Radio Resource Management in Broadband Wireless Access Networks

by

Dusit Niyato
M.Sc., University of Manitoba, 2005

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical and Computer Engineering

We accept this thesis as conforming
to the required standard

Professor E. Hossain, Supervisor, Department of Electrical & Computer Engineering

---

Professor P. Yahampath, Department of Electrical & Computer Engineering

---

Professor S. Noghanian, Department of Electrical & Computer Engineering

---

Professor Y. E. Liu, Outside Member, Department of Computer Science

---

Professor E. Elmallah, External Examiner

Radio Resource Management in Broadband Access Wireless Networks

BY

Dusit Niyato

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

Doctor of Philosophy

Supervisor: Professor E. Hossain

ABSTRACT

Broadband wireless access (BWA) technology such as IEEE 802.16-based WiMAX (Worldwide Interoperability for Microwave Access) systems, IEEE 802.11-based wireless local area networks (WLANs), i.e., WiFi networks, and beyond 3G cellular systems are being developed to provide high speed wireless connectivity and seamless mobility to users. Integration of these different technologies will give rise to a heterogeneous wireless access environment. Although the WiMAX standard defines the signaling messages for medium access control (MAC) mechanisms, radio resource management protocols for dynamic bandwidth allocation, subcarrier allocation, connection admission control and many other aspects are left open for innovations. Also, issues related to an integration of WiMAX networks with 3G and WiFi systems need to be resolved. Efficient protocol engineering, which is the theme of the research results presented in this report, would be critical for cost-effective deployment and operation of BWA technologies. In this research, resource management protocols are designed and optimized for WiMAX broadband networks and integrated WiMAX-WiFi networks.

The problem of radio resource management for WiMAX networks is considered at both subscriber stations (SS) and base stations (BS). Specifically, queue-aware bandwidth allocation and rate control mechanisms are proposed for WiMAX subscriber stations. While bandwidth allocation is used to allocate radio resource at the SS, rate control is used to limit the transmission rate of the traffic source to maintain the target quality of service (QoS) performance. A queueing analytical model is proposed to investigate the performance of these bandwidth allocation and rate control mechanisms. Afterwards, the resource management problem at the WiMAX base station is addressed. A queueing mode is formulated to obtain the QoS performance measures which are used by the bandwidth allocation algorithm at the WiMAX BS to allocate available bandwidth among the different connections. Two bandwidth allocation algorithms, namely, the optimal and the iterative algorithms, are proposed. While the optimal algorithm provides the best solution of resource allocation, the iterative algorithm incurs much less computational overhead.

A radio resource allocation framework is proposed for an integrated WiMAX-WiFi network where the WiMAX network serves as a multihop backhaul network for relaying Internet traffic to/from WiFi networks. For such a network, the problem of bandwidth allocation among local and relay traffic at a WiMAX base station (i.e., mesh router) is considered. Then, the resource allocation problem in an integrated WiMAX-cellular-WiFi network is considered where a mobile user is able to connect to the different access networks simultaneously. For such a network, the resource management problem is solved considering a cooperative environment where all available networks offer bandwidth to users to satisfy their QoS requirements. Then, this problem is solved considering a noncooperative environment where all networks are operated by different rational service providers.

Examiners:

Professor E. Hossain, Supervisor, Department of Electrical & Computer Engineering

---

Professor P. Yahampath, Department of Electrical & Computer Engineering

---

Professor S. Noghanian, Department of Electrical & Computer Engineering

---

Professor Y. E. Liu, Outside Member, Department of Computer Science

---

Professor E. Elmallah, External Examiner

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AMC	Adaptive Modulation and Coding
AP	Access Point
ARQ	Automatic repeat-ReQuest
ATL	Adaptive Transport Layer
BA	Bandwidth Allocation
BE	Best-Effort
BMAP	Batch Markovian Arrival Process
BRAN	Broadband Radio Access Networks
BS	Base Station
BWA	Broadband Wireless Access
CAC	Connection/Call Admission Control
CDMA	Code Division Multiple Access
CID	Connection Identifier
CP	Complete Partition
CS	Complete Sharing
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTMC	Continuous-Time Markov Chain
DCF	Distributed Coordination Function
DL-MAP	Downlink Map Messages
dMMPP	Discrete Time MMPP
DSL	Digital Subscriber Line
DSSS	Direct Sequence Spread-Spectrum
EBA	Early Backoff Announcement
EMA	Exponential Moving Average
ESRA	Enhanced Staggered Resource Allocation
ETSI	European Telecommunications Standards Institute
ETX	Expected Transmission Count
FDD	Frequency-Division Duplex
FSMC	Finite State Markov Chain
GPC	Grant Per Connection
GPSS	Grant-Per-Subscriber Station
HiperMAN	High Performance Radio Metropolitan Area Network
IEEE	Institute of Electrical and Electronics Engineers

MAC	Medium Access Control
MANET	Mobile Ad Hoc Networks
MIMO	Multiple-Input Multiple-Output
MMPP	Markov Modulated Poisson Process
MSS	Mobile Subscriber Stations
NLOS	Non-Line-Of-Sight
nrtPS	Non-Real-Time Polling Service
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal-Frequency Division Multiple Access
PCF	Point Coordination Function
PDU	Protocol Data Unit
PER	PDU Error Rate
PS	Polling Service
QoS	Quality of Service
RC-EDF	Enhanced Staggered Resource Allocation
RED	Random Early Detection
RRM	Radio Resource Management
RS	Reed-Solomon
rtPS	Real-Time Polling Service
RTT	Round-Trip Time
SIR	Signal-to-Interference Ratio
SNR	Signal-to-Noise Ratio
SS	Subscriber Stations
TCP	Transmission Control Protocol
TDD	Time-Division Duplex
TDMA	Time-Division Multiple Access
TU	Transferable Utility
UGS	Unsolicite Granted Service
UL-MAP	Uplink Map Messages
UMTS	Universal Mobile Telecommunications System
VoIP	Voice over IP
WiFi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Networks
WMAN	Wireless Metropolitan Area Network
WPAN	wireless Personal Area Network
WRAN	Wireless Regional Area Network

Acknowledgement

First and foremost, I would like to express my profound gratitude and appreciation to my supervisor, Professor Ekram Hossain, for his continuous support and invaluable guidance. I am grateful to Professor Pradeepa Yahampath, Professor Sima Noghanian, and Professor Yanni Ellen Liu for being in the examination committee as well as Professor Ehab Elmallah for being the7 external examiner of my Ph.D. oral examination.

I would like to recognize Telecommunications Research Laboratory (TR*Labs*), Natural Sciences and Engineering Research Council of Canada (NSERC), and University of Manitoba for financial supports. I would like to thank people from TR*Labs*, Sergio Camorlinga, Jeff Diamond, Nicole Alexander, etc. for various supports.

I would like to thank all my friends in Winnipeg, Teerawat (Mee) Issariyakul, Wannasorn (Golf) Kruahongs, Wattamon (KooK) Srisakuldee, Taweewat (Nueng) and Sudarat (Ae) Deemagarn, etc. for their numerous help.

Last but not the least, I would like to thank all my family members, my mom Usa, my dad Tawat, and my brother Wittaya Niyato for their support.

Chapter 1

Introduction

1.1 Broadband Wireless Access Networks

Due to the growing demand of mobile services and applications, different wireless technologies are being developed to support different coverage, speed, and reliability requirements for wireless transmission (Fig. 1.1). The wireless technology with the smallest coverage area is referred to as the wireless personal area network (WPAN) technology which supports short-range transmission (e.g., ten meters). Wireless local area network (WLAN) technology, e.g., IEEE 802.11-based WiFi technology, is currently one of the most popular wireless technologies. The transmission range of a WLAN is about hundred meters, and it supports data rates from few Mbps to hundred Mbps. Wireless metropolitan area network (WMAN) and wireless wide area network (WWAN) technologies have a longer transmission range (e.g., up to ten kilometers). Two major WMAN and WWAN technologies are IEEE 802.16-based WiMAX and IEEE 802.20-based MobileFi technologies which are designed primarily to support stationary and mobile users, respectively. The wireless technology with the longest transmission range is wireless regional area network (WRAN) which is currently being designed to support wireless transmission within hundred kilometers. IEEE 802.22 is one of the standards which is being developed for WRANs.

Recently, broadband wireless access networking technology based on the IEEE 802.16 standard for WMAN environment has caught much attention. Also known as the WiMAX forum (World Interoperability for Microwave Access), the IEEE 802.16-based technology is a promising alternative for last mile access in crowded urban and metropolitan areas, and also in sub-urban areas where installation of cable-based infrastructure is economically or technically infeasible. With an evolution of wireless

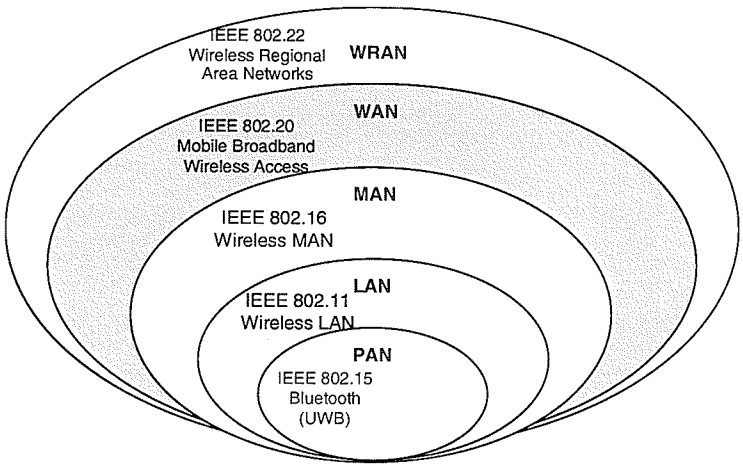


Figure 1.1. *Current wireless technologies.*

technology to support high speed connectivity and seamless mobility, WiMAX networks will be integrated with other wireless technologies such as the WiFi and 3G cellular technologies. This will give rise to a heterogeneous wireless access environment in which a user can access different wireless technologies simultaneously. Although the details of the physical and the medium access control (MAC) layers in the WiMAX standard are well defined, the radio resource management protocols (e.g., bandwidth allocation, scheduling, and admission control) remain as open problems. Protocol engineering (i.e., design, analysis, and optimization) for resource management is crucial to deliver wireless services to the users efficiently and reliably. Also, the radio resource management framework for an integrated/heterogeneous broadband wireless access network needs to be carefully designed and engineered to achieve the desired objectives of both the service providers and the users.

1.2 IEEE 802.16-Based Broadband Wireless Access Networks

IEEE 802.16 standard-based WiMAX networks are designed to provide high speed broadband wireless connectivity with quality of service (QoS) support. WiMAX is

an alternative to cable-based broadband access, e.g., digital subscriber line (DSL), for the areas where it is technically or economically infeasible to install the cable infrastructure. While the initial WiMAX standard was developed to support stationary users, the new standard (i.e., IEEE 802.16e-based mobile WiMAX) can support mobile users.

1.2.1 Deployment Scenarios

WiMAX technology intends to provide broadband connectivity to both fixed and mobile users in a wireless metropolitan area network (WMAN) environment. To provide flexibility for different applications, the standard supports two major deployment scenarios (Fig. 1.2) as follows:

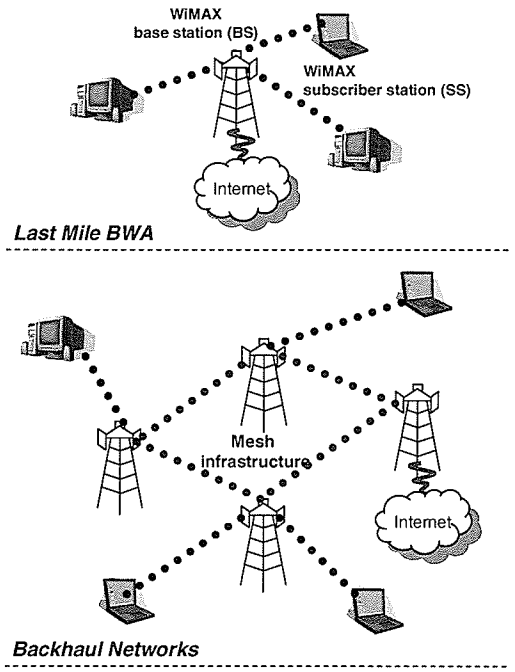


Figure 1.2. *WiMAX-based broadband wireless access networks.*

- *Last Mile BWA*: In this scenario, broadband wireless connectivity is provided to home and business users in a WMAN environment. The operation is based

on a point-to-multipoint single-hop transmission between a single base station (BS) and multiple subscriber stations (SSs).

- *Backhaul Networks*: This is a multihop (or mesh) scenario where a WiMAX network works as a backhaul for cellular networks to transport data/voice traffic from the cellular edge to the core network (Internet) through meshing among the WiMAX base stations.

1.2.2 Physical and MAC Layer Overview

1.2.2.1 Physical Layer

The physical layer of the WiMAX air interface operates at either a 10-66 GHz (i.e., IEEE 802.16) or a 2-11 GHz band (i.e., IEEE 802.16a), and it supports data rates in the range of 32-130 Mbps, depending on the operation bandwidth (e.g., 20, 25, or 28 MHz) as well as the modulation and coding schemes used. The IEEE 802.16 standard specifies different air interfaces for different frequency bands. In the 10-66 GHz band, the signal propagation between a BS and a subscriber station (SS) should be line-of-sight (LOS) and the air interface for this band is Wireless-SC (single carrier). In the 2-11 GHz band, three different air interfaces supporting non-line-of-sight (NLOS) communication can be used as follows:

- *WirelessMAN-SCa* for single-carrier modulation.
- *WirelessMAN-OFDM* for OFDM-based (orthogonal frequency-division multiplexing) transmission using 256 subcarriers. For this air interface, the MAC scheme for the SSs is based on time-division multiple access (TDMA).
- *WirelessMAN-OFDMA* for OFDMA-based (orthogonal frequency division multiple access) transmission using 2048 subcarriers. The MAC algorithm is based on orthogonal-frequency division multiple access (OFDMA) in which different groups of subcarriers are assigned to different SSs.

The frame structure of IEEE 802.16 is shown in Fig. 1.3.

To enhance a data transmission rate, an adaptive modulation and coding (AMC) technique is supported in WiMAX. Since the quality of the wireless link between a BS and an SS depends on the channel fading and interference conditions, through AMC,

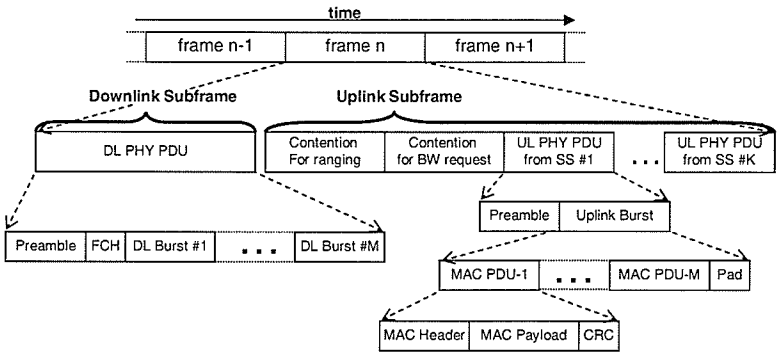


Figure 1.3. *IEEE 802.16 frame structure.*

Table 1.1. *Modulation and coding schemes for WiMAX.*

Rate ID	Modulation level (coding)	Information bits/symbol	Required SNR (dB)
0	BPSK (1/2)	0.5	6.4
1	QPSK (1/2)	1	9.4
2	QPSK (3/4)	1.5	11.2
3	16QAM (1/2)	2	16.4
4	16QAM (3/4)	3	18.2
5	64QAM (2/3)	4	22.7
6	64QAM (3/4)	4.5	24.4

the radio transceiver is able to adjust the transmission rate according to channel quality (i.e., signal-to-noise ratio (SNR) at the receiver). The Reed-Solomon (RS) code concatenated with an inner convolution code is used for error correction. However, advanced coding techniques (e.g., turbo codes and space time block codes) can be used as well.

Adaptive modulation and coding is used to adjust the transmission rate adaptively in each frame according to the channel quality. The modulation and coding schemes for the WiMAX air-interface are shown in Table 1.1.

IEEE 802.16d (802.16-2004) and IEEE 802.16e, which have evolved from 802.16a, use advanced physical layer techniques to support NLOS communication. IEEE 802.16e is specifically designed to support user mobility. The network model in this

standard has a single BS that serves mobile subscriber stations (MSSs) in the predefined coverage area. Both physical and MAC layers are enhanced to support IP mobility requirements. The IEEE 802.16g standard (under development) aims to support mobility at higher layers (transport and application) and across the backhaul network for multinetwork operation.

1.2.2.2 Medium Access Control (MAC) Layer

WiMAX uses a connection-oriented MAC protocol which provides a mechanism for the SSs to request bandwidth from the BS. Although each SS has a standard 48-bit MAC address, the main purpose of this address is for hardware identification. Therefore, a 16-bit connection identifier (CID) is used primarily to identify each connection to the BS. On the downlink, the BS broadcasts data to all SSs in the same network. Each SS processes only the MAC protocol data units (PDUs) containing its own CID and discards the other PDUs. WiMAX MAC supports the grant-per-SS (GPSS) mode of bandwidth allocation in which a portion of the available bandwidth is granted to each of the SSs and each SS is responsible for allocating the bandwidth among the corresponding connections.

WiMAX standard supports both frequency-division duplex (FDD) and time-division duplex (TDD) transmission modes. For TDD, a MAC frame is divided into uplink and downlink subframes. The lengths of these subframes are determined dynamically by the BS and broadcast to the SSs through downlink and uplink map messages (UL-MAP and DL-MAP) at the beginning of each frame. Therefore, each SS knows when and how long to receive and transmit data to the BS. In the uplink direction, a subframe also contains ranging information to identify an SS, information on the requested bandwidth, and data PDUs for each SS.

The MAC protocol in the standard supports dynamic bandwidth allocation. In this case, each SS can request bandwidth from the BS by using a BW-request PDU. There are two modes to transmit BW-request PDUs: contention mode and contention-free mode (e.g., polling). In the contention mode, an SS transmits BW-request PDUs during the contention period in a frame, and a backoff mechanism is used to resolve the contention among the BW-request PDUs from multiple SSs. In the contention-free mode, each SS is polled by the BS and after receiving the polling signal from

BS, an SS responds by sending the BW-request PDU. Due to predictable delay, the contention-free mode is suitable for QoS-sensitive applications. To provide access control and confidentiality in data transmission, WiMAX provides a full set of security features [1] which are implemented as a MAC sublayer functionalities.

In addition to the single-hop point-to-multipoint operation scenario, the IEEE 802.16a standard also defines signaling flows and message formats for multihop mesh networking among the BSs (i.e., infrastructure mesh). In this scenario, several BSs can communicate with each other, and an SS connects to the corresponding parent BS. Data traffic from an SS is transmitted through several BSs along the route in the mesh network to the destination BS or an Internet gateway.

1.2.3 QoS Framework and Service Types in WiMAX

WiMAX standard defines a QoS framework for different classes of services. The following three major types of services are supported, each of which has different QoS requirements [2]:

- *Unsolicited Grant Service (UGS)*: This service type supports constant-bit-rate (CBR) traffic. In this case, the BS allocates a fixed amount of bandwidth to each of the connections in a static manner, and therefore, delay and jitter can be minimized. UGS service is suitable for traffic with very strict QoS constraints for which delay and loss need to be minimized.
- *Polling Service (PS)*: This service supports traffic for which some level of QoS guarantee is required. It can be divided into two sub-types: real-time polling service (rtPS) and non-real-time polling service (nrtPS). The difference between these sub-types lies in the tightness of the QoS requirements (i.e., rtPS is more delay-sensitive than nrtPS). Not only delay-sensitive traffic but also non-real-time Internet traffic can use polling service to achieve a certain throughput guarantee. The amount of bandwidth required for this type of service is determined dynamically based on the required QoS performance and the dynamic traffic arrivals for the corresponding connections.
- *Best-Effort Service (BE)*: This is for traffic with no QoS guarantee (e.g, web and e-mail traffic). The amount of bandwidth allocated to BE service depends

on the bandwidth allocation policies for the other two types of service. In particular, the bandwidth left after serving UGS and PS traffic is allocated to BE service.

1.2.4 Mesh Mode of Operation

In addition to the single-hop point-to-multipoint operation scenario, the WiMAX standard (e.g., IEEE 802.16a) also defines signaling flows and message formats for the multihop mesh networking scenario among the subscriber stations (i.e., client meshing). Although meshing among the base stations (i.e., infrastructure meshing) has not been standardized yet, we envision that this will be adopted in the standard in the near future. In fact, Task Group 802.16j established by the WiMAX mobile multihop relay (MMR) study group is working on the standardizaton of relay-based (both fixed and mobile) infrastructure meshing [3]. Such an infrastructure mesh would be suitable as a wireless backhaul to serve IEEE 802.11-based WLAN hotspots.

In a WiMAX-based infrastructure mesh network, several base stations/mesh routers communicate with each other, and a subscriber station connects to a base station/mesh router. Data traffic from a subscriber station is transmitted through several base stations along a multihop route in the mesh network to the destination base station or an Internet gateway.

1.3 Other Broadband Wireless Access Technologies

1.3.1 WiBRO and HiperMAN

Korea Telecom developed the broadband wireless Internet technology, known as WiBRO, to operate in the licensed 2.3 GHz frequency band, which can support both fixed and mobile users. The channel bandwidth is 9 MHz which is used in a TDD mode. The MAC frame size is 5 ms and AMC is used to achieve an enhanced transmission rate. The MAC scheme is based on OFDMA and the QoS framework supports four service types as in the WiMAX standard. In Europe, the High Performance Radio Metropolitan Area Network (HiperMAN) standard was proposed by the

Table 1.2. *Comparison among 3G technologies.*

Technology	**WCDMA**	**CDMA2000**
Spectrum size	5 MHz	1.25 MHz
Data rate	384 Kbps/mobile	144 Kbps/mobile

Broadband Radio Access Networks (BRAN) group of the European Telecommunications Standards Institute (ETSI). HiperMAN is designed to operate in the 2-10 GHz (mainly in the 3.5 GHz) band. HiperMAN has a QoS framework, and supports AMC and dynamic power allocation for NLOS communications. Also, the mesh configuration is included in the standard. The WiBRO technology and HiperMAN standard are compatible with the IEEE 802.16a and 802.16-2004.

1.3.2 3G Networks

Third-generation (3G) wireless systems based on the code division multiple access (CDMA) technology support data rates of 384 Kbps for mobile users and 2 Mbps for stationary users. While 3G systems are designed primarily for mobile voice and data users, WiMAX systems are optimized to provide high-rate wireless connectivity for a large set of services and applications (e.g., with multimedia traffic), which require QoS guarantee. In addition, WiMAX systems can be used along with 3G wireless systems to provide QoS to the wireless Internet users in a cost-effective manner.

WiMAX networks can serve as backhauls for 3G networks [7]. Since such a network can provide high bandwidth with a large coverage area, 3G BSs can be easily and flexibly deployed to extend the cellular coverage area. However, one challenge here is efficient topology management aimed to minimize the network deployment cost while satisfying the QoS requirements for the cellular services. In [7], an optimal solution for designing WiMAX-based backhaul topology was presented. The problem was formulated as an integer programming problem, and a heuristic solution to this problem was presented. With this solution, the number of WiMAX links in the backhaul network can be reduced significantly compared to that for a ring topology.

Table 1.3. *Comparison among 3G, IEEE 802.16e, and IEEE 802.20 networks.*

	3G Networks	**WiMAX**	**MobileFi**
Objective	To provide voice and data services to mobile users	To provide BWA to fixed and mobile users	To provide mobile broadband connections to mobile users
Frequency band	2 GHz	2-10 GHz	3.5 GHz
Channel bandwidth	< 5 MHz	> 5 MHz	< 20 MHz
Transmission rate	Up to 10 Mbps (HSDPA from 3GPP)	10-50 Mbps	> 16 Mbps
Cell radius	Up to 20 km	Up to 50 km	-
Mobility	Full mobility functions (IP mobility, roaming, handoff, paging)	IP mobility	Full mobility functions and inter-technology handoff
Mobile speed	up to 120 km/hr	60 km/hr	Up to 250 km/hr
Multiple access	CDMA	TDMA or OFDMA	-
MAC Frame size	10 ms	< 10 ms	< 10 ms

1.3.3 IEEE 802.20/MobileFi

IEEE 802.20 (also called the MobileFi standard) is being designed specifically for mobile BWA (MBWA) services. The transmission range is 50 kilometers. This standard will be optimized to provide IP services for fixed and mobile users. IEEE 802.20 will operate in the licensed bands below 3.5 GHz and provide data transmission rates over 20 Mbps for a user speed up to 250 km/hour. OFDM is used in the physical layer and transmission can be non-line-of-sight.

Comparison among 3G networks, IEEE 802.16e-based WiMAX networks, and IEEE 802.20-based MobileFi networks is shown in Table 1.3 [4], [5], [6]. Specifically designed for mobile users, IEEE 802.16e can be an alternative to 3G cellular networks while IEEE 802.20 is being developed.

1.3.4 IEEE 802.11/WiFi WLAN

The WiFi technology based on the IEEE 802.11 standards has become very popular recently. There are several IEEE 802.11 standards which use different frequencies and

Table 1.4. *Comparison among IEEE 802.11 standards.*

Standard	Frequency	Data rate	Throughput
802.11	2.4 GHz	2 Mbps	0.7 Mbps
802.11a	5 GHz	54 Mbps	23 Mbps
802.11b	2.4 GHz	11 Mbps	11 Mbps
802.11g	2.4 GHz	54 Mbps	19 Mbps
802.11n	2.4 or 5 GHz	248 Mbps (with 2×2 MIMO)	74 Mbps

support different data rates. The IEEE 802.11a standard was developed to operate on a 5 GHz band, and it supports data rate up to 11 Mbps. The IEEE 802.11b utilizes a lower frequency spectrum (i.e., 2.4 GHz) with the same maximum data rate. The IEEE 802.11g operates on the 2.4 GHz band but provides maximum data rate of 54 Mbps. The new IEEE 802.11n standard utilizes advanced antenna technology (i.e., multiple-input multiple-output (MIMO)) to support data rates up to hundred Mbps. Comparison among these IEEE 802.11 standards is shown in Table 1.4.

The IEEE 802.11 standards support two different communication modes between WLAN nodes and an access point, i.e., the distributed coordination function (DCF) and point coordination function (PCF). In the DCF mode, carrier sense multiple access with collision avoidance (CSMA/CA), which is a contention-based MAC protocol, is used. However, in the PCF mode, a polling-based (and hence contention-free) MAC protocol is used. While DCF is simple to implement, PCF can provide performance guarantee for wireless transmission.

1.3.5 Heterogeneous Broadband Wireless Access Networks

In a heterogeneous environment, different wireless technologies (e.g., cellular, WiFi, and WiMAX networks) are expected to coexist and collaborate with each other to provide Internet services to the mobile users in a seamless manner [8]. While WLANs are more suitable for stationary/quasi-stationary users requiring high throughput connections, cellular networks are more efficient for voice-oriented and limited throughput mobile data services. On the other hand, WiMAX networks can provide very high speed wireless connectivity in presence of mobility. However, since the coverage area

of an WiMAX BS is larger, compared to that of a WLAN access point (AP) or a cellular BS, the bandwidth per area becomes limited. Therefore, an efficient load balancing mechanism among these three different wireless systems will be required to provide wireless access services in such a heterogeneous network.

1.4 Radio Resource Management in Broadband Wireless Access Networks

1.4.1 Resource Allocation and Connection Admission Control (CAC) in WiMAX Networks

The objective of radio resource management in a wireless network is to control resource allocation to the ongoing and the incoming connections so that the desired performance objectives (e.g., maximization of resource utilization, provisioning of QoS to the users) can be achieved. The main components of radio resource management are queue management, scheduling, and connection admission control (CAC). Queue management deals with the admission of incoming packets into a buffer. With a finite buffer, the queue management mechanism will be responsible for selectively dropping packets depending on the availability of buffer space for incoming packets. The scheduler allocates available radio resource (e.g., time slot, subchannel) for transmission of buffered packets. Connection admission control is used to decide whether incoming user/connection can be accepted to receive the service of the system or not. The queue management and the scheduling mechanisms must ensure that the QoS performance for the connections can be guaranteed. Also, the available radio resources need to be allocated in a fair manner among the connections. A CAC mechanism is required to ensure that, upon admission of new connections, the QoS performance of the ongoing connections from the different SSs do not deteriorate below an acceptable level, and also the radio resources are efficiently utilized.

The specific requirements for radio resource management in a WiMAX network can be summarized as follows.

- *Bandwidth allocation at both BS and SS:* In a WiMAX network, there are two modes of bandwidth allocation, namely, grant per connection (GPC) and grant

per subscriber station (GPSS) modes, which work on a per connection basis and on a per subscriber station basis, respectively. Therefore, the radio resource allocation mechanism is required to be designed at both SSs and BSs. It is typical for an SS to serve a number of local users, and these users might use applications with different QoS requirements. Therefore, resource allocation is needed locally at each SS. Also, to allocate radio resource to several SSs in a cell, efficient resource scheduling and admission control mechanisms are required at a BS.

- *Supporting delay and throughput-sensitive traffic:* In a WiMAX network, delay and throughput sensitive traffic are supported through rtPS and nrtPS service classes, respectively. To provide this support, radio resource management has to be optimized for different QoS metrics so that the requirements of all users can be satisfied.
- *Utility and revenue maximization:* From users' perspective, the network should maximize the satisfaction of users. This satisfaction can be quantified by the utility which is a function of the observed QoS performance. On the other hand, from service providers' perspective, the network revenue has to be maximized. Therefore, radio resource management mechanisms need to be designed to achieve the highest user utility while at the same time to maximize radio resource utilization in order to maximize service providers' revenue [9].

1.4.2 Resource Allocation for Multihop Mesh Networking and Internetworking with Other Networks

For WiMAX-based mesh networks, efficient radio resource management protocols need to be devised to guarantee user QoS performance requirements. Such a protocol should exploit the radio link and the physical layer information. Also, channel allocation among the different BSs must be optimally designed so that the highest resource utilization can be achieved. Resource allocation mechanism in such a network would impact the higher layer protocol (e.g., routing and transport layer) protocol performance. In [10], an interference-aware routing mechanism for WiMAX mesh networks was proposed. To reduce congestion in a relay BS (i.e., a BS responsible for

relaying Internet traffic), this routing scheme uses interference information from the physical layer to find the optimal route from the source BS to the gateway BS. Since the routing protocol performance strongly depends on the resource allocation scheme used at each BS, a cross-layer optimization approach should be used.

WiMAX mesh networks would be suitable for backhauling WiFi hotspots (e.g., in remote localities where wired infrastructure is not available). In this case, Internet traffic to/from WiFi APs are relayed through the WiMAX BSs. Therefore, protocol issues related to internetworking of these two systems need to be resolved. As an example, resource reservation for service flows can be done in a collaborative way in such a heterogeneous networking scenario.

1.4.3 Resource Allocation in Integrated WiMAX/WiFi/3G Cellular Networks

Future generation wireless terminals are expected to be able to access different wireless networks to provide ubiquitous connectivity as well as high throughput performance to mobile users. Integration of the diverse types of wireless access networks such as WiMAX, WiFi, and 3G cellular networks would give rise to new challenges for radio resource management. Although the problem of radio resource allocation and admission control was extensively studied in the literature, it has not been studied thoroughly in a heterogeneous wireless access setting considering both the user-centric and network-centric viewpoints. The challenges in designing radio resource allocation in this heterogeneous wireless network can be summarized as follows.

- *Load balancing:* When multiple wireless access networks are available, traffic load can be balanced among the different networks to provide better QoS performance (e.g., higher throughput) to the wireless users. For example, a user's traffic can be divided into multiple streams which can be transmitted over cellular network and WLAN simultaneously. In addition, load balancing can improve radio resource utilization by avoiding network congestion. In an integrated WiMAX/3G cellular network, when congestion occurs, some of the cellular users can be handed over to the WiMAX network to reduce the effect of call blocking.

- *Cooperative and noncooperative behaviors of network service providers:* In heterogeneous wireless networks, each network can be operated by different network service provider. These network service providers can cooperate or compete among each other to provide wireless access service to the users, so that their utility/revenue is maximized. In this case, game theory can be applied to analyze these cooperative and noncooperative behaviors. A game-theoretic solution is preferred since it ensures that all of the service providers are satisfied with the solution (i.e., utility of each service provider is maximized given the actions from other service providers).

1.5 Scope and Significance of This Research

The scope of the research presented in this thesis can be summarized as follows:

- *Radio resource management framework for WiMAX broadband wireless access networks:* The major components of this radio resource management framework are bandwidth allocation, rate control, and admission control. While bandwidth allocation is used to allocate available resource (i.e., time slot or subchannel) to the users, rate control is used to limit the traffic arrival so that the performance can be maintained at the target level. These components impact the packet-level performances (e.g., delay and loss). Analytical models are required to investigate the network performance. In a radio resource management framework, admission control is used to decide whether a new connection can be admitted or not. This is performed based on the available radio resource and the QoS requirements of both ongoing connections and an incoming connection. Admission control will affect the system performance in terms of network utilization and connection blocking probability.
- *Integration of WiMAX with other broadband wireless technologies:* WiMAX will complement existing wireless technologies (e.g., cellular and WiFi). There are two possible scenarios for integration, namely, a multihop communication scenario and a heterogeneous wireless networking scenario. In the multihop communication scenario, WiMAX networks can be used as backhauls to relay traffic from WLAN access point and cellular base stations to the Internet.

A heterogeneous wireless access environment (e.g., integrated WiMAX/WiFi/3G cellular environment) can enhance users' throughput performance since the mobile terminal can access multiple networks simultaneously. In such an environment, different network could be operated by different service providers. Therefore, radio resource management (e.g., bandwidth allocation, capacity reservation, and admission control.) has to be designed considering noncooperative behavior of the service providers.

In this research, the problem of radio resource management in a WiMAX network is addressed. Radio resource allocation, scheduling, and admission control methods are designed for WiMAX networks and novel analytical models are developed to investigate the performance of these methods. These analytical models consider the physical and MAC layer specifics of WiMAX standard. These analytical models can be used to optimize the system parameters under given performance objectives/constraints.

Also, the problem of radio resource management in a heterogeneous wireless access network is addressed. To solve the radio resource allocation problem in such a network, novel game-theoretic models are developed. The solutions obtained from these game-theoretic models are optimal from the service providers' perspective in the sense that they satisfy all the wireless service providers. Using game theory techniques, radio resource management solutions are obtained considering both cooperation and competition among the service providers.

In summary, the radio resource management models and solutions for broadband wireless access networks developed in this thesis are novel, mathematically rigorous, and provide interesting insights into the system performance. These are important tools which can be used to optimize the network performance under given performance objectives and resource constraints.

1.6 Organization of the Thesis

The rest of the thesis is organized as follows (Fig. 1.4):

- **Chapter 2:** In this chapter, a radio resource management framework is presented for the WiMAX subscriber stations. This framework is composed of a queue-aware uplink bandwidth allocation and a rate control mechanism. At a

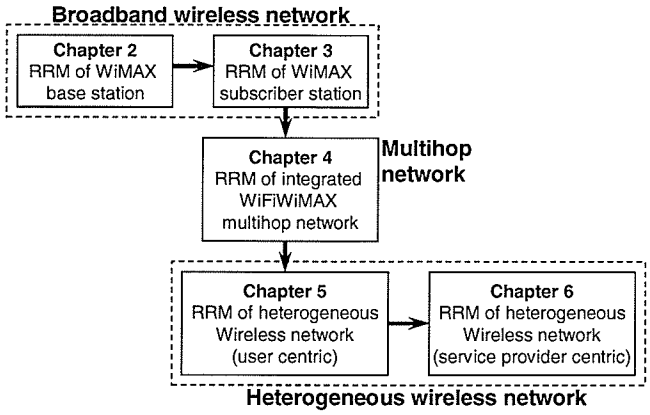


Figure 1.4. *Organization of the thesis.*

subscriber station, by utilizing the queue state information, bandwidth allocated to QoS-sensitive connections (i.e., polling service) can be adjusted adaptively. Also, the rate control mechanism can dynamically limit the transmission rate of the traffic source to maintain the QoS performance experienced by the subscriber station at the target level. A queueing model is developed to analyze the proposed radio resource management model, from which various performance measures can be obtained.

- **Chapter 3:** In this chapter, the radio resource management problem at the WiMAX base station is addressed. A queueing theoretic and optimization-based resource allocation model is developed which considers both packet-level and connection-level QoS constraints. The queueing model is used to obtain the packet-level QoS performance measures. Based on this QoS information, a joint bandwidth allocation and connection admission control algorithm are developed. Another queueing model is used to analyze the connection-level performance measures such as connection blocking probability and average number of ongoing connections. Then, an optimization formulation is used to obtain the optimal threshold settings for complete partitioning of the available bandwidth resources. With this resource allocation model, the connection-level QoS for the different types of connections can be maintained at the target level while maximizing the average system revenue.

- **Chapter 4:** In this chapter, a network architecture is presented for integrating WiFi WLANs with WiMAX-based multihop wireless mesh infrastructure to relay WLAN traffic to the Internet. A game-theoretic radio resource management is developed for this integrated network. In particular, a multi-player bargaining game formulation is used for fair bandwidth allocation and optimal admission control of different types of connections (e.g., WLAN connections, relay connections, connections from standalone WiMAX SSs) in a WiMAX base station/mesh router.
- **Chapter 5:** In this chapter, a bandwidth allocation and admission control framework is developed for an integrated WiMAX, WiFi, and cellular network architecture. This framework is developed based on a *bankruptcy game* which is a special type of an N-person cooperative game. A coalition among the different wireless access networks is formed to offer bandwidth to a new connection. *Shapley value* is considered as the solution for allocating bandwidth to a new connection. Subsequently, an admission control algorithm is proposed to ensure that the QoS performance of the admitted connections in the network can be maintained at the target level.
- **Chapter 6:** In this chapter, a game-theoretic framework is proposed for radio resource management in a heterogeneous wireless network considering the spatial and temporal variations in the traffic demand in the network. In this framework, a long-term bandwidth allocation method is used to assign the available bandwidth from different networks to the different service areas. For this long-term bandwidth allocation, a noncooperative game is formulated and its solution (i.e., Nash equilibrium) is obtained. Based on this long-term allocated bandwidth, a resource reservation method is used to prioritize vertical and horizontal handoff connections over new connections. A bargaining game is formulated to obtain the solution for the reservation threshold. Next, a short-term bandwidth allocation scheme is used to dynamically allocate bandwidth to the different connections. This short-term allocation is formulated as a noncooperative game for which two algorithms are proposed to obtain the solution. Then, based on the short-term allocated bandwidth and the reservation threshold, an admission control method is presented to ensure the QoS requirements of the

admitted users in the network.

- **Chapter 7:** This chapter provides a summary of the results presented in this thesis and outlines a few issues which can be pursued as an extension of this research.

The flow of the thesis is as follows (Fig. 1.4). In chapter 2, given the bandwidth assigned by the WiMAX BS, the SS allocates the bandwidth to the different service classes adaptively. In this case, the bandwidth assigned to each SS can be obtained from the radio resource management in chapter 3. This radio resource management is optimized for the optimal allocation for all connections in a cell. This single-hop WiMAX network is extended in chapter 4 by integrating WiMAX BSs and WiFi-based WLAN to relay traffic from different sources. Alternatively, WiMAX network can be integrated with other wireless technologies to provide heterogeneous wireless service. In chapter 5 and 6, the radio resource management frameworks for such heterogeneous wireless networks are proposed.

Chapter 2

Radio Resource Management in WiMAX: Part I

2.1 Introduction

2.1.1 Problem Statement

In this chapter, the problem of radio resource management framework for the WiMAX subscriber station (SS) is considered. This SS accommodates three types of connections (i.e., UGS, PS, and BE service). The first input of the framework is the amount of bandwidth assigned by the WiMAX base station (BS). The second input is the number of connections and their traffic descriptions (e.g., average packet generation rates). The outputs of this framework are the amount of allocation bandwidth and traffic shaping parameters of the connections. The framework should be able to efficiently utilize the available bandwidth. In this case, the framework must satisfy transmission requirements the UGS connections while minimizing packet delay of PS connections.

2.1.2 Contribution

According to the above requirements, a queue-aware uplink bandwidth allocation and rate control schemes are proposed for a subscriber station. These schemes can be applied for both real-time and non-real-time polling service (PS) as defined in the WiMAX specifications. Under the proposed bandwidth allocation scheme, the amount of bandwidth allocated for polling service can be adjusted dynamically ac-

cording to the variations in traffic load and/or channel quality (and hence the queue length) so that the packet-level QoS performances such as protocol data unit (PDU) delay[1] and PDU dropping probability can be maintained at the desired level. Also, rate control is used to limit the transmission rate of the connections under the polling service class so that the QoS performances can be controlled. The proposed queue-aware rate control scheme can be applied to each connection separately so that service differentiation (i.e., prioritization) among the connections can be achieved through different parameter settings.

A queueing analytical framework is presented to evaluate the performances of the proposed schemes. This is based on a discrete-time Markov chain which is formulated by considering a Markov modulated Poisson process (MMPP) as the traffic sources under polling service. The advantages of using the MMPP are two-fold: first, the MMPP is able to capture the burstiness in the traffic arrival pattern, which is a common characteristic for multimedia and real-time traffic such as voice over IP (VoIP) and MPEG video [14] as well as Internet traffic [15]. Second, for multiplexed traffic sources, MMPP model can be analytically obtained. The method will be shown later in this chapter.

The proposed radio resource management model for PS considers the impact of higher-priority traffic corresponding to the unsolicited granted service (UGS) class for which the bandwidth can be statically or dynamically allocated according to the connections' transmission rate requirements. An approximate queueing analytical model for best-effort (BE) service is also presented. With this model, the basic performance measures (e.g., average delay) for BE traffic can be obtained as well as the impact of polling service on BE service can be investigated. The simulations is used to validate the correctness of the analytical model.

The major contributions of this chapter can be summarized as follows:

- A queue-state aware bandwidth allocation mechanism is proposed for reserving transmission bandwidth at a subscriber station for polling service. Also, a queue-state-based rate control method (both on aggregate and per-flow basis) is presented to limit the packet generation rate for connections under polling

[1]WiMAX medium access control (MAC) uses a variable length protocol data unit (PDU) along with a number of other concepts that greatly increase the efficiency of the standard. Multiple MAC PDUs may be concatenated into a single burst to save physical layer (PHY) overhead.

service.

