoming connection request. The BS can make use of the output from the queueing model to furnish the counter offer.

- The queueing model can be utilized in a model-based admission controller design. When a CR user wants to establish a new connection for an application or service, the admission controller at the BS can be assisted by the output from the queueing model to make an admission decision. The model output can determine whether the required QoS of the new and existing users can be met given the number of active users in the system. If the QoS of the requested and existing users cannot be met, the connection request may be refused, or can be accepted otherwise. For example, in Fig. 5.8, with the given parameter values used to find the relationship between packet loss rate and the number of users in the system, only 10 CR users can be served with a target packet loss rate of 5%. If the system is already serving 10 active users at some point in time, the admission controller could refuse admission of a new connection on the basis of the achievable packet loss rate determined by the queueing model.

5.6 Conclusions

In this chapter, we have developed a queueing analytic framework to study the performance of opportunistic spectrum access by CR users in a dynamic spectrum access network. We have derived important QoS measures by modeling primary users' activity as a two state Markov chain and CR users' channel quality variation as a finite state Markov chain. In order to allocate available spectrum opportunities among the CR users, a channel-quality-

based opportunistic scheduling scheme has been considered. The proposed framework has important applications including cross-layer analysis and design for system engineering, and model-based admission control in CR-based broadband wireless access networks. A number of practical issues could be considered in extending the model presented in this chapter. Examples of such issues include more practical modeling of primary user activity, considering heterogeneous channel characteristics across users, and modeling for different multiuser diversity and channel access scheduling schemes.

5.7 Bibliography

[1] R. W. Broderson, A. Wolisz, D. Cabric, S. M. Mishra, and D. Willkomm, "CORVUS: A cognitive radio approach for usage of virtual unlicensed spectrum," *white paper* submitted at the University of Berkeley, CA, Jul. 2004.

[2] Federal Communications Commission, "Spectrum Policy Task Force," Rep. ET Docket no. 02-135, Nov. 2002.

[3] J. Mitola, "The software radio architecture," *IEEE Communications Magazine*, vol. 33, pp. 26-38, May 1995.

[4] N. Devroye, P. Mitran and V. Tarokh, "Achievable rates in cognitive radio," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 1813-1827, May 2006.

[5] S. A. Jafar and S. Srinivasa, "Capacity Limits of Cognitive Radio with Distributed and Dynamic Spectral Activity," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 529-537, Apr. 2007.

[6] S. Srinivasa and S. A. Jafar, "The Throughput Potential of Cognitive Radio: A Theoretical Perspective," *IEEE Communications Magazine*, vol. 45, pp. 73-79, May 2007.

[7] Z. Ji and K. J. R. Liu, "Dynamic Spectrum Sharing: A Game Theoretical Overview," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 88-94, May 2007.

[8] W.-Y. Lee and I. F. Akyildiz, "Optimal Spectrum Sensing Framework for Cognitive Radio," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3845-3857, Oct. 2008.

[9] H. Jiang, L. Lai, R. Fan, and H. V. Poor, "Optimal Selection of Channel Sensing Order in Cognitive Radio," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 297–307, Jan. 2009.

[10] L. B. Le and E. Hossain, "OSA-MAC: A Multi-Channel MAC Protocol for Opportunistic Spectrum Access in Cognitive Wireless Networks," *Proc. IEEE Wireless Communications and Networking Conference*, Las Vegas, Nevada, USA, Apr. 2008.

[11] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589-600, Apr. 2007.

[12] J. Jia, Q. Zhang, and X. Shen, "HC-MAC: A Hardware-Constrained Cognitive MAC for Efficeint Spectrum Management," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 106-117, Jan. 2008.

[13] H. Su and X. Zhang, "Cross-Layer Based Opportunistic MAC Protocols for QoS Provisioning over Cognitive Radio Wireless Networks," *IEEE Journal Selected Areas in Communications*, vol. 26, no. 1, pp. 118-129, Jan. 2008.

