Vertical handoff and resource management in heterogeneous wireless networks

Ma, Dong

2012

Ma, D. (2012). Vertical handoff and resource management in heterogeneous wireless networks. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/51044>

<https://doi.org/10.32657/10356/51044>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**VERTICAL HANDOFF AND RESOURCE
MANAGEMENT IN HETEROGENEOUS
WIRELESS NETWORKS**

**MA DONG
SCHOOL OF ELECTRICAL AND ELECTRONIC
ENGINEERING
2012**

**VERTICAL HANDOFF AND RESOURCE
MANAGEMENT IN HETEROGENEOUS
WIRELESS NETWORKS**

MA DONG

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

2012

Acknowledgements

First of all, I wish to express my sincere gratitude and appreciation to my supervisor, Associate Professor Ma Maode, for his guidance, help and encouragement throughout my PhD studies. He has patiently led me to the academic world step by step. The kind understanding and support from Professor Ma are invaluable for me to get through those unexpected troubles in my research and life.

I would like to thank my former flatmates, Dr. Ma Shouzhi, Dr. Liang Houkun, Ms. Qu Ting and Dr. Yang Mingfei, for the past unforgettable several years. Dr. Sun Xiaofeng and Dr. Chen Siwen have shared happiness with me in every dinner time. I am very grateful to Dr. Mattia and Dr. Mehdi for the discussions about our research and their hometowns. Appreciation also goes to Dr. Liu Zheng and Lu Qi for their stimulating discussions and constructive advices on simulation. I am also very grateful to Dr. Gao Peixia who spent her precious time to understand my research topics on queueing theory and helped me find the solution.

Special thanks are given to Nanyang Technological University, where I spent such a fantastic period of my life in such a beautiful campus. My deep gratitude and thanks go to my families and husband for their everlasting love and support.

Table of Contents

Acknowledgements.....	iii
Table of Contents	iv
List of Figures	vii
List of Tables.....	ix
Abstract.....	x
Chapter 1 Introduction.....	1
1.1 The Interworking of Heterogeneous Networks	1
1.2 Key Issues for the Design of Interworking System	2
1.3 Motivation and Research Scope	4
1.4 Organization of the Thesis	6
Chapter 2 Related Works and Background.....	8
2.1 Current Wireless Access Technologies.....	8
2.1.1 Cellular networks	8
2.1.2 WLAN.....	9
2.1.3 WiMAX.....	10
2.2 Interworking Architecture of Heterogeneous Wireless Systems	11
2.2.1 Loosely coupled	12
2.2.2 Tightly coupled	12
2.3 Vertical Handoff	14
2.3.1 Handoff initiation and objective.....	15
2.3.2 Handoff decision and network selection.....	16
2.3.3 Handoff execution.....	18
2.3.4 Existing works for the WiMAX/WLAN interworking system	19
2.4 Resource Management.....	20

2.4.1	Joint radio resource management.....	21
2.4.2	Admission control	22
2.4.3	Load balancing.....	24
2.5	Multicasting and broadcasting.....	25
2.6	Summary	28
Chapter 3	Interworking Architecture of WiMAX/WLAN.....	29
3.1	Background	29
3.1.1	The WLAN technologies.....	29
3.1.2	The WiMAX technologies.....	30
3.2	Interworking Architecture	31
3.3	Mechanisms for the Tightly Coupled Interworking.....	34
3.3.1	Data path function.....	34
3.3.2	Security issues.....	35
3.4	Summary	37
Chapter 4	Vertical Handoff Scheme for WiMAX/WLAN.....	38
4.1	Proposed Vertical Handoff Solution.....	38
4.1.1	Vertical handoff manager (VHOM).....	39
4.1.2	Vertical handoff procedure	40
4.1.3	Performance analysis and implementation details	43
4.2	Proposed Network Condition Estimation Algorithms	45
4.2.1	Estimation in WiMAX	46
4.2.2	Estimation in WLAN.....	52
4.3	Performance Evaluation	53
4.3.1	Evaluation of network condition estimation algorithms.....	54
4.3.2	Evaluation of the proposed vertical handoff scheme.....	57
4.4	Summary	61
Chapter 5	Proactive Load Balancing for Heterogeneous Overlay Networks.....	62
5.1	System Model.....	63
5.2	Call Admission Control Scheme.....	65
5.3	Load Balancing Algorithm.....	68

5.3.1	The request call is located in the double-coverage area	69
5.3.2	The request call is located in the single-coverage area.....	70
5.4	Performance Evaluation	71
5.4.1	System configurations	72
5.4.2	Performance comparison	73
5.5	Summary	78
Chapter 6 Network Selection & Resource Allocation for Multicasting/Broadcasting.....		79
6.1	Design of the NSRA Scheme	81
6.1.1	System model and notation.....	81
6.1.2	Problem specification.....	83
6.2	Lagrangian Algorithm.....	85
6.2.1	Lagrangian relaxation.....	85
6.2.2	Lagrangian heuristic algorithm	86
6.3	Implementation Details	89
6.4	Performance Evaluation	90
6.4.1	System configurations	91
6.4.2	Evaluation of the NSRA scheme.....	92
6.4.3	Evaluation of the LR algorithm.....	98
6.5	Summary	99
Chapter 7 Summary and Future Works		101
7.1	Summary of Contributions	101
7.2	Future Works.....	103
List of Publications.....		106
Bibliography		107
List of Abbreviation		115

List of Figures

Figure 2.1 Vertical handoff vs. horizontal handoff.....	14
Figure 3.1 The interworking architecture for WiMAX/WLAN overlay system.....	32
Figure 3.2 An example of the protocol stack of data plane.	35
Figure 3.3 Combined authentication architecture.	36
Figure 4.1 Components of VHOM.....	40
Figure 4.2 The QoS-triggered vertical handoff decision algorithm.	42
Figure 4.3 A sample of OFDMA-TDD frame structure.....	46
Figure 4.4 An example of uplink packet transmission.	48
Figure 4.5 Available bandwidth in WiMAX: (a) downlink; (b) uplink.	54
Figure 4.6 Mean packet delay in WiMAX: (a) downlink; (b) uplink.	55
Figure 4.7 WLAN estimation: (a) available bandwidth; (b) mean packet delay.	57
Figure 4.8 Simulation topology of WiMAX/WLAN interworking system.	58
Figure 4.9 Throughput comparison in the movement scenario.	59
Figure 5.1 The system model.....	64
Figure 5.2 Call admission control algorithm.	67
Figure 5.3 The simulation topology.	72
Figure 5.4 New call blocking probability.....	75
Figure 5.5 Mobility-triggered handoff call dropping probability.	75
Figure 5.6 Average QoS satisfaction ratio of calls.....	76
Figure 5.7 Average throughput of the integrated system.	77
Figure 5.8 Standard deviation of network utilizations.	77
Figure 6.1 Comparison of different network selection strategies.	81
Figure 6.2 Two-tier model with resident application.	93
Figure 6.3 Two-tier model with general mobile application.	94
Figure 6.4 Three-tier model with resident application.	95
Figure 6.5 Three-tier model with general mobile application.	96
Figure 6.6 Comparison with greedy algorithm and CPLEX.....	98

Figure 6.7 Comparison with the algorithm in [12]..... 100

List of Tables

Table 1.1 Comparison of different wireless technologies.	2
Table 4.1 Parameters used in the algorithm.....	41
Table 5.1 Key parameters used in the simulation.	73
Table 6.1 Notations used in the NSRA scheme.	83
Table 6.2 Lagrangian algorithm.....	87
Table 6.3 Key parameters used in the simulation.	91
Table 6.4 Greedy algorithm for the NSRA scheme.	98

Abstract

The trend of the fourth-generation wireless networks (4G) is the coexistence of heterogeneous technologies. With an overlay architecture, advantages of different technologies can be well integrated into one system, allowing subscribers to profit the best service anytime and anywhere, i.e., the so-called “always best connected” (ABC). To efficiently integrate heterogeneous networks, several key issues should be investigated. Among them, the most fundamental issue is the interworking architecture, which provides the solution on the platform to combine the overlay networks together. Under different architectures, the integration mechanisms in terms of mobility management, data transmission, quality-of-service (QoS) provision, and security might be quite different. The handoff management is another essential aspect for interworking. The handoff between heterogeneous networks is referred to as vertical handoff. Unlike traditional horizontal handoff which is mainly triggered by signal strength, more factors should be considered in a vertical handoff decision, such as monetary cost, network conditions, user preference, and so on. The objective of a vertical handoff will be no longer the continuous connection but the best connection. On the other hand, because the radio resources in the interworking system can be viewed as being shared in a resource pool, the complete and efficient use of these pooled resources is also an important topic.