- A queueing analytical model is developed to investigate the performances (under both steady state and transient state) of the queue-aware bandwidth allocation and the rate control mechanisms for polling service.
- An approximate queueing model is developed for analyzing the performance of best-effort traffic in presence of polling service.

2.2 Related Work

Radio resource scheduling (i.e., bandwidth allocation) and admission control are crucial for provisioning QoS in a 802.16 network. In [16], QoS-aware packet scheduling schemes were proposed for different types of traffic at the 802.16 base station. A resource allocation strategy, namely, enhanced staggered resource allocation (ESRA) method, was proposed in [17] with an objective to maximize the number of concurrent transmissions so that the throughput can be maximized. However, the buffer dynamics at the radio link level queue (and hence the queueing performance) was not analyzed.

In [18], an admission control scheme for broadband multi-services wireless networks was presented to limit the number of ongoing connections so that the QoS for each connection can be maintained at the desired level. A dynamic resource allocation scheme for broadband orthogonal frequency division multiple access (OFDMA) networks was presented in [19], where the allocation is performed in two steps, namely, bandwidth allocation and channel assignment. Also, an M/G/1/K queueing model was used to estimate the packet blocking probability based upon which dynamic bandwidth allocation can be performed. The QoS differentiation was not considered in this work. In [20], an adaptive call admission control method using stochastic control was proposed for BWA.

Although the general problem of radio resource management was studied extensively in the literature (e.g., in [21]-[24]), the radio link level queueing aspects were ignored in most of the cases and the queueing dynamics (and hence the packet-level QoS) was not exploited for resource management and transmission rate control in wireless networks. The problem of optimal polling among several queues was studied

in the literature. In [25], an optimal policy for polling (scheduling) was obtained to minimize stochastically the amount of unfinished work and the number of customers in the queues.

Rate control has been widely used in the wired-network environment to limit the transmission rate of the traffic sources. The performance of rate control mechanism in ATM networks was studied in [26] by using a queueing model, and the throughput degradation was quantified. Rate control can be implemented through random early drop (RED) [27] mechanism to block the incoming packets gradually to avoid congestion. A proportional rate control mechanism for wireless networks was proposed in [27] to stabilize traffic oscillations. In [28], a theoretical model for wireless traffic control was proposed considering the impacts of congestion and error in the wireless channel. The model was developed based on the rate-controlled earliest deadline first (RC-EDF) scheduling framework. However, these works did not consider multiple classes of connections with different QoS requirements.

The problem of analyzing radio link level queueing under wireless packet-transmission was addressed in the literature. In [30], a Markov-based model was presented to analyze the radio link level packet dropping process under ARQ-based error control. In [31], an analytical model for deriving packet loss rate, average throughput and average spectral efficiency under adaptive modulation and coding (AMC) was presented. The radio link level delay statistics for selective repeat ARQ were analyzed in [32]. Also, a heuristic algorithm was proposed to analyze the application layer delay performance. However, these queueing models considered only a single user transmission scenario.

2.3 System Model and Assumptions

2.3.1 System Description

We consider an uplink transmission scenario from an SS to a BS (Fig. 2.1) through the time-division multiple access (TDMA)/time-division duplex (TDD) access mode and single carrier modulation (e.g., as in WirelessMAN-SC) for three traffic types, namely, UGS, PS, and BE traffic (Fig. 2.2). For PS, a dedicated queue is used to buffer the PDUs from the corresponding connections. An SS of type GPSS, for which

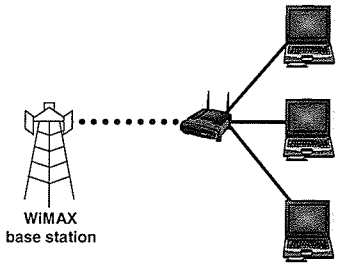


Figure 2.1. *Connection between a subscriber station and the base station.*

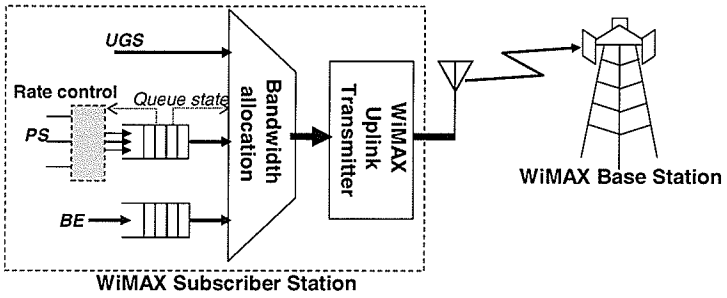


Figure 2.2. *System model.*

a certain amount of bandwidth, is reserved by the BS is considered. This allocated bandwidth is shared among the different service types in the same SS with UGS having the highest priority and the BE service having the lowest priority.

For better scalability, the PDUs from all the PS connections are aggregated into a single queue of size X PDUs. For the PS queue, rate control can be applied to control traffic at the packet-level and at the connection-level, respectively. If the rate control parameters for each of the connections are identical, all PS connections experience the same QoS performance. Since there is no performance guarantee for best-effort traffic, the queue size for the best-effort traffic is assumed to be infinity.

The key notations are listed in Table. 2.1.

Table 2.1. *List of key notations.*

Notation	Description
b_{max}	The maximum number of MAC PDUs that can be transmitted per subframe
b_{ugs}	The bandwidth allocated to UGS
ψ_b	Threshold of bandwidth allocation
$\mathcal{B}(x)$	Bandwidth allocated to polling service
τ_{min}, τ_{max}	The rate control thresholds for the number of PDUs
λ_o, λ_{min}	PDU arrival rate, minimum guarantee PDU arrival rate
N	Number of polling-service connections
ρ	Average PDU arrival rate at the polling-service queue
$\overline{x}$, $\overline{y}$	Average queue length of polling-service and best-effort service
P_{bl}	PDU blocking probability
η	Queue throughput
$\overline{b}$	Average allocated bandwidth to the polling-service queue
μ	Bandwidth utilization
$\overline{d}$	Average queueing delay
X	Queue size
θ	PDU error rate (PER)
$\mathbf{U}$, $\mathbf{\Lambda}$	Probability transition and Poisson arrival rate matrices of MMPP
$\mathbf{P}$, $\mathbf{Q}$	Probability transition matrices of polling-service and best-effort service queues
$\boldsymbol{\pi}_m$, $\boldsymbol{\pi}_{st}$	Steady state probability of MMPP and polling-service queue

2.3.2 Queue-Aware Bandwidth Allocation

We denote by b_{max} ($b_{max} \in \mathbb{N}$) the maximum number of MAC PDUs that an SS can transmit per uplink transmission subframe. We consider two modes of bandwidth allocation for PS, namely, complete partitioning (CP) and complete sharing (CS). With complete partitioning, a fixed amount of bandwidth b_{ugs} (from the total bandwidth allocated to an SS) is statically allocated for UGS while the remaining bandwidth (i.e., $b_{max} - b_{ugs}$) is allocated for PS and BE services. In case of complete sharing, when the bandwidth requirement for UGS traffic is less than b_{ugs}, the remaining available bandwidth will be available for PS. After the required amount of bandwidth has been allocated to UGS and PS traffic, the left-over bandwidth is allocated to BE traffic.

We propose an uplink bandwidth allocation scheme for PS, which takes the current number of PDUs in the PS queue into account. The allocation is done on a frame-by-frame basis in which the amount of bandwidth is determined for each transmission frame individually. In this scheme, the set of thresholds for bandwidth allocation is defined as follows:

$$\Psi = \{\psi_1, \psi_2, \cdots, \psi_b, \cdots, \psi_{b_{max}-b_{ugs}}\} \tag{2.1}$$

where $\psi_b \in \{1, \cdots, X\}$, $\psi_b < \psi_{b+1}$, and $b \in \{1, \cdots, b_{max} - b_{ugs}\}$. This set of thresholds is used to indicate the amount of bandwidth required in each uplink subframe. In particular, the amount of bandwidth allocated to polling service is calculated as a function of the number of PDUs in the PS queue, for complete partitioning and complete sharing schemes, respectively, as follows:

$$\mathcal{B}_{CP}(x) = \begin{cases} 0, & x = 0 \\ b, & \psi_b \leq x < \psi_{b+1} \\ b_{max} - b_{ugs}, & \psi_{b_{max}-b_{ugs}} \leq x \end{cases} \qquad \mathcal{B}_{CS}(x) = \begin{cases} 0, & x = 0 \\ b, & \psi_b \leq x < \psi_{b+1} \\ b_{max}, & \psi_{b_{max}-b_{ugs}} \leq x. \end{cases} \tag{2.2}$$

2.3.3 Queue-Aware Rate Control

We propose a queue-aware rate control mechanism for PS connections, in which the PDU arrival rate is controlled according to the number of PDUs in the queue. This rate control can be implemented either at the traffic source or at the PS queue. In the former case, the SS informs the traffic sources of the queue status. Note that since

the SS and the traffic sources are in the same local network, the delay incurred for updating queue status is ignored. In the latter case, rate control can be implemented similarly to the random early detection (RED) mechanism [27] in an Internet router, in which some PDUs received at the PS queue are randomly dropped.

Let $\tau_{min}, \tau_{max} \in \mathbb{N}$ denote the rate control thresholds for the number of PDUs in the queue and λ_{min} denote the minimum guaranteed arrival rate. Specifically, the transmission rate of traffic source under PS cannot be reduced below λ_{min}. Then, with a PDU arrival rate of λ_o, the rate control policy can be expressed as a function of the number of PDUs in the PS queue (x) as follows:

$$\tilde{\lambda}(x, \lambda_o, \lambda_{min}) = \begin{cases} \lambda_o, & x < \tau_{min} \\ \mathcal{F}(\lambda_o, x), & \tau_{min} \leq x < \tau_{max} \\ \lambda_{min}, & \tau_{max} \leq x \end{cases} \tag{2.3}$$

where $\tilde{\lambda}(.)$ is the controlled arrival rate, and $\mathcal{F}(\lambda, x)$ is a non-increasing function of x with the constraint $\lambda_{min} \leq \mathcal{F}(\lambda, x) \leq \lambda_o$.

The rate control mechanism can be applied on either an aggregate or a per-flow basis. In the former case, PDU arrival rates for all connections under PS are controlled using the same values of τ_{min}, τ_{max}, and λ_{min}. In the latter case, different parameter settings for rate control are used for each connection (i.e., $\tau_{min}(i)$, $\tau_{max}(i)$, and $\lambda_{min}(i)$ for connection i). While per-flow rate control is able to differentiate the QoS among different connections, aggregate rate control is simpler to implement and applicable when all connections have the same QoS requirements.

2.3.4 Error Control

To ensure the reliability of PDU transmission from SS to BS, an infinite persistent ARQ-based error recovery is used. That is, the erroneous PDUs will be re-transmitted until they are successfully received at the BS. If θ denotes the PDU error rate (PER), assuming an independent error process, the probability that n PDUs out of m transmitted PDUs are successfully received can be obtained as follows:

$$\theta_{n,m} = \binom{m}{n} (1-\theta)^n (\theta)^{m-n}, \quad n \in \{0, 1, \cdots, m\}. \tag{2.4}$$

We also assume that the transmission status for the PDUs transmitted in the previous frame time is made available to the transmitter before transmissions in the current frame time start.

2.4 Queueing Analytical Model for Polling Service (PS)

2.4.1 PDU Arrival Process for PS Connections

We assume that the PDU arrival process for each PS connection follows an MMPP model. An MMPP model is more general than a traditional Poisson model and is able to capture burstiness in the traffic arrival process. With MMPP, the PDU arrival rate λ_s is determined by the state s of the Markov chain, and the total number of states is S (i.e., $s = 1, 2, \cdots, S$). The MMPP process for connection i can be represented by $\mathbf{U}(i)$ and $\mathbf{\Lambda}(i)$, in which the former is the transition probability matrix of the modulating Markov chain, and the latter is the matrix of Poisson arrival rate. These matrices are defined as follows:

$$\mathbf{U}(i) = \begin{bmatrix} u_{1,1} & \cdots & u_{1,S} \\ \cdots & \cdots & \cdots \\ u_{S,1} & \cdots & u_{S,S} \end{bmatrix}, \mathbf{\Lambda}(i) = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_S \end{bmatrix}. \tag{2.5}$$

A discrete time MMPP (dMMPP) [33] is equivalent to an MMPP in the continuous time. In this case, the rate matrix $\mathbf{\Lambda}(i)$ is represented by diagonal probability matrix $\mathbf{\Lambda}_a(i)$ when the number of PDUs arriving in one frame is a. Note that $a \in \{0, 1, \cdots, A\}$, in which A is the maximum batch size for PDU arrival (e.g., $1 \leq A \leq \infty$) . Each diagonal element of $\mathbf{\Lambda}_a(i)$ can be obtained from

$$f_a(\lambda_s) = \frac{e^{-\lambda_s T}(\lambda_s T)^a}{a!} \tag{2.6}$$

where $f_a(\lambda_s)$ is the probability that a Poisson events occur during time interval T (i.e., frame length) with mean rate λ_s.

In the case of aggregated traffic from two users (e.g., user 1 and user 2), the matrices corresponding to state transition and PDU arrival probability for this multiplexed

source can be calculated as follows:

$$\mathbf{U} = \mathbf{U}(1) \otimes \mathbf{U}(2) \tag{2.7}$$

$$\mathbf{\Lambda}_a = \sum_{i+j=a} \mathbf{\Lambda}_i(1) \otimes \mathbf{\Lambda}_j(2), \quad i,j \in \{0,1,\cdots,A\} \tag{2.8}$$

for $a = 0, 1, \cdots, 2A$, where $\otimes$ denotes the Kronecker product. For the case with more than two users, these two matrices can be obtained in a similar way. The average PDU arrival rate for connection i is obtained as follows:

$$\rho(i) = \boldsymbol{\pi}_m(i) \left(\sum_{a=0}^{A} a\mathbf{\Lambda}_a(i) \right) \mathbf{1} \tag{2.9}$$

where $\boldsymbol{\pi}_m(i)$ is obtained by solving $\boldsymbol{\pi}_m(i)\mathbf{U}(i) = \boldsymbol{\pi}_m(i)$ and $\boldsymbol{\pi}_m(i)\mathbf{1} = 1$. Note that $\mathbf{1}$ is a column matrix of ones. Therefore, with a total of N connections the total average PDU arrival rate at the PS queue can be obtained as follows:

$$\rho = \boldsymbol{\pi}_m \left(\sum_{a=0}^{NA} a\mathbf{\Lambda}_a \right) \mathbf{1} \tag{2.10}$$

and $\boldsymbol{\pi}_m$ is obtained from $\boldsymbol{\pi}_m\mathbf{U} = \boldsymbol{\pi}_m$ and $\boldsymbol{\pi}_m\mathbf{1} = 1$.

2.4.2 PDU Arrival Process for UGS Connections

For modeling the PDU arrival process for UGS connections, a multistate on-off model which is a special type of dMMPP, is considered. The maximum number of states for each connection is C, and the number of PDU arrivals when the source is in state c is c. While the state transition matrix $\mathbf{V}(i)$ of the multistate on-off model for connection i is similar to that of MMPP, the PDU arrival probability matrices $\mathbf{\Gamma}_c(i)$ are different. In particular, the maximum batch size is C (i.e., $A = C$), and the diagonal elements of these matrices are defined as follows:

$$[\mathbf{\Gamma}_c(i)]_{j,j} = \begin{cases} 1, & j = c+1 \\ 0, & \text{otherwise} \end{cases} \tag{2.11}$$

for $c \in \{0, 1, \cdots, C\}$ where the first row corresponds to the case of no PDU arrival, and $[\mathbf{\Gamma}_c(i)]_{j,k}$ denotes the element at row j and column k of matrix $\mathbf{\Gamma}_c(i)$. If there are multiple UGS connections, the state transition matrix $\mathbf{V}$ and the PDU arrival

probability matrices $\mathbf{\Gamma}_c$ of the multiplexed connection can be obtained from (2.7) and (2.8). Note that b_{ugs} denotes the maximum total bandwidth for UGS, where $b_{ugs} = MC$ for a total of M multistate on-off sources.

2.4.3 Formulation of the Queueing Model for Polling Service

In our queueing model, the state of the PS queue (i.e., the number of PDUs in the polling service queue) is observed at the beginning of each frame time. A PDU arriving during frame time f will not be transmitted until the next frame time $f+1$ at the earliest. The state space of the queue can be defined as follows:

$$\Delta = \{(\mathscr{S}, \mathscr{C}, \mathscr{X}), \ , 1 \leq \mathscr{S} \leq NS, 1 \leq \mathscr{C} \leq M, 0 \leq \mathscr{X} \leq X\}, \tag{2.12}$$

where $\mathscr{S}$ is the state of dMMPP traffic sources, $\mathscr{C}$ is the state of multistate on-off sources, and $\mathscr{X}$ is the number of PDUs in the PS queue. While the states of dMMPP and multistate on-off models are independent for all connections, the number of PDUs in the queue depends on the dMMPP arrival probabilities, the bandwidth usage for UGS connections, and the service rate at the PS queue (and hence the amount of allocated bandwidth). Also, the amount of allocated bandwidth depends on the number of PDUs in the PS queue and the set of thresholds Ψ. Note that in case of complete partitioning, the model does not need to maintain the state of any multistate on-off source, and therefore, $\mathscr{C} = \{\emptyset\}$.

The transition matrix $\mathbf{P}$ of the queue can be expressed as in (2.13) where the rows of matrix $\mathbf{P}$ correspond to the number of PDUs in the PS queue (i.e., $\mathscr{X}$).

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_{0,0} & \cdots & \mathbf{p}_{0,AN} & & \\ \vdots & \vdots & \ddots & \ddots & \\ \mathbf{p}_{b_{max},0} & \cdots & \cdots & \mathbf{p}_{b_{max},b_{max}+AN} & \\ & \ddots & \ddots & \ddots & \ddots \\ & \mathbf{p}_{y,y-b_{max}} & \cdots & \cdots & \mathbf{p}_{y,y+AN} \\ & & \ddots & \ddots & \ddots \end{bmatrix} \tag{2.13}$$

Matrices $\mathbf{p}_{x,x'}$ represent the changes in the number of PDUs in the queue (i.e., there are x PDUs during the current frame time and it will be x' during the next frame time).

2.4.3.1 Arrival Process under Rate Control

With queue-aware rate control, the matrix for the Poisson arrival process $\mathbf{\Lambda}(i)$ in the MMPP model for connection i depends on the number of PDUs in the PS queue. Therefore, this matrix can be expressed as follows:

$$\mathbf{\Lambda}^{(x)}(i) = \begin{bmatrix} \tilde{\lambda}(x, \lambda_1, \lambda_{min}) & & \\ & \ddots & \\ & & \tilde{\lambda}(x, \lambda_S, \lambda_{min}) \end{bmatrix}. \tag{2.14}$$

Then, the matrix $\mathbf{\Lambda}_a^{(x)}(i)$ is obtained by using (2.6). If there are multiple traffic sources, (2.7) and (2.8) are used to obtain the complete PDU arrival process (i.e., $\mathbf{U}$ and $\mathbf{\Lambda}_a^{(x)}$) at the PS queue. Note that (2.14) can be used for both aggregate and per-flow-based rate control.

2.4.3.2 Transition Matrix for Complete Partitioning (CP) Model

In case of complete partitioning, the PDU arrival probability and dMMPP state transitions are given by $\mathbf{U}$ and $\mathbf{\Lambda}_a^{(x)}$. However, the PDU departure probabilities corresponding to all arrival states of dMMPP are identical and depend only on the number of PDUs in the queue and the PDU transmission error rate. Therefore, the probability of departure of n PDUs ($n \in \{0, 1, \cdots, b_{max} - b_{ugs}\}$) when there are x PDUs ($x \in \{0, 1, \cdots, X\}$) in the queue is obtained as follows:

$$\left[\mathbf{D}_n^{(x)}\right]_{j,j} = \theta_{n,\mathcal{B}_{CP}(x)} \tag{2.15}$$

where $j \in \{1, 2, \cdots, SN\}$. Note that every matrix $\mathbf{D}_n^{(x)}$ has the same size as that of $\mathbf{U}$. Each element $\mathbf{p}_{x,x'}$ of matrix $\mathbf{P}$ in case of complete partitioning is obtained as follows:

$$\mathbf{p}_{x,x-g} = \mathbf{U} \sum_{\{n,a|n-a=g\}} \left(\mathbf{\Lambda}_a^{(x)} \times \mathbf{D}_n^{(x)}\right) \tag{2.16}$$

$$\mathbf{p}_{x,x+h} = \mathbf{U} \sum_{\{n,a|a-n=h\}} \left(\mathbf{\Lambda}_a^{(x)} \times \mathbf{D}_n^{(x)}\right) \tag{2.17}$$

$$\mathbf{p}_{x,x} = \mathbf{U} \sum_{\{n,a|n=a\}} \left(\mathbf{\Lambda}_a^{(x)} \times \mathbf{D}_n^{(x)}\right) \tag{2.18}$$

for $g = \{1, \cdots, G\}$ and $h = \{1, \cdots, AN\}$ where $n \in \{0, \cdots, G\}$ and $a \in \{0, \cdots, AN\}$ represent the number of departed PDUs and the number of PDU arrivals, respectively.

Considering both the arrival and the departure events, (2.16), (2.17), and (2.18) above represent the transition probability matrices for the cases when the number of PDUs in the queue decreases by g, increases by h, and remains unchanged, respectively. Since the maximum total allocated bandwidth can be greater than the number of PDUs in the queue, and the decrease in the number of PDUs cannot be less than the number of PDUs in the queue, the maximum amount, by which the number of PDUs in the queue can decrease, is obtained from $G = \min(b_{max} - b_{ugs}, x)$.

2.4.3.3 Transition Matrix for the Complete Sharing (CS) Model

In this case, transmission of multistate on-off traffic for UGS connections, which have higher priority and affect bandwidth allocation for the PS traffic, must be considered. The departure probability matrix for the multistate on-off sources (corresponding to the UGS connections) can be established as follows:

$$\left[\mathbf{E}_n^{(x)}\right]_{c+1,c+1} = \begin{cases} \theta_{n,m}, & m = \min\left(\mathcal{B}_{CS}(x), b_{max} - c\right) \\ 0, & \text{otherwise} \end{cases} \tag{2.19}$$

where $c \in \{0, 1, \cdots, b_{ugs}\}$. Note that every matrix $\mathbf{E}_n^{(x)}$ has the same size as that of $\mathbf{V}$. For the CS case, each element $\mathbf{p}_{x,x'}$ of matrix $\mathbf{P}$ is obtained as follows:

$$\mathbf{p}_{x,x-g} = \mathbf{U} \otimes \mathbf{V} \sum_{\{n,a|n-a=g\}} \left(\mathbf{\Lambda}_a^{(x)} \otimes \mathbf{E}_n^{(x)}\right) \tag{2.20}$$

$$\mathbf{p}_{x,x+h} = \mathbf{U} \otimes \mathbf{V} \sum_{\{n,a|a-n=h\}} \left(\mathbf{\Lambda}_a^{(x)} \otimes \mathbf{E}_n^{(x)}\right) \tag{2.21}$$

$$\mathbf{p}_{x,x} = \mathbf{U} \otimes \mathbf{V} \sum_{\{n,a|n=a\}} \left(\mathbf{\Lambda}_a^{(x)} \otimes \mathbf{E}_n^{(x)}\right) \tag{2.22}$$

where $n \in \{0, 1, 2, \cdots, G\}$, $a \in \{0, 1, 2, \cdots, AN\}$, and $G = \min(b_{max}, x)$.

2.4.3.4 PDU Blocking Process

If the PS queue does not have enough space to accommodate all of the incoming PDUs, some of the PDUs will be blocked. In this case, the bottom part (i.e., the rows corresponding to the condition $(AN) + x > X$) of matrix $\mathbf{P}$ has to capture the

PDU blocking effect. Therefore, (2.17) and (2.21), which correspond to the CP and the CS cases, respectively, become

$$\mathbf{p}_{x,x+h} = \sum_{i=h}^{AN} \hat{\mathbf{p}}_{x,x+i} \quad \text{for } x+h \geq X \tag{2.23}$$

and for $x = X$, (2.18) and (2.22) become

$$\mathbf{p}_{x,x}\hat{\mathbf{p}}_{x,x} + \sum_{i=1}^{AN} \hat{\mathbf{p}}_{x,x+i} \quad \text{for } x = X \tag{2.24}$$

where $\hat{\mathbf{p}}_{x,x}$ is obtained from the case without PDU dropping.

2.4.3.5 Steady State Probabilities

The queueing performance measures for the PS traffic can be obtained from the steady state probability matrix $\boldsymbol{\pi}_{st}$ which is obtained by solving the equations

$$\boldsymbol{\pi}_{st}\mathbf{P} = \boldsymbol{\pi}_{st}, \quad \boldsymbol{\pi}_{st}\mathbf{1} = 1 \tag{2.25}$$

where $\mathbf{1}$ is a column matrix of ones. The matrix $\boldsymbol{\pi}_{st}$ contains steady state probabilities for the feasible combinations of the state variables $\mathscr{S}$, $\mathscr{C}$, and $\mathscr{X}$. This matrix can be decomposed into $\pi_{st}^{(CS)}(s,c,x)$, which is the steady state probability that the dMMPP source is in state s, the multistate on-off source is in state c, and there are x PDUs in the PS queue. Note that $\boldsymbol{\pi}_{st}$ is a row matrix and $[\boldsymbol{\pi}_{st}]_i$ indicates the element at column i of matrix $\boldsymbol{\pi}_{st}$. Since in the case of complete sharing, the system state does not keep track of multistate on-off sources, this steady state probability is reduced to $\pi_{st}^{(CP)}(s,x)$.

2.4.3.6 Transient State Probabilities

In this section, the system behavior in the transient state is investigated. A system exhibits transient behavior when it is not in the steady state, i.e., during the transition period until the system reaches an equilibrium state [29]. Transient analysis is important to observe the system behavior with changes in inputs or in system parameters, especially in a time-varying system which may rarely reach the steady state.

Based on the *Chapman-Kolmogorov* equations, the probability matrix of system states during frame time f can be obtained from

$$\boldsymbol{\pi}_{tr}(f) = \boldsymbol{\pi}_{tr}(f-1)\mathbf{P}(f) \tag{2.26}$$

where $\mathbf{P}(f)$ is the transition matrix during frame time f. The transient state probabilities $\pi_{tr}^{(CS)}(s,c,x,f)$ and $\pi_{tr}^{(CP)}(s,x,f)$ can be obtained in the same way as that for the steady state probabilities.

2.4.4 QoS Measures for Polling Service

Since the QoS measures for PS in both steady and transient states can be obtained in the similar way, $\pi^{(CS)}(s,c,x)$ and $\pi^{(CP)}(s,x)$ denote the complete sharing and the complete partitioning cases, respectively, representing the general probability that the dMMPP is in state s, the on-off source is in state c, and there are x PDUs in the PS queue.

2.4.4.1 Average Queue Length

The average queue length for the CP and the CS cases can be obtained, respectively, as follows:

$$\overline{x}^{(CP)} = \sum_{x=0}^{X} x \left(\sum_{s=1}^{SN} \pi^{(CP)}(s,x) \right) \tag{2.27}$$

$$\overline{x}^{(CS)} = \sum_{x=0}^{X} x \left(\sum_{s=1}^{SN} \sum_{c=1}^{(b_{ugs}+1)} \pi^{(CS)}(s,c,x) \right). \tag{2.28}$$

2.4.4.2 Average PDU Arrival Rate

For the CP and the CS cases, this can be calculated, respectively, for connection i as follows:

$$\overline{\rho}^{(CP)}(i) = \sum_{x=0}^{X} \left(\boldsymbol{\pi}_m(i) \left(\sum_{a=0}^{A} a\boldsymbol{\Lambda}_a^{(x)}(i) \right) \mathbf{1} \right) \sum_{s=1}^{SN} \pi^{(CP)}(s,x) \tag{2.29}$$

$$\overline{\rho}^{(CS)}(i) = \sum_{x=0}^{X} \left(\boldsymbol{\pi}_m(i) \left(\sum_{a=0}^{A} a\boldsymbol{\Lambda}_a^{(x)}(i) \right) \mathbf{1} \right) \sum_{s=1}^{SN} \sum_{c=1}^{(b_{ugs}+1)} \pi^{(CS)}(s,c,x). \tag{2.30}$$

The total average PDU arrival rate at the PS queue for the CP and the CS cases are calculated, respectively, as follows:

$$\overline{\rho}^{(CP)} = \sum_{i=1}^{N} \overline{\rho}(i)^{(CP)}, \quad \overline{\rho}^{(CP)} = \sum_{i=1}^{N} \overline{\rho}(i)^{(CP)}. \tag{2.31}$$

2.4.4.3 PDU Blocking Probability

To obtain the PDU blocking probability, the average number of blocked PDUs per frame time [30] is first calculated. Given that there are x PDUs in the PS queue and the queue size increases by h, if $h + x > X$, the number of blocked PDUs during one frame time is $h - (X - x)$, and zero otherwise. The average number of blocked PDUs per frame time for the complete partitioning and the complete sharing cases are obtained, respectively, as follows:

$$\overline{x}_{bl}^{(CP)} = \sum_{s=1}^{SN} \sum_{x=0}^{X} \sum_{h=X-x+1}^{SN-\mathcal{B}_{CP}(x)} \pi^{(CP)}(s,x) \left(\sum_{j=1}^{SN} \left[\mathbf{p}_{x,x+h}\right]_{s,j} \right) (h - (X-x)) \tag{2.32}$$

$$\overline{x}_{bl}^{(CS)} = \sum_{s=1}^{SN} \sum_{c=1}^{(b_{ugs}+1)} \sum_{x=0}^{X} \sum_{h=X-x+1}^{SN-\mathcal{B}_{CS}(x)} \pi^{(CS)}(s,c,x) \left(\sum_{j=1}^{SN+(b_{ugs}+1)} \left[\mathbf{p}_{x,x+h}\right]_{s,j} \right) \cdot (h - (X-x)) .$$

The terms $\left(\sum_{j=1}^{SN} \left[\mathbf{p}_{x,x+h}\right]_{s,j}\right)$ in (2.32) and $\left(\sum_{j=1}^{SN+(b_{ugs}+1)} \left[\mathbf{p}_{x,x+h}\right]_{s,j}\right)$ indicate the total probability that the number of PDUs in the queue increases by h at every state of the dMMPP and the multistate on-off sources. After calculating the average number of blocked PDUs per frame time, the probability that an incoming PDU is blocked can be obtained, for the CP and the CS cases, respectively, as follows:

$$P_{bl}^{(CP)} = \frac{\overline{x}_{bl}^{(CP)}}{\overline{\rho}^{(CP)}}, \quad P_{bl}^{(CS)} = \frac{\overline{x}_{bl}^{(CS)}}{\overline{\rho}^{(CS)}}. \tag{2.33}$$

2.4.4.4 Queue Throughput

This gives the average number of transmitted PDUs per frame time. We calculate the throughput by using the fact that if a PDU is not blocked upon its arrival, it will be transmitted eventually. Hence, the queue throughput (number of PDUs/frame

interval) for the complete partitioning and the complete sharing cases can be obtained, respectively, from

$$\eta^{(CP)} = \overline{\rho}^{(CP)}(1 - P_{bl}^{(CP)}), \quad \eta^{(CS)} = \overline{\rho}^{(CS)}(1 - P_{bl}^{(CS)}). \tag{2.34}$$

2.4.4.5 Average Allocated Bandwidth

For the proposed adaptive queue-aware bandwidth allocation, the average bandwidth allocated for the complete partitioning and the complete sharing cases can be obtained, respectively, from

$$\overline{b}^{(CP)} = \sum_{x=0}^{X} (\mathcal{B}_{CP}(x)) \left(\sum_{s=1}^{SN} \pi^{(CP)}(s,x) \right) \tag{2.35}$$

$$\overline{b}^{(CS)} = \sum_{x=0}^{X} \sum_{c=1}^{(b_{ugs}+1)} \left(\sum_{s=1}^{SN} (\mathcal{B}_{CS}(x) - (c-1))\pi^{(CS)}(s,c,x) \right). \tag{2.36}$$

2.4.4.6 Bandwidth Utilization

This performance measure indicates the utilization of allocated bandwidth and can be obtained from

$$\mu^{(CP)} = \frac{\eta^{(CP)}}{\overline{b}^{(CP)}}, \quad \mu^{(CS)} = \frac{\eta^{(CS)}}{\overline{b}^{(CS)}}. \tag{2.37}$$

2.4.4.7 Delay Statistics

Delay for a PDU is defined as the time interval (in terms of the number of frames) since the PDU has arrived at the queue until it is successfully transmitted. By applying the Little's law, average delay is obtained from

$$\overline{d}^{(CP)} = \frac{\overline{x}^{(CP)}}{\eta^{(CP)}}, \quad \overline{d}^{(CS)} = \frac{\overline{x}^{(CS)}}{\eta^{(CS)}}. \tag{2.38}$$

2.5 Queueing Model for Best-Effort (BE) Service

In this section, we present a model for approximate analysis of the basic performance measures (e.g., average queueing delay) for best-effort service, which has the least priority among the three service classes. Since the allocated bandwidth for the BE

queue depends on the state of the UGS and PS connections, and the number of PDUs in the PS queue, the state space for the BE queue can be expressed as follows:

$$\Delta_{BE} = \{(\mathscr{S}, \mathscr{C}, \mathscr{X}, \mathscr{Y}),\ 0 \leq \mathscr{X} \leq X, \mathscr{Y} \geq 0\} \tag{2.39}$$

where $\mathscr{Y}$ is the number of PDUs in the BE queue with infinite buffer size. However, maintaining all these states will make the model quite complicated. Therefore, we present an approximate model with the simplified state space for the BE queue as follows:

$$\Delta_{BE} = \{(\mathscr{Y}),\ \mathscr{Y} \geq 0\}, i > 0. \tag{2.40}$$

The model is approximate in the sense that the correlation among multistate on-off sources, dMMPP sources, and the number of PDUs in the PS queue is ignored. However, we will show later in this chapter that the basic performance measures obtained from this approximate model are close to those obtained from simulations. The presented model is for the complete partitioning case. However, the model for complete sharing can be developed in a similar way.

We assume that the PDU arrival process is Poisson with average rate λ_{BE}. The maximum bandwidth that can be allocated to the BE queue is denoted by $B = b_{max} - b_{ugs}$. The transition matrix $\mathbf{Q}$ for this model can be obtained as in (2.41).

$$\mathbf{Q} = \begin{bmatrix} q_{0,0} & \cdots & q_{0,A} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ q_{B,0} & \cdots & \cdots & q_{B,B+A} & & \\ & \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{2.41}$$

Note that since this matrix $\mathbf{Q}$ is used to represent the number of packets in the BE queue which is infinite, the structure of $\mathbf{Q}$ is different from $\mathbf{P}$ in (2.13).

Element $q_{y,y'}$ indicates the probability that the BE queue has y PDUs during the current frame time and it becomes y' in the next frame time. To obtain this probability, we calculate the probability of departure of a PDU from the BE queue based on the number of PDUs in the PS queue as follows:

$$k_n = \sum_{x=\psi_b}^{\psi_{b+1}-1} \left(\sum_{s=1}^{SN} \pi^{(CP)}(s,x) \right) \tag{2.42}$$

for $n = B - b$, $b \in \{0, 1, \cdots, B\}$ and zero otherwise. Then, each element $q_{y,y'}$ is obtained as follows:

$$q_{y,y-g} = \sum_{\{n,a|n-a=g\}} f_a(\lambda_{BE}) k_n \tag{2.43}$$

$$q_{y,y+h} = \sum_{\{n,a|a-n=h\}} f_a(\lambda_{BE}) k_n \tag{2.44}$$

$$q_{y,y} = \sum_{\{n,a|n=a\}} f_a(\lambda_{BE}) k_n \tag{2.45}$$

for $g = 1, 2, \cdots, G$ and $h = 1, 2, \cdots, A$, where $G = \min(B, y)$. Note that (2.43), (2.44), and (2.45) represent the transition probability matrices for the cases when the number of PDUs in the queue decreases by g, increases by h, and does not change, respectively.

Since the size of matrix $\mathbf{Q}$ is infinite, we apply the *matrix-geometric method* [34] to obtain the steady state probabilities. For this, we re-block matrix $\mathbf{Q}$ to obtain the transition probability matrix in the following form:

$$\mathbf{Q} \begin{bmatrix} \mathbf{K} & \mathbf{L} & & & \\ \mathbf{M} & \mathbf{N}_1 & \mathbf{N}_0 & & \\ & \mathbf{N}_2 & \mathbf{N}_1 & \mathbf{N}_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \tag{2.46}$$

When the stability condition, namely, $\boldsymbol{\delta}\mathbf{N}_2\mathbf{1} > \boldsymbol{\delta}\mathbf{N}_0\mathbf{1}$, where $\boldsymbol{\delta} = \boldsymbol{\delta}\mathbf{N}$, $\boldsymbol{\delta}\mathbf{1} = 1$, and $\mathbf{N} = \mathbf{N}_0 + \mathbf{N}_1 + \mathbf{N}_2$ is satisfied, then the matrix $\mathbf{R}$, which is the minimal non-negative solution of $\mathbf{R}\mathbf{N}_0 + \mathbf{R}\mathbf{N}_1 + \mathbf{R}^2\mathbf{N}_2$, can be determined such that $\boldsymbol{\zeta}_{i+1} = \boldsymbol{\zeta}_i\mathbf{R}$ where $\boldsymbol{\zeta}_i$ contains steady state probabilities corresponding to the number of PDUs in the BE queue. This matrix $\mathbf{R}$ can be obtained iteratively from

$$\mathbf{R}(k+1) = \mathbf{N}_0 + \mathbf{R}(k)\mathbf{N}_1 + \mathbf{R}^2(k)\mathbf{N}_2 \tag{2.47}$$

until $|\mathbf{R}(k+1) - \mathbf{R}(k)|_{i,j} < \epsilon$, $\forall i, j$ (e.g., $\epsilon = 10^{-9}$). Next, we calculate $\boldsymbol{\zeta}_0$ and $\boldsymbol{\zeta}_1$ by solving the following equations:

$$\mathbf{B}[\mathbf{R}] = \begin{bmatrix} \mathbf{K} & \mathbf{L} \\ \mathbf{M} & \mathbf{N}_1 + \mathbf{R}\mathbf{N}_2 \end{bmatrix}, \quad [\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1] = [\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1]\,\mathbf{B}[\mathbf{R}] \tag{2.48}$$

$$\boldsymbol{\zeta}_0\mathbf{1} + \boldsymbol{\zeta}_1(\mathbf{I} - \mathbf{R})^{-1}\mathbf{1} = 1. \tag{2.49}$$

Since $\boldsymbol{\zeta}_i$ consists of $A-1$ states of different number of PDUs in the BE queue, the steady state probability of y PDUs in the BE queue $\zeta(y)$ can be extracted as follows:

$$\zeta(y) = [\boldsymbol{\zeta}_i]_{col(i,y)}, \quad \text{where} \quad col(i,y) = i(A-1) + y + 1. \tag{2.50}$$

In this case, the calculation needs to be truncated at Y_t PDUs such that $1 - \sum_{y=0}^{Y_t} \zeta(y) < \epsilon$.

Then, the average number of PDUs in the BE queue $\overline{y}_{BE}$ and the average delay $\overline{d}_{BE}$ for a PDU in the BE queue can be simply obtained from

$$\overline{y}_{BE} = \sum_{y=0}^{Y_t} y\zeta(y), \quad \overline{d}_{BE} = \frac{\overline{y}_{BE}}{\lambda_{BE}}. \tag{2.51}$$

2.6 Performance Evaluation

2.6.1 Parameter Setting

We consider a TDMA/TDD-based uplink transmission scenario from a particular SS to the BS. The SS under consideration is stationary and works in GPSS mode. The communication between SS and BS uses *rate ID 0* [35] (i.e., QPSK with code rate 1/2). The PDU arrival process for each PS connection is assumed to be identical and follows a two-state MMPP model (i.e., $S = 2$) with the following parameters

$$\mathbf{U}(i) = \begin{bmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \end{bmatrix}, \boldsymbol{\Lambda}(i) = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 2.2 \end{bmatrix}, \quad i = 1, \cdots, N \tag{2.52}$$

where α indicates the traffic intensity, and the maximum batch size of PDU arrival is 20 (i.e., $A = 20$). When $\alpha = 1$, the average PDU arrival rate of this dMMPP connection is $\rho(i) = 1.9818$. The PDUs from all PS connections are aggregated into the PS queue, and the size of this queue is assumed to be 100 PDUs (i.e., $X = 100$). The transmitter serves the PS queue in a first-in-first-out fashion.

In our performance evaluation, we use $\alpha = 1.5$, bandwidth allocated to SS is 12 units (i.e., $b_{max} = 12$), the number of connections under PS is 2 (i.e., $N = 2$), and probability of successful transmission of a PDU is 0.995 (i.e., $\theta = 0.005$). For UGS

traffic, we use a 3-state on-off source with the transition matrix defined as follows:

$$\mathbf{V} = \begin{bmatrix} 0.3 & 0.7 & 0.0 \\ 0.2 & 0.5 & 0.3 \\ 0.0 & 0.5 & 0.5 \end{bmatrix}. \tag{2.53}$$

Therefore, this source requires bandwidth of 1.1667 units on average and b_{ugs} is set to 2. Note that we vary some of these parameters according to the evaluation scenarios, while the rest remain fixed according to the aforementioned setting.

For queue-aware bandwidth allocation, we consider sets of thresholds which are uniformly located over the range of buffer size. We use the notation $e_1 : (e)$ for the set of thresholds $\Psi\{e_1, e_1 + e, e_1 + 2e, \cdots, e_1 + (b_{max} - b_{ugs})e\}$. For example, $1 : (e = 5)$ represents the set $\Psi = \{1, 6, 11, 16, 21, 26\}$, where $(b_{max} - b_{ugs}) = 6$. For queue-aware rate control, we assume that the minimum guaranteed rate is a function of the original rate and is defined as follows: $\lambda_{min} = \lambda_o/2$. Also, we assume $\mathcal{F}(\lambda_o, x) = \lambda_o - \frac{\lambda_o(x-\tau_{min})}{2(\tau_{max}-\tau_{min})}$.