[14] C. Doerr, M. Neufeld, J. Fifield, T. Weingart, D. C. Sicker, and D. Grunwald, "MultiMAC: An Adaptive MAC Framework for Dynamic Radio Networking," *Proc. IEEE International Dynamic Spectrum Access Networks*, Baltimore, MD, USA, Nov. 2005.

[15] O. Simeone, Y. Bar-Ness and U. Spagnolini, "Stable Throughput of Cognitive Radios with and Without Relaying Capability," *IEEE Transactions on Communications*, vol. 55, no. 12, pp. 2351-2360, Dec. 2007.

[16] S. Shankar, "Squeezing The Most out of Cognitive Radio: A Joint MAC/PHY Perspective," *Proc. IEEE ICASSP'07*, Honolulu, Hawaii, USA, Apr. 2007.

[17] L.-C. Wang, Y.-C. Lu, C.-W. Wang, and D. S. L. Wei, "Latency Analysis for Dynamic Spectrum Access in Cognitive Radio: Dedicated or Embedded Control Channel?," *Proc. Personal, Indoor and Mobile Radio Communications Symposium* , Athens, Greece, Sept. 2007.

[18] S. Keshavamurthy and K. Chandra, "Multiplexing Analysis for Dynamic Spectrum Access," *Proc. IEEE* , Washington, DC, USA, Oct. 2006.

[19] S. Tang and B. L. Mark, "Performance Analysis of a Wireless Network with Opportunistic Spectrum Sharing," *Proc. IEEE Global Communications Conference*, Washington, DC, USA, Nov. 2007.

[20] D. Niyato and E. Hossain, "Medium Access Control Protocols for Dynamic Spectrum Access in Cognitive Radio Networks: A Survey," invited chapter in *Cognitive Radio Networks*, (Eds. Y. Xiao and F. Hu), Auerbach Publications, CRC Press, 2008.

[21] N. Baldo and M. Zorzi, "Fuzzy Logic for Cross-Layer Optimization in Cognitive Radio Networks," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 64-71, Apr. 2008.

[22] X. Liu, E.K.P. Chong, and N.B. Shroff, "A Framework for Opportunistic Scheduling in Wireless Networks," *Computer Networks*, vol. 41, no. 4, pp. 451474, Mar. 2003.

[23] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with Adaptive Modulation and Coding over Wireless Link: Cross-Layer Analysis and Design," *IEEE Transactions on Wireless Communications*, vol. 4 , pp. 1142-1153, May 2005.

[24] Z. Han and H. Jiang, "Replacement of Spectrum Sensing and Avoidance of Hidden Terminal for Cognitive Radio," *Proc. IEEE Wireless Communications and Networking Conference*, Las vegas, NV, USA, Mar.-Apr. 2008.

[25] M. Nakagami, "The m-Distribution–A General Formula of Intensity Distribution of Rapid Fading," *Statistical Methods in Radio Wave Propagation*, Oxford, UK: Pergamon, pp. 3-36, 1960.

[26] H. S. Wang and N. Moayeri, "Finite-State Markov channel-A Useful Model for Radio Communication Channels," *IEEE Transactions on Vehicular Technology*, vol. 44, pp. 163-171, Feb. 1995.

[27] R. Knopp and P. A. Humblet "INformation Capacity and Power Control in Single-Cell Multiuser Communications," *Proc. IEEE International Conference on Communications*, Seattle, WA, USA, Jun. 1995.

[28] T. Issariyakul and E. Hossain, "ORCA-MRT: An Optimization-Based Approach for Fair Scheduling in Multi-Rate Tdma Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2823-2835, Nov. 2005.

[29] M. F. Neuts, "Matrix Geometric Solutions in Stochastic Models–An Algorithmic Approach," John Hopkins Univ. Press, Baltimore, MD, 1981.