In this thesis, we first investigate the integration mechanisms as well as vertical handoff issues in the WiMAX and WLAN overlay system, which is recently developed and believed to be a promising approach toward 4G. A tightly coupled interworking architecture is addressed and the related important mechanisms for the integration are discussed. Based on this architecture, a seamless and proactive vertical handoff scheme is designed, targeting to provide always the best services for both mobile and fixed users. Both the performance of applications and network conditions are considered in the handoff process. In order to achieve proactive handoffs, the algorithms for the network condition detection are derived for stations to estimate the available bandwidth and

average packet delay of the two networks.

We then study the resource management of general heterogeneous overlay systems from the viewpoint of network providers. To effectively utilize the overall radio resources pooled in the interworking system, an efficient load balancing scheme is proposed which is applied to the call admission control process. When the available networks cannot provide enough resource for the request call without degrading the QoS obtained by the ongoing calls, the system will perform load balancing operations by dynamically initiating vertical handoffs among networks to create more room for the request call.

Finally, we make considerable research efforts on the radio resource management for the multicasting and broadcasting service in the heterogeneous wireless networks. A novel network selection and resource allocation scheme is designed to minimize the overall system resources consumed by the multicast or broadcast stream, which is formulated as an optimization problem and solved by the derived algorithm. By the scheme, the member stations at the same network region can be effectively converged together, meanwhile the stream can be transmitted with higher modulation and coding rates so that the consumed resources can be reduced.

Chapter 1

Introduction

In recent years, wireless communication systems have experienced an impressive evolution leading to the deployments of a great variety of technologies such as second and third-generation (3G) cellular networks, wireless local area networks (WLANs) and worldwide interoperability for microwave access (WiMAX).

1.1 The Interworking of Heterogeneous Networks

In Table 1.1, the key features of 3G cellular networks, WLAN and WiMAX are summarized and compared (their evolutions will be introduced in Section 2.1). We can see from the table that these wireless access technologies exhibit complementary characteristics. Cellular systems provide wide coverage areas, full mobility and roaming, but usually offer low bandwidth connectivity and limited support for data traffic. WLANs provide high data rate at low cost, but only within a small area, while WiMAX can supply mobile broadband over a metropolitan area. Each of these technologies has been tailored to reach a particular market or a particular group of users during these years. Not a single wireless access technology is able to optimally cover all the wireless communication scenarios.

Therefore, the fourth-generation (4G) wireless network will be the coexistence and interworking of these diverse but complementary wireless technologies. A typical 4G network architecture can be viewed as many overlapping wireless access domains, and is called a heterogeneous overlay network. With this network architecture, users will benefit from the different coverage and capacity characteristics of each network throughout the integrated networks. A large set of available resources will allow them

Table 1.1 Comparison of different wireless technologies.

	Advantages	Disadvantages
3G cellular	Wide coverage High mobility and high security Widest ally of Telecom operation	Comparatively low data rate Large investments Network upgrading at high cost
WLAN	High capacity under large bandwidth Cost-effective Unlicensed frequency band	Small coverage Low mobility Bad security
WiMAX	Metropolitan coverage High mobility and high capacity Comparatively low cost	Immature technique and user equipment; unstable performance in long distance transmission

to seamlessly connect to the most suitable network at any time and any place, which is known as always best connected (ABC) [4]. Meanwhile, for network operators, the integration of heterogeneous technologies provides more efficient usage of the network resources, and is expected to be the most economic and technologically diversified means of implementing the future ABC versions and providing both universal coverage and broadband access.

1.2 Key Issues for the Design of Interworking System

With the interworking of heterogeneous wireless technologies, enhanced services can be provided, while new challenges also arise from the heterogeneity in the process of achieving interoperability among different technologies.

- Interworking mechanisms

The fundamental issue for the design of heterogeneous wireless network is the interworking architecture, which provides the solution to combine different networks efficiently. Under different architectures, the integration mechanisms in terms of mobility management, data transmission, QoS provision, and security will be quite different. Both the business and technical issues should be taken into account. The interworking mechanisms may vary as well with the ownership, provided services, and the target

penetration level between different networks [5]. The corresponding architecture is expected to be capable of making the most of existing network components and protocols in order to minimize the implementation complexity and cost for network operators. Meanwhile, the combined system should always provide the advantages of heterogeneous networks for users.

- Mobility management

Mobility management remains to be an essential topic for the design of wireless networks because the performance of handoff operations significantly influences the user's experience on the service continuity and fluency. In the heterogeneous wireless interworking networks, both intra-technology handoff and inter-technology handoff may take place. Intra-technology handoff is the traditional horizontal handoff (HHO) process where the mobile station roams between two access points (APs) or base stations (BSs) using the same radio access technology. Inter-technology handoff is also called vertical handoff (VHO), which occurs when the station changes its point of attachment to a radio access network using a different technology. Horizontal handoff is inherently supported as a functionality of the mobility management in a homogeneous wireless network and the handoff procedure has been well defined via network specifications.

However, due to the use of different mobility management protocols in individual networks within the interworking system, vertical handoff becomes a challenging issue. It is expected that the vertical handoff operations introduce the minimum modifications to existing networks. On the other hand, within the overlapped areas in a heterogeneous interworking system, more than one network is available to stations. This provides an opportunity to dynamically change the connection via vertical handoffs, so that the most suitable network can be selected to provide the service demanded by the users or network operators. Consequently, unlike traditional horizontal handoffs which are mainly triggered by the signal quality, more factors have to be considered in vertical handoffs such as monetary cost, power consumption, user preference, application, available bandwidth, packet latency, and so on. In this context, being substantially different from traditional cases, vertical handoff is *exploited* as a tool of network selection for the benefits of both the mobile and fixed users. Therefore, the mobility in 4G networks is no longer only a physical concept but also a logical one [6], and the objective

of a handoff is no longer simply for connection but the best connection. To achieve this goal, choosing the correct time to initiate a vertical handoff and selecting the best network to be connected are crucially important but very challenging.

- Radio resource management

It is known that radio spectrum is one of the most important resources in wireless networks. Therefore, the radio resource management plays a key role in effectively provisioning QoS guarantee and effectively utilizing the scarce radio resources. The functions of radio resource management include radio bear control, call admission control, connection mobility control, dynamic resource allocation, packet scheduling, and load balancing.

In heterogeneous interworking systems, the radio resources of different networks can be seen as being combined in a resource pool. To achieve high utilization efficiency, a joint radio resource management is desired. Meanwhile, the unique characteristics of an individual network should be taken into account. If a call comes from an overlapped area, the controller cannot simply decide whether to admit or reject the call as that in the traditional way. A comprehensive decision considering which available network is most suitable for the call has to be made. Various factors could be taken into account before making such a decision including the QoS requirements of users and the traffic load in each network. The controller even has an opportunity to proactively transfer connections among networks to balance their traffic load. Moreover, if it is a multicasting or broadcasting call which has a large group of members distributing in many networks with different channel qualities, the selection of networks to serve its members and the allocation of resource for such call will significantly influence the overall system resource utilization. Therefore, more intelligent and effective resource allocation, network selection, and admission control algorithms are strongly needed in heterogeneous interworking systems.

1.3 Motivation and Research Scope

From the late 1990s, the interworking between cellular and WLAN has been studied extensively. There are many literatures documented on the integration mechanisms and vertical handoff schemes for the cellular/WLAN overlay system. Compared to cellular systems, WiMAX is a newly developed wireless access technology and has seldom

been proposed to be integrated with WLAN. Moreover, the existing schemes for the integration of cellular/WLAN may not be applicable for WiMAX/WLAN systems due to the heterogeneity of these networks. For instance, WiMAX and WLAN are developed for the packet-switched networking from the beginning, while cellular technologies are evolved from voice communication-oriented circuit-switched networking. Meanwhile, WiMAX provides a comparable data rate with WLAN. This weakens the preference on WLAN as in cellular/WLAN interworking systems. Therefore, the interworking of WiMAX and WLAN networks is an interesting topic and deserves further research efforts.