2.6.2 Simulation Environment

A time-driven simulator, developed in *MATLAB*, is used to evaluate the performance of the proposed queue-aware uplink bandwidth allocation and rate control algorithms and also to validate the correctness of the analytical model.

The PDU arrival and transmission events occur on a time-slotted fashion in which the length of a time-slot is equal to one frame interval. In the simulator, information on the states of a multistate on-off source (i.e., dMMPP for each connection) is maintained separately for each connection. The number of PDUs in the PS queue is calculated by considering the number of incoming PDUs for every time slot according to the state of dMMPP sources. In one time slot, the amount of bandwidth allocated to UGS service is determined based on the state of multistate on-off source. The remaining bandwidth is allocated to PS. Then, according to the threshold for bandwidth allocation setting (i.e., Ψ), the bandwidth left from PS will be allotted to BE service.

For queue-aware rate control of traffic arriving into the PS queue, some of the arriving PDUs are randomly blocked so that the arrival rate for the PS connections conforms to the desired setting (i.e., $\tilde{\lambda}(x, \lambda_o, \lambda_{min})$).

An independent packet error process is simulated for each wireless transmission. We replicate each simulation 10 times and for each replication the length of the simulation time is 200,000 time slots. We obtain the performance results for the bandwidth allocation (i.e., CP and CS schemes) and rate control scheme under varying traffic intensity, different settings of the bandwidth adaptation thresholds, and different rate control thresholds. Also, the performances of the proposed queue-aware bandwidth allocation schemes are compared with those of static allocation.

2.6.3 Numerical and Simulation Results

2.6.3.1 Queue-Length Distribution and Average Delay

Typical results on queue length distributions and average queueing delay for both the CS and the CP cases are shown in Fig. 2.3. As expected, the length of the PS queue grows with the number of connections. Also, since the PS traffic can use the unused bandwidth from UGS traffic (e.g., when the multistate on-off source is in the off state), with the same number of PS connections, the queue length for the CS scheme is smaller than that for the CP case. However, for a small number of PS connections (e.g., $N = 2$), the queue-length distributions become very close to each other (Fig. 2.3(a)) since the transmission rate is high enough to accommodate arriving PDUs. We observe that the simulation results follow the analytical results very closely, which confirms the correctness of the analytical model.

The average delay increases with the number of PS connections (Fig. 2.3(b)). The average delay of the CS scheme is better than that of the CP scheme since with CS the bandwidth which is not used by UGS will be yielded to polling service. We observe that when the traffic intensity is low and the number of PS connections is few, average delay remains constant since the transmission rate is high enough so that the delay remains constant over a range of values of traffic intensity.

However, when the traffic intensity reaches a certain point, which we call a *critical rate*, average delay increases rapidly to the maximum delay. This steep rise occurs when the queue status changes from stable to unstable since the PDU arrival rate becomes larger than the service rate.

As expected, the average PDU transmission delay for BE traffic increases as the PDU arrival rate at the PS queue increases (Fig. 2.4). With higher PDU arrival rate,

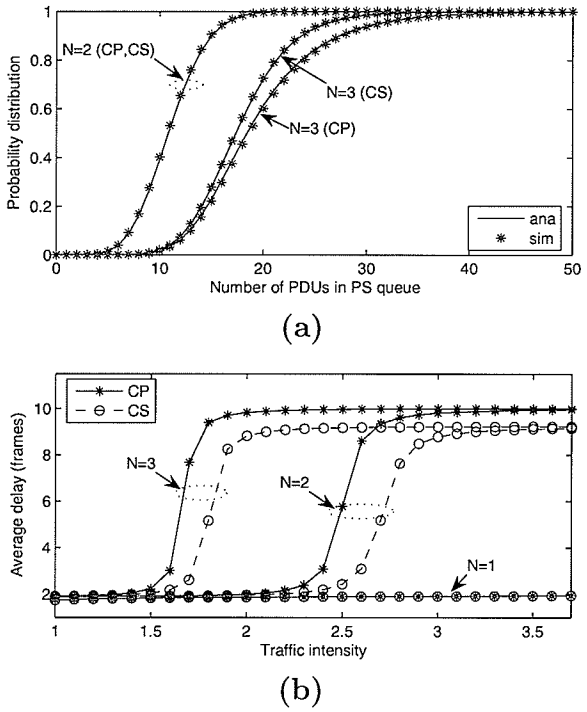


Figure 2.3. *(a) Queue distribution and (b) average delay for the PS queue.*

since the PS queue requires more transmission bandwidth, the bandwidth allocated to the BE queue becomes smaller. Again, the simulation results closely follow the numerical results.

2.6.3.2 Performance of Queue-Aware Dynamic Bandwidth Allocation

The probability distributions for the allocated bandwidth to PS under different settings of the bandwidth adaptation thresholds (i.e., e) are shown in Fig. 2.5. As is evident, the distribution with smaller e results in higher variance than that with larger e. The higher variance indicates more fluctuations in the allocated bandwidth for PS.

Fig. 2.6(a) illustrates how the different threshold settings for dynamic bandwidth adaptation impacts the average delay for the PDUS in the PS queue. Specifically, larger e leads to higher average delay when the traffic intensity is low. The results for static bandwidth allocation are also shown for comparison.

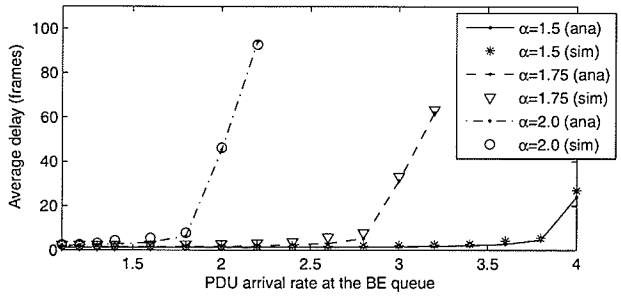


Figure 2.4. *Average delay for the BE queue.*

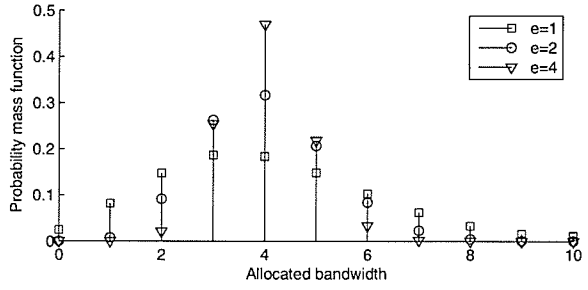


Figure 2.5. *Probability mass function for allocated bandwidth under different threshold settings.*

With static allocation, delay at low traffic intensity is always one. The critical rate and the maximum average delay (i.e., average delay when the queue becomes unstable) depend on the amount of allocated bandwidth (i.e., b_{max}) to the SS. Interestingly, the proposed bandwidth allocation scheme with complete partitioning can maintain constant delay when the traffic intensity is low, and the critical rates as well as the maximum average delay are equal to those for the case of static allocation when the traffic intensity is high. In case of complete sharing, the PS queue benefits from the off periods in the multistate on-off source, and therefore, the critical rate is higher and the maximum average delay is lower (e.g., 2.75 and 9 frames, respectively, in Fig. 2.6(a)). Also, we observe that the queue-aware allocation always achieves 100% utilization of the bandwidth (Fig. 2.6(b)).

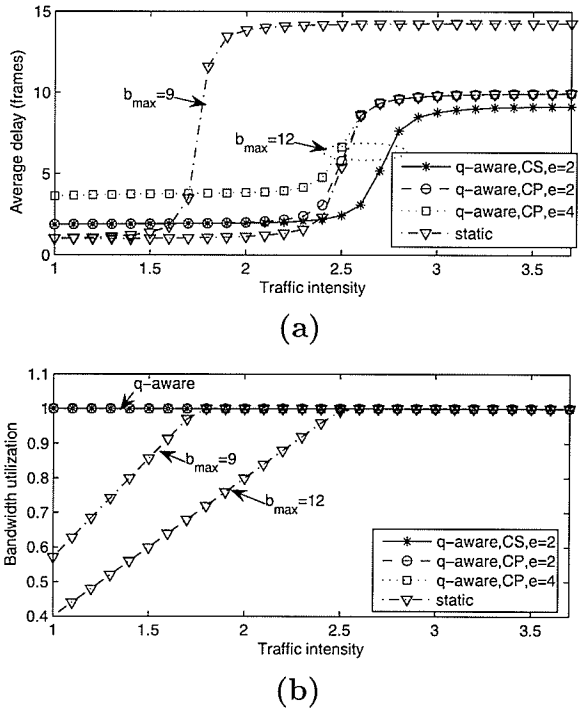


Figure 2.6. *Variations in (a) average delay and (b) bandwidth utilization under varying traffic intensity.*

Note that while selecting the thresholds for dynamic bandwidth allocation, the value of e should not be too large so that the average delay can be kept small, and

again, it should not be too small so that the high variability in the allocated bandwidth to the PS queue can be avoided (Fig. 2.6(a) and Fig. 2.5). The desired setting can be determined by using the analytical model.

2.6.3.3 Performance of the Queue-Aware Rate Control Scheme

Fig. 2.7(a) shows typical variations in the controlled PDU arrival rate for three different connections when the traffic intensity (per connection) increases. In this case we set $\tau_{max} = 70$ and vary τ_{min}. With variation in traffic intensity, the controlled arrival rate decreases when the queue length becomes larger than the threshold τ_{min}. However, according to the modeling assumption, the controlled arrival rate can not be reduced below the minimum guaranteed rate which is half of the traffic intensity in this case. This explains the "ripple"-like behavior of the controlled arrival rate. Note that the threshold settings determine the values of the traffic intensity at which the slopes of the envelope of the controlled arrival rate change and the minimum guaranteed rate is achieved.

Typical variation in average delay under different rate control threshold settings for the PDUs in the PS queue is shown in Fig. 2.7(b). Even though the average delay increases with increasing traffic intensity, due to rate control, the average delay does not approach maximum delay very rapidly as in the case without rate control. However, as the traffic intensity increases to a certain point (e.g., $\lambda = 5.5$ in Fig. 2.7(b)) there is no difference between any rate control threshold setting since the traffic sources reach their minimum guaranteed rates. Therefore, the average delay is close to the maximum delay which indicates that the queue is full most of the time. Also, smaller τ_{min} results in lower delay since the PDU arrival rate is controlled earlier compared to the case with larger τ_{min}.

2.6.3.4 Transient Analysis

For transient analysis of the QoS performances of the adaptive bandwidth allocation and rate control schemes, we assume that the PS queue is empty at time zero (i.e., $\pi_{tr}(0) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$). We vary the number of PS connections during different time periods (e.g., $N = 3, 4, 5, 6, 7, 5$ during time periods 1-40, 41-80, 81-120, 121-160, 161-200, and 201-240, respectively, in Fig. 2.8). We consider the complete partitioning

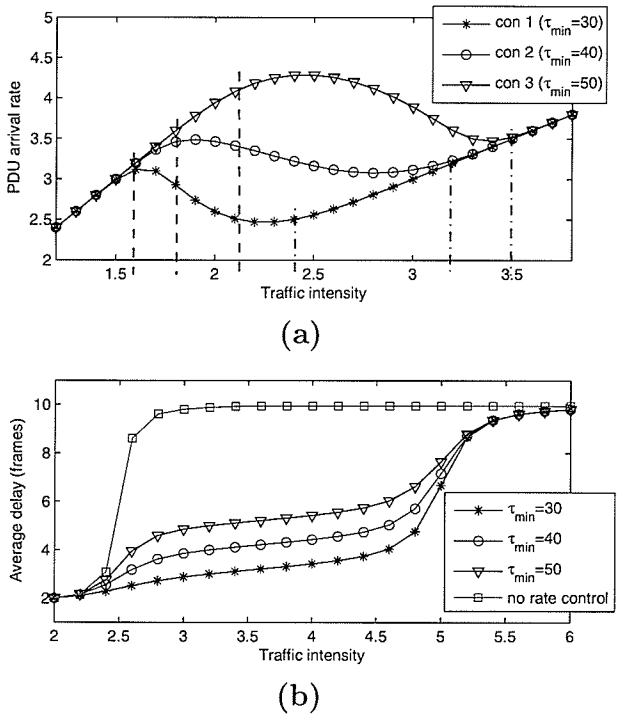


Figure 2.7. *Variations in (a) controlled PDU arrival rate for a PS connection and (b) average delay under different rate control threshold settings.*

case here with $b_{max} = 12$ and set the traffic intensity parameter to one (i.e., $\alpha = 1$).

Typical variations in queue length, amount of allocated bandwidth, controlled PDU arrival rate, and average delay with time are shown in Fig. 2.8. For controlled PDU arrival rate, we observe only first three connections each of which has a different threshold settings (i.e., $\tau_{min} = 30, 40, 50$ and $\tau_{max} = 70$). For the other connections, we assume $\tau_{min} = 40$ and $\tau_{max} = 70$.

The PS queue length increases asymptotically (towards the average number of PDUs at steady state) with increasing number of PS connections (Fig. 2.8(a)). With the queue-aware bandwidth allocation, when the number of PS connections becomes more than five (so that the sum of PDU arrival rates becomes larger than b_{max}), the allocated bandwidth reaches the maximum available bandwidth at which point the transmission rates for the connections are controlled (as shown in Fig. 2.8(b)). In this case, since different connections have different rate-control threshold settings,

the arrival rate is controlled differently for each conection.

We observe from Fig. 2.8(b) that, when the number of connections is less than five, the average delay remains constant. However, when the queue becomes unstable, average delay is less than maximum average delay since the arrival rate for each of the connections is controlled. Note that the discontinuities in the variation in average delay are due to the change in the number of PS connections which results in a sharp change in the PDU arrival rate into the queue. This causes transient variations in the amount of bandwidth allocation. Since the durations of these discontinuities are typically only a few frame intervals, the impact on overall performance would be negligible.

2.7 Chapter Summary

We have presented a queue-aware adaptive uplink bandwidth allocation and rate control mechanism for polling service in WiMAX broadband wireless access networks. This scheme is designed for a WiMAX SS. By utilizing the queue state information, the proposed mechanisms can maintain the packet-level QoS performances at the desired level. We have presented a comprehensive queueing analytical model to investigate the performances of the proposed schemes in both steady and transient states. An approximation model for the best-effort queue has been presented. Performance evaluation of the proposed radio resource management model has been carried out extensively which reveals the inter-relationships among the different performances measures. The correctness of the analytical model has been validated by simulations. Part of this chapter has been published in [9].

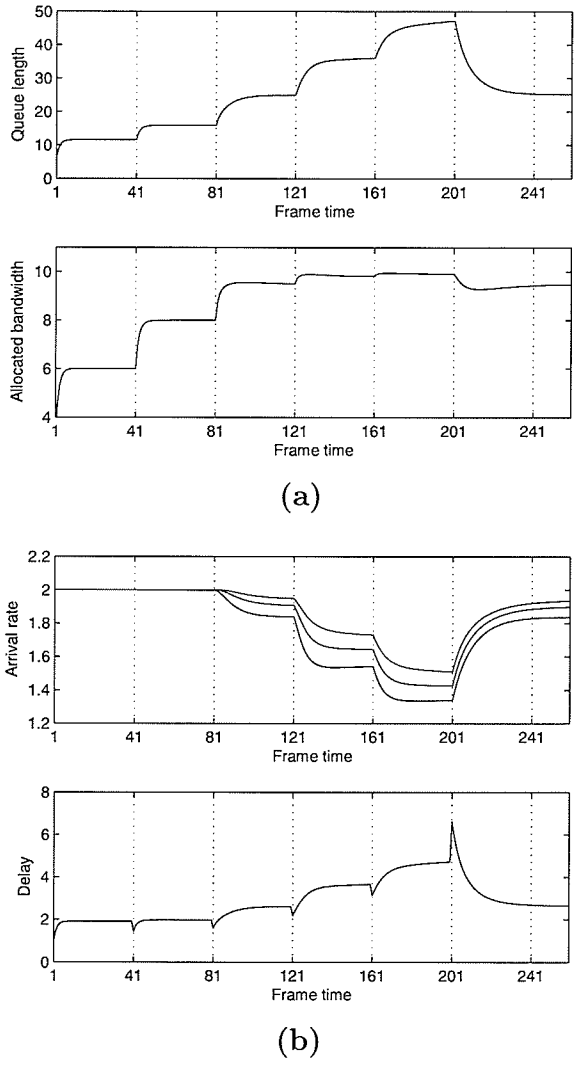


Figure 2.8. *(a) Queue length and allocated bandwidth for PS queue and (b) controlled arrival rate and average delay obtained from transient analysis.*

Chapter 3

Radio Resource Management in WiMAX: Part II

3.1 Introduction

3.1.1 Problem Statement

In this chapter, the problem of radio resource management framework for the WiMAX base station (BS) is considered. This BS accommodates multiple connections with different types (i.e., UGS, rtPS, nrtPS, and BE service). UGS connection requires fixed bandwidth for its transmission. rtPS and nrtPS connections have delay and throughput requirements. The first two inputs of this framework are the traffic description and the QoS requirement of the connection. The third input is the channel quality of the connection from the SS to BS. The first output of this framework is the decision on the accepting or rejecting new connection. If the connection is accepted, the second output of this framework is the amount of allocated bandwidth to the accepted connection. The objective of the framework is to maximize the satisfaction of the connections. The QoS requirements in both connection-level (e.g., new connection blocking probability) and packet-level (e.g., delay and throughput) must be satisfied.

3.1.2 Contribution

A joint bandwidth allocation (BA) and connection admission control (CAC) framework is proposed which can guarantee both the packet-level and the connection-level

QoS requirements for the different types of services (i.e., bandwidth, delay and transmission rate for UGS, rtPS and nrtPS, respectively, and connection blocking probability for BE service), and thereby, maximizes the system utility while at the same time maximizes the system revenue. We propose two approaches, namely, the optimal and the iterative approaches for joint bandwidth allocation and connection admission control. For the optimal approach, an assignment problem is formulated and solved by using binary integer linear programming. However, this optimal approach incurs a huge computational complexity, and therefore, may not be suitable for on-line execution. On the other hand, the iterative approach, which is based on the water-filling method, is an implementation-friendly one.

To analyze the connection-level performances (i.e., connection blocking probability and average number of ongoing connections), a queueing model is developed assuming a complete partitioning of the bandwidth resources among the different types of services. The optimal values of the partitioning thresholds are obtained by solving an optimization formulation with an objective to maximizing average system revenue under connection-level QoS (e.g., connection blocking probability) constraint. Note that this optimization formulation can be solved off-line (e.g., by an enumeration method) to obtain the optimal thresholds which are used in the joint BA and CAC algorithm. Another queueing analytical model is developed to analyze the packet-level performance measures (e.g., PDU dropping probability, delay statistic and throughput) for a connection in a particular service category under adaptive modulation and coding as specified in the WiMAX standard. Based on the queueing and the optimization models, performances of the proposed radio resource management approaches are evaluated and the analytical results are validated through extensive simulations.

3.2 Related Work

In [24], a bandwidth allocation and admission control scheme was proposed for TDMA and FDMA-based cellular wireless systems in which the amount of allocated bandwidth to an ongoing call is dynamically varied depending on the traffic load to accommodate more number of calls so that the call blocking probability can be minimized. Most of the CAC algorithms for cellular wireless networks proposed in the literature

analyzed the connection-level performances only without considering the packet-level performances resulting from the different radio resource management schemes. For example, in [36], a rate adaptation and admission control method was proposed for CDMA systems which maintains the signal-to-interference ratio (SIR) at the receiver at a target level while maximizing the transmission rate. To guarantee packet-level QoS, the CAC mechanism should take the packet-level performances into account. A QoS-aware scheduling scheme for WiMAX networks was presented in [37]. Specifically, a deficit fair priority queue scheduling algorithm was used to provide services to different types of flows in both uplink and downlink directions. The bandwidth allocation framework was organized in a hierarchical structure so that wireless resources can be managed efficiently, and QoS requirements can be met. An admission control strategy based on the available bandwidth was proposed as well.

For wireless mobile networks, the problem of providing packet-level QoS was studied quite extensively in the literature. A scheduling mechanism for downlink transmission was proposed in [38] to provide delay guarantee. In [39], a dynamic fair resource allocation scheme was proposed to support real-time and non-real-time traffic in cellular CDMA networks. Based on the principle of generalized processor sharing (GPS), the proposed traffic scheduler assigns rate and power resources to the mobiles according to their weights. Performances of the proposed scheme in terms of fairness and packet-level QoSs were evaluated in this chapter. An adaptive cross-layer scheduler was proposed in [40] for multiclass data services in wireless networks. The proposed scheduler uses the queueing information as well as takes the physical layer parameters into account so that the required QoS performances can be achieved. The capacity of TDMA and CDMA-based broadband cellular wireless systems was derived in [41] under constrained packet-level QoS.

3.3 System Model

We consider a single BS serving multiple connections (from SSs) through a TDMA/TDD access mode using single carrier modulation (e.g., as in WirelessMAN-SC). For each of the rtPS and nrtPS connections, a separate queue (with size of X PDUs) is used for buffering the PDUs (as shown in Fig. 3.1). In particular, for one connection there

are two queues for uplink and downlink transmissions from the SS and the BS, respectively. We consider an SS of type GPC. Therefore, during bandwidth allocation and connection admission control, a certain amount of bandwidth is reserved for each connection through that SS.

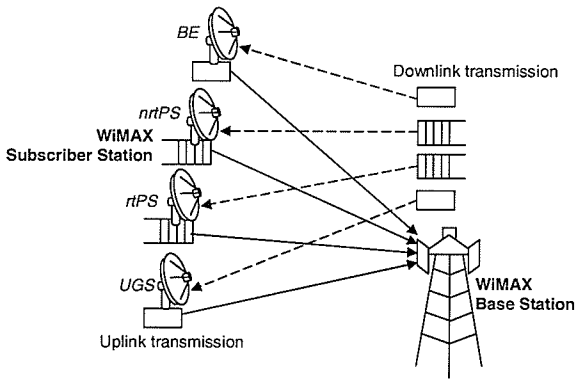


Figure 3.1. *WiMAX system model.*

The key notations are listed in Table. 3.1.

Adaptive modulation and coding is used to adjust the transmission rate adaptively in each frame according to the channel quality. The joint bandwidth allocation and admission control algorithm is executed at the BS in a centralized manner.

The radio resource management model with the proposed joint bandwidth allocation and admission control framework is shown in Fig. 3.2. In this model, there are two levels of optimization - one at the connection-level and the other at the packet-level. While the connection-level optimization is used to obtain the optimal setting for the complete partitioning thresholds for bandwidth allocation under connection-level QoS constraints, the packet-level optimization is used to allocate the available bandwidth among the ongoing and the newly arriving connections (when admitted) so that the corresponding packet-level QoS requirements can be satisfied. In this chapter, bandwidth b is defined as the number of PDUs that can be transmitted in one frame using rate ID $n = 0$. Queueing analytical models are used to estimate the average amount of allocated bandwidth per connection and to calculate connection-level and packet-level performance measures.

Table 3.1. *List of key notations.*

X	Queue size
b	Amount of bandwidth assigned to a connection
$b_{UGS}^{(req)}$, $d^{(req)}$, $\tau^{(req)}$	QoS requirements (i.e., bandwidth unsolicited granted service (UGS), average delay for real-time polling service (rtPS) and average throughput for non-realtime polling-service (nrtPS))
λ	PDU arrival rate
$\overline{\gamma}$	Average SNR of the receiver
g, h	Parameters of the Sigmoid utility function
C, M	Total bandwidth, total number of connections of subscriber station
$\mathbb{C}$	A set of connections of different services
$\mathcal{T}$	Amount of bandwidth reserved for different services
$\mathbf{F}$, $\mathbf{X}$	Column matrix of cost function, column matrix of bandwidth assignment
$\mathbf{G}$, $\mathbf{H}$, $\mathbf{J}$, $\mathbf{K}$	Matrices of the constraints of bandwidth assignment
$\mathbf{P}$	Probability transition matrix of the queue
$\boldsymbol{\pi}_t$, $\boldsymbol{\pi}_a$	Steady state probability of the connections and batch Markovian arrival process
P_{bl}, $\overline{c}$	Connection blocking probability, average number of ongoing connections
$\mathcal{R}$	System revenue from different services
τ	Average transmission rate
N	The total number of states of FSMC
$\boldsymbol{\pi}_\zeta$, $\boldsymbol{\pi}$	Steady state probability of FSMC and queue
$\overline{x}$, P_{drop}	Average number of PDUs in queue, PDU dropping probability
η, $\overline{w}$	Queue throughput, average queueing delay
$\tilde{b}$	Average allocated bandwidth per connection

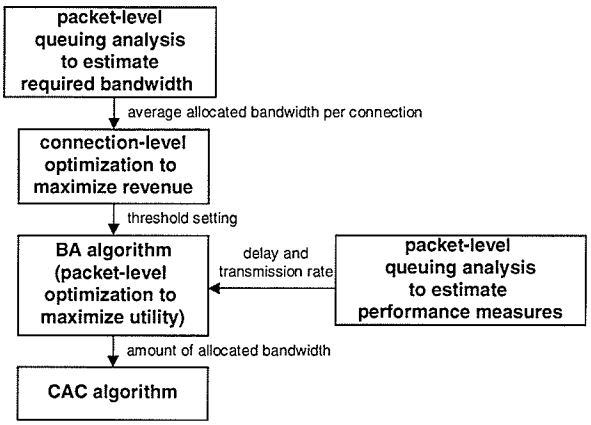


Figure 3.2. *Radio resource management model with the proposed joint bandwidth allocation and admission control.*

3.4 Radio Resource Management Framework for Joint Bandwidth Allocation (BA) and Connection Admission Control (CAC)

3.4.1 Methodology and System Parameters

The objective of the joint bandwidth allocation and connection admission control framework is to allocate the available bandwidth in a cell among the connections from different SSs and make decision on the admission of newly arriving connections such that the QoS requirements (i.e., bandwidth $b_{UGS}^{(req)}$, average delay $d^{(req)}$, average transmission rate $\tau^{(req)}$ for UGS, rtPS and nrtPS, respectively) for both uplink and downlink transmissions can be met. Also, since users' utility (i.e., level of satisfaction) should be maximized, an optimization problem is formulated and solved to obtain the amount of allocated bandwidth for all of the ongoing and the newly arriving connections (assuming that they are admitted into the system). Admission control decision is made based on the results of the optimization formulation. Specifically, a new connection is admitted if, upon admission of that connection, the QoS requirements of all the connections can be satisfied. Note that since the optimal approach incurs exponential time complexity, we present an iterative approach to obtain the solution which incurs significantly less computational complexity.

The bandwidth allocation is performed based on the in-connection level queueing performances of the rtPS and nrtPS connections. For an rtPS connection, the average delay requirement for uplink and downlink transmission is denoted by $d_i^{(up,req)}$ and $d_i^{(do,req)}$, respectively. Similarly, for an nrtPS connection, $\tau_i^{(up,req)}$ and $\tau_i^{(do,req)}$ denote the transmission rate (i.e., maximum queue throughput) requirements for uplink and downlink transmissions, respectively.

When a new connection arrives, the BS is informed of the traffic source descriptor (e.g., PDU arrival rates $\lambda_i^{(up)}$ and $\lambda_i^{(do)}$[1] for uplink and downlink, respectively) and the QoS requirement (i.e., delay requirement and transmission rate requirement for rtPS or nrtPS connections, respectively). Then, the BS measures the channel quality (i.e., average SNR at the receiver, $\overline{\gamma}_i$) corresponding to that incoming connection. These parameters are provided to the bandwidth allocation module (in Fig. 3.2) to compute the required amount of bandwidth as well as the user's utility for the incoming connection. Also, re-allocation of bandwidth among the ongoing connections is performed if necessary. The results of this computation is used to decide whether the new connection can be admitted or not.

Note that when a connection terminates, the bandwidth allocation algorithm is invoked again to re-allocate the released bandwidth among the ongoing connections.

3.4.2 Utility Functions

We use utility functions to represent the level of users' satisfaction on the perceived QoS for the different service types. In the system under consideration, utility for connection i depends on the amount of allocated bandwidth, delay statistics (e.g., average delay), throughput and admission control decision for the UGS, rtPS, nrtPS, and BE connections, respectively. Specifically, we use the following functions to represent the utility for UGS and BE connections:

$$U_{UGS}(b_i) = \begin{cases} 1, & b_i \geq b_{UGS}^{(req)} \\ 0, & \text{otherwise} \end{cases}, \quad U_{BE}(b_i) = \begin{cases} 1, & b_i \geq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

The utility for a UGS connection is the highest (i.e., one) if the amount of allocated bandwidth (b_i) for connection i is higher than or equal to the required bandwidth

[1]We use (up) and (do) to denote variables for uplink and downlink, respectively.

$(b_{UGS}^{(req)})$ while the utility for a BE connection is the highest if the connection is admitted into the network. For rtPS and nrtPS connections, we use the modified sigmoid function [42] to obtain utility as a function of the packet-level performance measures. The utility for rtPS and nrtPS connections (in both uplink and downlink) can be expressed as functions of the allocated bandwidth as follows:

$$U_{rtPS}(b_i) = 1 - \frac{1}{1 + \exp\left(-g_{rt}(d(\overline{\gamma}, \lambda, b_i) - d_i^{(req)} - h_{rt})\right)}$$

$$U_{nrtPS}(b_i) = \frac{1}{1 + \exp\left(-g_{nrt}(\tau(\overline{\gamma}, \lambda, b_i) - \tau_i^{(req)} - h_{nrt})\right)}$$

where $d(\overline{\gamma}, \lambda, b)$ and $\tau(\overline{\gamma}, \lambda, b)$ denote average delay and transmission rate as functions of PDU arrival rate (λ) and average SNR ($\overline{\gamma}$) when the amount of allocated bandwidth is b. A sample plot of this sigmoid utility function $U_{nrtPS}(b_i)$ is shown in Fig. 3.3. These utility functions can be calculated based on the queueing analysis to be presented later in this chapter.

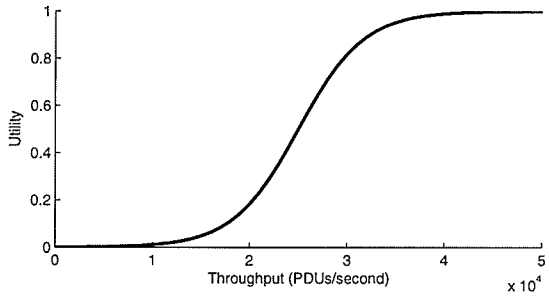


Figure 3.3. *Sigmoid utility function.*

Note that g_{rt}, g_{nrt}, h_{rt} and h_{nrt} are the parameters of the sigmoid function. Specifically, while g_{rt} and g_{nrt} determine the steepness (i.e., sensitivity of the utility function to delay or throughput requirement), h_{rt} and h_{nrt} represent the center of the utility function. From a service provider's perspective, these utilities represent the satisfaction level for the offered service. Therefore, the objective should be to maximize the sum of the utilities for all connections.

3.4.3 Optimization Formulation

We formulate the following optimization problem to allocate the total available bandwidth of C units among M (ongoing and incoming) connections:

$$
\begin{aligned}
\text{Maximize:} \quad & \sum_{i \in \mathbb{C}_{UGS}} \left(U_{UGS}(b_i^{(up)}) + U_{UGS}(b_i^{(do)}) \right) + \\
& \sum_{i \in \mathbb{C}_{rtPS}} \left(U_{rtPS}(b_i^{(up)}) + U_{rtPS}(b_i^{(do)}) \right) + \\
& \sum_{i \in \mathbb{C}_{nrtPS}} \left(U_{nrtPS}(b_i^{(up)}) + U_{nrtPS}(b_i^{(do)}) \right) + \\
& \sum_{i \in \mathbb{C}_{BE}} \left(U_{BE}(b_i^{(up)}) + U_{BE}(b_i^{(do)}) \right) \qquad (3.2) \\
\text{Subject to:} \quad & b_i^{(up)} = b_{UGS}^{(up,req)}, b_i^{(do)} = b_{UGS}^{(do,req)} \quad \text{for } i \in \mathbb{C}_{UGS} \\
& d(\overline{\gamma}_i, \lambda_i^{(up)}, b_i^{(up)}) \leq d_i^{(up,req)}, \\
& d(\overline{\gamma}_i, \lambda_i^{(do)}, b_i^{(do)}) \leq d_i^{(do,req)} \quad \text{for } i \in \mathbb{C}_{rtPS} \\
& \tau(\overline{\gamma}_i, \lambda_i^{(up)}, b_i^{(up)}) \geq \tau_i^{(up,req)}, \\
& \tau(\overline{\gamma}_i, \lambda_i^{(do)}, b_i^{(do)}) \geq \tau_i^{(do,req)} \quad \text{for } i \in \mathbb{C}_{nrtPS} \\
& b_i^{(up)} = b_i^{(do)} = 1 \quad \text{for } i \in \mathbb{C}_{BE} \\
& b_{min} \leq b_i^{(up)}, b_i^{(do)} \leq b_{max} \quad \forall i \\
& \sum_{\forall i} b_i^{(up)} + \sum_{\forall i} b_i^{(do)} \leq C \\
& \sum_{i \in \mathbb{C}_{UGS}} b_i \leq \mathcal{T}_{UGS}, \sum_{i \in \mathbb{C}_{rtPS}} b_i \leq \mathcal{T}_{rtPS}, \\
& \sum_{i \in \mathbb{C}_{nrtPS}} b_i \leq \mathcal{T}_{nrtPS}, \sum_{i \in \mathbb{C}_{BE}} b_i \leq \mathcal{T}_{BE}
\end{aligned}
$$

where $\mathbb{C}_{UGS}$, $\mathbb{C}_{rtPS}$, $\mathbb{C}_{nrtPS}$, and $\mathbb{C}_{BE}$ represent the sets of UGS, rtPS, nrtPS, and BE connections, respectively, b_{min} and b_{max} denote the minimum and the maximum amount of bandwidth that can be allocated to a connection, and $b_{UGS}^{(up,req)}$ and $b_{UGS}^{(do,req)}$ denote the bandwidth requirements for a UGS connection for uplink and downlink transmissions, respectively. The thresholds $\mathcal{T}_{UGS}$, $\mathcal{T}_{rtPS}$, $\mathcal{T}_{nrtPS}$, and $\mathcal{T}_{BE}$ represent the amount of bandwidth reserved for UGS, rtPS, nrtPS, and BE connections, respectively. The total available bandwidth is shared among the different services using a threshold-based complete partitioning approach. Note that prioritized band-

width allocation among the different types of services can be performed by setting these thresholds appropriately. An optimization-based scheme to obtain the optimal threshold setting under connection-level QoS constraints will be presented later in this chapter.

3.4.4 Optimal Approach for Bandwidth Allocation and Connection Admission Control

To solve the above optimization problem, we use binary integer programming by reformulating the problem as follows:

$$\begin{aligned} \text{Minimize:} \quad & \mathbf{F}^T\mathbf{X} \\ \text{Subject to:} \quad & \mathbf{GX} \leq \mathbf{H} \\ & \mathbf{JX} = \mathbf{K} \\ & [\mathbf{X}]_j \in \{0,1\} \quad \forall j \end{aligned} \tag{3.3}$$

where $\mathbf{F}$ is the column matrix of cost function which corresponds to negative value of users' utility, $\mathbf{X}$ is the column matrix of bandwidth assignment, $\mathbf{G}$ and $\mathbf{H}$ are the constraints on bandwidth, delay and transmission rate requirements for both uplink and downlink, and $\mathbf{J}$ and $\mathbf{K}$ represent the constraints on total amount of bandwidth and the thresholds to limit the amount of bandwidth allocated to each type of service. The solution of this binary integer programming formulation can be obtained by linear programming-based branch-and-bound algorithm [43]. Note that the gross upper bound time complexity of this algorithm is $O(2^{C^2(\Delta b)^2})$ [44] where $\Delta b = b_{max} - b_{min} + 1$. Note that the cost $\mathbf{F}$ is nonlinear function of the amount of bandwidth. Therefore, we transform this objective function in to the assignment of amount of bandwidth which is linear to the cost (i.e., negative of utility).

If the solution is feasible, matrix $\mathbf{X}$ will indicate the amount of bandwidth allocated to each connection. We can establish the bandwidth assignment matrix $\mathbf{Y}$ as follows:

$$\mathbf{Y} = \begin{bmatrix} [\mathbf{X}]_1 & \cdots & [\mathbf{X}]_{\Delta b} & \\ [\mathbf{X}]_{\Delta b+1} & \cdots & [\mathbf{X}]_{2\Delta b} & \\ \vdots & \vdots & \cdots & \vdots \\ [\mathbf{X}]_{((2M-1)\Delta b)+1} & \cdots & [\mathbf{X}]_{2M\Delta b} & \end{bmatrix} \tag{3.4}$$

where row i and column j of matrix $\mathbf{Y}$ correspond to a connection and the amount of bandwidth allocated to that connection, respectively. In particular, row i and $M+i$ correspond to uplink and downlink transmissions for connection i, respectively. The first and last columns of this matrix $\mathbf{Y}$ correspond to bandwidth b_{min} and b_{max}, respectively. The optimal amount of allocated bandwidth $\hat{b}_i$ to connection i is obtained from

$$\hat{b}_i = j + b_{min} - 1, \quad \text{if } [\mathbf{Y}]_{i,j} = 1 \tag{3.5}$$

where $[\mathbf{Y}]_{i,j}$ denotes the element at row i and column j of matrix $\mathbf{Y}$. Note that each row of $\mathbf{Y}$ contains only one non-zero element.

If the solution of the optimization problem is infeasible, there is no bandwidth allocation scheme for which the delay and the transmission rate requirements for the connections (upon admission of the new connection) can be satisfied. Therefore, the incoming connection is blocked; otherwise, the connection is accepted.

Since the joint BA and CAC algorithm is required to be executed in an on-line manner, the computational complexity of the above approach may be prohibitive from implementation point of view. Therefore, we propose an iterative approach which has less computational complexity, and therefore, more implementation-friendly.

3.4.5 Iterative Approach

The iterative approach (in Algorithm 1) is based on the water-filling method. In this case, the available bandwidth is first allocated to satisfy the target QoS (i.e., minimum bandwidth, delay and transmission rate) requirements. Then, the available bandwidth is allocated to the connection with the lowest utility such that the total utility can be increased in a fair manner. The algorithm terminates when either all available bandwidth is allocated or each of the connections receives the maximum possible bandwidth b_{max}. Note that the computational complexity of the algorithm is $O(C)$. This complexity can be determined from the Algorithm 1 for which the number of iterations depends on the total number of connections.

If the iterative algorithm is unable to find a feasible solution, the incoming connection is blocked; otherwise, the connection is accepted.

Algorithm 1 Iterative Algorithm

1: **for** $i \in \mathbb{C}_{UGS}$ **do**
2: $b_i^{(up)} \leftarrow b_{UGS}^{(up,req)}$, $b_i^{(do)} \leftarrow b_{UGS}^{(do,req)}$ // assign bandwidth to UGS connections first
3: **end for**
4: **for** $i \in \mathbb{C}_{rtPS}$ **do**
5: $b_i^{(up)} \leftarrow \min_b \left(d(\overline{\gamma}_i, \lambda_i^{(up)}, b) \leq d_i^{(up,req)}\right)$, $b_i^{(do)} \leftarrow \min_b(d(\overline{\gamma}_i, \lambda_i^{(do)}, b) \leq d_i^{(do,req)})$ // assign bandwidth to rtPS connections
6: **end for**
7: **for** $i \in \mathbb{C}_{nrtPS}$ **do**
8: $b_i^{(up)} \leftarrow \min_b(\tau(\overline{\gamma}_i, \lambda_i^{(up)}, b) \geq \tau_i^{(up,req)})$, $b_i^{(do)} \leftarrow \min_b(\tau(\overline{\gamma}_i, \lambda_i^{(do)}, b) \geq \tau_i^{(do,req)})$ // assign bandwidth to nrtPS connections
9: **end for**
10: **for** $i \in \mathbb{C}_{BE}$ **do**
11: $b_i^{(up)} \leftarrow 1$, $b_i^{(do)} \leftarrow 1$ // assign bandwidth to BE connections
12: **end for**
13: **if** $\left(\sum_{i\in\mathbb{C}_{UGS}} b_i > \mathcal{T}_{UGS}\right)$ or $\left(\sum_{i\in\mathbb{C}_{rtPS}} b_i > \mathcal{T}_{rtPS}\right)$ or $\left(\sum_{i\in\mathbb{C}_{nrtPS}} b_i > \mathcal{T}_{nrtPS}\right)$ or $\left(\sum_{i\in\mathbb{C}_{BE}} b_i > \mathcal{T}_{BE}\right)$ **then**
14: return (solution infeasible) // Reject new connection
15: **end if**
{// Allocate available bandwidth in order to maximize the utility}
16: $\mathbb{C}_{allo} \leftarrow \mathbb{C}_{UGS} \cup \mathbb{C}_{rtPS} \cup \mathbb{C}_{nrtPS} \cup \mathbb{C}_{BE}$
17: **while** $(\sum_i b_i \leq C)$ and $(b_i < b_{max} \ \exists i)$ and $(\mathbb{C}_{allo} \neq \emptyset)$ **do**
18: $i_{um} = \arg\min_i(U(b_i))$ // search for the connection with lowest utility
19: $b_{i_{um}} \leftarrow b_{i_{um}} + 1$ // increase bandwidth of that connection
20: **if** $(b_{i_{um}} == b_{max})$ **then**
21: $\mathbb{C}_{allo} \leftarrow \mathbb{C}_{allo} - i_{um}$ // discard that connection with maximum allocated bandwidth
22: **end if**
23: **if** $\left(\sum_{i\in\mathbb{C}_{UGS}} b_i == \mathcal{T}_{UGS}\right)$ **then**
24: $\mathbb{C}_{allo} \leftarrow \mathbb{C}_{allo} - \mathbb{C}_{UGS}$ // discard that connection if threshold for UGS service is reached
25: **end if**
26: **if** $\left(\sum_{i\in\mathbb{C}_{rtPS}} b_i == \mathcal{T}_{rtPS}\right)$ **then**
27: $\mathbb{C}_{allo} \leftarrow \mathbb{C}_{allo} - \mathbb{C}_{rtPS}$ // discard that connection if threshold for rtPS service is reached
28: **end if**
29: **if** $\left(\sum_{i\in\mathbb{C}_{nrtPS}} b_i == \mathcal{T}_{nrtPS}\right)$ **then**
30: $\mathbb{C}_{allo} \leftarrow \mathbb{C}_{allo} - \mathbb{C}_{nrtPS}$ // discard that connection if threshold for nrtPS service is reached
31: **end if**
32: **if** $\left(\sum_{i\in\mathbb{C}_{BE}} b_i == \mathcal{T}_{BE}\right)$ **then**
33: $\mathbb{C}_{allo} \leftarrow \mathbb{C}_{allo} - \mathbb{C}_{BE}$ // discard that connection if threshold for BE service is reached
34: **end if**
35: **end while**
36: return (feasible solution) // Accept new connection

3.5 Queueing Model for Connection-Level Performance Analysis

In order to obtain the connection-level performances (e.g., connection blocking probability and average number of connections) for each type of service, and subsequently, to obtain the optimal threshold settings for resource reservation under connection-level QoS constraints, we develop a queueing model[2]. Since the available bandwidth is to be shared based on a complete partitioning approach, we have

$$\mathcal{T}_{UGS} + \mathcal{T}_{rtPS} + \mathcal{T}_{nrtPS} + \mathcal{T}_{BE} = C. \tag{3.6}$$

For a given threshold $\mathcal{T}$ (i.e., $\mathcal{T} \in \{\mathcal{T}_{UGS}, \mathcal{T}_{rtPS}, \mathcal{T}_{nrtPS}, \mathcal{T}_{BE}\}$), the average amount of allocated bandwidth per connection $\tilde{b}$ in both uplink and downlink, connection arrival rate α, and connection holding time $1/\mu$ of a particular service type, a continuous-time Markov chain (CTMC) can be established for each service type to obtain the connection-level performance measures. Note that, the developed model can be applied to each service type separately as long as the condition in (3.6) holds.