Chapter 6

Conclusions and Future Research Directions

6.1 Conclusions

In this thesis, we have developed a number of useful analytical models for resource allocation schemes in emerging broadband wireless access networks. The models not only provide insight into the intricate relationship between the system as well as traffic parameters and the achievable QoS at user ends, but also provide framework for new resource allocation solutions for these networks. In particular, we have made following four major contributions in this thesis.

First, we have proposed a novel model-based resource allocation framework for downlink service flows in IEEE 802.16e mobile WiMAX systems. At first, we have developed a queueing analytical model that relates important performance measures of an admitted service flow to its packet arrival statistics and a parameter in the packet scheduling scheme. Referring to this parameter as the model parameter for the service flow, we have demon-

strated how the output from the queueing model can be utilized to set the appropriate value for this model parameter. This value of the model parameter essentially allows to compute the minimum number of packets that need to be scheduled from the queue of the service flow in order to meet its QoS requirement. Each admitted service flow is associated with a model parameter. Next, we have proposed a cross-layer adaptive and queue-aware resource allocation scheme that utilizes these parameter values along with the channel quality and backlog information to schedule packets and allocate OFDMA slots among downlink QoS connections. The proposed model-based approach offers simplicity of design for the on-line component, flexibility to the service providers in defining QoS parameters and better compatibility to the IEEE 802.16e standard. Instead of complexities involved with installing separate scheduling schemes within the same resource allocation framework for real-time and non-real-time traffic, our model-based approach offers an efficient unified way to handle both types of traffics and their diverse QoS requirements. We have run simulation experiments to evaluate the performance of our proposed scheme in providing QoS for VBR real-time and non-real-time multimedia and data services. Compared to EDF scheduling, which is specialized for allocating resource to delay-sensitive applications, our proposed scheme significantly reduces delay violation probability and packet loss rate for real-time services. As a result, the number of real-time service flows that can be supported by the system is also improved considerably. For non-real-time services, our proposed scheme achieves much better packet loss rate performance and can support significantly more service flows compared to throughput conscious WDRR scheduling. We have also shown that, even without priority-based or complex hierarchical approach, our proposed scheme can successfully protect QoS of individual service flows in a heterogenous traffic scenario involving both real-time and non-real-time services. The work has been submitted

for publication [1].

Second, we have developed a queuing analytic model for cross-layer analysis of downlink transmissions in a V-BLAST MIMO system exploiting multiuser diversity. The MIMO channel in the V-BLAST system is decomposed into subchannels associated with the transmit antennas and the users are then allocated the transmit antennas opportunistically based on the received SNR on the associated subchannels. The queuing model is developed considering the system dynamics resulting from the scheduling mechanism, fading in the MIMO channel, and the traffic arrival process. The transmit antenna correlation in the base station and the possibility of bursty traffic arrival are also factored into the model. After developing the queuing model, we elaborate the steps to compute important QoS performance measures including the delay distribution, throughput, and packet loss performance at the MAC layer. We also present selected numerical results that show cross-layer effects of the physical layer parameters as well as the effects of the traffic arrival statistics and the number of users in the system on the MAC layer performance. The queuing model has important applications in the QoS provisioning framework for a V-BLAST MIMO system as the model output can be used in admission control, QoS negotiation and setting proper values for QoS relevant parameters. This work has been published in [2].

Third, we have formulated an analytical model for the controlled channel access mechanism (also called HCCA) in IEEE 802.11e based WLANs considering VBR traffic applications. The model enables us to derive important performance measures in terms of the parameters used in the reference scheduler. The results obtained from the analytical model have shown that the QoS guarantees for VBR traffic cannot be met without adapting the transmission opportunities properly. Using the insights gained into the cause of the problems in the reference scheduler, we have developed a novel HCCA scheduler for IEEE

802.11e WLANs. The proposed scheduler adapts dynamically with dynamic traffic intensities that may arise from multimedia applications. The proposed PRO-HCCA scheduler utilizes simple but smart prediction mechanism along with an efficient online optimization process. Simulation experiments have shown that the PRO-HCCA scheduler achieves extremely robust and stable performance in meeting the QoS guarantees for VBR traffic. For related publications to this work, please refer to [3, 4, 5].