On the other hand, as previously discussed, vertical handoffs can be proactively initiated to transfer connections among networks in order to connect to the most suitable network according to the requirements of users or network operators. Although more factors have been considered in many existing vertical handoff approaches, the handoff processes in most of these approaches are still initiated at the border of networks, and thus cannot fully benefit from vertical handoffs. With the development of wireless technologies, the trend of 4G mobile terminals is the capability of being configured with multiple interfaces to connect to different networks simultaneously [7]. During the vertical handoff process, one interface can perform handoff operation while another interface maintains current connection. Handoff latency and packet loss are thus no longer as critical as those in traditional homogeneous networks. This significantly increases the feasibility of exploiting vertical handoff as a network selection tool. Moreover, as mentioned above, the network selection and resource allocation of multicasting/broadcasting services greatly influence the overall system resource utilization, while the works on this topic are still lacking.

In this thesis, we first investigate the integration mechanisms as well as vertical handoff issues in the WiMAX/WLAN overlay system [8, 9]. Interworking architectures of tightly coupled and loosely coupled are presented, respectively. By comparing the pros and cons of both options, we recommend the tightly coupled architecture as the preferred one. Important mechanisms for the integration under this approach are discussed in terms of data path function and security issues. Based on this architecture, a seamless and proactive vertical handoff scheme, considering both the performance of applications and network conditions, is designed to provide always the best services for both mobile and fixed users. Meanwhile, to achieve proactive handoffs, the network

condition detection algorithms are derived for stations to estimate the available bandwidth and average packet delay of both WiMAX and WLAN networks.

Furthermore, we investigate the resource management issue for general heterogeneous wireless overlay networks from the viewpoint of network operators [10]. To effectively utilize the overall radio resources pooled in the interworking system, an efficient load balancing scheme applied in the call admission control process is proposed. When the available networks cannot provide enough resource for the request call without degrading the QoS of the ongoing calls, the system will perform load balancing operations by proactively initiating vertical handoffs among networks in order to create more room for the request call.

Besides, we specially investigate the resource allocation issue for the multicasting and broadcasting service in heterogeneous overlay systems [11]. Although Yang and Chen [12] also studied this problem, they only considered the network selection without the consideration of resource allocation. Indeed, the members of a multicast or broadcast stream may consume different amount of radio resources in one network due to their distinct channel quality. Meanwhile, each member may consume different amount of radio resources in different available networks. Therefore, to effectively utilize the system resource, the resource allocation should be tightly integrated into the network selection process. Motivated by this, we propose an efficient network selection and resource allocation scheme for the multicast and broadcast streams. By enabling the member stations in the same network region to cluster together and deploying higher order modulation and coding schemes, the overall system bandwidth cost consumed by the stream can be minimized.

1.4 Organization of the Thesis

This thesis is organized as follows. Related research works on the integration of heterogeneous wireless networks including integration mechanisms, vertical handoff schemes, and radio resource management approaches are briefly reviewed in Chapter 2. In Chapter 3, we present possible interworking architectures for the WiMAX/WLAN overlay systems and address integration mechanisms for the tightly coupled interworking architecture. Based on this architecture, a vertical handoff scheme is proposed in Chapter 4. Network condition estimation algorithms in terms of available bandwidth

and packet delay are also derived. In Chapter 5, we propose a load balancing scheme for general heterogeneous overlay systems. Specially, the network selection and resource allocation for multicast and broadcast streams are investigated in Chapter 6. Finally, Chapter 7 summarizes the contributions of this thesis and outlines possible future works.

Chapter 2

Related Works and Background

The next-generation wireless network necessitates the coexistence of heterogeneous wireless access technologies. In recent years, some literatures have been published to investigate the integration of heterogeneous wireless technologies concerning the interworking mechanisms, vertical handoff, and radio resource management.

2.1 Current Wireless Access Technologies

2.1.1 Cellular networks

The most notable first-generation (1G) cellular system was the so-called advanced mobile phone system (AMPS), which was introduced by Bell Labs in 1980s for voice communication through 30 KHz modulated channels based on analog frequency-division multiple access (FDMA). 1G cellular systems had no uniform standardization, which throttled their wide usage in the market. Standardization was started since the second-generation (2G) cellular systems. The European global system for mobile communications (GSM) standard introduced digital communication with a combination of time-division multiple access (TDMA) and slow frequency hopping for the voice communication. In United States, two 2G cellular systems were standardized: Interim Standard-136 (IS) standard which considers TDMA and FDMA with phase-shift keyed modulation, and cdmaOne IS-95 standard which utilizes direct-sequence code-division multiple access (CDMA) with phase-shift keyed modulation and coding [1]. For more than a decade, the 2G systems have performed very well in voice communication. This has led to 400 million 2G mobile subscribers in the year 2000, and

this number has increased to about 4 billion for the year of 2010.

Despite their great success and market acceptance, 2G systems are fairly limited by their data transmission rate, e.g., only 9.6 kbps of GSM, a value which is sufficient for conventional voice services but extremely low to fulfil the requirement of multimedia applications and Internet services which become increasingly popular in recent years. In this context, 2.5G technologies including general packet radio system (GPRS) and enhanced data rates for GSM evolution (EDGE) have been developed; they offer packet-based data services. The peak data transmission rate is increased to 140 kbps by GPRS when a user aggregates all timeslots and to 384 kbps by EDGE with high-level modulation and coding.

In order to further improve the data rates to cover more applications, the 3G systems arose with newly assigned frequency bands and the deployments of new network elements, especially in the radio access part. 3G systems comprise several standards such as the universal mobile telecommunications system (UMTS), wideband CDMA (WCDMA) and CDMA2000. These systems can provide at least 144 kbps for full mobility applications in all cases, 384 kbps for limited mobility applications in macro- and micro-cellular environments, and 2 Mbps for low mobility applications particularly in the micro- and pico-cellular environments [2].

As the 2G and 3G systems coming into widespread use, the natural forthcoming step of cellular system is to integrate Internet-based services into the mobile devices, creating what is known as *mobile broadband* for future 4G. At this point, Long Term Evolution (LTE) and its enhanced version LTE-Advanced based on the UMTS technologies are under development. Theoretically, LTE is required to deliver a peak data rate of 300 Mbps in the downlink and 75 Mbps in the uplink. A more realistic requirement is to support a spectral efficiency three to four times greater than that of WCDMA in the downlink and two to three times greater in the uplink [148]. As LTE is still immature, existing works on the integration of cellular and other networks mainly focused on the 2G and 3G cellular systems. In this thesis, 3G system is selected to represent the cellular system.

2.1.2 WLAN

Along with the evolution of cellular systems, other wireless access technologies have

been developed and standardized. WLANs based on the IEEE 802.11 standard families have been widely deployed in hot spot areas providing Internet access at home, office, airports, and hotels. The concept of WLAN was introduced in early 1990s and has achieved great success due to its wide bandwidth with a much lower cost. The most popular WLAN standard IEEE 802.11b operates at the license-exempt frequency band from 2.4 GHz to 2.483 GHz and utilizes direct-spread spectrum to offer data rates of up to 11 Mbps within 100 m range. The subsequent version 802.11a operates at unlicensed 5 GHz band and provides up to 54 Mbps data rate within less than 100 m range, which is the first standard in the wireless domain to employ orthogonal frequency-division multiplexing (OFDM) technique. IEEE 802.11a also provides more channels than 802.11b does, and is able to accommodate users with higher data rates. To leverage this system design, 802.11g is further developed in the 2.4 GHz band with the same design as in 802.11a. So far, most 802.11 products are based on 802.11b and 802.11g standards. Due to their high market penetration, several amendments to the basic standard have also been developed [147]. Multi-input multi-output (MIMO) capability is introduced by 802.11n, in order to improve the data rate up to 600 Mbps. A flexible MIMO concept allows for arrays of up to four antennas that enable spatial multiplexing or beam forming. 802.11ac and 802.11ad standards under development target even higher throughput of more than 1Gbps.

Besides the high data rate, another key element to the success of 802.11 is its simple medium access control (MAC) operation based on the distributed coordination function (DCF) protocol [147]. Although the central resource management and quality of service (QoS) provisioning functions are not supported by DCF, the over-provisioning of capacity enables it to serve most of the original best effort applications. The later amendment 802.11e made efforts on the QoS provisioning; however, this standard failed to capture the market. On the other hand, WLAN is designed mainly for nomadic users who use this technology when staying mostly at a given place. Though some low mobility (e.g., walking speed) can also be supported, most of WLAN devices (such as PCs, laptops, etc.) are not optimized for mobility support.