The state space for this CTMC is $\Delta = \{\mathscr{M}; 0 \leq \mathscr{M} \leq \tilde{M}\}$ where $\mathscr{M}$ represents the number of ongoing connections in a particular service type and $\tilde{M} = \lfloor \mathcal{T}/\tilde{b} \rfloor$ is the maximum number of ongoing connections for threshold $\mathcal{T}$. The transition matrix of this Markov chain is defined as follows:

$$\mathbf{Q} = \begin{bmatrix} -\alpha & \alpha & & & \\ \mu & -\mu-\alpha & \alpha & & \\ \ddots & \ddots & \ddots & & \\ & c\mu & -c\mu-\alpha & \alpha & \\ & \ddots & \ddots & \ddots & \\ & & \tilde{M}\mu & -\tilde{M}\mu & \end{bmatrix}. \tag{3.7}$$

The steady state probability $\boldsymbol{\pi}_t$ of this Markov chain is obtained by solving $\boldsymbol{\pi}_t\mathbf{Q}\mathbf{0}$ and $\boldsymbol{\pi}_t\mathbf{1} = 1$ where

$$\boldsymbol{\pi}_t = \begin{bmatrix} \pi_t(0) & \cdots & \pi_t(c) & \cdots & \pi_t(\tilde{M}) \end{bmatrix}. \tag{3.8}$$

[2]Note that the threshold settings are required for the joint bandwidth allocation and admission control algorithm described before.

Note that $\mathbf{1}$ is a column matrix of ones and $\pi_t(c)$ represents the steady state probability that the number of ongoing connections is c. The connection blocking probability can be obtained from

$$P_{bl} = \pi_t(\tilde{M}) \tag{3.9}$$

and the average number of ongoing connections is calculated as follows:

$$\overline{c} = \sum_{c=1}^{\tilde{M}} c\pi_t(c). \tag{3.10}$$

We can formulate an optimization problem to maximize system revenue while the connection-level QoS (i.e., connection blocking probability) is maintained at the target level. By adjusting the thresholds $\mathcal{T}_{UGS}$, $\mathcal{T}_{rtPS}$, $\mathcal{T}_{nrtPS}$, and $\mathcal{T}_{BE}$, this objective can be achieved under given constraints. To calculate system revenue, we consider flat rate pricing in which the rates $\mathcal{R}_{UGS}$, $\mathcal{R}_{rtPS}$, $\mathcal{R}_{nrtPS}$, and $\mathcal{R}_{BE}$ apply for UGS, rtPS, nrtPS, and BE connections, respectively. In this case, the average number of ongoing connections and the connection blocking probability are defined as functions of the corresponding threshold (e.g., $\overline{c}_{UGS}(\mathcal{T}_{UGS})$ and $P_{bl}^{(UGS)}(\mathcal{T}_{UGS})$) and the optimization formulation can be expressed as follows:

$$\begin{aligned}
\text{Maximize:} \quad & \mathcal{R}_{UGS}\overline{c}_{UGS}(\mathcal{T}_{UGS}) + \mathcal{R}_{rtPS}\overline{c}_{rtPS}(\mathcal{T}_{rtPS}) + \\
& \mathcal{R}_{nrtPS}\overline{c}_{nrtPS}(\mathcal{T}_{nrtPS}) + \mathcal{R}_{BE}\overline{c}_{BE}(\mathcal{T}_{BE}) \\
\text{Subject To:} \quad & P_{bl}^{(UGS)}(\mathcal{T}_{UGS}) \leq \hat{P}_{bl}^{(UGS)} \\
& P_{bl}^{(rtPS)}(\mathcal{T}_{rtPS}) \leq \hat{P}_{bl}^{(rtPS)} \\
& P_{bl}^{(nrtPS)}(\mathcal{T}_{nrtPS}) \leq \hat{P}_{bl}^{(nrtPS)} \\
& P_{bl}^{(BE)}(\mathcal{T}_{BE}) \leq \hat{P}_{bl}^{(BE)}
\end{aligned} \tag{3.11}$$

where $\hat{P}_{bl}^{(UGS)}$, $\hat{P}_{bl}^{(rtPS)}$, $\hat{P}_{bl}^{(nrtPS)}$, and $\hat{P}_{bl}^{(BE)}$ are the target connection blocking probabilities for UGS, rtPS, nrtPS, and BE connections, respectively.

Note that the above optimization problem can be solved off-line and the threshold settings thus obtained could be used in the joint bandwidth allocation and admission control algorithm in an on-line fashion.

3.6 Queueing Analytical Model for Packet-Level Performance Analysis

3.6.1 Traffic Source and Arrival Probability Matrix

The PDU arrival process is modeled as a BMAP (Batch Markovian Arrival Process) [45]. The BMAP is associated with an S-state Markov chain and the probability of PDU arrival depends on the state of this Markov chain. A BMAP can be represented by matrix $\mathbf{A}_a$ (where $a \in \{0, 1, \ldots, A\}$) which is the transition probability matrix corresponding to arrival of a PDUs and A denotes the maximum arrival batch size. In particular, the element at row j and column j' of matrix $\mathbf{A}_a$ denotes the probability of arrival of a PDUs when the phase of the BMAP changes from j in the current frame period to j' in the next frame period. We can obtain the expected packet arrival probability from

$$\overline{\lambda} = \sum_{a=1}^{A} a\,(\boldsymbol{\pi}_a \mathbf{A}_a \mathbf{1}) \tag{3.12}$$

where $\boldsymbol{\pi}_a$ is obtained by solving $\boldsymbol{\pi}_a \mathbf{A} = \boldsymbol{\pi}_a$ and $\boldsymbol{\pi}_a \mathbf{1} = 1$, $\mathbf{A} = \sum_{a=0}^{A} \mathbf{A}_a$, and $\mathbf{1}$ is a column matrix of ones.

3.6.2 Channel Model and Transmission Probability Matrix

We consider a finite state Markov channel (FSMC) model which is a useful model for analyzing radio channel with non-independent fading (and hence bursty channel errors). A slowly varying Nakagami-m fading channel is represented by the FSMC model and each state of the FSMC corresponds to one transmission mode for AMC. With AMC, the SNR at the receiver is divided into multiple non-overlapping intervals (i.e., $N = 6$ denotes the highest rate ID in the WiMAX specifications) by thresholds Γ_n ($n \in \{-1, 0, 1, \ldots, N\}$)[3] where $\Gamma_{-1} = 0 < \Gamma_0 < \Gamma_1 < \ldots < \Gamma_{N+1} = \infty$. The channel is said to be in state n, if $\Gamma_n \leq \gamma < \Gamma_{n+1}$ (i.e., rate ID n will be used and I_n bits can be transmitted per symbol). To avoid possible transmission error, no PDU is transmitted when $\gamma < \Gamma_0$. Note that these thresholds correspond to the required SNR

[3]We use $n = -1$ to indicate the channel state when no transmission occurs (i.e., $I_{-1} = 0$).

at the receiver as specified in WiMAX standard, i.e., $\Gamma_0 = 6.4, \Gamma_1 = 9.4, \cdots, \Gamma_N = 24.4$. With Nakagami-$m$ fading, the probability of using rate ID n (i.e., $\Pr(n)$) is given by

$$\Pr(n) = \frac{\Gamma(m, m\Gamma_n/\overline{\gamma}) - \Gamma(m, m\Gamma_{n+1}/\overline{\gamma})}{\Gamma(m)} \tag{3.13}$$

where $\overline{\gamma}$ is the average SNR, m is the Nakagami fading parameter ($m \geq 0.5$), $\Gamma(m)$ is the Gamma function, and $\Gamma(m, \gamma)$ is the complementary incomplete Gamma function.

Assuming that the channel is slowly fading (i.e., transitions occur only between adjacent states), the state transition matrix for the FSMC can be expressed as follows [18]:

$$\boldsymbol{\zeta} = \begin{bmatrix} \zeta_{-1,-1} & \zeta_{-1,0} & & & & \\ \zeta_{0,-1} & \zeta_{0,0} & \zeta_{0,0} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \zeta_{N-1,N-2} & \zeta_{N-1,N-1} & \zeta_{N-1,N} & \\ & & & \zeta_{N,N-1} & \zeta_{N,N} & \end{bmatrix} \tag{3.14}$$

where each row of $\boldsymbol{\zeta}$ corresponds to a rate ID. Note that for the state corresponding to the row denoted by $n = -1$, since the SNR at the receiver is very low, to avoid possible transmision error, no PDU is transmitted (i.e., $I_n = 0,\ n < 0$).

Again, bandwidth b is defined as the number of PDUs that can be transmitted in one frame using rate ID $n = 0$ (i.e., with 0.5 bits per symbol). For a given amount of bandwidth and a transmission rate ID, the number of transmitted PDUs can be calculated from the number of information bits per symbol. For example, with $b = 1$, if rate ID $n = 0$, one PDU can be transmitted in one frame. if rate ID $n = 1$, two PDUs can be transmitted in one frame. Similarly, with the highest rate ID (i.e., $n = 6$), 9 PDUs (i.e., 2×4.5) can be transmitted in one frame. We assume that the channel condition for a connection (during both uplink and downlink transmissions) remains stationary over a frame interval ($\leq$ 2 ms) and all the PDUs corresponding to a connection transmitted during one frame period use the same rate ID.

We can define matrix $\mathbf{D}_k$ whose diagonal elements (at row $n+2$ and column $n+2$) correspond to the probability of transmitting k PDUs successfully during one frame when rate ID n is used. This matrix $\mathbf{D}_k$ can be defined as follows $[\mathbf{D}_k]_{n+2,n+2} = \theta_{2I_n b,k}$ where $k \in \{0, 1, \ldots, 2I_N b\}$, $[\mathbf{D}_k]_{j,j'}$ denotes the element at row j column j' of matrix

$\mathbf{D}_k$, and

$$\theta_{2I_nb,k} = \begin{pmatrix} 2I_nb \\ k \end{pmatrix} \theta^k (1-\theta)^{(2I_nb)-k} \tag{3.15}$$

for $k \leq 2I_nb$ and zero otherwise, where $2I_nb$ denotes the maximum number of transmitted PDUs and $\theta = 1 - PER_n$ is the probability that a PDU is successfully transmitted.

With b units of bandwidth, the average transmission rate for a connection can be obtained as follows:

$$\tau = \sum_{k=1}^{2I_nb} k \left(\boldsymbol{\pi}_\zeta \mathbf{D}_k \mathbf{1}\right) \tag{3.16}$$

where $\boldsymbol{\pi}_\zeta$ is obtained by solving $\boldsymbol{\pi}_\zeta \boldsymbol{\zeta} = \boldsymbol{\pi}_\zeta$ and $\boldsymbol{\pi}_\zeta \mathbf{1} = 1$.

3.6.3 State Space and Transition Matrix

For rtPS and nrtPS connections, the state of the queue is observed at the beginning of each frame. We assume that connection i is allocated with b_i units of bandwidth and a PDU arriving during frame period f will not be transmitted until frame period $f+1$ at the earliest. The state space of the queue for a tagged connection can be defined as follows:

$$\Phi = \{(\mathscr{X}, \mathscr{A}, \mathscr{F});\ 0 \leq \mathscr{X} \leq X, 1 \leq \mathscr{A} \leq S, -1 \leq \mathscr{F} \leq N\} \tag{3.17}$$

where $\mathscr{X}$, $\mathscr{A}$, $\mathscr{F}$ represent the number of PDUs in the queue, state of the BMAP, and channel state of FSMC, respectively. The transition matrix $\mathbf{P}$ for the queue can be expressed as follows:

$$\mathbf{P} = \left[\begin{array}{ccccccc}
\mathbf{p}_{0,0} & \cdots & \mathbf{p}_{0,A} & & & & \\
\vdots & \ddots & \ddots & \ddots & & & \\
\hline
\mathbf{p}_{U,0} & \cdots & \mathbf{p}_{U,U} & \cdots & \mathbf{p}_{U,U+A} & & \\
 & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & \mathbf{p}_{x,x-U} & \cdots & \mathbf{p}_{x,x} & \cdots & \mathbf{p}_{x,x+A} & \\
\hline
 & & \ddots & \ddots & \ddots & \ddots & \\
 & & & \mathbf{p}_{X,X-U} & \cdots & \mathbf{p}_{X,X} &
\end{array}\right]. \tag{3.18}$$

The rows of matrix $\mathbf{P}$ represent the number of PDUs in the queue and element $\mathbf{p}_{x,x'}$ inside this matrix denotes the transition probability for the case when the number of

PDUs in the queue changes from x in the current frame to x' in the next frame. This element $\mathbf{p}_{x,x'}$ also represents the transition of BMAP and FSMC states.

Since in one frame several PDUs can arrive and be transmitted, this matrix $\mathbf{P}$ is divided into three parts. The first part, from row 0 to $U-1$, where U is the maximum total PDU transmission rate with b_i units of bandwidth ($U = 2I_N b_i$ in our model), indicates the case that the maximum total transmission rate is greater than the number of PDUs in the queue and none of the incoming PDUs is dropped. The second part, from row U to $X-A$, represents the case in which the maximum PDU transmission rate is equal to or less than the number of PDUs in the queue and none of the incoming PDUs is dropped. The third part, from row $X-A+1$ to X, indicates the case that some of the incoming PDUs are dropped due to the lack of space in the queue. Let $\mathbf{D}_k^{(x)}$ denote the transmission probability when there are x PDUs in queue which can be obtained from

$$\mathbf{D}_k^{(x)} = \begin{cases} \mathbf{D}_k, & k < U' \\ \sum_{k=U'}^{U} \mathbf{D}_k, & k = U' \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{3.19}$$

where $U' = \min(x, U)$. Note that the maximum number of transmitted PDUs cannot be larger than the available number of PDUs in the queue.

The elements in the first and the second part of matrix $\mathbf{P}$ can be obtained as follows:

$$\mathbf{p}_{x,x-u} = \sum_{\{k,a|k-a=u\}} \mathbf{A}_a \otimes \left(\boldsymbol{\zeta} \times \mathbf{D}_b^{(x)}\right) \tag{3.20}$$

$$\mathbf{p}_{x,x+v} = \sum_{\{k,a|a-k=v\}} \mathbf{A}_a \otimes \left(\boldsymbol{\zeta} \times \mathbf{D}_b^{(x)}\right) \tag{3.21}$$

$$\mathbf{p}_{x,x} = \sum_{\{k,a|k=a\}} \mathbf{A}_a \otimes \left(\boldsymbol{\zeta} \times \mathbf{D}_b^{(x)}\right) \tag{3.22}$$

for $u = 1, \ldots, U'$ and $v = 1, \ldots, A$ where $k \in \{0, 1, 2, \ldots, U'\}$ and $a \in \{0, 1, 2, \ldots, A\}$ represent the number of departed PDUs and the number of PDU arrivals, respectively, and $\otimes$ denotes Kronecker product.

Considering both the PDU arrival and the PDU departure events, (3.20), (3.21), and (3.22) above represent the transition probability matrices for the cases when the

number of PDUs in the queue decreases by u, increases by v, and does not change, respectively.

The third part of matrix $\mathbf{P}$ ($\{x = X - A + 1, X - A + 2, \ldots, X\}$) has to capture the PDU dropping effect. Therefore, for $x + v \geq X$, (3.21) becomes

$$\mathbf{p}_{x,x+v} = \sum_{a=v}^{A} \hat{\mathbf{p}}_{x,x+a} \quad \text{for } x + v \geq X \tag{3.23}$$

and for $x = X$, (3.22) becomes

$$\mathbf{p}_{x,x} = \hat{\mathbf{p}}_{x,x} + \sum_{a=1}^{A} \hat{\mathbf{p}}_{x,x+a} \quad \text{for } x = X \tag{3.24}$$

where $\hat{\mathbf{p}}_{x,x'}$ is obtained for the case without any PDU dropping. Eqs. (3.23) and (3.24) indicate the case that the queue will be full if the number of incoming PDUs is greater than the available space in the queue. In other words, the transition probability to the state that the queue is full can be calculated as the sum of all the probabilities that make the number of PDUs in queue equal to or larger than the queue size X.

3.6.4 QoS Measures

To obtain the performance measures, the steady state probabilities for the queue would be required. Since the size of the queue is finite, the probability matrix $\boldsymbol{\pi}$ is obtained by solving the equations $\boldsymbol{\pi}\mathbf{P}\boldsymbol{\pi}$ and $\boldsymbol{\pi}\mathbf{1} = 1$, where $\mathbf{1}$ is a column matrix of ones. The matrix $\boldsymbol{\pi}$ contains the steady state probabilities corresponding to the number of PDUs in the queue, the state of the BMAP, and the channel state. The steady state probabilities $\pi(x, b, n)$ corresponding to the number of PDUs in queue is x, phase of BMAP is b and channel state is n can be extracted from a matrix $\boldsymbol{\pi}$. Using the steady state probabilities, the various performance measures can be obtained.

3.6.4.1 Average number of PDUs in the queue

For a connection, the average number of PDUs in the transmission queue is obtained as follows:

$$\overline{x} = \sum_{x=1}^{X} x \sum_{\forall b, \forall n} \pi(x, b, n). \tag{3.25}$$

3.6.4.2 PDU Dropping Probability

It refers to the probability that an incoming PDU will be dropped due to insufficient buffer space. We first calculate the average number of dropped PDUs per frame and then the PDU dropping probability can be obtained following the procedure in [30]. Given that there are x PDUs in the queue and the number of PDUs in the queue increases by v, the number of dropped PDUs is $v-(X-x)$ for $v > X-x$, and zero otherwise. The average number of dropped PDUs per frame is obtained as follows:

$$\overline{x}_{drop} = \sum_{j=1}^{p} \sum_{z=X-x+1}^{A} [\boldsymbol{\pi}]_j \left(\sum_{k=1}^{r} \left[\mathbf{P}_{\lfloor j/r \rfloor, \lfloor j/r \rfloor + z} \right]_{j-\lfloor j/r \rfloor r, k} \right) (z-(x-X)) \quad (3.26)$$

where $p = (X+1)S(N+2)$ and $r = S(N+2)$ denote the sizes of the matrices $\mathbf{P}$ and $\mathbf{p}_{x,x'}$, respectively. Note that we consider probability $\mathbf{p}_{x,x+v}$ rather than the probability of PDU arrival since we have to consider PDU transmissions during the same frame as well. After calculating the average number of dropped PDUs per frame, we can obtain the probability that an incoming PDU is dropped as follows:

$$P_{drop} = \frac{\overline{x}_{drop}}{\overline{\lambda}} \quad (3.27)$$

where $\overline{\lambda}$ is the average number of PDU arrivals per frame (as obtained from (3.12)).

3.6.4.3 Queue Throughput

This measures the number of PDUs transmitted in one frame, which is calculated based on the fact that if a PDU is not dropped upon its arrival, it will be transmitted eventually. Hence, the queue throughput (PDUs/frame) can be obtained from

$$\eta = \overline{\lambda}(1 - P_{drop}). \quad (3.28)$$

3.6.4.4 Average Delay

The average delay is defined as the number of frames that a PDU waits in the queue since its arrival before it is transmitted. This is obtained as follows:

$$\overline{w} = \frac{\overline{x}}{\eta} \quad (3.29)$$

where η is the throughput (same as the effective arrival rate at the queue) and $\overline{x}$ is the average queue length.

3.6.5 Average Amount of Allocated Bandwidth Per Connection

The average amount of allocated bandwidth per connection is required to estimate the connection-level performance measures by using the queueing analytical model presented in Section V. In this case, we consider rtPS and nrtPS connections for which the amount of allocated bandwidth is not fixed and it depends on the channel quality and the corresponding PDU arrival rate. Let $\tilde{n}_b$ denote the probability that a new connection requires b units of bandwidth to satisfy the corresponding QoS requirements. The average allocated bandwidth per connection is then obtained as follows:

$$\tilde{b} = \sum_{b=b_{min}}^{b_{max}} b\tilde{n}_b. \tag{3.30}$$

3.7 Parameter Setting and Simulation Environment

3.7.1 Parameter Setting

3.7.1.1 Wireless Channel and Radio Transmission

We consider a TDMA/TDD-based transmission scenario from multiple SSs to a BS. That is, multiple SSs access the uplink channel in TDMA mode and the downlink transmissions share the same frequency channel in TDD mode. The SSs work in GPC mode. The transmission bandwidth is 25 MHz, the transmission frame size is 2 ms and the length of a MAC PDU is fixed at 100 bits. AMC is used in which the modulation level and the coding rate is increased if the channel quality permits. The maximum number of PDUs that can be transmitted (i.e., total amount of bandwidth) in one frame period is 200 units per frame. The average SNR at the receiver is 15 dB and Doppler frequency is 15 Hz (i.e., $\overline{\gamma} = 15$ and $f_d = 15$). We vary some of these parameters according to the evaluation scenarios while the rest remain fixed.

3.7.1.2 Traffic Source

The PDU arrival process for each of the polling service connections follows a Markov Modulated Poisson Process (MMPP) and the maximum batch size of arrival is 50 (i.e.,

$A = 50$). The PDUs from rtPS and nrtPS connections are buffered into separate queues (in both uplink and downlink) and the queue size for each connection is assumed to be 200 PDUs (i.e., $X = 200$).

For performance evaluation, we use

$$\mathbf{U} = \begin{bmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \end{bmatrix}, \quad \mathbf{\Lambda} = \kappa \begin{bmatrix} 1 & \\ & 3 \end{bmatrix} \tag{3.31}$$

where κ indicates the PDU traffic intensity and we vary this parameter to observe the packet-level queueing performances.

3.7.1.3 QoS Constraints and Utility Functions

The QoS constraints for UGS, rtPS and nrtPS connections are assumed as follows: $b_{UGS}^{(up)} = b_{UGS}^{(do)} = 2$ units per frame, $d_i^{(up,req)} = d_i^{(do,req)} = 5$ frames $\forall i \in \mathbb{C}_{rtPS}$ and $\tau_i^{(up,req)} = \tau_i^{(do,req)} = 15$ PDUs per frame $\forall i \in \mathbb{C}_{nrtPS}$ (i.e., 15,000 PDUs per second). The parameters for the sigmoid utility function are set as follows: $g_{rt} = g_{nrt} = 2$ and $h_{rt} = h_{nrt} = 0$. The minimum and the maximum amount of bandwidth allocated per connection are 1 and 10 units (i.e., $b_{min} = 1$ and $b_{max} = 10$), respectively.

3.7.2 Simulation Environment

We use an event-driven simulator to evaluate the network performance under the proposed joint BA and CAC framework. The connection inter-arrival time and the connection holding time are assumed to be exponentially distributed. In particular, the average connection holding time for UGS, rtPS, nrtPS and BE connections is assumed to be 10, 15, 20, and 25 minutes, respectively. We vary the connection arrival rate to observe the system performance under different traffic load scenarios. The joint bandwidth allocation and admission control algorithm (optimal or iterative) is invoked when a connection arrives or departs. For each data point, we run the simulation for 5000 connections. A separate queue is maintained for each of the rtPS and nrtPS connections and the queue state is observed at the beginning of each frame.

We compare the performance of the proposed scheme with that of each of the static bandwidth allocation and the adaptive bandwidth allocation schemes. For the static scheme, the amount of bandwidth allocated to each of the rtPS and nrtPS connections

is 3 units per frame (i.e., $b = 3$) for both uplink and downlink transmission and 2 and 1 units for UGS and BE connections, respectively. For the adaptive scheme, we consider bandwidth allocation and admission control similar to that in [24] in which the amount of allocated bandwidth is dynamically adjusted according to the number of ongoing connections. Specifically, after dividing the available bandwidth equally among the connections, the remaining bandwidth (i.e., $C - \lfloor C/M \rfloor$) is randomly allocated to the ongoing connections. For the adaptive scheme, we set the minimum amount of allocated bandwidth to a connection to one unit per frame (i.e., for uplink and downlink transmissions), and therefore, the maximum possible number of ongoing connections is $200/2 = 100$. With these static and adaptive algorithms, there is no QoS guarantee for the different types of connections in the network.

Note that for all the bandwidth allocation and connection admission control schemes, an incoming connection is blocked if the average SNR at the receiver for that connection is below 7 dB.

3.8 Numerical and Simulation Results

3.8.1 Connection-Level Performance and Impact of Threshold Setting

With the bandwidth reservation thresholds[4] $\mathcal{T}_{UGS} = 20$, $\mathcal{T}_{rtPS} = 40$, $\mathcal{T}_{nrtPS} = 25$, and $\mathcal{T}_{BE} = 15$, variations in average number of ongoing connections and connection blocking probability with connection arrival rate are shown in Figs. 3.4(a) and (b), respectively. Based on measurement, the average amount of allocated bandwidth for the different types of connections is observed to be as follows: $\tilde{b}_{UGS} = 2$, $\tilde{b}_{rtPS} = 3.5$, $\tilde{b}_{nrtPS} = 3$, and $\tilde{b}_{BE} = 1$. It is evident that the numerical measures obtained from the analysis follow the simulation results very closely.

To demonstrate the impact of threshold setting, we fix the connection arrival rate to 0.4 connections per minute and the bandwidth reservation thresholds for UGS and BE connections are set as follows: $\mathcal{T}_{UGS} = 20$ and $\mathcal{T}_{BE} = 15$. We vary the thresholds for rtPS and nrtPS such that $\mathcal{T}_{rtPS} + \mathcal{T}_{nrtPS} = 65$. The average number

[4]These thresholds are used for uplink transmission.

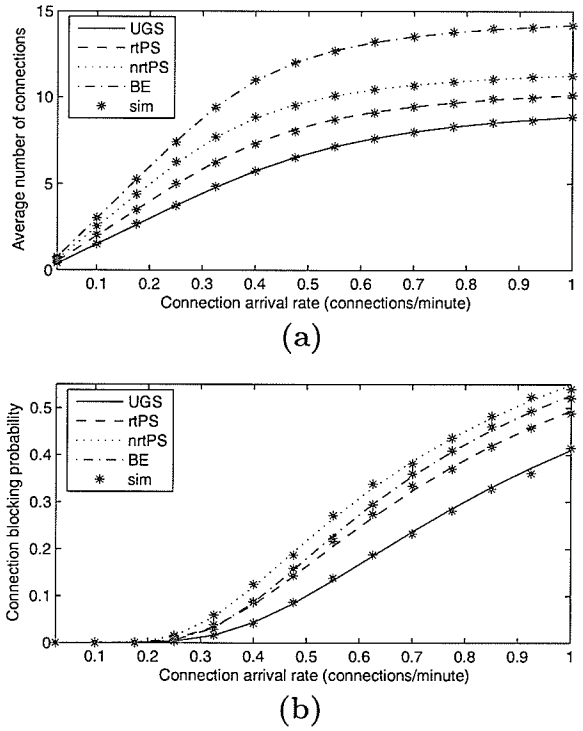


Figure 3.4. *Variation in (a) average number of ongoing connections and (b) connection blocking probability with connection arrival rate.*

of connections for rtPS, nrtPS and the sum of them are shown in Fig. 3.5. When the threshold for rtPS increases, the average number of rtPS connections increases while that of nrtPS decreases. We observe that there are many local maximum points for the sum of the average number of connections. Therefore, a general optimization technique would not be efficient to obtain the global maximum. Fortunately, the set of feasible solutions for the threshold settings is not too large and the computational complexity of the proposed queuing model is small. Also, this optimization problem can be solved off-line. Therefore, the solution of the optimization problem defined in (3.11) can be obtained by enumeration.

To illustrate this, we vary the arrival rate of UGS connections while fixing the arrival rate of rtPS, nrtPS and BE connections to 0.4 per minute. The connection-level QoS constraints are set as follows: $\hat{P}_{bl}^{(UGS)} = 0.1$, $\hat{P}_{bl}^{(rtPS)} = 0.2$, $\hat{P}_{bl}^{(nrtPS)} = 0.2$ and $\hat{P}_{bl}^{(BE)} = 0.5$. The amount of revenue per connection for the different services is

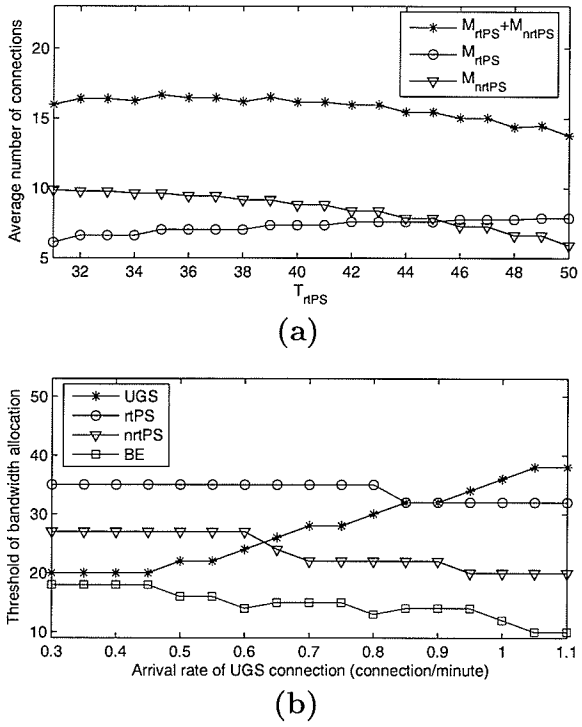


Figure 3.5. *(a) Total number of ongoing connections under different threshold setting and (b) threshold adaptation.*

assumed to be: $\mathcal{R}_{UGS} = 1$, $\mathcal{R}_{rtPS} = 1.5$, $\mathcal{R}_{nrtPS} = 1$, and $\mathcal{R}_{BE} = 0.5$.

With the above setting, the variations in bandwidth reservation threshold, connection blocking probability, and average revenue are shown in Fig. 3.5(b), Fig. 3.6(a) and (b), respectively. As expected, when the traffic load due to UGS connections increases, the joint BA and CAC algorithm requires a larger value of $\mathcal{T}_{UGS}$ (Fig. 3.5(b)) to satisfy the connection blocking probability constraint (Fig. 3.6(a)). At the same time, the value of $\mathcal{T}_{BE}$ decreases since the revenue per connection is the smallest for BE service.

Fig. 3.6(b) shows the variations in average revenue under different constraints (e.g., indicated by the numbers in the legend which correspond to UGS, rtPS, nrtPS and BE connections, respectively). As expected, as the QoS requirements become tighter, the average revenue becomes smaller. Also, with tighter QoS requirements, there is no feasible solution for the formulated optimization problem when the connection

arrival rate of UGS is higher than 0.9 connections per minute.

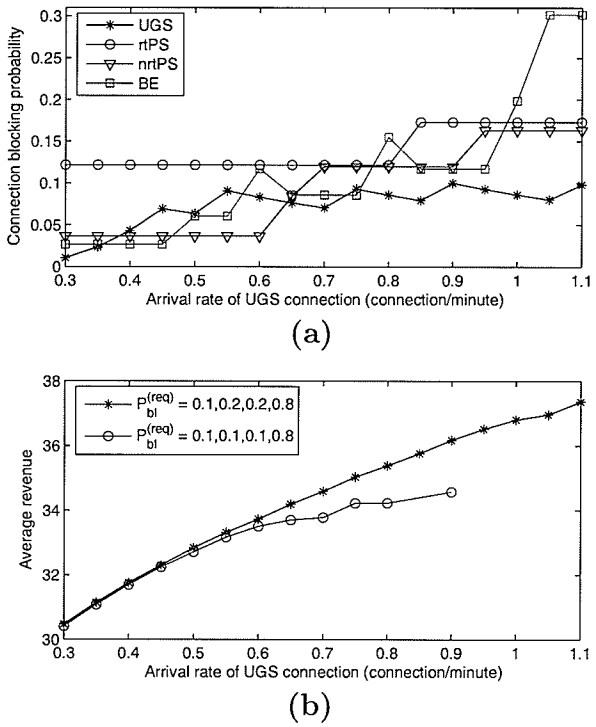


Figure 3.6. *(a) Connection blocking probability and (b) average revenue under threshold adaptation.*

3.8.2 Packet-Level Queueing Performances

Fig. 3.7 shows the impacts of wireless channel quality on the packet-level queueing performances of an rtPS connection. The queue throughput increases as the channel quality improves (Fig. 3.7(a)). Consequently, the PDU dropping probability decreases (Fig. 3.7(b)). Also, when the channel fading is more correlated (i.e., smaller f_d), the dropping probability is higher, since the probability that the wireless channel undergoes deep fading for a long period of time is higher in this case. In contrast, when the channel fading is less correlated (i.e., higher f_d), the probability that the queue is full becomes smaller.

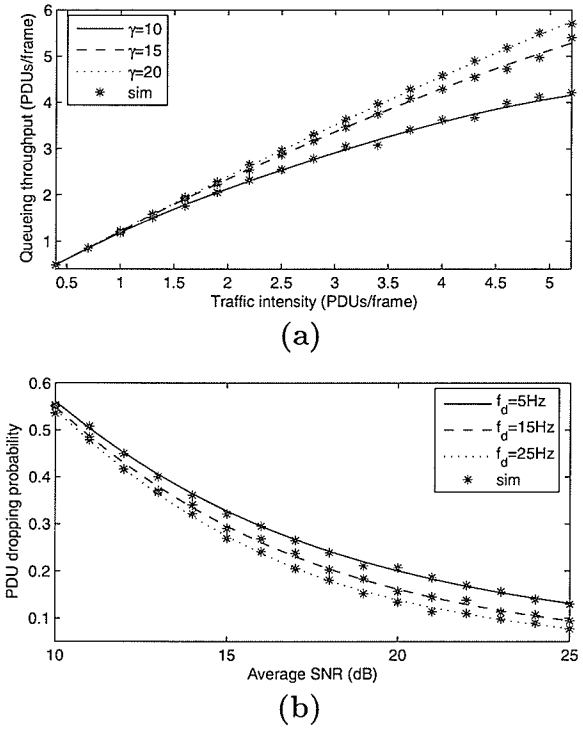


Figure 3.7. *(a) Average delay under different traffic intensities and (b) PDU dropping probability under different channel qualities.*

3.8.3 Performance of the Joint BA and CAC Algorithm

3.8.3.1 Comparison Between the Optimal and the Iterative Approaches– System Utility and Computational Complexity

Fig. 3.8(a) shows the variations in total system utility with the number of connections for both the optimal and the iterative approaches. It is evident that, the total system utility is pretty much the same for both the algorithms. The computation time[5] for these two algorithms is shown in Fig. 3.8(b). We observe that the computation time for the optimal approach increases exponentially with the number of connections while that for the iterative approach increases only linearly. The proposed iterative approach would be suitable to perform bandwidth allocation and admission control in an online fashion. We also observe that optimizing the system with only rtPS and

[5]Using Matlab in a Pentium III 2.0 GHz PC with 512 MB RAM.

nrtPS services (i.e., indicated by legend "Optimal (without UGS and BE)") incurs higher computation time. Since the utility for a UGS and a BE connection can be either one or zero, assignment of bandwidth to UGS and BE services is simpler than that of rtPS and nrtPS services.

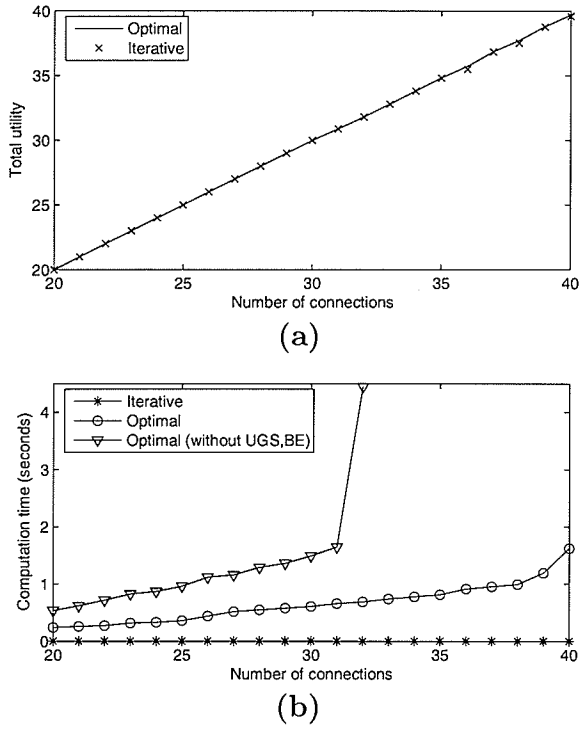


Figure 3.8. *(a) Total utility and (b) computation time for optimal and iterative bandwidth allocation approaches under varying number of connections.*

3.8.3.2 Comparison Among the Iterative, Static and Dynamic Algorithms– Connection-Level Performance

The variations in the average number of ongoing connections for UGS, BE, rtPS and nrtPS are shown in Fig. 3.9. As expected, the average number of ongoing connections increases as the connection arrival rate increases. However, at some point, due to admission control, the average number of ongoing connections saturates.

Figs. 3.10(a) and (b) show the connection blocking probability for the proposed iterative joint BA and CAC algorithm under different connection arrival rate. In

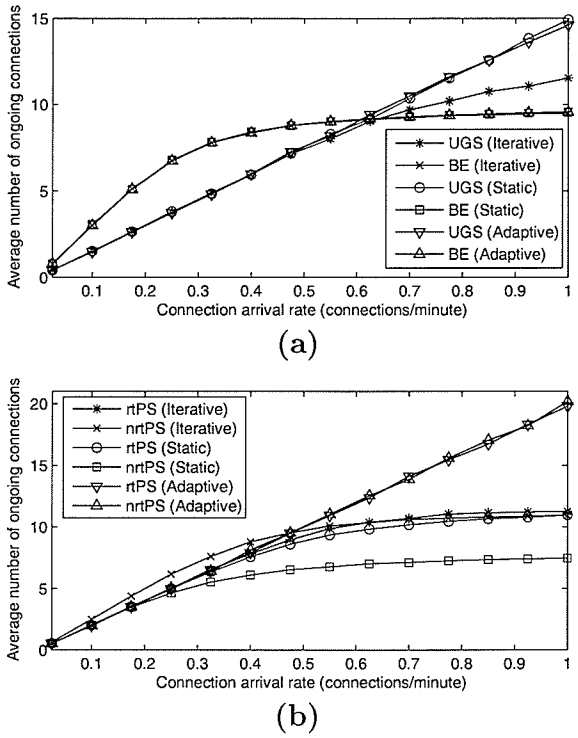


Figure 3.9. *(a) Average number of ongoing UGS and BE connections and (b) average number of ongoing PS connections.*

this case, we set $\mathcal{T}_{UGS} = 25$, $\mathcal{T}_{rtPS} = 40$, $\mathcal{T}_{UGS} = 25$ and $\mathcal{T}_{BE} = 10$ and we assume that the connection arrival rate is the same for all types of connections and the PDU arrival rates for uplink and downlink transmissions are symmetric. As expected, the connection blocking probability increases with increasing connection arrival rate.

The connection blocking probability for BE connections is the same for both the algorithms since we assume that this service type has the lowest priority (e.g., achieved via threshold setting). Again, the connection blocking probability in this case is similar to that for each of the static and adaptive schemes. However, with the chosen threshold values, for the iterative algorithm, the blocking probability for UGS is observed to be the highest.

For rtPS and nrtPS connections, the blocking probability is the highest with the static algorithm (Fig. 3.10(b)). This is due to the fact that the static algorithm always allocates a fixed amount of bandwidth to rtPS/nrtPS connections without any

adaptation based on channel quality and PDU arrival rate. However, the adaptive algorithm is able to adjust the amount of bandwidth allocated to each connection according to the traffic load in the cell. Therefore, the blocking probability is the lowest for the adaptive case. For the iterative algorithm, since the delay and the transmission rate requirements are to be satisfied for rtPS and nrtPS, respectively, the blocking probabilities are higher compared to the adaptive case which does not provide any packet-level QoS guarantee. In particular, to guarantee the QoS performances (i.e., delay and transmission rate), the number of ongoing connections must be limited. As a result, some of the incoming connections must be blocked to avoid deterioration in the QoS of the ongoing connections.