In the final contribution of this thesis, we expanded our analytical modeling to cognitive radio technology that has the potential to significantly increase spectrum utilization in wireless broadband access networks. The cross-layer analytic model captures dynamic resource availability due to opportunistic access of radio spectrum by cognitive users in presence of primary users activity in the spectrum. It allows to relate channel characteristics, traffic parameters and other important system parameters with the achievable QoS performance for cognitive users. Later, we show a number of application of the model including cross-layer system engineering and model-based admission control. For publications related to this work, please refer to [6, 7].

6.2 Future Research

The analytical models and resource allocation solutions developed in this thesis not only have far reaching important applications for system designers of emerging broadband wireless services, but also open up interesting possibilities of future research. In the following, we list some important future research directions that can follow from this thesis.

6.2.1 Models for End-to-End QoS in Multihop Wireless Broadband Networks

In this thesis, we focused on single-hop wireless broadband systems; a natural extension would be to consider multihop architectures. An interesting future research will be to extend our analytical models to wireless mesh networks [8], which can form broadband wireless access with radio nodes organized in a mesh topology. Such networks can be built with various broadband wireless technology including 802.11, 802.16, cellular or combinations of more than one type. Another very widespread scenario of multihop broadbnad wireless access networks would be where WMAN will be used to bring broadband service to a premise where WLAN will be the technology used for distributing the access across the premise. In such systems, a multihop modeling for end-to-end QoS will be required. One idea would be to use tandem queue modeling in which output process from the queue representing one hop is modeled as the input process to the queue representing the next hop in the route connecting the communicating nodes [9]. Besides providing important inputs in designing resource allocation schemes for these multihop systems, these models can be useful in assessing the performance and selecting paramer values of the interoperation modules in heterogeneous systems. Besides, the models can be utilized to design novel model–based schemes to map QoS parameters between one part of the multihop network to the other part that implementing a different broadband wireless technology. The ultimate goal will be devise better resource managment schemes for providing end-to-end QoS in multihop broadband wireless access networks.

6.2.2 Cross-Layer Models for Various Multiuser MIMO Systems

In Chapter 3, we analyzed a multiuser V-BLAST MIMO system with the assumption of a zero-forcing receiver, that offered several benefits in terms of performance in multiuser schduling as well as in our analytical modeling. As an extension to this work, we plan to consider other receiver schemes such as MMSE and SIC-based detection in our cross-layer analytical model. One of the challenges will be the fact that there is no closed-form expression available for post detection SNR for these receiver schemes. We will investigate how we can address this issue in our modeling. In the transmitter side, it will also be interesting to explore the effect of power control instead of adaptive modulation and coding. This is because power control is more widely employed in many practical wireless systems. Aside from these, we plan to explore other multiuser scheduling and antenna assignment schemes than the independent stream scheduler assumed in the analysis done in Chapter 3. Besides the V-BLAST system considered in our work, there is opportunity for analyzing other multiuser MIMO systems employing various transmitter and receiver techniques. Therefore, the work presented in this thesis on V-BLAST MIMO systems may serve as a good starting point for many exciting avenues for future research.

6.2.3 Analyzing Effects of CQI Feedback Issues

In this thesis, we have assumed perfect CQI and zero feedback delay in the queueing models. Such assumptions have very limited impact in slow fading and/or slow mobility scenarios. However, the effects of imperfect CQI and feedback delay may become pronounced in high mobility or fast fading scenarios [10] and need to be considered in the model. To minimize potential negative effects, many feedback mechanism and link adaptation schemes have been developed [11]. An exciting future research will be to analyze the cross-layer

effects of imperfect CQI and delayed feedback on these schemes and to use the findings to improve their designs as well as to develop better schemes. Another interesting research issue to consider would be to analyze the effects of CQI feedback overhead on potential throughput degradation in different broadband wireless access networks and to utilize the findings to design low-overhead CQI feedback mechanisms.