2.1.3 WiMAX

In order to fill the gaps on coverage, data rates, and mobility between cellular networks

and WLANs, WiMAX has been developed by the IEEE 802.16 working group. The version 802.16d concluded in 2004 provides fixed wireless broadband access. Its later version 802.16e concluded in 2005 has included mobility support and allowed fixed wireless and mobile non-line-of-sight (NLOS) applications. It is also called “Mobile WiMAX”. The first system development and commercial service were launched in 2007 in Korea. At the physical (PHY) layer, Mobile WiMAX incorporates many strong technologies, such as orthogonal frequency division multiple access (OFDMA), time-division duplexing (TDD), MIMO, adaptive modulation and coding (AMC), that can be combined to produce high spectral efficiency and resilient channels resulting in high-rate and wide-area mobile multimedia services [3]. Mobile WiMAX supports roaming service in metropolitan and regional networks, and then allows mobile connectivity to mobile users. The target mobility is 120 km/h and peak throughput is 18.7 Mbps for downlink (DL) and 5.0 Mbps for uplink (UL) in the case of DL/UL ratio = 29:18 and 10 MHz bandwidth. Theoretically, a WiMAX BS can provide broadband wireless access in a range up to 50 km for fixed stations and 5 to 15 km for mobile stations with a maximum shared data rate of up to 70 Mbps. In addition, for effective provision of real-time multimedia services, WiMAX adopts a connection-oriented bandwidth request and allocation MAC. It is designed to be capable of providing high-rate, high-quality data services to mobile stations in the medium over metropolitan areas.

2.2 Interworking Architecture of Heterogeneous Wireless Systems

So far, various interworking architectures have been proposed, and most of them focus on the cellular and WLAN interworking systems such as WCDMA/WLAN [13], CDMA2000/WLAN [14], and UMTS/WLAN [15]. According to the inter-dependence between different networks, these interworking architectures can be classified into two categories: loosely coupled and tightly coupled [16]. Loosely coupled indicates a means of interconnecting both networks independently with only a common subscription, while tightly coupled suggests that the WLAN works just as a cellular access network and connects to the cellular core network.

2.2.1 Loosely coupled

With the loosely coupled approach, the cellular and WLAN networks are integrated beyond the core networks, usually through an external IP network such as Internet [5]. There are no direct links between the cellular networks and WLANs. Usually, pervasive IP technology is employed by the loosely coupled architecture to combine together the heterogeneous networks, following the standards of the Internet community such as Mobile IP [17] and authentication, authorization and accounting (AAA) [18] framework. The roaming between two networks is achieved with the aid of Mobile IP [152]. The user access is controlled by the AAA service, which enables the cellular provider to collect the WLAN accounting records and generate a unified billing statement that indicates usage and price schemes for both networks. The data traffic of WLANs within a loosely coupled system goes directly to Internet without passing the cellular core network and different mechanisms of two networks can be implemented independently to handle user mobility, authentication and billing.

The main advantage of loosely coupled approach is the independent deployment of heterogeneous networks that allows the network operators to take advantage of the implemented networks from other providers with minimal enhancements in network devices. Subscribers can use all integrated wireless access networks while subscribing to only one network operator. For example, a UMTS subscriber can reuse its subscriber identity module (SIM) or user service identity module (USIM) for WLAN services [19]. Owing to such flexibility, the loosely coupled approach has attracted the most research and application attention at the initial stage of heterogeneous wireless networks [20, 21]. However, this approach is relatively inefficient because of its long signaling path and redundant processing during roaming process as well as a large number of network elements involved in management operations. As measured in [22], the average handoff latency reaches 400 ms, which is not suitable for real-time multimedia services. More discussions on its mobility performance will be given in Section 2.3.3.

2.2.2 Tightly coupled

In the tightly coupled architecture, WLANs connect to the cellular core network just as a cellular radio access network and emulate functions that are natively available in the cellular networks. The WLAN gateway hides the details of WLANs to the core of the

cellular network, and implements all the protocols (mobility management, authentication, etc.) required in the cellular network. Mobile terminals in this approach are required to implement the corresponding cellular protocol stack on top of their standard WLAN network cards. All the traffic generated by the terminals in WLANs is injected into the cellular core network.

Compared to the loosely coupled approaches, the tightly coupled ones could work in a more efficient way as follows:

- The tightly coupled architecture reuses the existing cellular infrastructure to a large extent while the cellular radio is simply replaced with WLAN radio. The cellular radio access networks and WLANs share the functionalities in the cellular core network and then the system implementation cost can be reduced [150].
- In the heterogeneous wireless network, the radio resources of networks can be seen as combined in a pool. In order to efficiently exploit the pooled radio resources, the joint or common management of all the radio resources in the integrated networks is desired. It is believed that the tighter the coupling between networks is, the better the resources are utilized jointly leading to improved performance [23].
- Moreover, under the tightly coupled approaches, WLANs and cellular networks could be integrated below the IP layer, and their associated mobile stations have IP addresses assigned from the same pool. The user roaming across the two domains follows the cellular mobility management protocol with no need to change its IP address. Consequently, Mobile IP is no longer necessary in the tightly coupled systems, although it is a common feature of loosely coupled ones. As a result, the considerable signaling overhead as well as redundant processes introduced by Mobile IP can be avoided, which makes seamless vertical handoff, i.e., a strongly expected function for heterogeneous interworking networks to provide continuous QoS guarantee for mobile users, to be possible in the tightly coupled systems [24].

Despite their advantages, the tightly coupled approaches have not been widely deployed in the cellular/WLAN integrated systems. This is mainly caused by the fact that cellular network and the WLAN network are heterogeneous in nature (i.e., cellular is originally circuit-based while WLAN is packet-based). A gateway is needed by WLANs to implement all cellular protocols and hide the details of WLANs. Modifica-

tions are also introduced to cellular networks which require long-term standardizations. Moreover, the capacity of the cellular core network may become a bottleneck of the injected WLAN traffic.

Nevertheless, the tightly coupled architectures are believed to be the next logical step toward future seamless integrated network for their enhanced performance [24]. The future wireless network is expected to be a converged network with a common IP-based core network shared by a variety of access networks [25]. The access technology specific functions only propagate till the gateway to the shared common core network as that in the tightly coupled architecture. Moreover, current wireless technology such as WiMAX evolves as a packet-switched network from the beginning and can be seen as an extent of WLAN. The tight coupling of WiMAX and WLAN will be much easier than the coupling of cellular and WALN due to the homogeneous nature. Therefore, efforts in both the definition and implementation of heterogeneous interworking systems are developed toward the tightly coupled direction. Currently, there is a standardization activity within the third generation partnership project (3GPP) focusing on this topic, which is known as generic access network (GAN) or unlicensed mobile access (UMA) [26]. Issues for the tight coupling of 3GPP networks with WiMAX [27] or WLANs [28] are discussed in literatures. Meanwhile, more and more solutions based on the tightly coupled systems have also been proposed for the seamless vertical handoff support [29-31] or joint radio resource management [23].

2.3 Vertical Handoff

In heterogeneous wireless interworking networks, there are normally two types of

Figure 2.1 Vertical handoff vs. horizontal handoff.

handoffs. One is horizontal handoff that is happening as traditional handoffs between adjacent cells with the same wireless access technology, and the other is vertical handoff which is the switch of ongoing connections between networks with different wireless access technologies. Horizontal handoff within a homogeneous wireless network is inherently supported as a functionality of the mobility management, and the handoff protocol has well defined operations in each handoff stage (i.e., initiation, decision, and execution). In the integrated systems, however, networks are heterogeneous while complementary in various aspects. The environment of a vertical handoff is much more complex than traditional cases; thus, many new challenges are introduced by the network heterogeneity.

2.3.1 Handoff initiation and objective

Traditional horizontal handoffs are mainly triggered by the channel quality which may be measured with the metrics of received signal strength (RSS), signal-to-noise ratio (SNR), bit error rate (BER), and packet loss rate. When a mobile station approaches the network border with a metric violating the given threshold, the mobile station needs to connect to another possible network. The main objective of horizontal handoffs is to maintain the connectivity and the relevant researches focus mainly on how to reduce the handoff latency and packet loss.