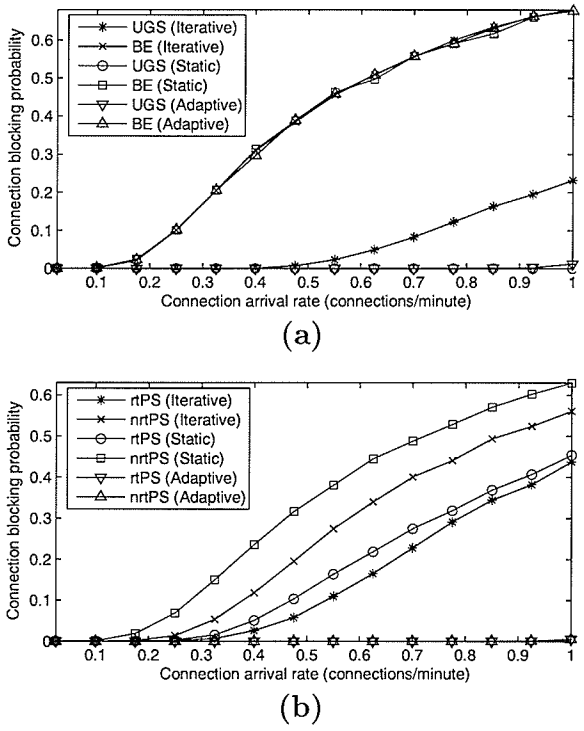


Figure 3.10. *Connection blocking probability for (a) UGS and BE connections and (b) rtPS and nrtPS connections.*

We observe the variations in the number of bandwidth relocations per minute for the different schemes (Fig. 3.11). In particular, for iterative and static schemes, the number of bandwidth relocations increases as the connection arrival rate increases.

However, at high arrival rate (i.e., > 0.5 connections per minute), since more number of connections are rejected, the rate of increase in number of bandwidth relocations decreases. Again, since the adaptive scheme provides the smallest blocking probability, the number of bandwidth relocations is the highest compared with that for each of the iterative and static schemes.

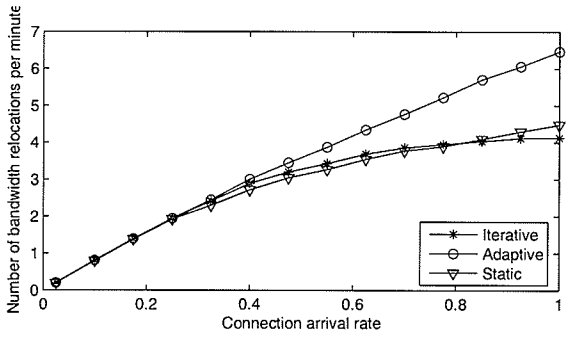


Figure 3.11. *Variation in the number of bandwidth relocations under different connection arrival rates.*

3.8.3.3 Comparison Among the Iterative, Static and Dynamic Algorithms– Packet-Level Performance

Fig. 3.12(a) shows the average delay performance for rtPS connections[6]. As expected, the static and the iterative algorithms can maintain the average PDU delay below the target requirement (i.e., 5 frames). However, for the adaptive algorithm, since the amount of allocated bandwidth is dynamically adjusted, when the load in the network becomes high, the amount of allocated bandwidth is reduced which results in larger queueing delay. Consequently, the delay requirement is violated. We observe similar effect on the transmission rate performance of nrtPS connections (Fig. 3.12(b)). However, since the iterative algorithm aims at maximizing user utility, when traffic load is low, the available bandwidth is completely allocated to the ongoing connections. Consequently, the transmission rate becomes high. Also, the iterative algorithm is able to maintain the transmission rate for an nrtPS connection higher than the requirement (i.e., 15,000 PDUs/second).

[6]The performance results shown are for uplink transmission, however, the results for downlink transmission are expected to be similar.

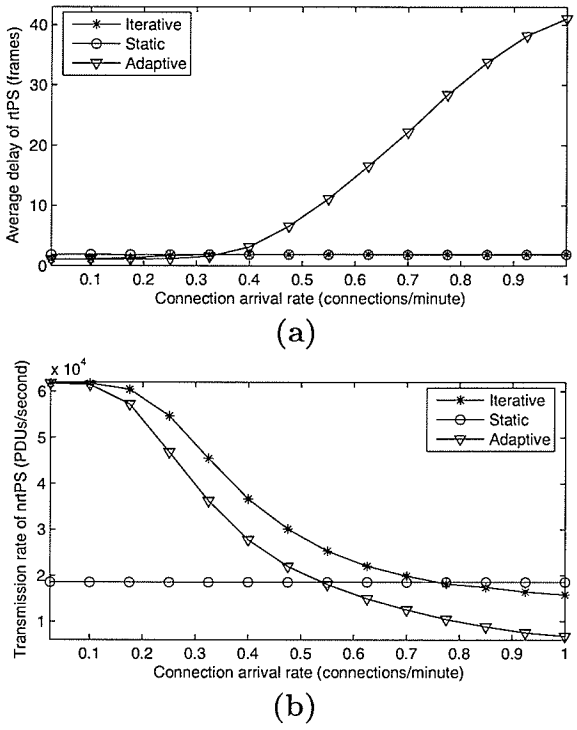


Figure 3.12. *(a) Average delay for rtPS and (b) transmission rate for nrtPS.*

3.8.3.4 Comparison Among the Iterative, Static and Dynamic Algorithms– Total System Utility

Fig. 3.13 shows variation in total system utility with traffic load. With the proposed iterative scheme for joint BA and CAC allocation, the highest total utility is achieved (compared to the static and the adaptive schemes) under varying connection arrival rates. This is simply due to the fact that unlike the two other schemes, the iterative scheme aims at maximizing the total system utility while satisfying the QoS requirements.

3.9 Chapter Summary

We have presented a joint bandwidth allocation (BA) and connection admission control (CAC) framework for WiMAX-based broadband wireless access) networks. Based on an assignment problem formulation, the optimal approach has been devised

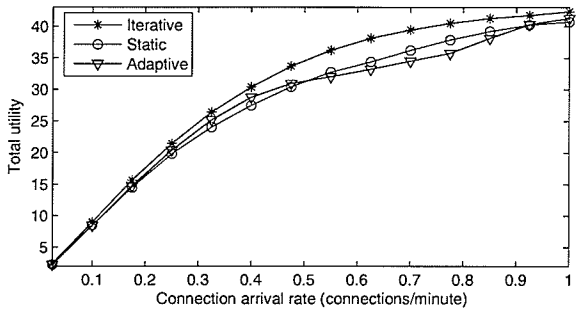


Figure 3.13. *Variations in total utility with connection arrival rate.*

for which the bandwidth allocations for the different connections can be obtained by using the integer binary linear programming technique. A water-filling based iterative scheme (with significantly lower computational complexity) has also been proposed which performs as efficiently as the optimal scheme. For both of these schemes, a complete partitioning approach for bandwidth reservation among the different service types (i.e., UGS, rtPS, nrtPS, and BE) has been used and the schemes provide packet-level QoS guarantee to the nrtPs and rtPS types of connections while maximizing the total system utility.

A queueing analytical model for connection-level performance evaluation under the proposed radio resource management framework has been presented. Based on an optimization formulation, using this queueing model, the optimal values (which maximize the average system revenue) for the bandwidth reservation thresholds have been obtained under constrained connection-level QoS requirement. Also, to ananlyze the packet-level performance, a queueing analytical model has been presented considering adaptive modulation and coding at the physical layer. In summary, the joint bandwidth allocation and connection admission control framework provides a unified radio resource management solution to provide both packet-level and connection-level QoS for the different service types in WiMAX-based broadband wireless access networks. At the same time, it maximizes network utility. Part of this chapter has been published in [46].

Chapter 4

Radio Resource Management Framework for Integrated WiFi/WiMAX Multihop Mesh/Relay Networks

4.1 Introduction

4.1.1 Problem Statement

In this chapter, the problem of radio resource management in a WiFi/WiMAX multihop relay network is considered. This multihop relay network utilizes WiMAX base stations (BSs) to serve WiFi and WiMAX users. The radio resource management framework aims to provide fair and efficient bandwidth allocation of BS to different types of connections. The first input of this framework is the number of connections and their transmission rates. The second input is the transmission rate between the BSs in the network. The output of this framework is the burst size in MAC frame of WiMAX BS.

4.1.2 Contribution

An integrated WLAN/WMAN multihop relay architecture is presented for mobile hotspots. The related research issues are described. To this end, based on a game-theoretic model, we present a bandwidth management and admission control frame-

work for WiMAX base stations to allocate bandwidth among out-going connections from standalone subscriber stations and WLAN access points as well as relay traffic from the upstream base stations. The admission control method is designed to limit the number of ongoing connections at a mesh router so that the total utility for the ongoing connections is maximized in that router.

4.2 An Integrated WMAN/WLAN Architecture

An integrated WMAN/WLAN architecture to provide remote hotspot services is shown in Fig. 4.1. The network architecture basically consists of two parts - the backhaul multihop mesh infrastructure consisting of the WiMAX base stations/mesh routers[1] and the interface between a WLAN access point and a WiMAX base station.

4.2.1 Mesh Infrastructure

Each of the WiMAX base stations in the mesh infrastructure serves standalone subscriber stations and WiFi access points/edge routers in which each of the edge routers has a dual radio (WiFi and WiMAX) interface. In Fig. 4.1, the mesh infrastructure consists of three WiMAX base stations/mesh routers. BS-1 serves multiple standalone SSs and one WiFi access point/edge router and it is connected with the gateway base station. We assume that each of the subscriber stations and the edge routers uses the grant per subscriber station (GPSS) service class in which the WiMAX base station allocates bandwidth to each of the subscriber stations and edge routers separately. Also, we assume that, to avoid co-channel interference adjacent base stations use different frequency bands.

For the mesh infrastructure, we consider an WiMAX OFDM/TDMA-TDD-based air interface (i.e., WirelessMAN-OFDM) between two base stations. With OFDM/TDMA all subchannels are allocated to one connection at a time. For uplink and downlink transmission using OFDM, each of the WiMAX base stations uses 50 subchannels each having a bandwidth of 200 KHz. The total bandwidth required (including the guard bands) is 20 MHz. The frame size is assumed to be 2 ms. Adaptive mod-

[1]We use the term "WiMAX" in a generic sense without referring to any particular version of this standard.

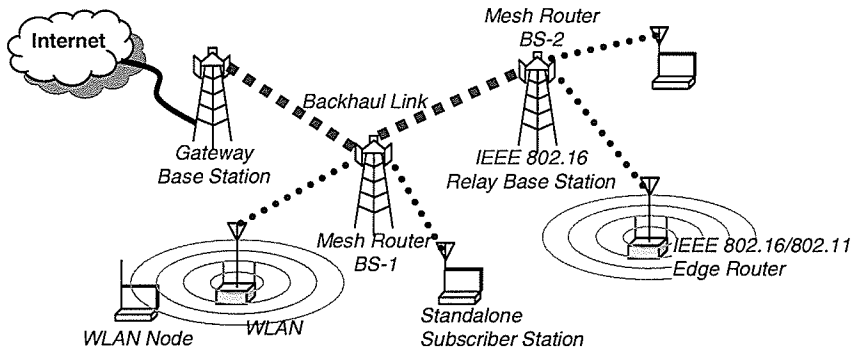


Figure 4.1. *Integration of WiFi WLANs with WiMAX mesh networks.*

ulation and coding (AMC) with 7 transmission modes is used in each subchannel independently based on the subchannel quality. The transmission rate $T_i^{(h)}(D_i^{(h)}) = \sum_k \tau(\overline{\gamma}_{i,k,h}, D_i^{(h)})b$ for connection i at base station h is a function of the burst size $D_i^{(h)}$ and average signal-to-interference and noise ratio (SINR) $\overline{\gamma}_{i,k,h}$ for subchannel k, where b is the bandwidth of each subchannel. Note that $\tau(\overline{\gamma}_{i,k,h}, D_i^{(h)}) = \sum_{n=1}^{7} I_n \mathrm{Pr}_n(\overline{\gamma}_{i,k,h}) D_i^{(h)}$ is the transmission rate (per frame) on subchannel k, where I_n is the number of transmitted bits per symbol for AMC state n, and Pr_n is the probability of using mode n which can be obtained as in [31] for Nakagami-m fading channels.

4.2.2 Air Interface Between Edge Router and Mesh Router

The dual radio interface at WiFi access point/edge router uses two different frequency bands. Data packets corresponding to *local* and *Internet traffic* (which can be distinguished based on the IP packet header) are stored in separate queues (Fig. 4.2). The local traffic is due to the connections among nodes in the coverage area of a WLAN and Internet (or relay) traffic is due to connections traversing the mesh backbone to an Internet gateway. Packets from the Internet traffic queue are fragmented and reformatted into WiMAX frames to be transmitted to the mesh router using the WiMAX radio interface. This protocol adaptation is performed in the MAC layer, where the IEEE 802.11 header is removed and then the data unit (including header of higher layer protocol such as IP) is fragmented into protocol data units (PDUs) for the WiMAX uplink subframe.

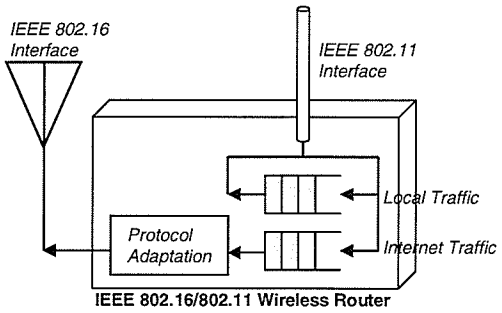


Figure 4.2. *Edge router with IEEE 802.16a and IEEE 802.11 air interfaces.*

4.2.3 Model for WiFi WLAN

We consider WiFi WLANs with direct sequence spread-spectrum (DSSS)-based physical layer and Distributed Coordination Function (DCF) as the MAC scheme. The length of a time slot is 20 μs, the minimum and the maximum values of the backoff window size are $CW_{min} = 32$ and $CW_{max} = 1024$ time slots, respectively, and the packet size is 8000 bits.

The traffic load condition (e.g., unsaturated or saturated) in a WLAN is estimated by the following two parameters at a WLAN node: probability of successful transmission P_s and probability of collision P_c. In the unsaturated case, the amount of load from all active nodes is less than the network capacity and the collision probability is low. On the other hand, the collision probability is high when the network is saturated. To determine whether the WLAN is in unsaturated or saturated condition, we use a threshold τ_{col} (e.g., $\tau_{col} = 0.2$) for collision probability. In particular, if the estimated collision probability is less than τ_{col}, the network is considered to be unsaturated, and saturated otherwise.

In the unsaturated case, the estimated received bandwidth $\tilde{S}_i$ by node i in a WLAN is assumed to be equal to the transmission rate λ_i of that node. However, in the saturated case, it is proportional to the ratio of the user transmission rate to the maximum achievable transmission rate and a function of the successful packet transmission probability P_s. The packet transmission probability and packet collision probability are estimated by using an exponential moving average (EMA) with weight

β (e.g., $\beta = 0.95$).

4.3 Research Issues in an Integrated WLAN/WMAN Network

4.3.1 Topology Management for the Mesh Infrastructure

One major issue is efficient topology management of the mesh infrastructure to minimize network deployment cost while satisfying the quality of service (QoS) requirements for the local and the relay connections. In [7], the problem of WiMAX-based backhaul topology design was formulated as an integer programming problem and then a greedy algorithm was presented to obtain a near-optimal solution. With this solution, the number of WiMAX links in the backhaul network can be reduced significantly compared to that for a ring topology. In an integrated WLAN/WMAN environment, traffic load at the hotspots and the user mobility patterns should be considered for optimal topology design for the mesh infrastructure.

4.3.2 Radio Resource Management

Efficient radio resource management at the mesh routers can be achieved by using intelligent bandwidth allocation, channel assignment, and admission control schemes for different types of connections (e.g., connections from WLAN access points, standalone subscriber stations, and relay connections). Also, fairness between local and relay traffic and prioritization among different types of traffic (e.g., through scheduling) according to their QoS requirements must be considered. A radio resource management framework for subchannel allocation and connection admission control in WiMAX-based OFDMA wireless mesh networks was presented in [49]. The objective of the framework is to guarantee the QoS requirements on a per-connection basis for both relay and local connections. Also, an admission control policy for the relay connections at a mesh router was presented based on the packet-level QoS measures. A channel assignment scheme based on carrier-to-interference information was proposed in [50] to enhance transmission rate in a multihop relay network. To achieve high network capacity, radio resource management techniques for WiMAX-based mesh in-

frastructure should be designed considering advanced physical layer techniques such as multiple-input multiple-output (MIMO) combined with OFDM. Again, resource utilization of the mesh network can be improved by balancing and sharing load among the mesh routers through efficient routing mechanisms.

Since WiMAX networks (e.g., based on 802.16a) and WLANs (e.g., based on 802.11b) may operate on overlapped frequency spectrum (i.e., 2-11 GHz), this can result in severe interference in an integrated WLAN/WMAN network. Therefore, dynamic adaptation for frequency spectrum allocation would be required. A cognitive radio approach for sharing radio resources in frequency, space, and time domain was proposed in [51] and a dynamic frequency selection scheme was presented to minimize interference. Also, power control was used to minimize transmit power and time agility was exploited to adjust traffic pattern to avoid interference. However, MAC and higher layer protocol performances were not considered in this work.

4.3.3 Link Level Error Control and End-to-End QoS

In a multihop WiMAX mesh infrastructure, the radio link control mechanisms should be designed considering the end-to-end QoS (e.g., packet reliability and packet delay) requirements. In [52], an analytical model based on an absorbing Markov chain was presented to obtain the various end-to-end performance measures in a static multihop network under different link-level error control strategies. This model considered wireless transmission with AMC which is a standard feature in WiMAX. However, no specific MAC scheme was considered and also the impact of local traffic at a mesh router was ignored.

Space diversity technique such as cooperative diversity [53] can improve the transmission performances in a multihop network. Cooperative diversity relies on transmissions by several nodes and each node acts as a virtual transmission antenna for the receiver. Since these nodes transmit from different locations, the spatial diversity of independent multipath fading can be exploited to improve the transmission quality. In cooperative diversity, intermediate nodes can amplify-and-forward or decode-and-forward packets to the destination node. While the former is able to achieve a full diversity, the latter can prevent error propagation. Integration of error control and error recovery as well as packet scheduling and routing schemes with cooperative

diversity are interesting research issues for multihop wireless networks.

4.3.4 Routing Strategies

A routing algorithm for a WiMAX mesh infrastructure should consider the quality of wireless links along different routes and the QoS requirements for the corresponding connections. Performance evaluation of three different link quality metrics for routing in a static multihop network was carried out in [54]. The metrics are as follows: (a) "expected transmission count (ETX)" which is based on the loss rate of broadcast packets between pairs of neighboring nodes, (b) "per-hop round trip time (RTT)" based on the round trip delay observed by unicast probes between neighboring nodes, and (c) "per-hop packet pair delay" based on the delay between a pair of back-to-back probes to the neighboring nodes.

In [10], an interference-aware routing mechanism for 802.16 multihop mesh networks was proposed. To reduce congestion in a mesh router responsible for relaying Internet traffic, this routing scheme uses interference information from the physical layer to find the optimal route from a source base station to the gateway base station (i.e., base station connected directly to the Internet). Since the routing protocol performances would strongly depend on the resource allocation scheme used at each base station, a cross-layer optimization approach should be used.

In [55], congestion-based routing strategies based on opportunity driven multiple access (ODMA) for multihop TDD-CDMA networks was proposed. This routing strategy was designed to minimize transmit power of all base stations while the error and the transmission rate performance requirements are met. Also, time slot assignment (i.e., dynamic channel allocation) was integrated into the routing algorithm to maximize system performance.

In a WiMAX infrastructure mesh network, the routing protocol should be optimized considering the MAC dynamics as well as subchannel allocation and other radio resource management techniques.

4.3.5 Protocol Adaptation and QoS Support

In an integrated WLAN/WMAN network, protocol adaptation at the edge router would be required to provide QoS support to WLAN connections. In [47], a heterogeneous two-hop architecture was proposed for mobile hotspots exploiting Universal Mobile Telecommunications System (UMTS) services. In this architecture, WLAN and 4G cellular networks cooperate to relay users' traffic to the destination. Protocol adaptation and QoS support mechanisms were also proposed to support real-time traffic such as voice, video, and interactive applications.

WiMAX standard has a predefined QoS framework. Also, the IEEE 802.11e standard was designed specifically for traffic with QoS guarantees. However, the approaches to QoS provisioning are different in these two standards. In particular, WiMAX supports three major different traffic types (i.e., unsolicited granted service, polling service, and best-effort) while IEEE 802.11e supports two major traffic types (i.e., low and high priority traffic). Also, the MAC protocols are different in WiMAX and 802.11e networks. Therefore, a unified QoS framework is required for an integrated WMAN/WLAN network. In [56], a QoS framework for 802.16/802.11e internetworking was proposed based on the mapping of the QoS requirements of an application and the necessary messaging procedures were defined. However, the mechanisms to ensure the QoS requirements (e.g., bandwidth assignment, scheduling, admission control) were not considered.

4.3.6 Optimizing Transport Layer Protocol Performance in an Integrated WLAN/WMAN Network

Multihop transmission in an integrated WLAN/WMAN network affects the error recovery and the congestion control performances at the transport layer. Performance modeling, analysis, and optimization of transport layer protocol such as TCP (Transmission Control Protocol) in such a heterogeneous environment are challenging research problems. In [57], performance of TCP in an WiFi-based multihop network was investigated. It was observed that for a specific network topology and flow pattern there exists an optimal TCP window size to achieve the highest throughput. However, only a single TCP flow was considered. In an integrated WLAN/WMAN

environment, radio link and routing protocols should be designed to optimize TCP performance by exploiting the cross-layer interactions into account.

4.4 Bandwidth Management and Admission Control in an WiMAX Mesh Router in an Integrated WLAN/WMAN Network: A Game-Theoretic Model

Developed mainly for use in the field of economics, game theory has been used for radio resource management and protocol engineering (e.g., in [58]). A game is described by a set of rational players, the strategies associated with the players, and the payoffs for the players. A rational player has his own interest, and therefore, will act by choosing an available strategy to achieve his interest. In this case, a player is assumed to be able to evaluate exactly or probabilistically the outcome or payoff of the game which depends not only on his action but also on other players' actions. Two important characteristics of a game are individualism and mutual independence. While individualism influences the rationality (i.e., self-interest) and the cooperation among the players, mutual independence determines the actions of the players in response to those of other players.

In an integrated WLAN/WMAN multihop network (as shown in Fig. 4.1), mobile users (or connections) with different requirements and channel quality share the available radio resource in the mesh routers. Each of the users is assumed to be rational to achieve the highest performance. Therefore, a game-theoretic model can be used for efficient resource allocation among the different connections.

Bargaining game is one of the game models proposed to analyze the situation in which the players cooperatively try to make an agreement and the players have a choice to *bargain* with each other so that they can gain maximum benefit which is higher than that they could have obtained by playing the game without cooperation. The amount of resource allocated to each player affects the payoff of the other player. Therefore, all players seek for the optimal and fair portion of resource through negotiation.

We present a bargaining game model for distributed bandwidth management and admission control for a mesh router in an integrated WLAN/WMAN multihop network. We consider three different types of traffic, i.e., local traffic from standalone subscriber stations, WLAN traffic, and relay traffic from upstream routers. A bargaining game is formulated to allocate bandwidth to these traffic types in a fair manner. Then, an admission control algorithm is proposed with a view to optimizing the total utility of the system. Both connection-level and in-connection level performances are analyzed for these bandwidth management and admission control schemes.

The motivation of using bargaining game is that the solution is fair and efficient [59]. In particular, the allocation is efficient due to *Pareto optimality* [62] while fairness is achieved by satisfying the concept of equilibrium. In economics, Pareto optimality defines an agreement (i.e., strategy) for which one player cannot increase his utility without decreasing the utility (payoff) of the other player(s). Conversely, if an agreement is not Pareto optimal, there exists another strategy which provides better payoff to the players.

4.4.1 Bandwidth Allocation and Admission Control Process

The process of bandwidth allocation and admission control for a particular out-going connection from a WLAN node in an integrated WiMAX/WiF wireless mesh network is shown in Fig. 4.3. When a connection is initiated, the corresponding node in the WLAN sends a *connection request* message to the edge router. Upon receiving this message, the edge router performs bandwidth estimation based on the estimated successful packet transmission and packet collision probabilities (as described earlier). Then, the admission control is performed.

If the edge router decides to accept the new out-going connection, it sends a *resource request* message to the corresponding mesh router which executes the bandwidth allocation algorithm locally. Based on the allocated bandwidth, the admission control algorithm is invoked. If the mesh router decides to accept the new connection, it sends a *resource request* message to the next mesh router. This bandwidth allocation and admission control process continues in each of the mesh routers along the path to the gateway router. The connection is not admitted if the new connection is rejected at any one of the intermediate routers. For standalone subscriber stations, a

new connection can use a *join* message to communicate with the corresponding router which performs the bandwidth allocation and admission control process accordingly.

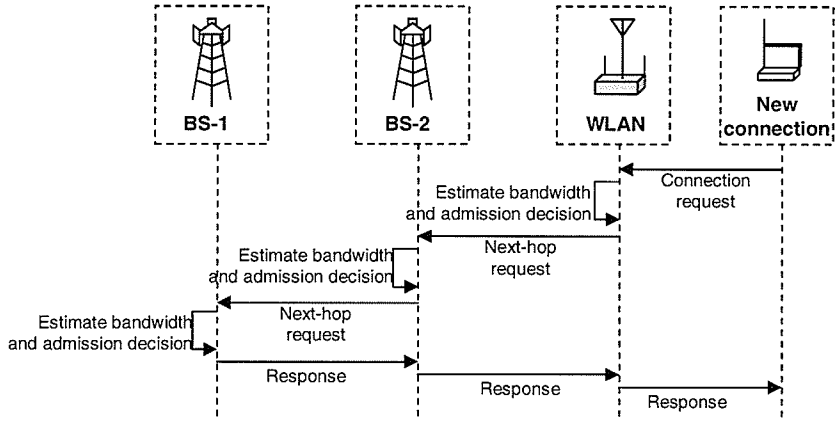


Figure 4.3. *Flow of control messages for bandwidth allocation and admission control.*

To measure the revenue gained from a connection we use the concept of utility function. The utility for an admitted connection with transmission rate T is given as follows [60]:

$$U(T) = w \log(1 + \alpha T) \tag{4.1}$$

where w and α are constants indicating the scale and the shape of the utility function.

At base station h, bandwidth allocation is required to reserve available transmission time (i.e., burst size) for three different types of traffic, i.e., WLAN traffic, traffic from local standalone subscriber stations, and relay traffic. Let $\mathbb{C}_{wl}$, $\mathbb{C}_{ss}$, and $\mathbb{C}_{re}$ denote the set of connections from WLAN, connections from standalone subscriber stations, and relay connections, respectively. The total utility for traffic type j ($j \in \{wl, ss, re\}$) can be obtained from

$$U_j^{(h)}(B_j) = \sum_{i \in \mathbb{C}_j} w_i \log\left(1 + \alpha_i T_i^{(h)}(D_i^{(h)})\right) \tag{4.2}$$

where $B_j = \sum_{i \in \mathbb{C}_j} D_i^{(h)}$ denotes the total burst size allocated to all connections i of type j (i.e., from set $\mathbb{C}_j$); recall that $D_i^{(h)}$ denotes the burst size allocated by base station h to its ith connection. To allocate bandwidth to each type of connection, we use a bargaining game formulation which will be described in the next section.

The admission control mechanism can be established based on the utility and the allocated burst size. In particular, when a new connection of service type j arrives, every router decides whether this connection can be accepted or not by considering the change in total utility. The total utility at router h and at the WLAN access point for connections of service type j and for the new connection can be computed in a similar way as in (4.2).

The total utility of a WiMAX base station and a WLAN access point increases as the number of connections increases. However, at a certain point, it will decrease since the utility gained from a new connection cannot compensate the performance degradation of the ongoing connections. We take advantage of this behavior to make the admission control decision. In particular, a new connection is accepted only when the total utility increases, and rejected otherwise. In this multihop environment, a new connection is accepted only if all the routers along the route to the Internet and the corresponding edge router decide to accept the connection.

4.4.2 Bargaining Game Formulation

Different types of connections (i.e., WLAN connections, connections from standalone subscriber stations, and relay connections) have different preferences on bandwidth allocation. In order to allocate bandwidth in a fair manner, we use a bargaining game formulation in which different types of connections negotiate with each other to obtain their share of bandwidth at a mesh router. The optimal allocation of bandwidth which maximizes system utility can be achieved from the Pareto optimality [62]. Second, bandwidth allocation must be fair to all types of traffic. A fair allocation of bandwidth can be achieved from the equilibrium of the game based on the payoff metrics.

The game-theoretic formulation for bandwidth allocation at a mesh router can be described as follows:

- **Players:** In this game, the players are the traffic from WLAN, standalone subscriber station, and relay traffic, which are denoted by subscript $j \in \{wl, ss, re\}$.
- **Strategy:** The strategy for player j is the total burst size for traffic type j in a transmission frame.
- **Payoff:** The payoff for player j is the total utility U_j gained from the achievable transmission rate.

The process of choosing strategies can be modeled as a bargaining game. In a multi-player game [61], the players try to make an agreement on trading a limited amount of resource. The players have a choice to *bargain* with each other so that they can gain benefit higher than that they could have obtained by playing the game without cooperation. The payoff (i.e., utility) for the players is given by $\Omega = \{(U_{wl}(B_{wl}), U_{ss}(B_{ss}), U_{re}(B_{re})) : 0 \leq U_{wl}(B_{wl}), U_{ss}(B_{ss}), U_{re}(B_{re})\}$ (i.e., feasible set), where B_{wl}, B_{ss}, and B_{re} denote total burst size allocated to WLAN connections, connections from subscriber stations, and relay connections, respectively. If an agreement among the players cannot be reached, the utility that the players will receive is given by the threat point $(U'_{wl}(0), U'_{ss}(0), U'_{re}(0))$. In particular, $(U'_{wl}(0), U'_{ss}(0), U'_{re}(0)) = (0, 0, 0)$ is the threat point for this game. A threat point represents the payoff for each player when the solution of the game cannot be reached.

The bargaining game model is formulated as

$\mathcal{F}(\Omega, U'_{wl}(0), U'_{ss}(0), U'_{re}(0)) = (U^*_{wl}(B^*_{wl})U^*_{ss}(B^*_{ss}), U^*_{re}(B^*_{re}))$,

where $(U^*_{wl}(B^*_{wl}), U^*_{ss}(B^*_{ss}), U^*_{re}(B^*_{re}))$ denotes the solution (equilibrium) [61] of the game $\mathcal{F}(.)$. The Pareto optimality can provide the candidate strategies (i.e., B_{wl}, B_{ss}, and B_{re}) for which one of the players can achieve the highest utility. In particular to this bandwidth allocation game, the solution $(U_{wl}(B_{wl}), U_{ss}(B_{ss}), U_{re}(B_{re}))$ must be Pareto optimal (i.e., $B_{wl} + B_{ss} + B_{re} = F$, where F is the total frame size) to ensure the efficiency of the allocation. Then, we need the equilibrium of the game such that all the players are satisfied with the utilities they receive. That is, the Nash bargaining solution of this game is the utility triplet $(U^*_{wl}, U^*_{ss}, U^*_{re})$ such that [61]

$$(U^*_{wl}, U^*_{ss}, U^*_{re}) = \arg \max_{U_{wl}, U_{ss}, U_{re}} (U_{wl} - U'_{wl})(U_{ss} - U'_{ss})(U_{re} - U'_{re}). \quad (4.3)$$

This solution can be obtained by using a search method. In this chapter, we use simplex method [63] to optimize the objective function defined in (4.3). The decision variables are B_{wl}, B_{ss}, and B_{re} which denote the total burst size allocated to WLAN connections, connections from subscriber stations, and relay connections, respectively. Again, since the candidate strategy for the solution must be Pareto optimal, the search space for the decision variables is constrained by the condition $B_{wl} + B_{ss} + B_{re} = F$.

The amount of bandwidth assigned to connection i of type j at router h is deter-

mined based on the weight w_i (in the utility function in (4.1)) as follows:

$$D_i^{(h)} = \frac{w_i B_j}{\sum_{i \in \mathbb{C}_j} w_i} \tag{4.4}$$

where $D_i^{(h)}$ is the burst size for connection i, B_j is the total burst size allocated to connections of type j, and $\sum_{i \in \mathbb{C}_j} w_i$ is the sum of weights of connections of type j.

Note that if the solution does not exist, the burst size which is allocated to each type of connection is proportional to the number of ongoing connections of that type and the corresponding weights. This ensures that the resource allocation is fair to all connection types even though the game formulation cannot obtain the solution.

4.4.3 Performance Evaluation

4.4.3.1 Parameter Setting

We evaluate the performances of the proposed bandwidth allocation and admission control framework for the network topology shown in Fig. 4.1. The average SNR at the receiver for connections between BS-1 and the gateway BS and connections between BS-2 and BS-1 is assumed to be 12.5 dB and 8.5 dB, respectively. The average SNR for connections between a standalone subscriber station and a mesh router and connections between an edge router and a mesh router is assumed to be in the range of 10-20 dB. All of the WLAN nodes are assumed to use the same transmission rate. The parameters to evaluate the utility functions are set as follows: $w_{wl} = w_{ss} = w_{re} = 1$, $\alpha_{wl} = \alpha_{re} = 1/100$, and $\alpha_{ss} = 1/70$ (i.e., traffic from standalone subscriber stations has less priority than WLAN traffic and relay traffic).

4.4.3.2 Pareto Optimality and the Solution of the Bargaining Game Solution

We show the Pareto optimality for BS-1 which serves 10 WLAN connections, 10 connections from standalone subscriber stations, and 30 relay connections (from BS-2). The Pareto optimality and the solution of the bargaining game obtained from the analytical model for bandwidth allocation are shown in Fig. 4.4. Note that this solution is obtained by using local search method. The solution is located at $(U_{wl}^*, U_{ss}^*, U_{re}^*) = (17.49, 19.67, 33.78)$ and the corresponding burst-size is 0.287, 0.225,

and 0.463 ms, respectively. As expected, BS-1 assigns the largest amount of bandwidth to relay traffic. However, at the same time BS-1 needs to satisfy the WLAN connections and those from standalone subscriber stations, and therefore, BS-1 reserves some bandwidth for these connections. We observe that, even though the total utility for WLAN connections is smaller, BS-1 assigns larger burst-size to WLAN connections than that for connections from standalone subscriber stations. In fact, even though the bargaining game attempts to achieve fair total utility, since $\alpha_{wl} < \alpha_{ss}$, BS-1 needs to assign a larger burst-size to prioritize WLAN connections over connections from standalone subscriber stations. Since the bandwidth requirement of WLAN connection increases, solution of bargaining game with fairness property has to adjust the burst-size accordingly.

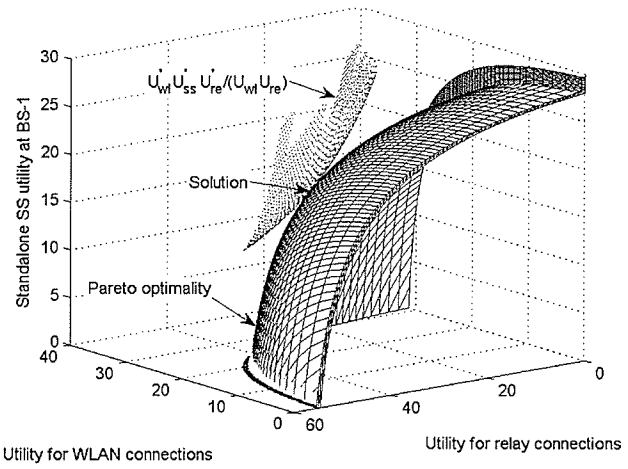


Figure 4.4. *Pareto optimality and solution of bandwidth sharing at BS-1 (from analysis).*

4.4.3.3 Bandwidth Adaptation Under Varying Number of Connections

Fig. 4.5 shows bandwidth adaptation (due to the bargaining game) at BS-1 and BS-2 under varying number of connections obtained from bargaining game model. Since WLAN connections have higher priority than the connections from standalone subscriber stations, at BS-2, bandwidth allocated to WLAN connections increases.

Again, at BS-1, bandwidth allocated to relay connections increases (due to the traffic relayed from BS-2) as the total number of connections becomes large. From Fig. 4.5, we observe that the simplex method used to obtain the solution is numerically stable in which the bandwidth adaptation functions are linear.

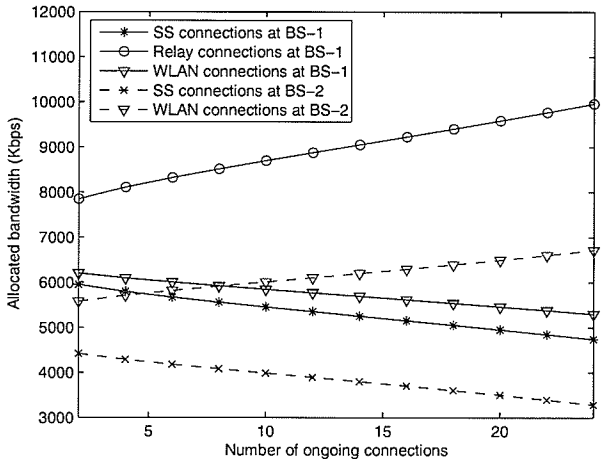


Figure 4.5. *Bandwidth adaptation under different number of ongoing connections (from analysis).*

4.4.3.4 Connection-Level Performances Under Varying Connection Arrival Rate

We assume that connection arrivals follow a Poisson process and connection holding time is exponentially distributed with an average of 20 minutes. Arrival rates for WLAN connections and connections from standalone subscriber stations are varied at BS-1 and BS-2. The average number of ongoing connections and connection blocking probability due to admission control and average achievable bandwidth due to bandwidth allocation algorithm obtained from simulation are shown in Figs. 4.6-4.8, respectively.

As expected, the average number of ongoing connections and connection blocking probability increase as traffic intensity (i.e., connection arrival rate) increases. In this case, since BS-2 serves two edge routers, the number of ongoing WLAN connections

at this router is much higher than the other types of connections. However, as traffic load becomes high (e.g., arrival rate is higher than 0.6 connections per minute), connections from BS-2 experience high blocking probability due to the bottleneck at BS-1. Since BS-1 needs to allocate bandwidth to local WLAN connections and connections from standalone subscriber stations, it cannot allocate a large amount of bandwidth to relay connections from BS-2. Consequently, the admission control method at BS-1 rejects more relay connections. Since WLAN connections are prioritized over connections from standalone subscriber stations, at BS-1, the local WLAN connections experience lower blocking probability. On the other hand, even though BS-2 prioritizes WLAN connections, since it has to serve connections from two edge routers, blocking probability for WLAN connections is very high.

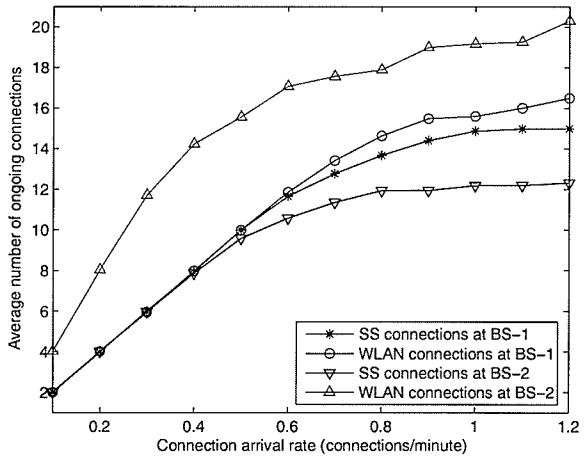


Figure 4.6. *Average number of ongoing connections under varying connection arrival rate (from simulation).*

As expected, the amount of bandwidth assigned to a connection decreases as traffic intensity increases. Local connections at BS-1 receive larger amount of bandwidth than relay connections from BS-2 (Fig. 4.8). Again, WLAN connections receive higher amount of bandwidth than connections from standalone subscriber stations.

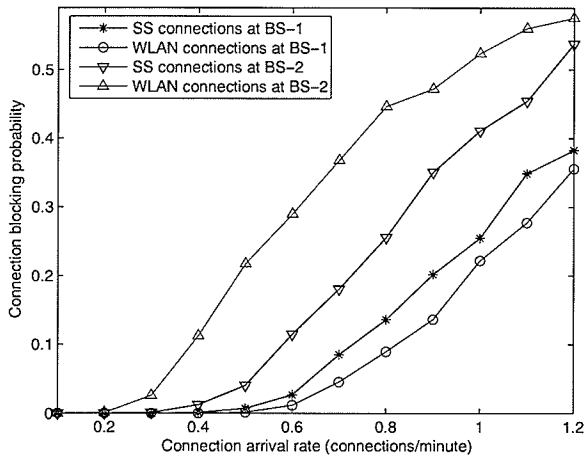


Figure 4.7. *Connection blocking probability under varying connection arrival rate (from simulation).*

4.4.3.5 Variation in Total Utility for Different Types of Connections Under Varying Connection Arrival Rates

Fig. 4.9 shows typical results on the variation in total utility obtained from the bargaining game analysis and simulations for different types of connections. As expected, when the connection arrival rate increases, the total utility (obtained from the analysis) first increases (since sufficient amount of bandwidth is available for each of the connections) and it decreases afterwards (since the bandwidth share for each connection decreases due to a large number of connections). However, in the simulation results, since admission control is applied to reject an incoming connection if admission of that connection reduces total utility, the total utility increases as traffic load increases. In this case, the utility is the highest for the relay connections since they are the largest in number (i.e., connections from standalone subscriber station and WLAN at BS-2).

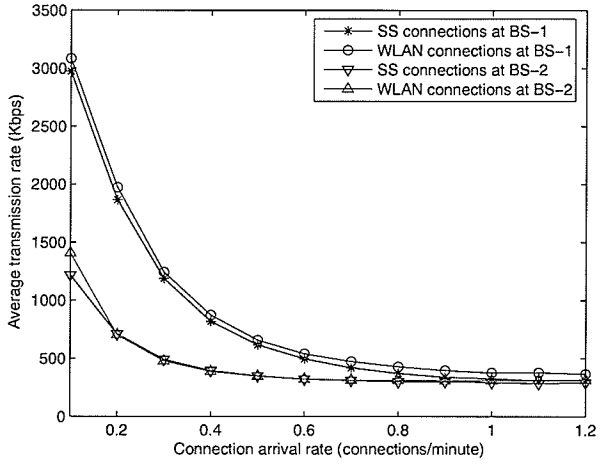


Figure 4.8. *Average amount of allocated bandwidth under varying connection arrival rate (from simulation).*

4.5 Chapter Summary

In this chapter, we have presented an architecture for integrating WLAN hotspots with WiMAX-based multihop broadband wireless mesh networks. The research issues related to protocol design have been outlined and some of the solution approaches proposed in the literature have been reviewed. To this end, for this integrated architecture we have presented a game-theoretic framework for radio resource management in the mesh routers. In particular, based on a bargaining game formulation, a bandwidth allocation scheme has been presented for fair resource allocation and an admission control policy has been proposed to maximize the utilities for the different types of connections. Typical numerical and simulation results have been presented to demonstrate the performances of the proposed radio resource management framework. Part of this chapter has been published in [64].