6.2.4 Enhanced Queueing and Game Theoretical Models for Spectrum Sharing in CR-based Wireless Broadband Systems

Sharing of spectrum resources by primary user networks with cognitive users is often dictated by perceived utility from such sharing. From a cognitive users' perspective, the more spectrum that it can opportunistically access, the more it has the potential to improve the QoS of its applications. However, for primary user networks, although allowing more access might mean more potential revenue (if cognitive access is given for a price), increased interference to primary users might lead to user dissatisfaction and lost revenue. In such situations, where the primary user network and cognitive users are not likely to be centrally controlled by a single entity, game theoretic models can be useful to devise new spectrum sharing strategies. In devising these strategies, the queueing model developed in this thesis could be useful at the CR network end as the derivation of perceived utility can utilize the model output on performance measures. In this thesis, however, we developed the queueing model from the perspective of cognitive users only. Another avenue of research would be to develop queuing models from PU network perspective, where the effect of spectrum sharing with secondary users can be analyzed given certain access behaviour of cognitive users. These models, in turn, can be used in game theoretic analysis.

6.3 Bibliography

[1] M. M. Rashid and V. K. Bhargava, "A Model-Based Downlink Resource Allocation Scheme for IEEE 802.16e Mobile WiMAX," Submitted to *IEEE Transactions on Vehicular Technology*, Nov. 2009.

[2] M. M. Rashid, E. Hossain, and V. K. Bhargava, "Cross-Layer Analysis of Downlink V-BLAST MIMO Transmission Exploiting Multiuser Diversity," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4568–4579, Sep. 2009.

[3] M. M. Rashid, E. Hossain, and V. K. Bhargava, "Controlled Channel Access Scheduling for Guaranteed QoS in 802.11e-Based WLANs," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1287-1297, Apr. 2008.

[4] M. M. Rashid, E. Hossain, and V. K. Bhargava, "HCCA Scheduler Design for Guaranteed QoS in 802.11e Based WLANs," *Proc. IEEE International Wireless Communications and Networking Conference*, Hong Kong, Mar. 2007.

[5] M. M. Rashid, E. Hossain, and V. K. Bhargava, "Queueing Analysis of 802.11e HCCA with Variable Bit Rate Traffic," *Proc. IEEE International Conference on Communications*, Istanbul, Turkey, Jun. 2006.

[6] M. M. Rashid, J. Hossain, E. Hossain, and V. K. Bhargava, "Opportunistic Spectrum Scheduling for Multiuser Cognitive Radio: A Queueing Analysis," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 5259-5269, Oct. 2009.

[7] M. M. Rashid, J. Hossain, E. Hossain, and V. K. Bhargava, "Opportunistic Spectrum Access in Cognitive Radio Networks: A Queueing Analytic Model and Admis-

sion Controller Design," *Proc. International Global Telecommunications Conference*, Washington, DC, USA, Nov. 2007.

[8] I. Akyildiz and X.Wang, "A Survey on Wireless Mesh Networks," *IEEE Communications Magazine*, vol. 43, no. 9, pp. s23-s30, Sep. 2005.

[9] L. Le and E. Hossain, "Tandem Queue Models with Applications to QoS Routing in Multihop Wireless Networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 8, pp. 1025–1040, Aug. 2008.

[10] D. Tse and P. Viswanath, *Fundamental of Wireless Communications*, Cambridge University Press, 2005.

[11] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B.D. Rao, and M. Andrews, "An Overview of Limited Feedback in Wireless Communication Systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341-1365, Oct. 2008