In heterogeneous interworking systems, multiple networks with complementary characteristics are available at overlapped areas. Therefore, besides the basic function of the connectivity maintenance, vertical handoff has an opportunity to transfer connections between networks even under the acceptable channel quality. For instance, if the service provided by the serving network cannot satisfy the QoS requirements, a vertical handoff may be initiated to connect to another network with better conditions so that the performance of current applications can be improved. If a mobile station moves into a network with lower price or lower power consumption, a vertical handoff may also be initiated to reduce monetary cost or save power. The handoff operations in such cases are usually controlled by users. On the other hand, if a certain area in an integrated system becomes hot or congested, the network operator can also transfer some calls from the hot area to cold areas via vertical handoffs. In such case, the vertical handoff is typically network-controlled. Therefore, various objectives could be

achieved by exploiting vertical handoffs.

Although previous works have indeed taken various factors into vertical handoff decisions, the handoffs are mainly initiated at the border of networks. One example is the *WLAN-first* scheme which is usually deployed in cellular/WLAN interworking system [32-34], where WLANs are taken as the preferred network due to their much higher bandwidth and lower price. Once a mobile station moves into the coverage of an AP, it initiates a handoff to WLAN immediately and attempts to stay in it as long as possible although the cellular network is still available. In this approach, stations may suffer from saturated WLANs when an excessive number of stations access the network. An improvement is introduced in [35] by estimating the WLAN traffic load before making a handoff decision. However, this operation is still only performed at the border of WLAN.

With the development of wireless technologies, the multimode terminals are allowed to connect to multiple networks simultaneously [7, 36]. During vertical handoff process, one interface of the station may perform handoff operations with the target network while another interface maintains current application through the serving network, and thus the packet loss can be avoided. This significantly improves the feasibility of exploiting vertical handoffs at anytime within overlapped areas. The exploitation of vertical handoffs for users or network operators will be the trend of future research and implementations.

2.3.2 Handoff decision and network selection

As a vertical handoff happens between different networks with complementary characteristics, intelligent handoff decision algorithms are needed to select the most suitable network for service and various metrics could be taken into account. The selection of these metrics depends on the predefined handoff policy.

Certainly, the basic vertical handoff policy is still to maintain the connectivity when the channel quality of serving network is not acceptable. The main metrics in this case include RSS and SNR. One challenge arising in the heterogeneous interworking system is that the parameter such as RSS from different networks can vary significantly due to different techniques used in the physical layer and cannot be easily compared with each other. Other techniques are thus usually needed along with RSS to make a

handoff decision. Zhang et al. [37] proposed a fast Fourier transform (FFT) approach to further detect the signal decay quickly and accurately when the sampled RSS is consistently less than a threshold. In [38], some previous RSSs are brought into a polynomial function to predict the next RSS, which will initiate the handoff before RSS becomes too weak. On the other hand, another issue that needs to be paid more attention to in such approaches is to avoid the ping-pong effect. Traditional hysteresis schemes cannot be used directly because RSSs from different networks are difficult to be compared. The velocity and moving direction of the mobile station is measured in [39] in order to estimate the residual time in the current network and then avoid unnecessary handoff.

More advanced policies could be deployed in vertical handoff decisions, which can consider various metrics such as application type, network condition parameters, monetary cost, power consumption, mobile speed, user preference, and security. In [40, 41], power consumption is selected as the main metric. To maximize the battery life, the mobile station selects the uplink or downlink that has the lowest power consumption from all available networks. In [42], the user satisfaction is evaluated mainly by monetary cost. The network selection is based on the predicted consumer surplus value (difference between the monetary value of the data transferred and the actual price charged) and transfer completion time, with an aim to always select the cheapest network for user.

Also, there are many works that combine multiple metrics in the vertical handoff decisions. The challenge in such case is how to make an intelligent decision based on various metrics. A popular way is *weighted-based* which sets different weights over different factors and a handoff is made to the network yielding the highest magnitude of a weighted sum [43-45]. Both the weighted product [43] and a weighted logarithm of attributes [44] can be classified into this type. As a specific weight reflects the relative importance of the corresponding attribute, it is important to adapt the weights toward achieving the desired behaviour of vertical handoff. A *policy-based* scheme is addressed in [46], by which a network with attributes best matching the specified policy is selected as the target network. In [47, 48], the *utility-based* strategy is deployed for network selection, where multiple factors are considered in a utility function. A handoff is made to the candidate network with the highest utility value. A *cost function-based* scheme is designed in [49] by taking into account both the network perfor-

mance and the user needs. These literatures mainly focus on how to make a handoff decision. The technical details on obtaining the parameters used in the decision, especially the network condition related parameters, are not provided.

Moreover, some more complex vertical handoff decision algorithms are also designed [50-53]. In [52], the handoff decision problem is formulated as a Markov decision process and solved by dynamic programming. It considers the movement and location information, which is provided by the location service server. In [50], artificial neural networks (ANNs) are used. Various collected user preference parameters and network information are input into a middleware, which utilizes the ANN to find the best match between the user needs and the network performance. In [51], the vertical handoff decision is formulated as a fuzzy multiple attribute decision-making (MADM) problem. Fuzzy logic is used to represent the imprecise information of some attributes of the networks and the preferences of the user. In [53], different vertical handoff strategies are enabled between two extremes: the former triggers a handoff once a channel is interrupted and the latter waits for a fixed time interval after channel interruption before executing a handoff. Fuzzy logic based method is used to model input uncertainty and then decides which handoff strategy to be performed. It is important to mention that the complexity of such algorithms is very high, and their implementations in stations with limited computing and storage capability may not be possible [54].

2.3.3 Handoff execution

Once a decision of handoff is made, the handoff operations should be performed in a fast, smooth, and seamless way [55]. In a loosely coupled network, Mobile IP is usually deployed to relay the connection of the handoff station from the source network to the target. This usually introduces a significant increase on the system load, handoff latency, and packet loss [56]. The weaknesses of Mobile IP led to enhancements standardized by Internet engineering task force (IETF), such as hierarchical Mobile IP (HMIP) [154] and fast handovers for Mobile IP (FMIP) [155]. HMIP locally handles handoff through a special node which acts as a local home agent in the visited network. By FMIP, the link layer information (L2 trigger) is used either to predict or to rapidly respond to handoff events. Specific techniques based on these ideas are also addressed in literatures to improve the vertical handoff performance in heterogeneous wireless

systems. Hierarchical-structured foreign agents are deployed in [57] to localize the registration messages, and the link layer handoff events are used in [58, 59] to trigger the Mobile IP operations. Although these enhancements pave the way for the improvement of original Mobile IP in terms of reducing handoff latency, they are still hindered by several problems. They do not effectively reduce the handoff latency and packet loss to a level meeting the requirements of time-sensitive traffic [6]. Meanwhile, these solutions are client-based, i.e. the mobile station controls the handoff procedure. Hence the signaling overhead over wireless links is still a problem. Moreover, the network components (such as foreign agents) in the visited and home networks need to exchange the contexts or buffered data packets of the handoff stations. The cooperation between them is thus required, which weakens the benefit of independent deployment of networks provided by the loosely coupled architecture.

It is worthy to mention that the most recent enhanced version of Mobile IP is a network-based approach - Proxy Mobile IP (PMIP). With PMIP, the visited network handles mobility management on behalf of the mobile station. By doing so, the tunneling overhead and the signaling message exchanges via the wireless links can be reduced [60]. On the other hand, as multimode stations can connect to multiple networks simultaneously (see Section 1.3), the handoff latency and packet loss in heterogeneous wireless networks are not as critical as those in legacy homogeneous networks. The reduced signaling overhead increases the applicability of PMIP in heterogeneous interworking systems. However, it has to be noted that with PMIP the visited network components need to perform the registration to the home network on behalf of the mobile station; therefore, the cooperation between different networks is still required by this solution.

In a tightly coupled network, vertical handoff is internal to the core network and the IP address of the mobile station needs not change. The handoff is executed based on inherent mobility management protocols of the overlay network such as cellular network. As a result, the handoff latency and signaling overhead in this case is much lower as compared to loosely coupled cases, and seamless vertical handoff can be expected [24].