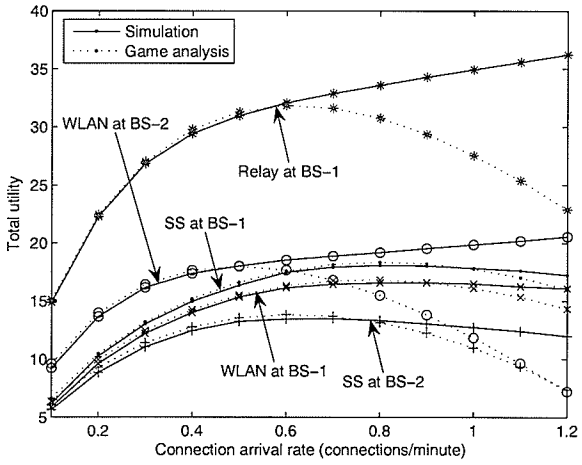


Figure 4.9. *Variation in total utility under varying connection arrival rate (from simulation).*

Chapter 5

A Cooperative Game Framework for Bandwidth Allocation in Heterogeneous Wireless Networks

5.1 Introduction

5.1.1 Problem Statement

In this chapter, the problems of an integration of different wireless technologies (i.e., heterogeneous wireless network) and its radio resource management framework are considered. This heterogeneous wireless network utilizes WiMAX, cellular, and WiFi networks to provide high speed wireless connectivity to mobile users. These networks are operated by different service providers which are cooperative to provide wireless connections to the users. The objective of the radio resource management framework is to satisfy the bandwidth requirement of the user, while the service providers are satisfied with the bandwidth assignment strategy. Also, the bandwidth assignment should be adaptable to the traffic load in the network. The inputs of this framework are the number of ongoing connections, and the bandwidth requirement of the new connection. The first output of this framework is accepting or rejecting the new connection. If the new connection is accepted, the second output is the amount of its allocated bandwidth from different networks.

5.1.2 Contribution

The bandwidth allocation and admission control algorithms are presented for wireless access in a heterogeneous network environment. In such an environment, a mobile station is assumed to have three different radio interfaces, namely, WiFi WLAN, CDMA cellular, and WiMAX WMAN radio interface. The objective of the proposed bandwidth allocation is to allocate the requested bandwidth to a new connection/session based on the available bandwidth in each network and the subscription level for that connection to each of the wireless access networks. This problem is formulated as *bankruptcy game* which is a special type of N-person cooperative game. A coalition is formed among the networks to ensure that the allocation satisfies all the networks in the system. A standard method in game theory, namely, *the core* is used to obtain the feasible bandwidth allocation scenarios. Then, to obtain the solution (i.e., the amount of allocated bandwidth in each network for a new connection/session), *Shapley value* is used. Based on the bandwidth allocation algorithm, an admission control method is proposed to ensure that the amount of bandwidth allocated (from all the networks) to the new connection is large enough to satisfy the corresponding user's requirement.

5.2 Related Work

The issues related to integration of diverse wireless access technologies such as cellular, WLAN and mobile ad hoc networks (MANET) with a view to providing quality-of-service (QoS) to the mobile users were studied in [65]. In [66], an adaptive transport layer (ATL) was proposed for heterogeneous wireless networks with the capabilities of adaptive congestion control, multimedia support, and providing fairness of transmission. In [67], a network selection mechanism in heterogeneous wireless networks was proposed. Specifically, *gray relation analysis* was used to decide which network should be used for each mobile. This decision is based on users' preference, service application and network condition. However, the problem of bandwidth allocation was not considered.

Game-theoretic framework (e.g., N-person cooperative game) was used to solve the resource management problem in wireless networks. In [68], a cooperative routing

protocol for MANETs was proposed. In this protocol, a coalition among the nodes in the network is formed to reduce energy consumption due to data transmission. The nodes are rewarded according to the contribution in the coalition. The payment for each node in the coalition is determined by Shapley value. In [69], a MAC scheme based on cooperative game was presented for wireless ad hoc networks. A bandwidth allocation scheme was proposed to achieve fairness in IEEE 802.11 distributed coordination function (DCF) network.

5.3 System Model

We consider a heterogeneous wireless access environment consisting of WiFi wireless LAN (WLAN), CDMA cellular network and WiMAX wireless MAN (WMAN) radio interfaces as shown in Fig. 5.1. A mobile with multiple radio transceivers (e.g., software radio) is able to connect to these radio access networks simultaneously. Each mobile in this system has different interfaces to connect to these networks.

We consider a geographical area that is totally covered by a WMAN base station and partly covered by a cellular base station and partly by a WLAN access point (AP) as denoted in Fig. 5.1 by number 1, 2 and 3, respectively. We assume that a mobile is able to connect to each network if it is in the corresponding coverage area and perfect power control is assumed to ensure uniform available transmission rate across the coverage area.

In the system model under consideration, a mobile can subscribe to different service classes in which each class has different bandwidth requirements. The subscription class is determined when the mobile initiates the connection and we assume that an ongoing connection remains in the same class until it terminates.

5.3.1 IEEE 802.11 WLAN

We consider IEEE 802.11-based radio interface with DCF medium access control (MAC) which is based on the carrier sense multiple access with collision avoidance (CSMA/CA) protocol. However, since CSMA/CA is a contention-based MAC protocol, we adopt a distributed reservation-based MAC protocol, namely, *early backoff announcement (EBA)* [70], which is an enhanced version of DCF. EBA is also back-

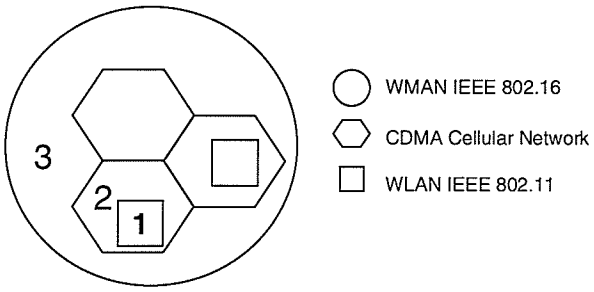


Figure 5.1. *Service area in a heterogeneous wireless network.*

ward compatibile with IEEE 802.11 DCF. By incorporating backoff information into MAC header, mobiles can completely avoid collisions.

5.3.2 CDMA Cellular Wireless Access

We consider a wideband CDMA cellular wireless access system [39]. For a certain number of active users in a cell, the signal-to-interference-plus-noise ratio (SINR) of uplink transmission for each mobile is determined from the transmit power of all other mobiles, background noise power at the base station, and intercell interference power.

5.3.3 WiMAX WMAN

We consider a WiMAX-based WMAN radio interface operating at 10-66 GHz band which supports data rate in the range of 32-130 Mbps depending on the bandwidth of operation as well as the modulation and the coding schemes. In the 10-66 GHz band, the signal propagation between a base station and a mobile is line-of-sight and single-carrier modulation is used. *WirelessMAN-SC* is the air interface specification for WiMAX operating in this frequency band. Depending on the channel quality (i.e., signal-to-noise ratio (SNR) at the receiver), different modulation schemes such as QPSK, 16-QAM and 64-QAM can be used.

We only consider *non-real-time polling service (nrtPS)* type of connections here. The amount of bandwidth required for such a connection is determined dynamically based on the required QoS performances of the corresponding connection. A 802.16 subscriber station (SS)/mobile uses contention-free (polling) mode to request bandwidth by using request PDU (BW-request) message.

5.4 Bandwidth Allocation and Admission Control in Heterogeneous Wireless Access Environment

The objective of the proposed admission control is to guarantee the total transmission rate requested by the new connection. The bandwidth allocation algorithm tries to allocate bandwidth from each network in a fair manner. In other words, each of these networks cooperates with each other to provide high bandwidth service to the new connection. Therefore, we use a *bankruptcy game* formulation which is a special type of N-person cooperative game to obtain the solution of the bandwidth allocation problem in a heterogeneous wireless access network.

In this section, we first describe a standard bankruptcy game. To obtain the solution of this game, the *coalition form* and the *characteristic function* for an N-person cooperative game are presented. Then, the stability of the game is analyzed through *the core*. Next, the solution of the bankruptcy game formulation is obtained by *Shapley value*. Finally, the bandwidth allocation and the admission control algorithms are presented.

5.4.1 Bankruptcy Game

To illustrate a bankruptcy game, let us assume that a company becomes bankrupt. This company owns money to N creditors, and therefore, this money is needed to be divided among these creditors. Typically, the sum of the claims from the creditors is larger than the money of the bankrupt company. This conflicting situation introduces an N-person game where the players of the game are seeking for the equilibrium point to divide the money. A detailed study and extensive analysis on this bankruptcy game was presented in [71].

The standard bankruptcy game can be expressed [72] by a finite set of agents $\mathbb{A}$, a real positive number M which denotes the amount of money and a nonnegative vector $\mathbf{d} \in \mathbb{R}^N$ of claims where the condition $\sum_{i \in \mathbb{A}} d_i \geq M$. To satisfy every agent, the solution of the bankruptcy game must have the following two properties:

- the money must be completely distributed and
- each agent has to obtain nonnegative money not exceeding the demand.

If x_i denotes the solution (i.e., amount of money distributed to agent i), the rule of this game can be expressed as follows:

$$0 \leq x_i \leq d_i, \qquad \forall i \in \mathbb{A} \tag{5.1}$$

$$\sum_{i \in \mathbb{N}} x_i = M. \tag{5.2}$$

5.4.2 Coalition Form and Characteristic Function

The bankruptcy game is an N-person cooperative game with *transferable utility (TU)* which allows *side payments* to be made among the players [73]. This side payment might be used by the players to reach the best strategy. Also, a coalition always exists in a bankruptcy game so that the agents (i.e., players) can cooperate with each other to gain better benefit. Since the number of players in such a game is larger than two, using strategic form of the game is cumbersome. Instead, the *coalition form* is preferred to represent such a game. Also, the payoff of coalition is expressed by the *characteristic function.*

A *coalition* $\mathbb{S}$ is defined as a subset of $\mathbb{A}$, $\mathbb{S} \subset \mathbb{A}$. In this case, $\emptyset$ and $\mathbb{A}$ denote an *empty coalition* and a *grand coalition*, respectively. The *coalition form* of an N-person game is defined by the pair $(\mathbb{A}, \nu)$ where ν is a *characteristic function* of the game. Two important properties of a characteristic function are:

1. $\nu(\emptyset) = 0$
2. if $\mathbb{S} \cap \mathbb{T} = \emptyset$, then $\nu(\mathbb{S}) + \nu(\mathbb{T}) \leq \nu(\mathbb{S} \cup \mathbb{T})$. This refers to the *superadditivity* property of the characteristic function.

The characteristic function can be obtained from [73]

$$\nu(\mathbb{S}) = \text{Value}_2 \left(\sum_{i \in \mathbb{S}} u_i(x_1, \ldots, x_n) \right) \tag{5.3}$$

for $x_i \in \mathbb{X}_i$, where $\mathbb{X}_i$ is the set of pure strategies of player i and $u_i(x_1, \ldots, x_n)$ is the payoff function for player i if player 1 chooses strategy x_1, player 2 chooses strategy x_2 and so on, and $\text{Value}_2(.)$ is the value of the 2-person game in which the first player is the set $\mathbb{S}$ and the second player is $\overline{\mathbb{S}} = \mathbb{A} - \mathbb{S}$.

In particular, for the bankruptcy game that we are considering here, the characteristic function can be defined as follows [72]:

$$\nu(\mathbb{S}) = \max\left(0, M - \sum_{j \notin \mathbb{S}} d_j\right) \tag{5.4}$$

for all possible coalition $\mathbb{S}$.

5.4.3 The Core

The core is generally used to obtain stability region for the solution of an N-person cooperative game. In this case, the concept of *imputation* must be established. Let the payoff vector $\mathbf{x} = [x_1, \ldots, x_i, \ldots, x_n]$ denote the amount received by agent i. This payoff vector is *group rational* if $\sum_{i=1}^{n} x_i = \nu(\mathbb{A})$. In particular, the highest total payoff can be achieved by forming a coalition among all agents. Also, the payoff vector is *individually rational* if $x_i \geq \nu(\{i\})$. That is, an agent will not agree to receive money less than that the agent could obtain without coalition. Then, the imputation is defined as the payoff vectors that is both group rational and individually rational, namely,

$$\mathbb{P} = \left\{ \mathbf{x} = [x_1, \ldots, x_n] \,\middle|\, \sum_{i \in \mathbb{A}} x_i = \nu(\mathbb{A}), \text{ and } x_i \geq \nu(\{i\}), \forall i \in \mathbb{A} \right\}. \tag{5.5}$$

An imputation $\mathbf{x}$ is unstable with coalition $\mathbb{S}$ if $\nu(\mathbb{S}) > \sum_{i \in \mathbb{S}} x_i$. Specifically, if the imputation is unstable, there is at least one agent who is unsatisfied due to the coalition. Then, the core is defined as the set $\mathbb{C}$ of stable imputations and can be expressed mathematically as follows [73]:

$$\mathbb{C} = \left\{ \mathbf{x} = [x_1, \ldots, x_n] \,\middle|\, \mathbf{x} \in \mathbb{P} \text{ and } \sum_{i \in \mathbb{S}} x_i \geq \nu(\mathbb{S}), \forall \mathbb{S} \subset \mathbb{A} \right\}. \tag{5.6}$$

The *core* is useful to obtain the stability condition of the game. However, it may contain several points and in some cases it could be empty. Therefore, the solution that provides the most preferable distribution strategy is required. In this chapter, we apply *Shapley value* which is one of the methods to obtain the solution of an N-person cooperative game.

5.4.4 Shapley Value

The solution of an N-person game can be obtained by several methods proposed in the literature (e.g., Shapley value, *nucleolus* and τ-value). However, we choose Shapley value for the solution of the bandwidth allocation problem since the computational complexity of this method is small and also we observe from simulations that the Shapley value provides relatively fair solution compared with other methods.

To compute Shapley value, let us define the *value function* $\phi(\nu)$ as the worth or value of agent i in the game with characteristic function ν, i.e., $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_i, \ldots, \phi_n]$. The Shapley value can be obtained by considering the money that an agent receives depending on the order that agent joins the coalition. In particular, the Shapley value is the average payoff to an agent if the agents enter into the coalition in a completely random order [73]. The Shapley value $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_i, \ldots, \phi_n]$ can be computed as follows:

$$\phi_i(\nu) = \sum_{\mathbb{S} \subset \mathbb{A}, i \in \mathbb{A}} \frac{(|\mathbb{S}| - 1)!\,(n - |\mathbb{S}|)!}{n!} \left(\nu(\mathbb{S}) - \nu(\mathbb{S} - \{i\})\right) \tag{5.7}$$

where $|\mathbb{S}|$ indicates the number of elements in the set $\mathbb{S}$.

5.4.5 Bandwidth Allocation Algorithm

Based on a standard bankruptcy game as described before, we propose a bandwidth allocation algorithm for a new connection which can be served simultaneously by three different wireless access networks, i.e., WLAN, cellular network and WMAN. Here, the mobile initiating the connection is analogous to the bankrupt company and the requested bandwidth is the money (estate) that has to be distributed among the different networks (i.e., agents). This situation leads to the similar conflict as in the bankruptcy problem in which each network tries to offer bandwidth as much as possible to gain revenue from new connection. Therefore, in this case, the total number of agents is $N = 3$ and the set of agents is defined as $\mathbb{A} = \{wl, ce, wm\}$ for WLAN, cellular wireless network and WMAN, respectively.

When a new connection requests for bandwidth M, a central controller (e.g., radio network controller (RNC)) determines the amount of offered bandwidth to this connection from each network. This offered bandwidth is a function of the

Table 5.1. *Notations and descriptions of the variables for bankruptcy game and proposed bandwidth allocation algorithm.*

Variable	Bankruptcy game	Bandwidth allocation
n	total number of agents	total number of networks
M	money (estate)	requested bandwidth
$\mathbb{A}$	set of agents	set of networks
d_i	claims of agent i	offered bandwidth by network i
x_i	solution of money distributed to agent i	bandwidth allocated to the new connection in network i

subscription class for that connection/mobile and the available bandwidth in each network. In particular, the offered bandwidth can be defined as follows:

$$d_i = \begin{cases} \tilde{b}_{k,i}, & \tilde{b}_{k,i} < \left(B_i^{(a)}\right)^r \\ \left(B_i^{(a)}\right)^r + \mathcal{N}\left(B_i^{(a)} - \left(B_i^{(a)}\right)^r\right), & \tilde{b}_{k,i} \geq \left(B_i^{(a)}\right)^r \end{cases} \tag{5.8}$$

where $\tilde{b}_{k,i}$ is the predefined offered bandwidth by network i to a new connection (or the corresponding mobile) with subscription class k, $B_i^{(a)}$ is the available bandwidth in network i, $b_k^{(req)}$ is the amount of requested bandwidth by a new connection in class k, $\mathcal{N}$ is a uniform random number between zero and one, and r is a control parameter which will be referred to as the bandwidth shaping parameter (i.e., $0 < r \leq 1$).

Note that with the above definition of offered bandwidth, network i offers bandwidth $\tilde{b}_{k,i}$ to a new connection under normal traffic load situation. However, when the network becomes congested (i.e., defined by the condition $\tilde{b}_{k,i} > \left(B_i^{(a)}\right)^r$) the offered bandwidth is gradually shaped by the random number $\mathcal{N}$ and the shaping parameter r to ensure that the network does not offer too much bandwidth to the new connection.

In the proposed bandwidth allocation algorithm, the Shapley value becomes the amount of allocated bandwidth in each network i, i.e., $x_i = \phi_i(\nu)$, $\forall i \in \mathbb{A}$. The notations and the descriptions of the variables for the bankruptcy game and the bandwidth allocation algorithm are shown in Table 5.1.

5.4.6 Admission Control Algorithm

The admission control algorithm ensures that the requested bandwidth of a new connection can be satisfied. When a mobile initiates a new connection, the information on the required bandwidth is sent to the central controller, which computes the offered bandwidth by each network. Then, the Shapley value is obtained. The new connection is accepted if $\sum_{i\in\mathbb{A}} x_i \geq b_k^{(req)}$ and $x_i \in \mathbb{C}$, $\forall i \in \mathbb{A}$ (i.e., the Shapley value is in the core, namely, the solution is stable), and rejected otherwise.

5.5 Numerical Study

5.5.1 Parameter Setting

In case of IEEE 802.11 WLAN, the data transmission rate is 11 Mbps and the maximum saturation throughput of one access point achieved through EBA is 6.2 Mbps [70]. For the CDMA cellular wireless access, the transmission bandwidth is assumed to be 5 MHz. We assume SINR is 8.17 dB so that the bit-error-rate is less than 10^{-4}. The total transmission rate in a CDMA cell is 2 Mbps. For the WiMAX-based wireless access, WirelessMAN-SC radio interface is used in the frequency band of 10 GHz with bandwidth of 25 MHz and 16-QAM modulation scheme with coding rate 1/2 to achieve a transmission rate of 50 Mbps in a single cell.

In the system under consideration, we assume three classes of mobile subscription and the bandwidth corresponding to these subscription levels are 200, 350 and 500 Kbps (i.e., $b_1^{(req)} = 200$, $b_2^{(req)} = 350$, $b_3^{(req)} = 500$). To calculate the offered bandwidth, we assume $\tilde{b}_{k,wl} = 200$, $\tilde{b}_{k,ce} = 150$, and $\tilde{b}_{k,wm} = 250$ for all k and we assume $r = 0.85$. Also, we assume that 50, 30, and 20 percent of the new connections in all the coverage areas are in the subscription class 1, 2, and 3, respectively. The connection arrival process is assumed to be Poisson and the connection holding time is assumed to be exponentially distributed.

5.5.2 The Core and Shapley Value

In this section, we demonstrate the calculation of the core and the Shapley value. Let us assume that a new connection requests for a bandwidth of 500 Kbps. According

to (5.4), characteristic function of all coalitions are as follows:

$$\begin{array}{lll} \nu(\emptyset) = 0, & \nu(\{wl\}) = 100, & \nu(\{wl, ce\}) = 250 \\ \nu(\mathbb{A}) = 500, & \nu(\{ce\}) = 50, & \nu(\{wl, wm\}) = 350 \\ & \nu(\{wm\}) = 150, & \nu(\{ce, wm\}) = 300. \end{array} \tag{5.9}$$

Since the number of networks is three, the core can be presented by *barycentric coordinates* as in Fig. 5.2. With this representation, the plane of the plot is denoted by $x_{wl}+x_{ce}+x_{wm} = \nu(\mathbb{A}) = 500$ and the edges of the triangle represent the characteristic function $\nu(\{i\})$. For example, the bottommost edge represents $\nu(\{wl\}) = 100$. The constraint of the core (i.e., $\sum_{i\in\mathbb{S}} x_i \geq \nu(\mathbb{S})$) is the line drawn across the triangle. For instance, the horizontal line represents $x_{ce} + x_{wm} = \nu(\{ce, wm\}) = 300$. Based on these constraints, the shaded areas represent the unstable imputations. For example, the topmost shaded area corresponds to an unstable imputation where the satisfaction for the cellular and the WMAN access networks is not achieved (i.e., WLAN provides too much bandwidth to the new connection). There is an area (the unshaded area in Fig. 5.2) that refers to the core (i.e., the solution space that makes the game stable). Based on (5.7), the Shapley value is $\phi = [166, 116, 217]$ which is denoted by the cross symbol in Fig. 5.2.

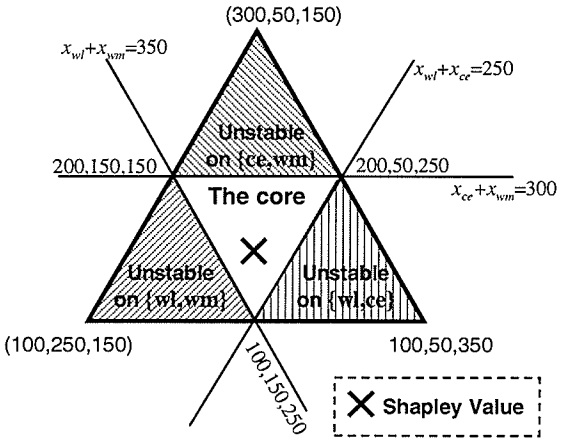


Figure 5.2. *Barycentric coordinates of the core and Shapley value for the numerical example.*

Next, we vary the amount of requested bandwidth and the resulting allocation in every network is shown in Figs. 5.3(a) and (b) for the normal case and for the case

when the cellular network becomes congested, respectively. As expected, to satisfy the bandwidth requirement for a new connection, the amount of allocated bandwidth increases as the requested bandwidth increases.

In Fig. 5.3(a), the allocation can be divided into four intervals according to the amount of requested bandwidth (i.e., $[0, 150)$, $[150, 400)$, $[400, 600)$ and $[600, \infty)$). In the first interval, the bandwidth allocation in every network is equal, since the entire amount of requested bandwidth can be accommodated by each network. Therefore, the fair way to allocate bandwidth is to allocate equal amount from each network. For the second interval, since the requested bandwidth becomes larger than that of offered from one of the networks, the bandwidth allocation in each network becomes different. In the third interval, the differences among the allocated bandwidth in each network become larger since the requested bandwidth is larger than the offered bandwidth in two networks. If the requested bandwidth becomes increasingly higher and becomes larger than the offered bandwidth in all of the networks, the allocated bandwidth becomes constant. Fig. 5.3(b) shows the case when the cellular network becomes congested (i.e., the bandwidth offered by this network becomes 50 Kbps). We observe that the trend of the allocated bandwidth in each network is still similar to that in Fig. 5.3(a).

5.5.3 Performances of Bandwidth Allocation and Admission Control Algorithms

In this section, we present the connection-level performances of the proposed bandwidth allocation and admission control algorithms. We consider the network as shown in Fig. 5.1. In this case, the mobiles are divided into three groups associated with service area 1 (i.e., mobiles under the coverage of WLAN, cellular network and WMAN), 2 (i.e., under the coverage of cellular network and WMAN), and 3 (i.e., under the coverage of WMAN only). The traffic intensity (i.e., connection arrival rate per minute) depends on the evaluation scenario and the connection holding time for the connections in area 1, 2 and 3 is assumed to be 20, 10, and 25 minutes, respectively. The average number of ongoing connections, bandwidth utilization and connection blocking probability under different traffic intensity are shown in Figs. 5.4-5.5. Moreover, we present the connection blocking probability obtained from the opportunistic

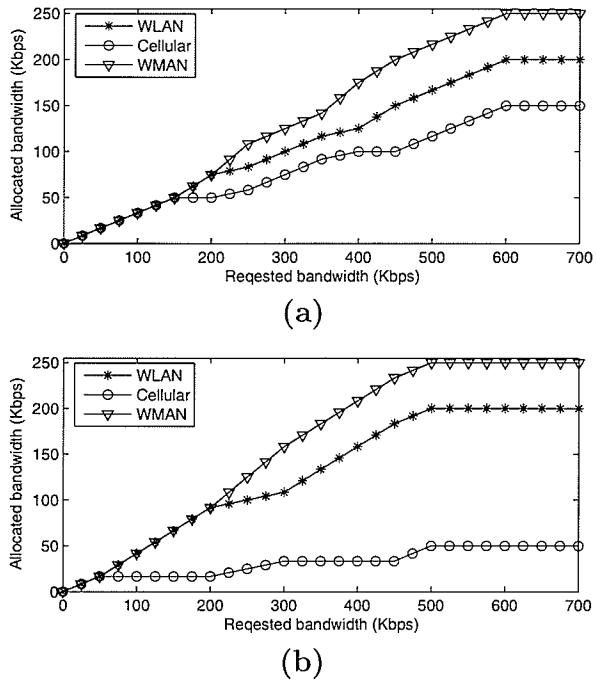


Figure 5.3. *Example of bandwidth allocation (a) in normal case and (b) when the cellular network becomes congested.*

network selection (i.e., when a mobile chooses the network with the largest available bandwidth) for the comparison purpose.

In particular, in Fig 5.4, the connection arrival rate in area 1 is equal to the traffic intensity as shown in x-axis while arrival rate in area 2 and 3 are half and one third of that traffic intensity, respectively. As expected, the average number of connections in each area, bandwidth utilization of each network, and connection blocking probability increase as the traffic intensity increases. However, the connection blocking probability for the proposed algorithm is smaller than that for the opportunistic scheme.

Fig. 5.5 shows the same performance measures for the case when the connection arrival rates in all the areas are equal to the traffic intensity. Note that this scenario is used to evaluate the performance when the traffic load is very high. Similar to Fig 5.4, all performance measures increase as the traffic intensity increases. However, we observe that when the traffic intensity is very high (i.e., traffic intensity is larger

than 6 connections per minute) the number of connections in area 3 decreases while that in area 1 still increases. When the traffic load in WMAN reaches a saturation point, the number of connections in this area cannot be increased anymore. However, the bandwidth allocation algorithm distributes the requested bandwidth (i.e., from WLAN and cellular network) to WMAN so that the bandwidth available to the connections in area 3 decreases while that in area 1 increases. Consequently, the number of connections in area 3 decreases. These observations are confirmed by the higher bandwidth utilization in the WMAN air interface in Fig. 5.5(b) compared with that in Fig. 5.4(b). Again, with the proposed algorithm, the connection blocking probability is smaller than that for the opportunistic scheme.

5.6 Chapter Summary

In this chapter, we have presented a bandwidth allocation and an admission control algorithm for heterogeneous wireless access networks in which a mobile can connect to multiple radio interfaces (i.e., WLAN, cellular network, and WMAN) simultaneously. To meet this requirement, a proper load balancing method is required. Therefore, we have formulated the problem of bandwidth allocation in such a system as a *bankruptcy game*. With a bankruptcy game, each network can cooperate to provide the requested bandwidth to a new connection. By using a well-developed game theory framework, the stability of the bandwidth allocation has been analyzed by using the concept of *the core* and the amount of the allocated bandwidth in each network has been obtained from the *Shapley value*. Based on this bandwidth allocation algorithm, an admission control scheme has been presented. We have presented numerical results to demonstrate the system performance under the proposed algorithms. Part of this chapter has been published in [75].

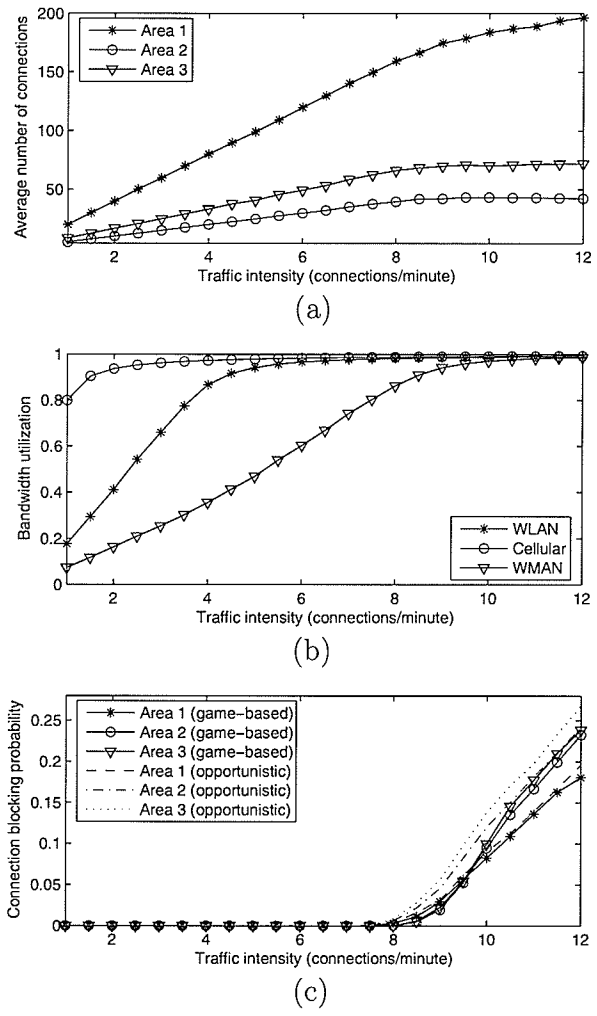


Figure 5.4. *(a) Average number of connections, (b) bandwidth utilization, and (c) connection blocking probability under unequal connection arrival rate.*

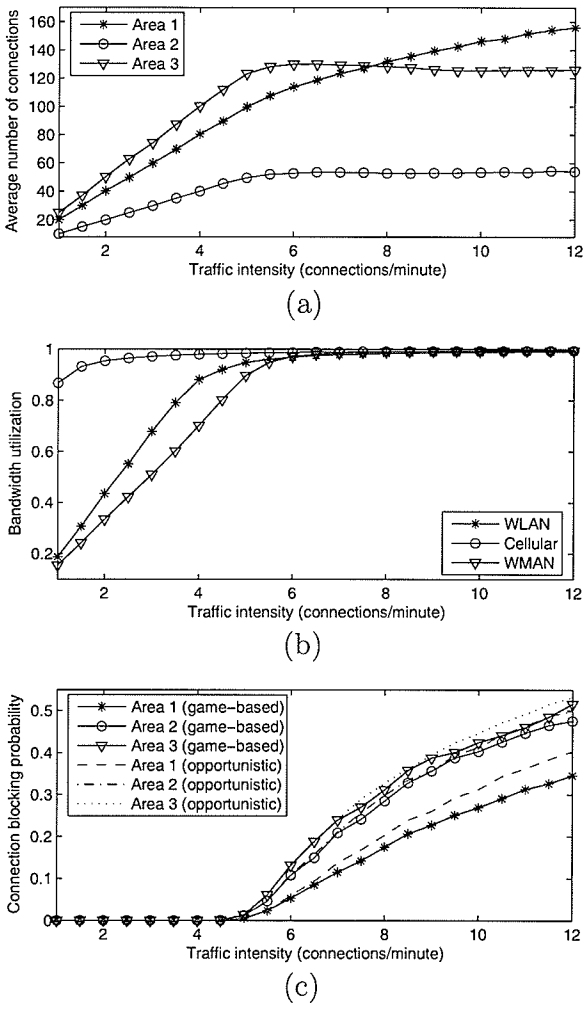


Figure 5.5. *(a) Average number of connections, (b) bandwidth utilization, and (c) connection blocking probability under equal connection arrival rate.*

Chapter 6

Radio Resource Management in Heterogeneous Wireless Access Networks

6.1 Introduction

6.1.1 Problem Statement

In this chapter, the problem of radio resource management in a heterogeneous wireless network is considered from service providers' point of view. This heterogeneous wireless network is composed of WiMAX-based WMAN, CDMA cellular network, and WiFi-based WLAN whose service providers are noncooperative. The objective of the radio resource management framework is to find an equilibrium point of bandwidth allocation such that all service providers are satisfied. The input of this framework is the number of users in a service area. The output is the bandwidth reserved for the different sub-areas and for different type of connections.

6.1.2 Contribution

A game-theoretic radio resource management (RRM) framework for wireless access in a heterogeneous network environment is presented. The objectives of the proposed framework are to maximize network utility through efficient resource allocation, achieve prioritization among different types of connections such as new connections and vertical and horizontal handoff connections, and ensure that the performance of

ongoing connections does not deteriorate due to accepting too many connections in a service area.

The problem of radio resource management is divided into two major parts, i.e., network-level and connection-level radio resource management. In the network-level, the limited available bandwidth of each network must be allocated to each service area so that the utilities of the different access networks, which are presumably operated by the different service providers, are maximized. In this competitive environment, therefore, we use a noncooperative game to obtain the solution of the bandwidth allocation by the different access networks in a service area. Again, for seamless mobility across the service areas, a portion of the radio resources in a service area needs to be reserved for handoff connections. Since new connections and handoff connections have to share the available bandwidth in a service area, an agreement on bandwidth reservation can be made so that the desired quality-of-service (QoS) performances (e.g., handoff connection dropping probability, new connection blocking probability) can be achieved. Therefore, we formulate a bargaining game to obtain the capacity (i.e., bandwidth) reservation thresholds for the different types of connections. Both network-level bandwidth allocation and capacity reservation can be performed on a long-term basis based on the statistics of the connections in the different service areas.

On the other hand, connection-level bandwidth allocation must be performed on a short-term basis and should be adapted upon arrival and departure of every connection in a service area. Again, each network (i.e., service provider) in a service area aims at maximizing it's utility while offering bandwidth to a new connection. Therefore, a connection-level noncooperative game is solved to obtain the offered bandwidth to the new connection.

6.2 Related Work

Recently, game theory has been widely used for resource management in wireless networks. In [79], the admission and rate control problem for CDMA systems was formulated as a noncooperative game. The formulation considered the choice of a user to churn from current provider to another. The decision on whether a new user can be

admitted or not and the allocated transmission are determined from the Nash equilibrium. An admission control game for CDMA networks was formulated in [80] to obtain an efficient and fair resource allocation for multiple classes of traffic. Game theory was also used to solve the power control problem in wireless networks [81][82][83]. However, all these works considered the radio resource management problem in a single wireless access network.

In a heterogeneous wireless access environment, a vertical handoff mechanism needs to consider not only the radio link and the physical layer parameters but also the network and the transport layer parameters. In [85], a framework for vertical handoff was presented where the handoff decision metrics include service type, data rate requirement, network condition, and cost of handoff. A dynamic optimization was proposed to provide QoS guarantee to the mobile users while maximizing the network utilization. In [86], a mobility management solution was proposed for heterogeneous wireless access networks to handle vertical handoff and network roaming. This solution was designed based on a formal policy representation model for decision-making process for inter-network mobility. However, maximization of network utility (from service providers' point of view) was not considered in these works.

The problem of integrating WLANs into the cellular wireless networks was investigated in the literature. In [87], a hierarchical radio resource management framework was designed to support seamless handoff between a WLAN and a cellular network. A hierarchical and distributed framework for seamless roaming across cellular networks and WLANs was proposed in [88]. The QoS mapping and internetwork message translation mechanisms were designed to support seamless handoff among multiple WLANs and cellular networks. However, a more general heterogeneous network architecture should be considered for radio resource management in the next generation wireless networks.

6.3 Model for Heterogeneous Wireless Access and the RRM Framework

6.3.1 System Model

We consider a heterogeneous wireless access environment consisting of IEEE 802.11 wireless LAN (WLAN), CDMA cellular network, and IEEE 802.16 wireless MAN (WMAN) radio interfaces as shown in Fig. 6.1. A mobile with multiple radio transceivers (e.g., implemented through software-defined radio) is able to connect to these radio access networks simultaneously.

We consider a geographic region which is entirely covered by a WMAN base station and partly covered by cellular base stations, and partly by WLAN access points (APs) as shown in Fig. 6.1. In Fig. 6.1, B denotes the amount of cellular network or WMAN. c_a and m_a are the amount of bandwidth assigned by the cellular network and WMAN respectively to the area a. Users in the different service areas in this region have access to different types and different numbers of wireless networks. In particular, in area 1, only WMAN service is available. In area 2 and area 4, services from cellular networks and WMAN are available. In area 3 and area 5, a mobile can connect to all three types of networks. Note that the RRM framework to be presented in this chapter can be applied to any other service area setting (different from the one shown in Fig. 6.1) in a considered geographic region.

Different wireless access networks are operated by different service providers. We assume that a mobile is able to connect to each of the networks in the corresponding service area and perfect power control is assumed to ensure uniform available transmission rate across the coverage area. A connection can be handed over between service areas with same types of wireless networks (e.g., from area 2 to area 4 both of which have services available from WMAN and cellular network) or between areas with different types of networks (e.g., from area 1 to area 2). We refer to the former as a horizontal handoff while the latter as a vertical handoff.

In this heterogeneous wireless access network, we assume that multi-interface mobile terminals are able to connect to three different wireless access networks simultaneously. These wireless access networks are IEEE 802.11 WLAN, CDMA cellular network, and IEEE 802.16 WMAN.

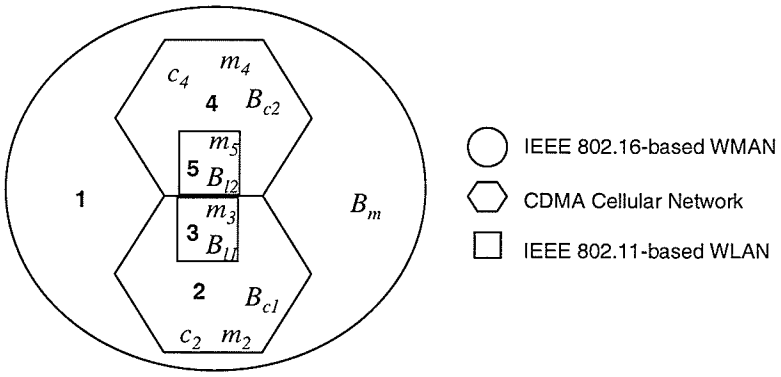


Figure 6.1. *Service areas under consideration in a heterogeneous wireless access environment.*

The key notations are listed in Table 6.1.

6.3.2 Radio Resource Management (RRM) Framework

The RRM framework is composed of four components: network-level bandwidth allocation for a service area, capacity reservation for the different types of connections, connection-level bandwidth allocation, and admission control (Fig. 6.2). In network-level bandwidth allocation, available bandwidth from different access networks are assigned to the service areas so that all of the service providers are satisfied with the allocation.

Based on the network-level bandwidth allocation, capacity reservation is used for service differentiation among new connections, vertical and horizontal handoff connections. In connection-level allocation, the required amount of bandwidth is allocated to an arriving connection in a service area from the different available access networks. Finally, the admission control component utilizes the results of capacity reservation together with connection-level allocation to decide whether an incoming connection can be accepted or not. Note that while network-level bandwidth allocation and capacity reservation can be performed in an off-line manner (on a long-term basis), connection-level allocation and admission control are performed in an on-line manner (on a short-term basis) upon arrival and departure of connections in a service area. The network-level bandwidth allocation and capacity reservation can be performed periodically (e.g., every hour) for which the statistics of the connections can be es-

Table 6.1. *List of key notations.*

λ, $1/\mu$	Average connection arrival rate, average connection holding time
N	Average number of ongoing connections
B	Total available bandwidth in these networks
m, c	Bandwidth allocated by WMAN and cellular network to a particular service area
$BR^{(net)}$	Best response of the network-level bandwidth allocation
A_a	Total bandwidth offered by all networks to service area a
C_a, R	Total capacity (i.e., number of connections) of service area a, bandwidth requirement for a connection
$c_a^{(v)}$, $c_a^{(h)}$	Capacities reserved for vertical and horizontal handoff connections
$\overline{n}_a$, $P_a^{(n)}$	Average number of ongoing connections, blocking probability for new connections
$P_a^{(v)}$, $P_a^{(h)}$	Dropping probabilities for vertical and horizontal handoff connections
$U^{(n)*}, U^{(v)*}, U^{(h)*}$	Equilibrium of the bargaining game
M	The number of services for each connection
p_i, ϕ_i	The amount of offered bandwidth to an incoming connection, profit of network i
$V_i^{(con)}$, $F_i(p_i)$	Revenue from connection gained by network i, cost of network i in offering bandwidth p_i
$\mathbb{P}$	Set containing the amount of bandwidth offered by all networks to an incoming connection
$U_{i,x}^{(con)}(.)$	Utility gained by network i from connection x
$\mathbb{Q}^{(x)}(p_i)$	Set the elements of which denote the amount of bandwidth allocated to connection x
$q_i^{(x)}$	Bandwidth currently allocated to ongoing connection x in network i
D_i	Bandwidth allocated by network i in a particular service area
$BR_i^{(con)}$	Best response of the connection-level bandwidth allocation
n	Current number of ongoing connections in a particular service area
Ψ	Difference between the best responses and the strategies adopted by other networks
p_j^{st}, p_j^{pr}, p_j^{cu}	Initial strategy, strategy in the previous iteration, strategy in the current iteration of network j

timated (e.g., connection arrival rate and connection holding time). In contrast, the connection-level allocation and admission control are performed are performed when the connection arrival and departure events occur (e.g., several seconds [89]).

In the system model under consideration, there are multiple service providers who noncooperatively offer wireless access services to the users. In particular, these service providers aim to maximize their utility while allocating bandwidth to the connections. Since each of the service providers has his own interest to maximize his payoff, the bandwidth allocation problem can be modeled as a noncooperative game the solution of which satisfies all of the service providers. Therefore, for network-level bandwidth allocation, assuming that the network service providers are rational, a noncooperative game is formulated and the solution is obtained from the Nash equilibrium (which will be shown to maximize the total network utility as well).