2.3.4 Existing works for the WiMAX/WLAN interworking system

In recent two years, the handoff management in WiMAX/WLAN interworking systems begins to attract more attention, but still with many open issues. References [61, 62] focus on the protocol related issues such as message exchanging during handoff. Both of them use IEEE 802.21 [63] to manage the handoff process, which is a newly emerging standard aiming at providing a framework for media-independent handoff (MIH) among heterogeneous networks. A movement-aware vertical handoff scheme is proposed in [39], in which the velocity and moving pattern of the station are predicted. The main objective is to avoid unnecessary handoff. With these approaches, WiMAX and WLANs are integrated in a loosely coupled manner although not explicitly stated. The work designed in [64] is based on a tightly coupled architecture while the integration mechanisms are not discussed.

References [65-69] focus on the vertical handoff decision algorithms. Traffic type is used in [69] as the main metric considered in the handoff decision. Real-time traffic tends to use WiMAX while non-real-time traffic tends to use WLAN. Service charge is considered in [68] with the objective to minimize the user payment. Practical service fee models, i.e., flat pricing for WLAN and partially-flat pricing for WiMAX, are considered in the decision algorithm. Reference [67] focuses on maximizing the delivered information bits during handoff by forcing the station to wait for an additional network selection time. On the other hand, QoS parameters are considered in [65, 66]. A mechanism that supports always the best connected service via vertical handoffs for the IEEE 802.11e WLAN and IEEE 802.16d WiMAX interworking systems is proposed in [66]. The major metrics taken into account in the handoff decisions include throughput, packet loss, and delay. However, the technical details on dynamically obtaining the values of these parameters are not provided in these literatures.

2.4 Resource Management

The radio resource management (RRM) in the wireless network is responsible for the efficient utilization of air interface resources to guarantee a certain QoS level for different users according to their traffic profiles. Radio resource management includes functions such as radio bear control, call admission control, connection mobility control, dynamic resource allocation, packet scheduling and load balancing.

2.4.1 Joint radio resource management

In heterogeneous wireless interworking systems, radio resources can be treated as being in a resource pool. It is crucial to effectively manage the pooled radio resources to ensure seamless interworking and optimize overall system resource utilization. Although individual resource management can be tuned to optimally utilize local radio resources within their respective access networks, they may not efficiently perform in heterogeneous interworking systems if the different resource management functions are not properly managed. Hence, a major issue is how to jointly utilize the resources of the different networks in an efficient manner while simultaneously achieve the desired QoS and minimize the service cost from both the user and network operator perspectives. Joint RRM (JRRM) [70], also known as common RRM (CRRM) [71], is thus strongly desired in heterogeneous wireless interworking systems.

3GPP has standardized a number of architectures for CRRM [72]. The functional model assumed in 3GPP for CRRM operation considers the total amount of resources available for a network operator divided into radio resource pools. Each radio resource pool consists of the resources available in a set of cells and is managed by a local resource management entity. A CRRM server (CRMS) is introduced to collect measurements and exchange management information from local resource management entities, and apply a common algorithm.

With JRRM, the information of individual networks including location, cost, mobility model, etc., can be collected by a central controller (e.g., CRMS), and then more intelligent resource allocation schemes can be designed to exploit the benefits of the interworking of heterogeneous networks from a complete system view. For example, upon encountering congestion at an individual network, its local resource management entity may inform the central controller, which then helps transfer the call to other possible networks where more resources are available [71]. The numerical examples in [73] illustrate that the network revenue earned by the joint admission control scheme is significantly larger than that when individual networks are optimized independently with no vertical handoffs between them.

Many research works on the radio resource management in heterogeneous interworking systems are based on JRRM, including joint scheduling, joint congestion control, and joint admission control. For example, a joint scheduling mechanism is de-

signed in [74], which allows the traffic to be split into multiple layers with each transmitted through different networks within a tightly coupled system, by taking into account the network available bandwidth and QoS requirements. The cellular network is selected as the default access network for its reliable coverage to serve the base-layer of the stream. If another network becomes available during the session, it is used to transport the enhancement-layer of the stream to improve the service quality. The drawback of this approach is the capacity bottleneck induced by the default cellular network. At this point, an improved solution is introduced in [75], which provides flexibility of rearranging other ongoing sessions among the interworking networks based on user mobility, traffic load, and preferences. However, this solution also selects cellular as the default access network for initial communication. Such static default network operation may introduce unnecessary handoff soon after the session initiation. To solve this problem, the CRRM scheme in [76] incorporates location prediction information with a service and mobility model to provide more stable communication.

2.4.2 Admission control

The first step of the system resource management in wireless networks is the admission control (AC), which aims to admit as many as possible calls into the system without degrading the QoS. Many researches and approaches of admission control for homogeneous wireless networks have been reported in literatures [77-79]. In homogeneous wireless networks, each mobile user with a single-mode terminal is confined to a single access network, and then the traditional admission control algorithms mainly decide whether to admit an incoming call into the network without violating the service level agreement of existing users.

In heterogeneous interworking systems, multiple access networks with different radio technologies are overlapped in the same area. If a multimode terminal generates a call within such an area, the joint admission control scheme is not only responsible for deciding whether the call may be accepted or not but it also needs to decide which of the available networks is most suitable to accommodate the incoming call. Consequently, the *network selection* is inherently included in the admission control operations for heterogeneous interworking systems. We have reviewed a number of network selection strategies for vertical handoff calls in Section 2.3.2; they mainly reflect individual us-

er's needs on monetary cost, user preference, and power consumption. Besides these factors, the network operator's requirements such as maximizing network revenue, minimizing bandwidth cost, and balancing traffic load, also need to be considered in admission control schemes.

For the admission control in the cellular/WLAN interworking systems, the simplest admission strategy is to have an unconditional preference to WLANs due to their much higher capacity and lower monetary cost [34]. Both the new calls and handoff calls within the overlapped region will be admitted into WLAN. One problem introduced by such strategy is the potential of unnecessary handoffs because mobile users with high mobility tend to spend a short time in the overlapped area and require an upward vertical handoff back to the cellular network shortly after a downward handoff to WLAN. An improved admission control scheme is designed in [80] by probabilistically rejecting vertical handoffs for highly mobile cellular users, where the vertical handoff processing load and blocking probability can be reduced while maintaining reasonable throughput in WLAN. This scheme is further developed in [81] based on the consideration that choosing WLAN with a fixed probability regardless of the traffic load in WLAN or the user's location, may not be beneficial for the cellular users in the dual-coverage area to connect to WLAN. The probability of migrating to WLAN is adaptively set in [81] based on the traffic load in WLAN and the location distribution of the cellular users.

The traffic type is considered as the main admission metric in [82-84]. In [84], an admission decision is made by comparing the amount of real-time traffic with a predetermined channel occupancy threshold as well as comparing the amount of non-real-time traffic with a buffer threshold. In [83], real-time voice calls are preferentially admitted into cellular networks for their ubiquitous coverage and fine QoS provisioning. The residual bandwidth unused by voice traffic in both the cellular and WLANs are then combined to effectively serve elastic data traffic for a large multiplexing gain. In [82], a randomized guard channel admission control policy is proposed in which a random number of channels are reserved for voice handoffs. Some channels are also randomly selected to be exclusively used by new voice connections. The remaining bandwidth is then shared by all data connections.

The admission control scheme proposed in [85] focuses on reducing the occurrence of hard handoffs from WLAN to UMTS on the real-time applications. Both the service

class and user's location are considered. A joint session admission control scheme for vertical handoff calls in integrated CDMA/WLAN networks is designed in [73], where both the WLAN and CDMA network conditions are used to formulate the admission control problem as a semi-Markov decision (SMDP). With a linear programming method, the admission decision is made according to the QoS state, but this approach does not consider the rapid state changes caused by the channel variation and user mobility. A fuzzy Q-learning admission control (FQAC) is introduced in [86] for the WCDMA/WLAN interworking networks, where various factors are taken into account including the QoS requirements, interferences from original cells and other adjacent cells, the numbers of real-time and non-real-time users in the system, and user's mobility.

2.4.3 Load balancing

In homogenous networks, the system typically needs to reject or block incoming calls if all of the bandwidth has been used up to provide the highest QoS to existing users. This operation leads to the high call dropping or blocking probability introduced in hot areas. Various approaches and algorithms have been proposed to reduce the blocking probability [87-89], but most of them are based on the idea of degrading the QoS of existing users to an acceptably lower level in order to free some bandwidth for the incoming calls. Although the system performance in terms of bandwidth utilization or network operator's revenue can be improved by these approaches, the resulted quality degradation on individual users is detrimental to QoS provisioning. The approach proposed in [149] can dynamically distribute traffic between different cells within the cellular system; however, additional devices named ad hoc relaying stations are required to relay the signals of mobile stations in congested cells to other cells with enough available bandwidth.