On the other hand, the problem of capacity reservation among new, horizontal handoff, and vertical handoff connections can be formulated as a resource sharing problem. In this case, a negotiation among the players can be performed to achieve efficient and fair sharing. This negotiation is practical in this case, since all of the connections (new, horizontal, and vertical handoff connections) are in the same service area. Therefore, a bargaining game formulation is used to obtain the fair allocation. In this bargaining game, new connections, vertical handoff and horizontal handoff connections negotiate with each other to obtain the reservation thresholds so that the connection-level QoS requirements (i.e., new connection blocking and handoff connection dropping probabilities) for the different types of connections are satisfied. In this case, we consider the equilibrium as the solution of the game. Both network-level bandwidth allocation and capacity reservation are performed based on the average number of ongoing connections in a service area which can be obtained from the network statistics at the steady state (i.e., long-term parameters).

For connection-level allocation, the problem of bandwidth allocation to an incoming connection in a service area is modeled as a trading market. In a service area, each network is assumed to be rational and self-interested to maximize its profit. All networks noncooperatively seek for the optimal strategy so that their profits are maximized. Also, these strategies must be stable in the sense that every firm is satisfied with the solution given other firms' strategies. We establish a revenue function for a

network offering bandwidth to an arriving connection by considering the utility that the network gains through allocation of bandwidth to the new connection and the cost function, which accounts for the loss of utility due to revocation of some bandwidth from the ongoing connections.

We formulate a noncooperative game and the Nash equilibrium is obtained as the solution of this game (i.e., the amount of bandwidth offered by each network to an incoming connection). We present two algorithms to obtain the solution, i.e., the search and the iterative algorithms. For the search algorithm, an optimization problem is formulated which is solved by a standard direct search method. On the other hand, the iterative algorithm takes advantage of the problem structure to obtain the solution iteratively. The rate of convergence of these algorithms is studied. We conduct extensive performance evaluation for the proposed radio resource management framework.

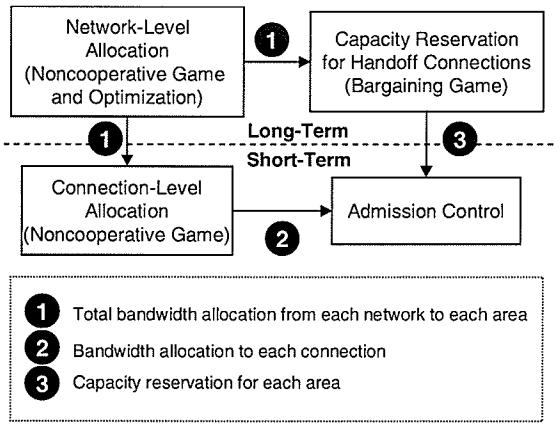


Figure 6.2. *Components of the proposed radio resource management framework.*

6.3.3 Cooperative and Noncooperative Approaches: Qualitative Comparison

In the previous chapter, the bandwidth allocation problem was formulated as a cooperative game (i.e., bankruptcy game) and the solution (i.e., the amount of bandwidth offered to a new connection) was obtained from Shapley value. This cooperative game approach is different from a noncooperative approach. In noncooperative approach,

each network is rational and selfish to maximize its own profit. Consequently, all networks compete with each other to achieve their objectives. More specifically, the difference between cooperative and noncooperative approaches lies in the fact that the former is group-oriented, while the latter is individual-oriented. In a cooperative approach, groups of players seek for fair resource allocation. On the other hand, in a noncooperative approach, allocation is performed based on the individual's profit gained from the resource. Also, the solution of noncooperative approach ensures that none of the players unilaterally deviates from the solution by changing the strategy to achieve higher payoff.

6.4 Network-Level Bandwidth Allocation and Capacity Reservation

6.4.1 Network-Level Bandwidth Allocation

The objective of network-level bandwidth allocation is to allocate bandwidth to a particular service area from each of the available networks in that service area so that all of the service providers are satisfied. This network-level allocation is required to ensure that a particular amount of bandwidth is reserved for each service area which is not affected by sudden traffic fluctuations in other service areas. To analyze network utility in a service area for the above bandwidth allocation, we use a utility function of throughput defined as follows [84]:

$$U_{i,x}^{(con)} = w \log(\alpha b) \tag{6.1}$$

where $U_{i,x}^{(con)}$ is the utility of network i for an allocated bandwidth of b to connection x, w and α are constants indicating the scale and the shape of the utility function.

We formulate a noncooperative game among the different access networks and the Nash equilibrium for pure strategy is obtained as the solution of this game. Then we show that the stable solution (i.e., Nash equilibrium) thus obtained also maximizes the long-term total utility of the entire network. For this, we require the average number of ongoing connections (in the steady state) in the different service areas which can be obtained either analytically or empirically. For simplicity, we use an

M/M/m/m queueing model to obtain the average number of ongoing connections in a service area. In particular, given the average connection arrival rate λ (i.e., sum of new connection and handoff connection arrival rates) and the average connection holding time $1/\mu$, the average number of ongoing connections can be obtained from

$$N = \sum_{i=1}^{T} i \left(\frac{(\rho)^i / i!}{\sum_{j=1}^{T} (\rho)^j / j!} \right) \tag{6.2}$$

where $\rho = \frac{\lambda}{\mu}$ and T is a large number (e.g., $T = 1000$).

For the service areas shown in Fig. 6.1, there are five wireless access networks, i.e., one WMAN, two cellular networks and two WLANs. Let the total available bandwidth in these networks be B_m, B_{c1}, B_{c2}, B_{l1}, and B_{l2}, respectively. Let N_i ($i\{1, \ldots, 5\}$) denote the average number of ongoing connections in area i, m_i denote the amount of bandwidth offered by WMAN to area i ($\sum_{i=1}^{5} m_i = B_m$), c_2 and c_3 denote bandwidth offered by cellular network to area 2 and area 3 ($c_2 + c_3 = B_{c1}$), and c_4 and c_5 denote bandwidth offered by cellular network to area 4 and area 5 ($c_4 + c_5 = B_{c2}$), respectively. Note that B_{l1} and B_{l2} denote the amount of bandwidth available from WLANs in area 3 and area 5, respectively.

6.4.1.1 Noncooperative Game for Network-Level Bandwidth Allocation

The formulation of the noncooperative game can be described as follows:

- **Players:** The two players of this game are the WMAN and the cellular network for a service area in Fig. 6.1[1].
- **Strategies:** The strategies for the WMAN are m_i ($i = \{2, \ldots, 5\}$) and those for the cellular network are c_j ($j = \{2, 4\}$) (Fig. 6.1).
- **Payoffs:** The payoff for the WMAN is the utility gained from offering bandwidth m_i to all service areas, and that for the cellular network is the utility gained from offering bandwidth c_2, c_3, c_4, and c_5 to service areas 2, 3, 4, and 5, respectively.

Specifically, the payoff for the WMAN is given as follows:

[1]Since a WLAN covers only one service area, WLANs are not considered in this game.

$$U_{wman}(m_i, c_j) =$$
$$w\left[\left(N_1 \log\left(\alpha\frac{B_m - \sum_{i=2}^{5} m_i}{N_1}\right)\right) + \left(N_2 \log\left(\alpha\frac{c_2 + m_2}{N_2}\right) - N_2 \log\left(\alpha\frac{c_2}{N_2}\right)\right) + \right.$$
$$\left(N_3 \log\left(\alpha\frac{B_{l1} + B_{c1} - c_2 + m3}{N_3}\right) - N_3 \log\left(\alpha\frac{B_{l1} + B_{c1} - c_2}{N_3}\right)\right) +$$
$$\left(N_4 \log\left(\alpha\frac{c_4 + m_4}{N_4}\right) - N_4 \log\left(\alpha\frac{c_4}{N_4}\right)\right) + \left(N_5 \log\left(\alpha\frac{B_{l2} + B_{c2} - c_4 + m_5}{N_5}\right) -\right.$$
$$\left.\left. N_5 \log\left(\alpha\frac{B_{l2} + B_{c1} - c_4}{N_5}\right)\right)\right] \tag{6.3}$$

and the payoff for a cellular network is given as follows:

$$U_{cell}(m_i, c_j) =$$
$$w\left[\left(N_2 \log\left(\alpha\frac{c_2 + m_2}{N_2}\right) - N_2 \log\left(\alpha\frac{m_2}{N_2}\right)\right) + \left(N_3 \log\left(\alpha\frac{B_{l1} + B_{c1} - c_2 + m_3}{N_3}\right)\right.\right.$$
$$\left.\left. - N_3 \log\left(\alpha\frac{B_{l1} + B_{c1} + m_3}{N_3}\right)\right) + \left(N_4 \log\left(\alpha\frac{c_4 + m_4}{N_4}\right) - N_4 \log\left(\alpha\frac{m_4}{N_4}\right)\right) +$$
$$\left.\left(N_5 \log\left(\alpha\frac{B_{l2} + B_{c2} - c_4 + m_5}{N_5}\right) - N_5 \log\left(\alpha\frac{B_{l1} + B_{c2} + m_5}{N_5}\right)\right)\right]. \tag{6.4}$$

Here, the payoff for an access network is determined from the surplus utility on that due to the allocations by other networks.

The *Nash equilibrium* of a game is a strategy profile (list of strategies - one for each player) with the property that no player can increase his payoff by choosing a different action given the other players' actions [91]. The pure strategy pair (m_i, c_j) is a *Nash equilibrium* if

$$\begin{aligned} U_{wman}(m_2^*, m_3^*, m_4^*, m_5^*, c_2^*, c_4^*) &\geq U_{wman}(m_2, m_3, m_4, m_5, c_2^*, c_4^*) \quad \forall m_2, m_3, m_4, m_5 \\ U_{cell}(m_2^*, m_3^*, m_4^*, m_5^*, c_2^*, c_4^*) &\geq U_{wman}(m_2^*, m_3^*, m_4^*, m_5^*, c_2, c_4) \quad \forall c_2, c_4. \end{aligned}$$

To determine the *Nash equilibrium*, we use *best response functions* for both players. The best response function for the WMAN at the network level $BR_{wman}^{(net)}(c_2', c_4')$, given that the cellular network chooses strategy (c_2', c_4'), is obtained by finding strategy m_i that maximizes the utility of the WMAN, i.e.,

$$(m_2, m_3, m_4, m_5) = \arg\max_{m_2, m_3, m_4, m_5} U_{wman}(m_2, m_3, m_4, m_5, c_2', c_4'). \tag{6.5}$$

Similarly, the best response function for the cellular network (at the network level) $BR_{cell}^{(net)}(m_2', m_3', m_4', m_5')$, given that the WMAN chooses strategy (m_2', m_3', m_4', m_5'), is expressed as follows:

$$(c_2, c_4) = \arg\max_{c_2,c_4} U_{cell}(m_2', m_3', m_4', m_5', c_2, c_4). \tag{6.6}$$

The set of strategies $(m_2^*, m_3^*, m_4^*, m_5^*, c_2^*, c_4^*)$ is a *Nash equilibrium* if and only if $(m_2^*, m_3^*, m_4^*, m_5^*) = BR_{wman}^{(net)}(c_2^*, c_4^*)$ and $(c_2^*, c_4^*) = BR_{cell}^{(net)}(m_2^*, m_3^*, m_4^*, m_5^*)$.

To obtain the best response function for WMAN at the network level, we differentiate U_{wman} with respect to m_2, m_3, m_4, and m_5 assuming that c_2 and c_4 are constants. Then, we have

$$\frac{\partial U_{wman}(m_i, c_j)}{\partial m_2} = 0 \quad \Rightarrow$$
$$N_1(m_2 + c_2) = N_2(B_m - m_2 - m_3 - m_4 - m_5) \tag{6.7}$$
$$\frac{\partial U_{wman}(m_i, c_j)}{\partial m_3} = 0 \quad \Rightarrow$$
$$N_1(m_3 + B_{c1} - c_2 + B_{l1}) = N_3(B_m - m_2 - m_3 - m_4 - m_5) \tag{6.8}$$
$$\frac{\partial U_{wman}(m_i, c_j)}{\partial m_4} = 0 \quad \Rightarrow$$
$$N_1(m_4 + c_4) = N_4(B_m - m_2 - m_3 - m_4 - m_5) \tag{6.9}$$
$$\frac{\partial U_{wman}(m_i, c_j)}{\partial m_5} = 0 \quad \Rightarrow$$
$$N_1(m_5 + B_{c2} - c_4 + B_{l2}) = N_5(B_m - m_2 - m_3 - m_4 - m_5) \tag{6.10}$$

and for the cellular network, we have

$$\frac{\partial U_{cell}(m_i, c_j)}{\partial c_2} = 0 \quad \Rightarrow \quad N_2\left(B_{l1} + B_{c1} - c_2 + m_3\right) = N_3\left(c_2 + m_2\right) \tag{6.11}$$
$$\frac{\partial U_{cell}(m_i, c_j)}{\partial c_4} = 0 \quad \Rightarrow \quad N_4\left(B_{l2} + B_{c2} - c_4 + m_5\right) = N_5\left(c_4 + m_4\right). \tag{6.12}$$

By solving the above system of equations (i.e., (6.7)-(6.12)), we can obtain the solution of the game which is the Nash equilibrium and satisfies both the WMAN and the cellular network.

6.4.1.2 Optimization of Total Network Utility

The total utility of the entire network (i.e., utility for connections in the entire service region) can be obtained from:

$$U_{tot}(m_i, c_i) =$$
$$w\left[N_1 \log\left(\alpha\frac{m_1}{N_1}\right) + N_2 \log\left(\alpha\frac{m_2+c_2}{N_2}\right) + N_3 \log\left(\alpha\frac{m_3+c_3+B_{l1}}{N_3}\right)\right.$$
$$\left. + N_4 \log\left(\alpha\frac{m_4+c_4}{N_4}\right) + N_5 \log\left(\alpha\frac{m_5+c_5+B_{l2}}{N_5}\right)\right]. \tag{6.13}$$

To maximize total utility, an optimization problem is formulated as follows:

$$\begin{aligned} &\text{Maximize:} && U_{tot}(m_i, c_i) \\ &\text{Subject to:} && \sum_{i=1}^{5} m_i = B_m \\ & && c_2 + c_3 = B_{c1}, \quad c_4 + c_5 = B_{c2} \end{aligned} \tag{6.14}$$

where the decision variables are m_2, m_3, m_4, m_5, c_2 and c_4. To obtain the optimal solution, we differentiate (6.13) with respect to each of the decision variables as follows:

$$\frac{\partial U_{tot}}{\partial m_2} = 0 \Rightarrow \frac{N_1}{B_m - m_2 - m_3 - m_4 - m_5} = \frac{N_2}{m_2 + c_2} \tag{6.15}$$
$$\frac{\partial U_{tot}}{\partial m_3} = 0 \Rightarrow \frac{N_1}{B_m - m_2 - m_3 - m_4 - m_5} = \frac{N_3}{m_3 + B_{c1} - c_2 + B_{l1}} \tag{6.16}$$
$$\frac{\partial U_{tot}}{\partial m_4} = 0 \Rightarrow \frac{N_1}{B_m - m_2 - m_3 - m_4 - m_5} = \frac{N_4}{m_4 + c_4} \tag{6.17}$$
$$\frac{\partial U_{tot}}{\partial m_5} = 0 \Rightarrow \frac{N_1}{B_m - m_2 - m_3 - m_4 - m_5} = \frac{N_5}{m_5 + B_{c2} - c_4 + B_{l2}} \tag{6.18}$$
$$\frac{\partial U_{tot}}{\partial c_2} = 0 \Rightarrow \frac{N_2}{m_2 + c_2} = \frac{N_3}{m_3 + B_{c1} - c_2 + B_{l1}} \tag{6.19}$$
$$\frac{\partial U_{tot}}{\partial c_4} = 0 \Rightarrow \frac{N_4}{m_4 + c_4} = \frac{N_5}{m_5 + B_{c2} - c_4 + B_{l2}}. \tag{6.20}$$

Then, the optimal values of m_2, m_3, m_4, m_5, c_2, and c_4 (i.e., m_2^*, m_3^*, m_4^*, m_5^*, c_2^*, and c_4^*) are obtained by solving the above system of equations.

Obviously, (6.7), (6.8), (6.9), (6.10), (6.11), and (6.12) are the same as (6.15), (6.16), (6.17), (6.18), (6.19), and (6.20). This shows that the solution of the optimization formulation, which maximizes total network utility, is basically the Nash equilibrium obtained from the noncooperative game formulation described before.

6.4.2 Capacity Reservation

6.4.2.1 Prioritization among Different Types of Connections and Connection-Level Performance Measures

In a heterogeneous wireless access network, connections can be handed off between service areas with same wireless access technologies (e.g., from area 2 to area 4 in Fig. 6.1) or between areas with different access technologies (e.g., from area 2 to area 3). While the former is referred to as a horizontal handoff, the latter is referred to as a vertical handoff. To achieve prioritization among horizontal handoff connections, vertical handoff connections, and new connections, a portion of system capacity in a service area needs to be reserved for high-priority connections (e.g., vertical and horizontal handoff connections). For this, a simple but effective guard channel [92] scheme was proposed in the literature. We adopt this guard channel concept to reserve a portion of system capacity for handoff connections. The parameters for this reservation scheme (i.e., the reservation thresholds) are obtained from a bargaining game formulation among the different types of connections. We consider a bargaining game for capacity reservation since all types of connections (i.e., new and handoff connections) in a particular service area need to share the limited bandwidth (i.e., capacity) offered by each network. The payoff for each type of connections is a function of connection blocking or dropping probability. In this case, the available capacity needs to be allocated to satisfy all types of connections in a service area. Note that if there are two types of connections (e.g., new and horizontal handoff connections), the same bargaining game formulation is still applicable in which the number of players is two.

If A_a denotes the total bandwidth offered by all networks to service area a (e.g., $A_2 = m_2 + c_2$), which is obtained from the network-level allocation, the total capacity (i.e., number of connections) of service area a can be obtained from $C_a = \lfloor \frac{A_a}{R} \rfloor$ where R is the bandwidth requirement for a connection. In order to prioritize handoff connections over new connections, capacity $c_a^{(v)}$ and $c_a^{(h)}$ are reserved for vertical and horizontal handoff connections, respectively. In particular, a new connection can be accepted if the current number of ongoing connections is less than $\min(C_a - c_a^{(v)}, C_a - c_a^{(h)})$. A vertical and a horizontal handoff connection can be accepted if the current number of ongoing connection is less than $C_a - c^{(v)}$ and $C_a - c^{(h)}$, respectively.

Based on an M/M/m/m queueing model, we can analytically obtain the average number of ongoing connections $\overline{n}_a$, blocking probability for new connections $P_a^{(n)}$, and dropping probabilities for vertical and horizontal handoff connections $P_a^{(v)}$ and $P_a^{(h)}$, respectively.

To quantify users' satisfaction as a function of the above connection-level performance measures, we use the following utility function [93]:

$$U = \frac{1}{1-\sigma} \max\left(0, 1 - \sigma \exp(P_a)\right) \tag{6.21}$$

where σ $(0 < \sigma < 1)$ is the parameter of the utility function. In particular, the larger the value of σ, the more sensitive is the utility to the connection blocking or connection dropping probability P_a.

6.4.2.2 Bargaining Game Formulation

To obtain the values of $c_a^{(v)} \geq 0$ and $c_a^{(h)} \geq 0$ which satisfy all the new connections, the vertical and horizontal handoff connections, we formulate a bargaining game in which different types of connections negotiate with each other to reach the equilibrium. A bargaining game can ensure fair payoffs for all the players while achieving prioritization among different types of connections. Note that a bargaining game is required for each service area.

The bargaining game formulation for capacity reservation can be described as follows:

- **Players:** The total number of players is three - an incoming new connection, a vertical handoff connection, and a horizontal handoff connection.
- **Strategies:** The strategies for the vertical handoff connection, the horizontal handoff connection, and the new connection are the reserved capacities, i.e., $c_a^{(v)}$, $c_a^{(h)}$, and $C_a - \max\left(c_a^{(v)}, c_a^{(h)}\right)$, respectively.
- **Payoffs:** The payoff for each connection (i.e., player) is the utility U obtained as a function of the corresponding connection-level performance. We use $U^{(n)}$, $U^{(v)}$, and $U^{(h)}$ to denote the payoffs for the new connection, the vertical and the horizontal handoff connections, respectively.

The payoff (i.e., utility) for the players is given by $\Omega = \{(U^{(n)}, U^{(v)}, U^{(h)}) : 0 \leq U^{(n)}, U^{(v)}, U^{(h)} \leq 1\}$ (i.e., feasible set). If an agreement among the players can-

not be reached, the utility that the players will receive is given by the threat point $(\tilde{U}^{(n)}, \tilde{U}^{(v)}, \tilde{U}^{(h)})$. In particular, $(\tilde{U}^{(n)}, \tilde{U}^{(v)}, \tilde{U}^{(h)}) = (0, 0, 0)$ is the threat point for this game. This threat point represents that if the game fails (and hence the resource reservations cannot be made), the players will start negotiation again.

The bargaining game is formulated as $\mathcal{F}(\Omega, \tilde{U}^{(n)}, \tilde{U}^{(v)}, \tilde{U}^{(h)}) = (U^{(n)*}, U^{(v)*}, U^{(h)*})$, where $(U^{(n)*}, U^{(v)*}, U^{(h)*})$ denotes the solution (equilibrium) [61] of the game $\mathcal{F}(.)$. The Pareto optimality of the game defines an agreement such that one player cannot increase his utility without decreasing the utility of anyone of the other players. Specifically, the utility triplet $(U^{(n)}, U^{(v)}, U^{(h)})$ is Pareto optimal if there exists a utility triplet $(\hat{U}^{(n)}, \hat{U}^{(v)}, \hat{U}^{(h)})$ such that if $U^{(n)} \geq \hat{U}^{(n)}$, $U^{(v)} \geq \hat{U}^{(v)}$, and $U^{(h)} \geq \hat{U}^{(h)}$, then $(\hat{U}^{(n)}, \hat{U}^{(v)}, \hat{U}^{(h)}) = (U^{(n)}, U^{(v)}, U^{(h)})$. The Pareto optimality can provide the candidate strategies (i.e., $c_a^{(h)}$ and $c_a^{(v)}$) for which one of the players can achieve the highest utility. Then, we need the equilibrium of the game such that all the players are satisfied with the utilities they receive. That is, the equilibrium of the bargaining game is the utility triplet $(U^{(n)*}, U^{(v)*}, U^{(h)*})$ such that [61]

$$(U^{(n)*}, U^{(v)*}, U^{(h)*}) = \arg\max_{U^{(n)}, U^{(v)}, U^{(h)}} \left(U^{(n)} - \tilde{U}^{(n)}\right)\left(U^{(v)} - \tilde{U}^{(v)}\right)\left(U^{(h)} - \tilde{U}^{(h)}\right). \tag{6.22}$$

Since the state space of the available strategies (i.e., $(c_a^{(h)}, c_a^{(v)})$) is small, we obtain the solution of this bargaining game by enumeration.

6.5 Connection-Level Bandwidth Allocation and Admission Control

While network-level bandwidth allocation and capacity reservation are performed on a long-term basis based on the average number of ongoing connections in a service area at the steady state, connection-level allocation is performed in each service area upon arrival/departure of a connection (i.e., on a short-term basis). When a connection arrives, the amount of bandwidth to be offered to an incoming connection is determined and also the admission control procedure is invoked. When a connection departs a service area, the bandwidth released from the connection is distributed among the ongoing connections so that the resource utilization can be maximized.

6.5.1 Noncooperative Game for Connection-Level Bandwidth Allocation

For connection-level bandwidth allocation, we formulate a noncooperative game among the networks available in a particular service area (e.g., WMAN and cellular network in area 2, and WMAN, cellular network and WLAN in area 4 in Fig. 6.1). These networks compete with each other to offer bandwidth to an incoming connection (i.e., new connection, vertical handoff connction, or horizontal handoff connection) in order to maximize their profits (i.e., utility), and also all the networks are satisfied with the solution of the game. The decision on the amount of offered bandwidth to a new connection by a network depends on the actions taken by other networks. In particular, one network will receive small utility from a new connection if other networks offer large amount of bandwidth to that connection. However, if a particular network has many ongoing connections, offering a large amount of bandwidth to a new connection will degrade the network utility. We consider this performance degradation as a cost of offering bandwidth to a new connection. This conflicting situation is modeled as a noncooperative game.

6.5.1.1 Formulation of the Game

The number of networks providing service to a new connection in a particular service area is M ($M \in \{2, 3\}$). The supplied service here is the bandwidth and a new connection is a customer in the market. Based on this model, the strategic form of the game can be defined as follows:

- **Players:** The players in this game are the networks available in a particular service area.
- **Strategies**: The strategy of each of the players is the amount of offered bandwidth to an incoming connection (denoted by p_i for network i) which in non-negative.
- **Payoffs:** The payoff for each player is the profit (i.e., revenue minus cost) of network i (denoted by ϕ_i) in offering bandwidth to an incoming connection.

The revenue (i.e., utility) of network i is computed from the surplus utility over that due to the allocations by other networks and can be obtained as follows:

$$V_i^{(con)}(\mathbb{P}) = w\left[\log\left(\alpha\sum_j p_j\right) - \log\left(\alpha\sum_{j\neq i} p_j\right)\right] \tag{6.23}$$

where $\mathbb{P}$ is the set containing the amount of bandwidth offered by all networks to an incoming connection (i.e., pure strategies of all networks). This set is defined as follows: $\mathbb{P} = \left\{ \cdots \quad p_j \quad \cdots \quad p_i \quad \cdots \right\}$.

Since the bandwidth offered to an incoming connection must be taken from the ongoing connections, for network i, the cost of offering bandwidth p_i can be considered as a loss in total utility. This cost can be calculated as follows:

$$F_i(p_i) = \sum_x \left(U_{i,x}^{(con)}\left(\mathbb{Q}^{(x)}(0)\right) - U_{i,x}^{(con)}\left(\mathbb{Q}^{(x)}(p_i)\right)\right) \tag{6.24}$$

where $U_{i,x}^{(con)}(.)$ is the utility gained by network i from connection x (as in (6.1)) and $\mathbb{Q}^{(x)}(p_i)$ represents a set the elements of which denote the amount of bandwidth allocated to connection x. Specifically, p_i is the amount of bandwidth offered to an incoming connection by network i. Note that the bandwidth allocated to the ongoing connections in a service area depends on the network-level bandwidth allocation. The set $\mathbb{Q}^{(x)}(p_i)$ can be defined as follows:

$$\mathbb{Q}^{(x)}(p_i) = \left\{ \cdots \quad \frac{D_k q_k^{(x)}}{\sum_x q_k^{(x)}} \quad \cdots \quad \frac{(D_i - p_i) q_i^{(x)}}{\sum_x q_i^{(x)}} \quad \cdots \right\} \tag{6.25}$$

where $q_i^{(x)}$ is the bandwidth currently allocated to connection x in network i, and D_i is the bandwidth allocated by network i in a particular service area (i.e., $D_i \in \{m_j, c_k, B_l\}$, B_l is the WLAN bandwidth in a service area). In particular, the first term of the cost function in (6.24) is the total utility of network i gained from the ongoing connections and the second term is the total utility of network i after offering bandwidth p_i to an incoming connection. The profit of network i in offering bandwidth p_i to an arriving connection is then defined as follows:

$$\phi_i(\mathbb{P}) = V_i^{(con)}(\mathbb{P}) - f_i F_i(p_i) \tag{6.26}$$

where f_i (which is a positive real number) denotes the weight of cost function for network i.

6.5.1.2 Nash Equilibrium of the Noncooperative Game

The Nash equilibrium is obtained by using the best response function which is the best strategy of one player given other players' strategies. The best response function for network i (in the connection-level) given the amount of bandwidth offered by other networks p'_j (for $j \neq i$) is defined as follows:

$$BR_i^{(con)}(\mathbb{P}') = \arg\max_{p_i} \phi_i(\mathbb{P}) \tag{6.27}$$

where set $\mathbb{P}'$ contains the amount of bandwidth offered by network j ($j \neq i$) and it can be defined as $\mathbb{P}' = \left\{ \cdots \quad p'_j \quad \cdots \right\}$. In this case, set $\mathbb{P}$ can be obtained from set $\mathbb{P}'$ as follows $\mathbb{P} = \mathbb{P}' \cup \{p_i\}$.

For a service area with three wireless access networks, we can express the best response function, given other networks' strategies p'_j and p'_k, as follows:

$$BR_i^{(con)}\left(\{p'_j, p'_k\}\right) = \arg\max_{p_i} \phi_i\left(\{p'_j, p'_k, p_i\}\right). \tag{6.28}$$

The set $\mathbb{P}^* = \{p_j^*, p_k^*, p_i^*\}$ denotes the Nash equilibrium of this game if and only if

$$p_i^* = BR_i^{(con)}\left(\{p_j^*, p_k^*\}\right), \quad p_j^* BR_j^{(con)}\left(\{p_i^*, p_k^*\}\right), \quad p_k^* BR_k^{(con)}\left(\{p_i^*, p_j^*\}\right). \tag{6.29}$$

Similarly, in presence of two networks, the best response function can be expressed as follows: $BR_i^{(con)}\left(\{p'_j\}\right) = \arg\max_{p_i} \phi_i\left(\{p'_j, p_i\}\right)$. The set $\mathbb{P}^* = \{p_j^*, p_i^*\}$ denotes Nash equilibrium of this bandwidth allocation game in the two-network case if $p_i^* = BR_i^{(con)}\left(\{p_j^*\}\right)$, $p_j^* = BR_j^{(con)}\left(\{p_i^*\}\right)$.

Observation 1 *The Nash equilibrium of the above noncooperative game is unique.*

Proof. To prove the uniqueness of Nash equilibrium, we consider particularly the case with two networks - WMAN and cellular network. First, we obtain the best response function by differentiating the profit function in (6.26) with respect to p_i given p'_j, where $\mathbb{P} = \{p_i, p'_j\}$. The profit function can be expressed in a general form as follows:

$$\begin{aligned}\phi_i(\{p_i, p'_j\}) \;=\; w\Bigg[&\log\left(\alpha(p_i + p'_j)\right) - \log\left(\alpha p'_j\right) - f_i\left(\sum_y \log\left(\alpha \frac{D_i q_i^{(y)}}{\sum_x q_i^{(x)}}\right) - \right. \\ &\left.\left. \sum_y \log\left(\alpha \frac{(D_i - p'_j) q_i^{(y)}}{\sum_x q_i^{(x)}}\right)\right)\right].\end{aligned} \tag{6.30}$$

Now

$$\frac{\partial \phi_i(\{p_i, p'_j\})}{\partial p_i} = 0 \tag{6.31}$$

gives

$$p_i^* = \frac{D_i - f_i n p'_j}{1 + f_i n} \tag{6.32}$$

where n is the current number of ongoing connections in a particular service area. Similarly, for the best response function of network j, we have

$$p_j^* = \frac{D_j - f_j n p'_i}{1 + f_j n}. \tag{6.33}$$

Therefore, the best response function for each of these two networks is a linear function of the strategy adopted by the other network.

The Nash equilibrium is unique if and only if the slopes of these linear best response functions are unequal. We show this by reformatting (6.33) as follows:

$$p'_i = \frac{D_j - p_j^*(1 + f_j n)}{f_j n}. \tag{6.34}$$

The slopes of p_i^* and p'_i, given p'_j and p_j^*, are $-\frac{f_i n}{1+f_i n}$ and $-\frac{1+f_j n}{f_j n}$, respectively. Since

$$\frac{f_i n}{1 + f_i n} = \frac{1 + f_j n}{f_j n} \tag{6.35}$$

gives $f_j n + f_i n = -1$, which is impossible (since f_j, f_i, and n must be positive to make the cost in (6.26) a negative number), by contradiction, the uniqueness of the Nash equilibrium is established. A similar procedure can be used to show that the Nash equilibrium is unique in a three-network case. This completes the proof. ■

6.5.1.3 A Heuristic Search Algorithm to Compute the Nash Equilibrium

Since obtaining the Nash equilibrium may be computationally intensive (depending on the number of connections involved and the type of the revenue and the cost functions), we apply a heuristic search algorithm which is easy to implement and also applicable for a wide range of utility functions. In this case, we minimize an objective function which depends on the strategies adopted by other networks.

Mathematically, the Nash equilibrium is given as follows: $\mathbb{P}^* = \{p_j^*, p_k^*, p_i^*\}$, where $\{p_j^*, p_k^*, p_i^*\} = \arg\min_{p_j, p_k, p_i} \Psi(p_j, p_k, p_i)$, where

$$\begin{aligned}\Psi(p_j, p_k, p_i) &= \left|p_i - BR_i^{(con)}(\{p_j, p_k\})\right| + \left|p_j - BR_j^{(con)}(\{p_i, p_k\})\right| + \\ &\quad \left|p_k - BR_k^{(con)}(\{p_i, p_j\})\right|\end{aligned}$$

in which the objective function $\Psi(.)$ is defined as the difference between the best responses and the strategies adopted by other networks.

However, in a three-network case, given the strategies of two networks, we can obtain the best response for the other network; therefore, the number of decision variables is reduced to two. Given strategies p_j and p_k, the objective function can be computed as in Algorithm 2. A standard search algorithm can be used to obtain the solution. We use *Nelder-Mead* direct search method [94] here.

Algorithm 2 Objective function for the search algorithm

Input: p_j and p_k

1: $p_i' \leftarrow BR_i^{(con)}(\{p_j, p_k\})$

2: $p_j' \leftarrow BR_j^{(con)}(\{p_i', p_k\})$

3: $p_k' \leftarrow BR_k^{(con)}(\{p_i', p_j'\})$

{compute objective function}

4: $\Psi \leftarrow \left|p_j' - p_j\right| + \left|p_k' - p_k\right|$

Output: Ψ

6.5.1.4 Iterative Algorithm to Compute the Nash Equilibrium

In Algorithm 2, we observe that the objective function can be computed from auxiliary variables p_j' and p_k'. Therefore, we introduce an iterative algorithm in which we use these auxiliary variables again to obtain new variables in the next iteration (Algorithm 3). With starting points p_j^{st} and p_k^{st}, the algorithm iterates until the difference between variables in the previous iteration (i.e., p_j^{pr} and p_k^{pr}) and in the current iteration (i.e., p_j^{cu} and p_k^{cu}) becomes less than the threshold th (e.g., $th = 10^{-4}$).

Observation 2 *Given a starting point, the iterative algorithm converges.*

Proof.

Algorithm 3 Iterative algorithm

Input: p_j^{st}, p_k^{st}

1: $p_j^{pr} \leftarrow p_j^{st}$, $p_k^{pr} \leftarrow p_k^{st}$ {initialize variables in the current iteration}
2: **repeat**
3: $\quad p_i^{cu} \leftarrow BR_i^{(con)}\left(\{p_j^{pr}, p_k^{pr}\}\right)$
4: $\quad p_j^{cu} \leftarrow BR_j^{(con)}\left(\{p_i^{cu}, p_k^{pr}\}\right)$
5: $\quad p_k^{cu} \leftarrow BR_k^{(con)}\left(\{p_i^{cu}, p_j^{cu}\}\right)$
6: $\quad \Psi \leftarrow \left|p_j^{cu} - p_j^{pr}\right| + \left|p_k^{cu} - p_k^{pr}\right|$ {compute the difference}
7: $\quad p_j^{pr} \leftarrow p_j^{cu}$, $p_k^{pr} \leftarrow p_k^{cu}$ {update variables}
8: **until** $\Psi \leq th$ {termination criteria}

Output: p_i^{cu}, p_j^{cu}, p_k^{cu}

Since the solutions of (6.32) and (6.33) can be either both non-zero positive (**Case I**) or one negative (or zero) and the other positive (**Case II**)[2], we provide proofs for these two separate cases.

Case I: The iterative algorithm $p_i(t) = \mathcal{F}(p_i(t-1))$, where $p_i(t)$ is the strategy at iteration t, converges if the following conditions are satisfied [95]. First, a solution point (i.e., Nash equilibrium) must exist. Second, the function $\mathcal{F}(p)$ should have the following three properties: positivity (i.e., $\mathcal{F}(p) > 0$), monotonicity (i.e., $p > p' \Rightarrow \mathcal{F}(p) > \mathcal{F}(p')$), and scalability (i.e., $\beta > 1 \Rightarrow \mathcal{F}(\beta p) < \beta \mathcal{F}(p)$).

For the proposed iterative algorithm, we have proved that the Nash equilibrium is unique (observation 1). For monotonicity, we rewrite (6.32) as follows:

$$p_i(t) = \mathcal{F}(p_i(t-1)) = \frac{1}{1+f_i n}\left(D_i - f_i n\left(\frac{D_j - f_j n p_i(t-1)}{1+f_j n}\right)\right). \tag{6.36}$$

It can show that, if $p > p'$, then

$$\begin{aligned} \frac{-D_j + f_j n p}{1+f_j n} &> \frac{-D_j + f_j n p'}{1+f_j n} \\ \Rightarrow \frac{1}{1+f_i n}\left(D_i - f_i n\left(\frac{D_j - f_j n p}{1+f_j n}\right)\right) &> \frac{1}{1+f_i n}\left(D_i - f_i n\left(\frac{D_j - f_j n p'}{1+f_j n}\right)\right) \\ \Rightarrow \mathcal{F}(p) &> \mathcal{F}(p'). \end{aligned} \tag{6.37}$$

[2]Both the solutions of (6.32) and (6.33) cannot be negative or zero.

For scalability, since $\beta > 1$, we have

$$\begin{aligned} D_i - \frac{f_i n D_j}{1 + f_j n} &< \beta \left(D_i - \frac{f_i n D_j}{1 + f_j n} \right) \\ \Rightarrow \frac{1}{1 + f_i n} \left(D_i - f_i n \left(\frac{D_j - f_j n \beta p}{1 + f_j n} \right) \right) &< \beta \frac{1}{1 + f_i n} \left(D_i - f_i n \left(\frac{D_j - f_j n p}{1 + f_j n} \right) \right) \\ \Rightarrow \mathcal{F}(\beta p) &< \beta \mathcal{F}(p). \end{aligned} \tag{6.38}$$

These properties also hold for $p_j(t) = \mathcal{G}(p_j(t-1))$.

Case II:

We show that if the amount of allocated bandwidth to one player becomes negative, the bandwidth allocated to the other player converges to a fixed positive solution. Since the amount of bandwidth cannot be negative in a real scenario, in the algorithm we set it to zero.

Without loss of generality, assuming that, $p_i(t-1) > 0$, $p_i(t) < 0$, and $p_j(t) > 0$, we have

$$p_i(t) = 0, \quad \text{when} \quad \frac{D_i - f_i n p_j(t)}{1 + f_i n} < 0. \tag{6.39}$$

Then

$$p_j(t+1) = \mathcal{G}(p_i(t) = 0) = \frac{D_j - f_j n \times 0}{1 + f_j n}. \tag{6.40}$$

Therefore, $p_j(t+1) > p_j(t)$, and in general, $p_i(t+z) < 0$, $\forall z > 0$. Hence, the iterative algorithm $\mathcal{G}(p_i(t))$ will converge to

$$p_j^* = p_j(t+1) = \mathcal{G}(p_i(t) = 0). \tag{6.41}$$

A similar procedure can be used to show that the algorithm converges in a three-network case. This completes the proof.

■

6.5.2 Bandwidth Distribution

When a connection departs a service area, the bandwidth released from the departing connection is distributed among the ongoing connections. The distribution of bandwidth is based on the current amount of allocated bandwidth to the ongoing connections as follows:

$$q_x = q_x + \frac{(\max(q_y) - q_x)\hat{q}}{\sum_x (\max(q_y) - q_x)} \tag{6.42}$$

where q_x is the bandwidth allocated to the ongoing connection x and $\hat{q}$ is the bandwidth released by the departing connection. Note that, if all of the connections have same bandwidth, $\hat{q}$ is distributed equally among the ongoing connections.

6.5.3 Admission Control Algorithm

When a mobile initiates a new connection, the information on the required bandwidth is sent to the central controller, which computes the offered bandwidth by each network. In this case, either the search algorithm or the iterative algorithm is used to obtain p_i^*, p_j^*, and p_k^* which denote the amount of bandwidth offered by network i, j, and k in a particular service area, respectively, to a new connection. Since the admission control algorithm ensures that the requested bandwidth of an incoming connection is honoured when it is admitted, the following condition is checked:

$$p_i^* + p_j^* + p_k^* \geq R \tag{6.43}$$

where R is the bandwidth requirement of a new connection.

If the incoming connection is a newly initiated connection, the admission control procedure checks whether the total number of ongoing connections in a particular area is less than $C_a - \min(c_a^{(v)}, c_a^{(h)})$. If this is true, an arriving connection is accepted. Similarly, the admission control procedure checks with threshold $c_a^{(v)}$ and $c_a^{(h)}$ for vertical and horizontal handoff connections, respectively.

6.6 Performance Evaluation

6.6.1 Parameter Setting

We consider the service areas shown in Fig. 6.1. In case of IEEE 802.11 WLAN, the channel rate is 11 Mbps and the maximum saturation throughput achieved through EBA in a WLAN is 7 Mbps [70]. For the CDMA cellular wireless access, the transmission bandwidth is assumed to be 5 MHz. We assume that the ratio of bit energy and noise-plus-interference power spectral density at the receiver is 8.17 dB which corresponds to a bit-error-rate of 10^{-4}. The total transmission rate in each CDMA cell is 2 Mbps. For the IEEE 802.16-based wireless access, the transmission rate is 10

Mbps in a single cell. Note that the subscripts wm, ce and wl are used to denote the parameters corresponding to WMAN, cellular network, and WLAN, respectively.

The parameters for the network utility function are set as follows: $w = 1$ and $\alpha = 0.7$. The parameter for the utility function for a connection (σ) is set to 0.7, 0.8, and 0.9, respectively, for a new connection, a vertical handoff connection, and a horizontal handoff connection. With these values of σ, the utilities for a new connection, a vertical handoff connection, and a horizontal handoff connection become zero if the corresponding blocking and dropping probabilities become higher than 0.3, 0.1, and 0.05, respectively.