In heterogeneous interworking networks, an opportunity is provided to redistribute traffic among different networks to reduce congestion in hot areas. The traffic load is usually taken as an important metric in admission control algorithms [49, 90-92]. Taking [49] as an example, a cost-function-based network selection strategy is proposed which combines the traffic load and received signal strength in a cost function. An originating call prefers to be accepted by a network with lower traffic load and stronger

received signal strength in order to achieve better traffic balance among heterogeneous networks and good service quality. The work in [90] performs load balancing by setting a price to each access network and then dynamically adjusting their service price to distribute traffic load to the network with lower price. In [93], game theory is deployed based on a linear pricing model. The authors took a user-centric viewpoint and selected the wireless access network based on their utility which is a function of allocated bandwidth and price per connection. When a coverage area becomes congested, network operators in that area charge a higher price to gain more revenue and load balancing. In [94], a load-balancing index is introduced to reduce the utility value of a heavily loaded network, and its utility function consists of four evaluation terms including data rate, delay, packet-error rate, and mobility, respectively. The utility-based network selection algorithm designed in [95] targets to achieve load balancing between the UMTS networks and WLANs. The fundamental consideration of these schemes is to direct an incoming call into a network with lower traffic load to decrease the congestion probability. They lack the mechanisms to attempt to admit the call when its available networks have already been congested.

To fully utilize the hierarchical architecture provided by the heterogeneous interworking networks, the proposal of exploiting vertical handoffs to proactively balance the traffic load among networks and then make room for the incoming call in hot area is introduced in [71] and [20]. But no technical details are provided in [71]. With the solution designed in [20], when the utilization of a WLAN network exceeds or falls below the given thresholds, vertical handoff is initiated to transfer ongoing calls to or from the overlapping cellular network to keep the WLAN working in effective states. However, the solution takes the perspective of stand-alone WLANs instead of balancing the traffic load over the entire system.

2.5 Multicasting and broadcasting

Multimedia applications (e.g., video conferencing, video-on-demand, distributed interactive simulations like large multiplayer games, peer-to-peer file sharing, etc.) have become an important part of current and next generation network applications. The rapid and widespread deployment of broadband wireless networks has raised the expectation of real-time multimedia services in mobile environments. However, support-

ing bandwidth-intensive multimedia applications requires efficient management of network resources. When many users attempt to simultaneously receive the same multimedia content, even high-bandwidth wireless-link resources would be rapidly drained if a separate point-to-point channel is required for each user. In this context, group communication (including multicasting and broadcasting), which is an efficient mechanism for one-to-many transmissions over wireless links, has attracted great attention in recent years in both academia and industry.

The 3GPP working group proposed the multimedia broadcast multicast service (MBMS) for multicast multimedia content in UMTS [96]. Users may receive MBMS data via a dedicated logical channel (by executing the radio bearer setup procedure) or via a common logical channel (by simply tuning to the common logical channel). A common logical channel can serve more than one MBMS call at the same time. However, the use of the common logical channel may cause interference to the dedicated logical channels serving the traditional calls. To more efficiently utilize the radio resource to serve both unicast and MBMS calls, two channel allocation algorithms are proposed in [97]: reserved resource for multicasting and unreserved resource for multicasting. If the number of MBMS calls in the system is larger than an interference threshold, the unreserved resource for multicasting algorithm turns on the common logical channel to serve the new incoming calls.

Multicast in the WLAN standard IEEE 802.11 is not well specified; only addressing and basic operations are given. A multicast packet is transmitted without request-to-send (RTS)/clear-to-send (CTS)/acknowledgement (ACK) handshaking, which provides naturally an unreliable transmission. Therefore, some works on the multicast in WLANs focus on improving the reliability of the transmission including the hidden-node problem and error-recovery process [98, 99]. The hidden-node problem disrupts a contention-free multicast transmission, and the error-recovery process enables the sender to know whether a multicast packet has been successfully received by all of the member stations. The error-recovery process, which requires feedback from the receivers, causes the wireless system to be less bandwidth efficient due to the overhead introduced by feedback. To overcome this problem, the scheme proposed in [100] uses one OFDM symbol for the ACKs from all the group member stations, and each member station indicates its packet reception status by transmitting one of the two binary phase shift keying (BPSK) symbols on a previously assigned subcarrier within the

OFDM symbol.

WiMAX Forum has defined mechanisms for the delivery of multicast broadcast service (MBS) content over the Mobile WiMAX interface [101]. The MBS can be supported by either constructing a separate MBS zone in the downlink frame with unicast service, or the entire frame can be dedicated to MBS for broadcast service (downlink only). The flexible duration of MBS zones permits the scalable assignment of radio resources to MBS traffic. Reference [102] presents an MBS architecture that is based on location-management areas, which can increase the sizes of MBS zones to reduce the average handoff delay without too much bandwidth waste. In WiMAX, adaptive coding and modulation is supported. To ensure that all the users in a multicast group can receive the stream with the same quality, the stream is usually encoded by BS with the most robust modulation scheme corresponding to the user(s) of the worst link channel quality. It is known that the more robust modulation and coding scheme deployed, the lower the data rate is, and hence the more radio resource consumed by the same amount of data [139]. Therefore, the stations with bad channel conditions become the bottleneck of the multicast transmission. At this point, a cooperative multicast scheme is proposed in [103] to efficiently exploit spatial diversity among multiple users, based on a two-phase cooperative transmission model. In the first phase, BS multicasts data at a high rate; and the users in good channel conditions help relay the received data to the remaining users in the second phase.

So far, there are still very few works on the multicasting and broadcasting in heterogeneous wireless interworking networks. Within an integrated system, the members of a multicast or broadcast stream are distributed in multiple networks; meanwhile these members perceive different channel conditions in the available networks. Issues arise naturally including how to select a network for the multicast or broadcast stream and how many radio resources should be allocated in each selected network. Reference [12] is supposed to be the first piece of work that considers the network selection of general multicast streams in heterogeneous wireless networks. By adaptively selecting the cell and wireless technology for each mobile station to cluster more stations together, the overall bandwidth cost consumed by the stream can be minimized. The authors further extended the scheme to be used for the layer-encoded video multicasting [104] and data broadcast with network coding [105]. However, in these schemes, only the network selection is taken into account. All the mobile stations in one network have the

same bandwidth cost, which is not suitable for advanced wireless technologies where mobile stations with different channel conditions consume different bandwidth cost due to different modulation and coding schemes deployed. Meanwhile, mobile stations are involved in the network selection process and iteratively exchange bandwidth cost information with the network controller. The signaling overhead over wireless links and the implementation complexity on mobile devices further lower the feasibility of the schemes.

2.6 Summary

In this chapter, we first introduced current wireless access technologies briefly, including cellular system, WLAN and WiMAX. Then we reviewed and analyzed existing works on the interworking architecture, vertical handoff and resource management issues for the integration of these heterogeneous networks. We can see that the integration of cellular/WLAN system has been studied relatively well, but the integration of WiMAX/WLAN system still needs further investigations. Meanwhile, the benefits of exploiting vertical handoffs have not been fully explored by previous works. Although many admission control and handoff algorithms take traffic load as an important metric, they cannot proactively initiate vertical handoffs to make more room for the request call when congestion happens. Moreover, the schemes concentrating on the network selection as well as resource allocation for the multicasting and broadcasting calls in heterogeneous wireless interworking systems are not well reported.

Chapter 3

Interworking Architecture of Wi-MAX/WLAN

As described in Section 1.1, WiMAX and WLAN networks present complementary strength in terms of mobility support, coverage, service quality and implementation cost. Their interworking will allow WiMAX network operators to economically off-load data traffic to WLAN in indoor locations, hotspots, and other areas with a high user density on one hand; on the other hand, the customers will benefit from the enhanced performance through the greater coverage, better QoS support, and lower overall cost of such a combined service. Therefore, the integration of WiMAX and WLAN networks has been seen as a promising approach toward 4G. As reviewed in Chapter 2, existing research works for heterogeneous wireless networks mainly focus on the interworking of cellular/WLAN, instead of WiMAX/WLAN due to the relatively new WiMAX technology. To realize integration, the primary step is to design the interworking architecture, based on which the functions including mobility management, data transmission, QoS provisioning, and security can be designed. Therefore, followed by a brief introduction of the key technologies deployed by WiMAX and WLAN networks related to our work, we investigate extensively the integration architecture and mechanisms of WiMAX/WLAN overlay systems in this chapter.