6.6.2 Network-Level Allocation

Fig. 6.3(a) shows the solution obtained from the network-level allocation, i.e., the amount of bandwidth allocated by each network to each of the service areas. In this case, the average number of connections in area 1, 2, 4, and 5 is 10, 5, 5 and 20, respectively, while that of area 3 is varied. As the number of connections in area 3 increases, the amount of bandwidth corresponding to this area (i.e., m_3 and c_3 offered by WMAN and cellular network) increases accordingly. Since the total bandwidth of the cellular network is limited, bandwidth allocated by the cellular network to area 2 (i.e., c_2) decreases significantly. However, to be fair to the connections in area 2, WMAN tries to allocate larger amount of bandwidth to this area (i.e., m_2 increases). On the other hand, WMAN needs to reduce the amount of bandwidth allocated to other areas (i.e., m_1, m_4 and m_5). In this case, since area 5 is serviced by WLAN for which the available bandwidth is large (e.g., 7 Mbps), m_5 decreases at a rate higher than that for each of m_1 and m_4.

Note that the bandwidth allocation by the cellular network in service area 4 is mostly unaffected even though the number of connections in area 3 increases. However, when the number of connections in area 3 becomes very large (e.g., more than 27 connections), since WMAN cannot contribute bandwidth to area 5 (i.e., m_5), the cellular network needs to alter its allocations for area 4 and area 5 (i.e., c_4 and c_5, respectively).

Fig. 6.3(b) shows the amount of bandwidth allocated to each area. As expected, the total amount of bandwidth allocated to area 3 increases as the number of con-

nections increases. In this case, since bandwidth from WMAN allocated to area 5 is taken away and given to area 3, total amount of bandwidth allocated to area 5 decreases. We observe that both area 2 and area 4 receive equal amount of bandwidth since they serve the same number of connections. Similarly, when the number of connections in both area 3 and area 5 is 20 (in Fig. 6.3(b)), same amount of bandwidth is allocated to each of these service areas. These results show that the noncooperative game provides fair bandwidth allocation at the network level.

6.6.3 Capacity Reservation

Fig. 6.4 shows the Pareto optimality and the equilibrium of the capacity reservation for area 3 in which the average arrival rate for new, vertical handoff and horizontal handoff connections is 2.4, 1.2, and 0.6, respectively, and the average connection holding time is 10 minutes. It is expected that while one player can increase his payoff, payoff of another player must be decreased, since the threshold setting will affect the connection blocking/dropping probabilities of all the players. In this case, at the equilibrium, the capacity reserved for vertical handoff and horizontal handoff connections is 2 and 1, respectively.

6.6.4 Connection-Level Allocation

6.6.4.1 Best Response Functions

Fig. 6.5(a) shows the best responses for WMAN and cellular network under different strategies in service area 2. The amount of bandwidth allocated by WMAN and cellular network to this area is 1400 and 1100 Kbps, respectively (which corresponds to an average number of connections of 16 in area 3 in Fig. 6.3(a)). The Nash equilibrium is located at the point where the best responses of WMAN and cellular network intersect. The equilibrium varies with the number of ongoing connections. Smaller number of ongoing connections result in a larger amount of bandwidth offered to an arriving connection and vice versa.

Fig. 6.5(b) shows the best response functions for WMAN, cellular network, and WLAN in service area 3. The amount of bandwidth allocated by WMAN, cellular network, and WLAN to this area is 2100, 1900, and 2400 Kbps, respectively, and the

number of ongoing connections is assumed to be 5 in this service area. In this case, the best response function of each player is a plane and the Nash equilibrium is located at the intersection of these planes. Since the WLAN and the cellular network have the largest and the smallest amount of available bandwidth, the bandwidth offered to an incoming connection from WLAN and cellular network is the highest and the lowest, respectively.

6.6.4.2 Iterative and Heuristic Search Algorithms

Fig. 6.6 shows the speed of convergence of the iterative and the heuristic search algorithms to obtain the Nash equilibrium in a three-player noncooperative game. The starting point for both the algorithms is set to 100 Kbps for both WMAN and cellular network. Although both the algorithms achieve the same solution, we observe that the iterative algorithm can reach the final solution much faster (i.e., within 15 iterations in Fig. 6.6) and more smoothly than the search algorithm. Therefore, the iterative algorithm is superior to the search algorithm in terms of both stability and efficiency.

6.6.4.3 Bandwidth Adaptation

Fig. 6.7(a) shows variations in the amount of bandwidth offered to a new connection in area 3 under different number of ongoing connections. As expected, when the number of connections in this area is small, an incoming connection will receive large amount of bandwidth. This bandwidth decreases as the number of ongoing connections increases. Also, bandwidth offered to an arriving connection by WLAN is the largest since WLAN has the highest available bandwidth. We observe that the amount of available bandwidth in WLAN has significant impact on the amount of bandwidth offered to an arriving connection (Fig. 6.7(b)).

6.6.5 Performance of Admission Control

We assume that the bandwidth requirement for every connection in the network is 200 Kbps. That is, if the connection-level bandwidth allocation cannot allocate bandwidth larger than 200 Kbps to an incoming connection, that connection is rejected.

The arrival rates of new, vertical handoff, horizontal handoff connections for area 1, 2, 4, and 5 are (1.4, 0.7, 0.35), (0.4, 0.2, 0.1), (0.6, 0.3, 0.15), and (1.2, 0.6, 0.3), respectively. The connection arrival rates in area 3 are denoted by $\gamma(2, 1, 0.5)$, where γ is the traffic intensity[3].

Fig. 6.8(a) shows variations in the average amount of bandwidth allocated to a connection in each service area. The bandwidth allocated to a connection in service area 3 becomes the highest in the network when there are small number of ongoing connections (e.g., when traffic intensity is 0.4-0.5). In this case, the WLAN can offer a large amount of bandwidth to a connection in service area 3. When the traffic intensity increases, connections in area 1 and area 3 receive slightly lower amount of bandwidth than other areas since traffic load in both these areas is higher than the load in other areas. When the traffic load in area 3 increases, the average amount of bandwidth allocated to a connection in most of the service areas decreases due to the load balancing achieved through the game-theoretic bandwidth allocation. However, the connection-level allocation in area 5 is not affected by traffic load in area 3 since in area 5 the WLAN contributes most of the bandwidth to the connections.

Figs. 6.8(b), 6.9(a), and 6.9(b) show variations in new connection blocking probability, and connection dropping probabilities for vertical and horizontal handoff connections. As expected, traffic load in area 3 impacts the connection-level performances in area 1, 2, and 4 (but not area 5). Due to the capacity reservation, in each service area, handoff connection blocking probability is smaller than the new connection blocking probability. In this case, connections in area 2 and area 4 experience high blocking and dropping probabilities since the cellular networks in these areas have to share the bandwidth with the connections in area 3 and area 5. Also, the WMAN cannot contribute a large amount of bandwidth to area 2 and area 4 since it needs to serve the connections in area 1 in which only WMAN service is available.

6.6.6 Summary of the Observations

The performance evaluation results can be summarized as follows:

- The network-level bandwidth allocation tries to allocate bandwidth to the ser-

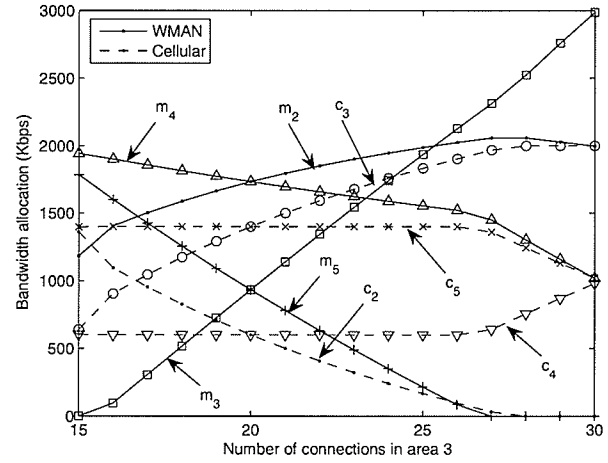
[3]Traffic load corresponding to new, vertical handoff, horizontal handoff connections is obtained by multiplying γ with 2, 1, 0.5, respectively.

vice areas from the different access networks in a fair manner.

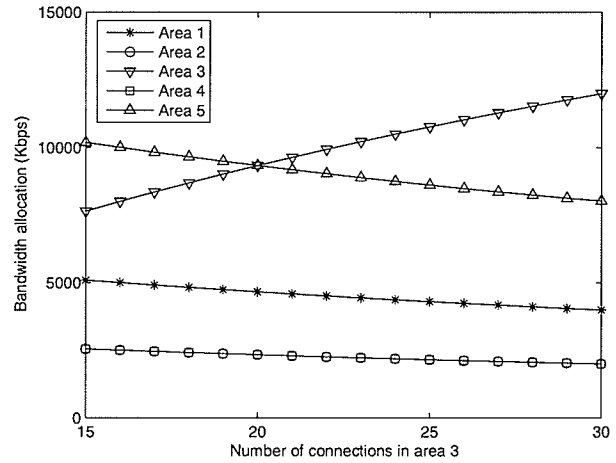
- Combined with network-level allocation, capacity reservation can be used to obtain the reservation thresholds for vertical and horizontal handoff connections so that connection blocking and dropping probabilities are maintained below the target level.
- Connection-level bandwidth allocation (i.e., bandwidth allocation in a short-term basis) is required to adapt with the traffic fluctuation in a service area as well as the variation in the available bandwidth from the different access networks.
- The iterative algorithm for connection-level bandwidth allocation can converge to the solution quickly, and therefore, would be more suitable for online execution.
- In a heterogeneous wireless access environment, admission control is required not only to maintain the performances of the ongoing connections at the desired level but also to prioritize the handoff connections.

6.7 Chapter Summary

In this chapter, we have presented a game-theoretic framework for radio resource management in heterogeneous wireless access networks consisting of WMAN, cellular networks, and WLANs. This framework provides a fair resource allocation in the different service areas while satisfying both the service providers' and the users' requirements. Also, it can adapt to both long-term and short-term variations of network resources and traffic load conditions. The performances of the different components of this framework, namely, network-level bandwidth allocation, capacity reservation, connection-level bandwidth allocation, and admission control have been analyzed. Part of this chapter has been published in [96].



(a)



(b)

Figure 6.3. *(a) Bandwidth allocated by different networks to each service area and (b) total amount of bandwidth allocated to each service area.*

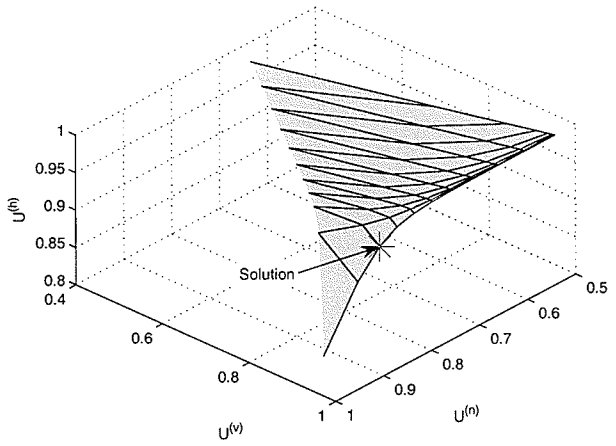
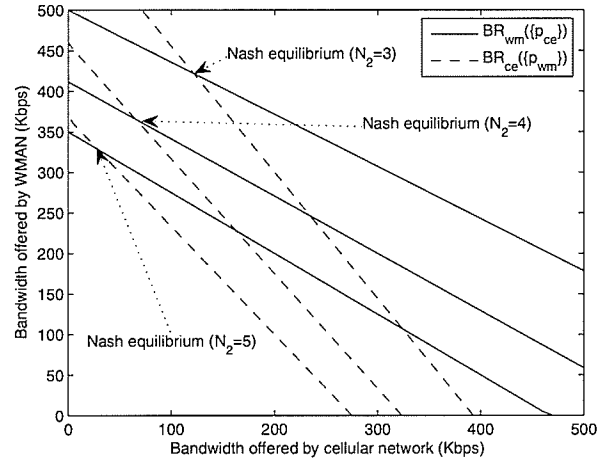
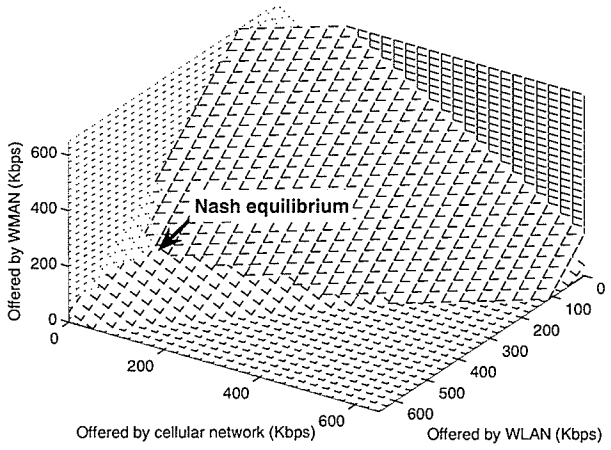


Figure 6.4. *Pareto optimality and equilibrium of the bargaining game for capacity reservation.*



(a)



(b)

Figure 6.5. *Best response functions of (a) WMAN and cellular network in service area 2, and (b) WMAN, cellular network, and WLAN in service area 3.*

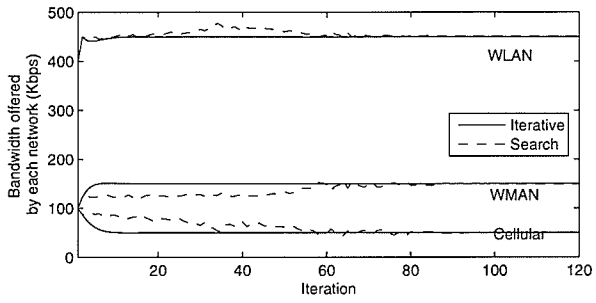


Figure 6.6. *Comparison of speed of convergence between the iterative and the search algorithms.*

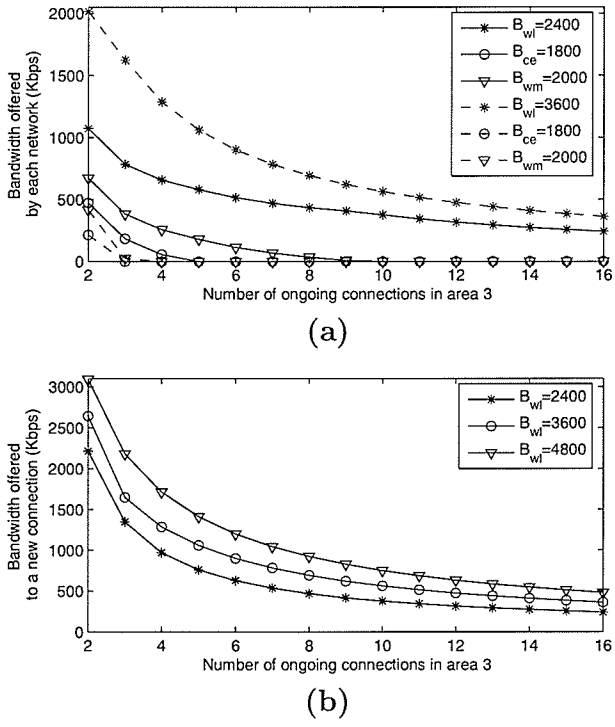


Figure 6.7. *(a) The amount of bandwidth offered by each network and (b) the total amount of bandwidth received by a new connection.*

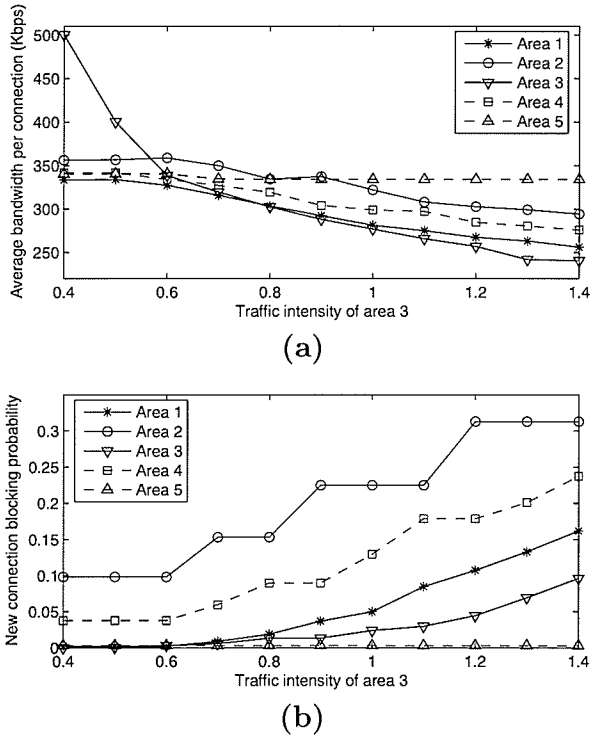


Figure 6.8. *(a) Average amount of allocated bandwidth per connection and (b) new connection blocking probability.*

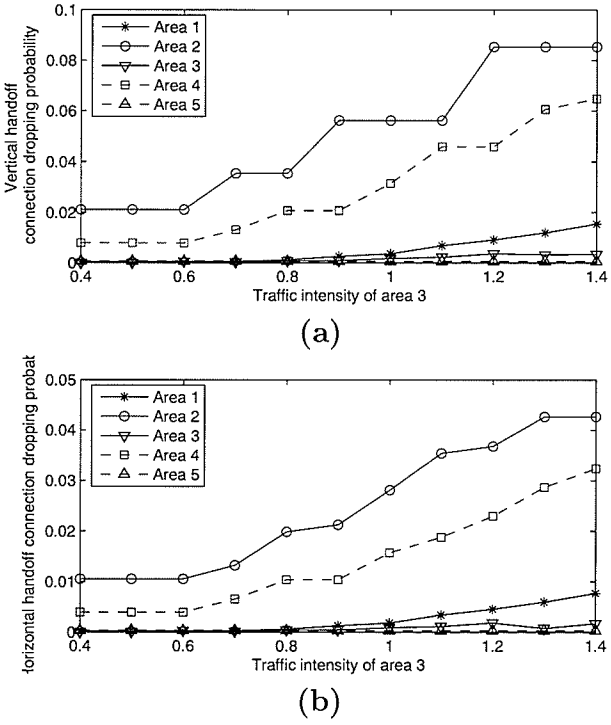


Figure 6.9. *(a) Horizontal and (b) vertical handoff connection dropping probability.*

Chapter 7

Summary and Future Works

7.1 Summary of Contributions

The research contributions presented in this thesis can be summarized as follows:

- *Chapter 2:* An adaptive queue-aware uplink bandwidth allocation and rate control mechanisms in a WiMAX SS for *polling service* have been proposed. While the bandwidth allocation mechanism adaptively allocates bandwidth for polling service in presence of higher priority *unsolicited grant service*, the rate control mechanism dynamically limits the transmission rate for the connections under polling service. Both of these schemes exploit the queue status information to guarantee the desired QoS performance for polling service. A queueing analytical framework has been developed to analyze the proposed resource management model from which various performance measures for polling service in both steady and transient states can be obtained. The performance of *best-effort service* has been analyzed in presence of unsolicited grant service and polling service. Analytical results have been validated by simulations, and typical numerical results have been presented.
- *Chapter 3:* A queueing theoretic and optimization-based model has been presented for radio resource management in WiMAX-based multi-service broadband wireless access (BWA) networks considering both packet-level and connection-level QoS constraints. Two bandwidth allocation approaches have been proposed. While for the optimal approach an assignment problem is formulated and solved, a water-filling mechanism is used for the iterative approach. The latter incurs significantly less computational complexity compared to the former

while providing similar system performances. To limit the amount of bandwidth allocated to each service type, the total available bandwidth is shared among the different types of services using a complete partitioning approach. To analyze the connection-level performance measures such as connection blocking probability and the average number of ongoing connections, a queueing model has been developed. Then, an optimization formulation has been used to obtain the optimal threshold settings for complete partitioning of the available bandwidth resources so that the connection-level QoS (e.g., connection blocking probability) for the different services can be maintained at the target level while maximizing the average system revenue. To analyze the packet-level performance measures such as the packet delay statistics and transmission rate (or throughput), a queueing analytical model has been developed, considering adaptive modulation and coding (AMC) at the physical/radio link layer. In summary, the queueing-theoretic and optimization-based model for joint BA and CAC provides a unified radio resource management solution for the WiMAX-based broadband wireless access networks.

- *Chapter 4:* An architecture for integrating WiFi WLANs with WiMAX-based multihop wireless mesh infrastructure to relay WLAN traffic to the Internet has been presented. The major research issues in this integrated architecture have been outlined and the related works have been reviewed. A game-theoretic model has been developed for radio resource management in this integrated network architecture. In particular, a multi-player bargaining game formulation has been used for fair bandwidth allocation and optimal admission control of different types of connections (e.g., WLAN connections, relay connections, connections from standalone subscriber stations) in a WiMAX base station/mesh router. Both connection-level and in-connection-level performances for this bandwidth management and admission control framework have been presented.
- *Chapter 5:* A bandwidth allocation and admission control algorithms based on *bankruptcy game*, which is a special type of an N-person cooperative game, has been presented. A coalition among the different wireless access networks is formed to offer bandwidth to a new connection. The stability of the allocation has been analyzed by using the concept of the *core* and the amount

of allocated bandwidth to a connection in each network is obtained by using the *Shapley value*. Subsequently, an admission control algorithm has been proposed. Numerical results have been presented to demonstrate the behaviors of the proposed algorithms.

- *Chapter 6:* A game-theoretic framework for radio resource management (i.e., bandwidth allocation and admission control) in a heterogeneous wireless access environment has been proposed. In this framework, first, a noncooperative game has been used to obtain the bandwidth allocations to a service area from the different access networks available in that service area (on a long-term basis). The Nash equilibrium for this game gives the optimal allocation, which maximizes the utilities of all the connections in the network (i.e., in all the service areas). Second, based on the obtained bandwidth allocation, to prioritize vertical and horizontal handoff connections over new connections, a bargaining game has been formulated to obtain the capacity reservation thresholds so that the connection-level QoS requirements can be satisfied for the different types of connections (on a long-term basis). Third, a noncooperative game to obtain the amount of bandwidth allocated to an arriving connection (in a service area) by the different access networks (on a short-term basis) has been formulated. Based on the allocated bandwidth and the capacity reservation thresholds, an admission control is used to limit the number of ongoing connections so that the QoS performances are maintained at the target level for the different types of connections.

7.2 Future Works

Some of the issues, which will be addressed in our future research, are as follows:

- *Alternative solutions of game formulations for radio resource management in heterogeneous wireless access networks:* In the bargaining game formulation for the bandwidth allocation of an integrated WiFi/WiMAX multihop relay network, alternative solutions (e.g., Kalai-Smorodinsky solution (KSS) and Egalitarian solution (ES) [97]) can be also considered which provide different types of fairness performance. These solutions can be compared with the Nash bar-

gaining solution (NBS). Similarly, in the noncooperative game formulation for the bandwidth allocation in a heterogeneous wireless network, correlated equilibrium [98] can be considered as an alternative solution. This correlated equilibrium can be compared with the Nash equilibrium.

- *Implementation of the radio resource management frameworks in a prototype system:* To evaluate the performances of the proposed RRM frameworks in a practical system, these can be implemented in a prototype system using off-the-shelf radio hardware.
- *Pricing in broadband wireless access networks:* Pricing is a very important issue for wireless service providers. Pricing in wireless networks has to be carefully designed since it impacts not only the revenue of the service providers, but also the satisfaction of the mobile users. There are two major factors which impact pricing in wireless network - user demand and competition among multiple service providers. If the price is high, even though a service provider can generate more revenue, user's satisfaction degrades and demand decreases. As a result, the revenue of the service provider may not be maximized. Also, if there are multiple service providers, competition among them will impact the price. A service provider may reduce the price to attract more users. Due to the heterogeneity of the networks, in which multiple wireless networks are operated by different service providers, pricing is crucial to maximize the revenue of the service providers. We will develop competitive pricing schemes for heterogeneous broadband wireless access networks.
- *Application of heterogeneous wireless broadband access networks to intelligent transportation systems:* Intelligent transportation system (ITS) application will improve the performance and safety of transportation by vehicles. Heterogeneous broadband wireless access technology can facilitate information exchange in vehicle-to-vehicle and vehicle-to-roadside communication (e.g., traffic information and warning system) environments. Application of heterogeneous broadband wireless access in ITS and the related protocol engineering issues for vehicle-to-vehicle and vehicle-to-roadside communications will be investigated.
- *Performance of higher layer protocols in heterogeneous wireless networks:* The performances of higher layer (e.g., routing and TCP) protocols in a heteroge-

neous wireless network need to be evaluated. With a heterogeneous wireless interface at a mobile device, data can be transmitted over multiple streams through different wireless interfaces. The routing of these streams must be optimized to achieve the best QoS performance. Also, when a user performs a vertical handoff between different wireless networks, the effect from the low level protocol (e.g., physical and MAC) to the transport layer (e.g., the handoff delay could be interpreted as congestion by TCP) must be investigated.

- *Cognitive radio in heterogeneous wireless networks:* Cognitive radio emerges as the new paradigm in wireless communications. A cognitive radio transceiver has the capability of observing, learning, optimizing, and adapting to the wireless environment. The cognitive radio concept can be applied to heterogeneous wireless networks to improve the utility of the users and service providers. For example, with multiple choices of available wireless networks, a user can observe and learn the performance of each network. Then, based on the knowledge of all network, a user can choose the best network to connect to.

Bibliography

[1] D. Johnston and J. Walker, "Overview of IEEE 802.16 security," *IEEE Security and Privacy Magazine*, vol. 2, no. 3, pp. 40-48, May-June 2004.

[2] 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, January 2007.

[3] http://ieee802.org/16/tgs.html

[4] A. Ghosh, D. R. Wolter, J. G. Andrews, and R. Chen, "Broadband wireless access with WiMax/802.16: Current performance benchmarks and future potential," *IEEE Communications Magazine*, vol. 43, no. 2, pp. 129-136, February 2005.

[5] Recommendation ITU-R M.1225, "*Guidelines for evaluation of Radio transmission technologies for IMT-2000*, 1997."

[6] M. Klerer, "Introduction to IEEE 802.20: Technical and procedural orientation," *IEEE 802.20 Working Group Permanent Documents.*

[7] T. Bu, M. C. Chan, and R. Ramjee, "Designing wireless radio access networks for third generation cellular networks," in *Proceedings of Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 1, pp. 68-78, March 2005.

[8] D. Niyato and E. Hossain, "Call admission control for QoS provisioning in 4G wireless networks: Issues and approaches," *IEEE Network*, vol. 19, no. 5, pp. 5-11, September-October 2005.

[9] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 6, pp. 668-679, June 2006.

[10] H. Y. Wei, S. Ganguly, R. Izmailov, and Z. J. Haas, "Interference-aware IEEE 802.16 WiMAX mesh networks," in *Proceedings of IEEE Vehicular Technology Conference (VTC) Spring*, vol. 5, pp. 3102-3106, May-June 2005.

[11] C. K. Chang, "A mobile-IP based mobility system for wireless metropolitan area networks," in *Proceedings of IEEE International Conference Workshops on Parallel Processing (ICPP)*, pp. 429-435, June 2005.

[12] A. F. Graves, B. Wallace, S. Periyalwar, and C. Riccardi, "Clinical grade - a foundation for healthcare communications networks," in *Proceedings of Interna-*

tional Workshop on Design of Reliable Communication Networks (DRCN), October 2005.

[13] IEEE 802.16 Standard - Local and Metropolitan Area Networks - Part 16, *IEEE Std 802.16a-2003.*

[14] T. V. J. Ganesh Babu, T. Le-Ngoc, and J. F. Hayes, "Performance of a priority-based dynamic capacity allocation scheme for wireless ATM systems," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 355-369, February 2001.

[15] L. Muscariello, M. Meillia, M. Meo, M. A. Marsan, and R. L. Cigno, "An MMPP-based hierarchical model of Internet traffic," in Proc. *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 4, pp. 2143-2147, June 2004.

[16] K. Wongthavarawat and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," *Journal of Communication Systems*, vol. 16, pp. 81-96, February 2003.

[17] K. K. Leung and A. Srivastava, "Dynamic allocation of downlink and uplink resource for broadband services in fixed wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 5, pp. 990-1006, May 1999.

[18] G. Liu, W. Lang, W. Wu, Y. Ruan, X. Shen, and G. Zhu, "QoS-guaranteed call admission scheme for broadband multi-services mobile wireless networks," in *Proceedings of IEEE International Symposium on Computers and Communications (ISCC)*, vol. 1, pp. 454-459, June-July 2004.

[19] G. Li and H. Liu, "Dynamic resource allocation with finite buffer constraint in broadband OFDMA networks," in *Proceedings of IEEE Wireless Communications and Networking (WCNC)*, vol. 2, pp. 1037-1042, March 2004.

[20] G. Liu, G. Zhu, and W. Wu, "An adaptive call admission policy for broadband wireless multimedia networks using stochastic control," in *Proceedings of IEEE Wireless Communications and Networking (WCNC)*, vol. 3, pp. 1324-1329, March 2004.

[21] M. Soleimanipor, W. Zhuang, and G. H. Freeman, "Optimal resource management in wireless multimedia wideband CDMA systems," *IEEE Transactions on Mobile Computing*, vol. 1, no. 2, pp. 143-160, April-June 2002.

[22] S. Baey, M. Dumas, and M.-C. Dumas, "QoS tuning and resource sharing for UMTS WCDMA ultiservice mobile," *IEEE Transactions on Mobile Computing*, vol. 1, no. 3, pp. 221-235, July-September 2002.

[23] J. Ye, J. Hou, and S. Papavassiliou, "A comprehensive resource management framework for next generation wireless networks," *IEEE Transactions on Mobile Computing*, vol. 1, no. 4, pp. 249-264, October-December 2002.

[24] C.-T. Chou and K. G. Shin, "Analysis of adaptive bandwidth allocation in wireless networks with multilevel degradable quality of service," *IEEE Transactions on Mobile Computing*, vol. 3, no. 1, pp. 5-17, January-March 2004.

[25] Z. Liu, P. Nain, and D. Towlsey, "On optimal polling policies," *Queueing Systems Theory and Applications*, vol. 11, no. 1-2, pp. 59-83, March 1992.

[26] L. Kalampoukas, A. Varma, and K. K. Ramakrishnan, "Two-way TCP traffic over rate controlled channels: effects and analysis," *IEEE/ACM Transactions on Networking*, vol. 6, no. 6, pp. 729-743, December 1998.

[27] H. Zhang, J. Cong, and O. W. Yang, "Rate control over RED with data loss and varying delays," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 6, pp. 3035-3040, December 2003.

[28] T. Inzerilli, "Design and performance modeling for traffic control in wireless links," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 4, pp. 230-2311, June 2004.

[29] D. W. Dormuth and A. S. Alfa, "Two finite-difference methods for solving MAP(t)/PH(t)/1/K queueing models," *Queueing Systems* (Kluwer), vol. 27, pp. 55-78, 1997.

[30] M. Zorzi, "Packet dropping statistics of a data-link protocol for wireless local communications," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 1, pp. 71-79, January 2003.

[31] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1142-1153, May 2005.

[32] M. Rossi and M. Zorzi, "Analysis and heuristics for the characterization of selective repeat ARQ delay statistics over wireless channels," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 5, pp. 1365-1377, September 2003.

[33] P. Salvador, R. Valadas, and A. Pacheco. "Multiscale fitting procedure using Markov modulated Poisson processes," *Telecommunication Systems*, vol. 23, pp. 123-148, 2003.

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

[35] I. Koffman and V. Roman, "Broadband wireless access solutions based on OFDM access in IEEE 802.16," *IEEE Communications Magazine*, vol. 40, no. 4, pp. 96-103, April 2002.

[36] E. Altman, "Capacity of multi-service cellular networks with transmission-rate control: A Queueing Analysis," in *Proceeding of ACM International Conference on Mobile Computing and Networking (MOBICOM)*, pp. 205-214, September 2002.

[37] J. Chen, W. Jiao, and H. Wang, "A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 5, pp. 3422-3426, May 2005.

[38] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assurance," *IEEE Transactions on Vehicular Technology.*, vol. 53, no. 5, pp. 1547-1557, September 2004.

[39] L. Xu, X. Shen, and J. W. Mark, "Fair resource allocation with guaranteed statistical QoS for multimedia traffic in wideband CDMA cellular Network," *IEEE Transactions on Mobile Computing*, vol. 4, no. 2, pp. 166-177, March-April 2005.

[40] K. B. Johnsson and D. C. Cox, "An adaptive cross-layer scheduler for improved QoS support of multiclass data services on wireless systems," *IEEE Journal on Selected Areas in Communications* , vol. 23, no. 2, pp. 334-343, February 2005.

[41] S. V. Krishnamurthy, A. S. Acampora, and M. Zorzi, "On the radio capacity of TDMA and CDMA for broadband wireless packet communications," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 1, pp. 60-70, January 2003.

[42] M. Xiao, N. B. Shroff, and E. K. P. Chong, "Utility-based power control in cellular wireless systems," in *Proceedings of Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 1, pp. 412-421, 2001.

[43] L. A. Wolsey, *Integer Programming*, John Wiley & Sons, 1998.

[44] C. Bastarrica, A. A. Shvartsman, and S. Demurjian, "A binary integer programming model for optimal object distribution," in *Proceedings of International Conference on Principles of Distributed Computing (OPODIS)*, pp. 91-105, 1998.

[45] W. Turin and M. Zorzi, "Performance analysis of delay-constrained communications over slow Rayleigh fading channels," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 801-807, October 2002.

[46] 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, November 2006.

[47] A. Iera, A Molinaro, S. Polito, and G. Ruggeri, "End-to-end QoS provisioning in 4G with mobile hotspots," *IEEE Network*, vol. 19, no. 5, pp. 26-34, September-October 2005.

[48] R. Bruno, M. Conti, and E. Gregori, "Mesh networks: Commodity multihop ad hoc networks," *IEEE Communications Magazine*, vol. 43, no. 3, pp. 123-131, March 2005.

[49] D. Niyato and E. Hossain, " A radio resource management framework for the IEEE 802.16-based OFDM/TDD wireless mesh networks," in *Proceedings of IEEE*

International Conference on Communications (ICC), Istanbul, Turkey, 11-15 June, 2006.

[50] E. Yanmaz and O. K. Tonguz, "Dynamic load balancing and sharing performance of integrated wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 862-872, June 2004.

[51] X. Jing, S. C. Mau, D. Raychaudhuri, and R. Matyas, "Reactive cognitive radio algorithms for co-existence between IEEE 802.11b and 802.16a networks," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 5, pp. 2465-2469, November-December 2005.

[52] T. Issariyakul, E. Hossain, and A. S. Alfa, "Analysis of latency for reliable end-to-end batch transmission in multi-rate multi-hop wireless networks," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 5, pp. 3494-3498, May 2005.

[53] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity – Part I: System description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927-1938, November 2003.

[54] R. Draves, J. Padhye, and B. Zill, "Comparison of routing metrics for static multi-hop wireless networks," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 133-144, August 2004.

[55] T. Rouse, S. McLaughlin, and I. Band, "Congestion-based routing strategies in multihop TDD-CDMA networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 3, pp. 668-681, March 2005.

[56] K. Gakhar, A. Gravey, and A. Leroy, "IROISE: A new QoS architecture for IEEE 802.16 and IEEE 802.11e interworking," in *Proceedings of International Conference on Broadband Networks (BROADNETS)*, pp. 607-612, October 2005.

[57] Z. Fu, H. Luo, P. Zerfos, S. Lu, L. Zhang and M. Gerla, "The impact of multihop wireless channel on TCP performance," *IEEE Transactions on Mobile Computing*, vol. 4, no. 2, pp. 209-221, March-April 2005.

[58] S. K. Das, H. Lin, and M. Chatterjee, "An econometric model for resource management in competitive wireless data networks," *IEEE Network*, vol. 18, no. 6, pp. 20-26, November-December 2004.

[59] X. R. Cao, H. X. Shen, R. Milito, and P. Wirth, "Internet pricing with a game theoretic approach: Concepts and examples," *IEEE/ACM Transactions on Networking*, vol. 10, pp. 208-216, April 2002.

[60] H. Shen and T. Basar, "Differentiated Internet pricing using a hierarchical network game model," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 3, pp. 2322-2327, June-July 2004.

[61] S. Chae and P. Heidhues, "A group bargaining solution," *Mathematical and Social Sciences* (Elsevier), vol. 48, no. 1, pp. 37-53, 2005.

[62] M. J. Osborne, *An Introduction to Game Theory*, Oxford University Press, 2003.

[63] E. K. P. Chong and S. H. Zak, *An Introduction to Optimization*, Wiley-Interscience, July 2001.

[64] D. Niyato and E. Hossain, "Integration of IEEE 802.11 WLANs with IEEE 802.16-based multihop infrastructure mesh/relay networks: A game-theoretic approach to radio resource management," *IEEE Network*, vol. 21, no. 3, pp. 6-14, May-June 2007.

[65] D. Cavalcanti, D. Agrawal, C. Cordeiro, B. Bin, and A. Kumar, "Issues in integrating cellular networks WLANs, and MANETs: A futuristic heterogeneous wireless network," *IEEE Wireless Communications*, vol. 12, no. 3, pp. 30-41, June 2005.

[66] O. B. Akan and I. F. Akyildiz, "ATL: An adaptive transport layer suite for next-generation wireless Internet," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 802-817, June 2004.

[67] Q. Song and A. Jamalipour, "A network selection mechanism for next generation networks," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 2, pp. 1418-1422, May 2005.

[68] J. Cai and U. Pooch, "Allocate fair payoff for cooperation in wireless ad hoc networks using Shapley Value," in *Proceeding of IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 219-226, April 2004.

[69] Z. Fang and B. Bensaou, "Design and implementation of a MAC scheme for wireless ad-hoc networks based on a cooperative game framework," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 7, pp. 4034-4038, June 2004.

[70] J. Choi, J. Yoo, C. Kim, and S. Choi "EBA: An enhancement of the IEEE 802.11 DCF via distributed reservation," *IEEE Transactions on Mobile Computing*, vol. 4, no. 4, pp. 378-390, July-August 2005.

[71] B. O'Neill, "A problem of rights arbitration from the Talmud," *Mathematical Social Sciences 2*, pp. 345-371, 1982.

[72] M. Pulido, J. S. Soriano, and N. Llorca, "Game theory techniques for university management: An extended bankruptcy model," *Operation Research*, vol. 109, pp. 129-142, 2002.

[73] T. S. Ferguson, *Game Theory Text*, Mathematics Department, UCLA.

[74] L. S. Shapley, "A value for N-Person game," *Annals of Mathematics Studies, Princeton University Press*, vol. 2, pp. 307-317, 1953.

[75] D. Niyato and E. Hossain, "A cooperative game framework for bandwidth allocation in 4G heterogeneous wireless networks," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 9, pp. 4357-4362, June 2006.

[76] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 2, pp. 371-382, March 2002.

[77] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 1-12, February 1997.

[78] C. J. Ho, J. A. Copeland, C. T. Lea, and G. L. Stuber, "On call admission control in DS/CDMA cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 6, pp. 1328-1343, November 2001.

[79] H. Lin, M. Chatterjee, S. K. Das, and K. Basu, "ARC: An integrated admission and rate control framework for competitive wireless CDMA data networks using noncooperative games," *IEEE Transactions on Mobile Computing*, vol. 4, no. 3, pp. 243-258, May-June 2005.

[80] J. Virapanicharoen, and W. Benjapolakul, "Fair-efficient threshold parameters selection in call admission control for CDMA mobile multimedia communications using game theoretic framework," in *Proceedings of IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 439-444, January 2005.

[81] S. Koskie and Z. Gajic, "A Nash game algorithm for SIR-based power control in 3G wireless CDMA networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 5, pp. 1017-1026, October 2005.

[82] T. Alpcan, T. Basar, and S. Dey, "A power control game based on outage probabilities for multicell wireless data networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 890-899, April 2006.

[83] F. Meshkati, M. Chiang, H. V. Poor, and S. C. Schwartz, "A game-theoretic approach to energy-efficient power control in multicarrier CDMA systems," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 6, pp. 1115-1129, June 2006.

[84] A. Tang, J. Wang, and S. H. Low, "Counter-intuitive throughput behaviors in networks under end-to-end control," *IEEE/ACM Transactions on Networking*, vol. 14, no. 2, pp. 355-368, April 2006.

[85] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multinetwork environments," *IEEE Wireless Communications*, vol. 11, no. 3, pp. 8-15, June 2004.

[86] P. Vidales, J. Baliosian, J. Serrat, G. Mapp, F. Stajano, and A. Hopper, "Au-

tonomic system for mobility support in 4G networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 12, pp. 2288-2304, December 2005.

[87] G. T. Karetsos, S. A. Kyriazakos, E. Groustiotis, F. D. Giandomenico, and I. Mura, "A hierarchical radio resource management framework for integrating WLANs in cellular networking environments," *IEEE Wireless Communications*, vol. 12, no. 6, pp. 11-17, December 2005.

[88] N. Shenoy and R. Montalvo, "A framework for seamless roaming across cellular and wireless local area networks," *IEEE Wireless Communications*, vol. 12, no. 3, pp. 50-57, June 2005.

[89] T.-C. Chau, K. Y. M. Wong, and B. Li, "Optimal call admission control with QoS guarantee in a voice/data integrated cellular network," *IEEE Transactions on Wireless Communications*, vol. 5, no. 5, pp. 1133-1141, May 2006.

[90] L. R. Christensen, D. W. Jorgenson, and L. J. Lau, "Transcendental Logarithmic Utility Functions," *The American Economic Review*, vol. 65, no. 3, pp. 367-383, June 1975.

[91] J. Nash, "Non-cooperative games," *The Annals of Mathematics*, vol. 54, 1951

[92] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 35, no. 3, pp. 77-92, August 1986.

[93] J. Hou, J. Yang, and S. Papavassiliou, "Integration of pricing with call admission control to meet QoS requirements in cellular networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 9, pp. 898-910, September 2002.

[94] J. A. Nelder and R. Mead, "A simplex method for function minimisation," *The Computer Journal*, vol. 7, pp. 308-313, 1965.

[95] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341-1347, September 1995.

[96] D. Niyato and E. Hossain, "A noncooperative game-theoretic framework for radio resource management in 4G heterogeneous wireless access networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 3, pp. 332-345, March 2008.

[97] D. Fudenberg and J. Tirole, *Game theory*, MIT Press, 1991.

[98] R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, vol. 1, no. 1, pp. 67-96, March 1974.