3.1 Background

3.1.1 The WLAN technologies

For WLAN systems, the very basic form is called basic service set (BSS). There are two types of BSS: infrastructure BSS and independent BSS. The former is composed of an AP, which works as the interface between the wired infrastructure and the wireless link, and a number of stations that are connected by AP. The latter is, however, composed of a number of stations, which communicate directly. This type of BSS is also referred to as an ad hoc mode. Our study focuses on the infrastructure BSS.

IEEE 802.11 standards [106] define the MAC and PHY layer specifications of the air interface between AP and stations. The fundamental access method of the IEEE 802.11 MAC is DCF known as carrier-sense multiple access with collision avoidance (CSMA/CA), which is connectionless and contention-based. The carrier sense at the MAC layer can be achieved by advertising reservation information announcing the impending use of the medium. The exchange of RTS/CTS frames is one means of this medium reservation. Prior to the data transmission, the source station sends RTS and the destination station replies with CTS. Other stations that hear RTS/CTS will update their network allocation vector (NAV) to the busy state for a period of time given in the RTS/CTS frames. During this period, the station will defer its own data transmission.

3.1.2 The WiMAX technologies

The PHY and MAC layer specifications over the air interface of Mobile WiMAX to be considered in this thesis are defined by IEEE 802.16e [107]. At the PHY layer, OFDMA is deployed. In each OFDMA frame, different modulations can be used in different groups of subcarriers (subchannels) allocated to different users. This way, the available system resources can be utilized more efficiently. Both TDD and frequency division duplexing (FDD) operations are supported. For the TDD mode, the system transmits the data frame-by-frame, and each frame consists of a downlink subframe and an uplink subframe to prevent collisions between the downlink and uplink transmission. At the MAC layer, IEEE 802.16 defines access in two modes: point-to-multipoint (PMP) and mesh. A WiMAX mesh network usually suits for constructing wide-area wireless backhaul networks such as citywide wireless coverage. A PMP network is used for providing the last-mile access to mobile users. Therefore, in this thesis, we focus on the PMP mode which organizes nodes into a cellular-like structure

consisting of a BS and a number of stations.

Downlink transmission from BS to stations is broadcast, and uplink transmissions from stations to BS are unicast. BS acts as the central point to facilitate downlink and uplink access for a particular mobile station. BS allocates burst regions within a frame for a mobile station to send or receive data. A station knows these allocated regions by decoding the DL-MAP and UL-MAP messages at the beginning of each downlink sub-frame, where it can learn the burst region that is destined to it in the downlink, and to find out the region in the uplink assigned for it to place a transmission.

Beyond the air interface, the network specifications of WiMAX are developed by WiMAX Forum. The network architecture under development is built on two logical network entities, the access serving network (ASN) and the connectivity serving network (CSN) [108]. ASN represents a complete set of network functions required to provide link layer connectivity and local mobility over the air interface. ASN is decomposed into a set of BSs connected to a central control and gateway instance called the ASN-gateway (ASN-GW). CSN is a set of network functions that provide IP connectivity services to stations, which may further comprises network elements such as routers, AAA proxy/servers, home agent, and user databases as well as interworking gateways or enhanced network servers to support location-based services.

3.2 Interworking Architecture

We present a tightly coupled architecture and a loosely coupled architecture, respectively, for the WiMAX/WLAN interworking systems as shown in Figure 3.1. In the loosely coupled architecture, AP connects to Internet independently and the data paths in WLAN and WiMAX are completely separated. Under this approach, Mobile IP is required to maintain the active connections of the station after a vertical handoff. The data from or to the station will be relayed by its home network and a longer transmission delay is introduced as a result. Each interface of the station is configured with a particular user profile which is consistent with its own network components such as the AAA server. Complete authentication and key distribution operations should be performed during the initialization process of each interface. Moreover, the AAA server/proxy in the newly attached network needs to relay accounting messages for the station to the AAA server in the home network after a vertical handoff. But the independ-

Figure 3.1 The interworking architecture for WiMAX/WLAN overlay system.

ence between WiMAX and WLAN networks enables the loosely coupled system to be easily extended. Such approach is suitable for the case with little cooperation between the WiMAX and WLAN networks.

In the tightly coupled architecture, AP connects to ASN GW of WiMAX directly just as an access network of WiMAX. Several network components, such as the AAA server and dynamic host configuration protocol (DHCP) server, are shared by WiMAX and WLANs in this approach. In the implementation, there could be two possible scenarios, i.e., AP and BS connect to the same or different Mobile IP foreign agent. For the former scenario, the handoffs between BS and AP can be limited to link layer operations since the IP address of the station needs not to be changed; then seamless handoffs can be expected. Moreover, the data of the station will be transmitted directly

via the newly attached network after a handoff in this scenario. The previous serving network will no longer be involved in the transmission. For the latter scenario, the IP address of the station needs to be changed after a handoff, and thus Mobile IP is needed to maintain the connectivity after a vertical handoff. However, it should be noted that the vertical handoffs in such case can still be faster than those in the loosely coupled networks. Within a tightly coupled system, the Mobile IP functions are performed by the WiMAX network components including AAA servers and ASN GWs. There are existing data path and cooperation between these entities; then it is easier to deploy the Mobile IP enhancements mentioned in Section 2.3.3. Furthermore, the security related operations can be greatly expedited during handoffs in such case. The authentication and key distribution between the authentication server (AAA server) and the client (mobile station) can even be omitted because both of them are not changed by the handoff.

Usually, the overlay system is topologically small owing to the limited coverage of BS; and then the first scenario is expected to be deployed with higher popularity. Moreover, its support on seamless vertical handoffs well exhibits the advantage of tightly coupled systems. Therefore, we suggest the first scenario as the preferred tightly coupled network architecture in this thesis. The major disadvantage of such architecture is its limitation on the network size, i.e., the system cannot be expanded arbitrarily. It is however especially suitable for the case that one operator deploys the overlay system at the populous area such as the city centre and business district, where WLANs are placed at hotspots to offload traffic from their covering WiMAX.

As discussed in Chapter 2, the tightly coupled architecture can work in a more efficient way than the loosely coupled ones. Besides the seamless vertical handoff support, the core network components are shared by heterogeneous networks in a tightly coupled system leading to a lower implementation cost, meanwhile the system capacity can also be optimized by joint system resource management. The tightly coupled architecture is viewed as a promising approach for the current and future heterogeneous interworking systems. Due to the heterogeneity between the cellular networks and WLANs, the implementation complexity (e.g., the modification of cellular protocols and the introduction of gateway for WLANs) hinders the deployment of tightly coupled architectures in the cellular/WLAN interworking systems. However, we have known that both WiMAX and WLAN are defined within the IEEE 802 family which

defines various access and connectivity protocols for the packet-switched networking from the beginning. WiMAX and WLAN are hence homogeneous in nature. Moreover, the core of WiMAX under development is built on an all-IP architecture [109], which can provide enough capacity for the injected WLAN traffic. This enables WiMAX to work as backhaul of WLANs in some applications [110, 111]. Therefore, the tightly coupled architecture seems applicable for the WiMAX/WLAN integration systems, and thus deserves further research efforts. Although other approaches (e.g., [64]) have been designed based on the assumption of tightly coupled WiMAX/WLAN, the interworking architecture as well as the deployed mechanisms have not been well designed. In the following section, we will address the detailed mechanisms for this tightly coupled approach.

3.3 Mechanisms for the Tightly Coupled Interworking

In the tightly coupled interworking, two important problems should be solved. First, we need to find a technology to integrate APs into WiMAX ASN and establish the data path between AP and ASN GW. Second, the security architectures of both networks should be combined together in an appropriate way.

3.3.1 Data path function

It is known that WLAN was originally developed to work in a similar way to that of IEEE 802.3 Ethernet. In order to multiplex multiple layer-3 protocols such as IP and address resolution protocol (ARP), IEEE 802.2 logical link control (LLC) sits on the top of the MAC layer of IEEE 802.11. On the other hand, Ethernet is also supported by WiMAX [112]. An Ethernet-specific part of the packet convergence sublayer (Ethernet-CS) is designed within the IEEE 802.16e MAC layer, which supports the transport of Ethernet frames with IEEE 802.16e MAC frames and extends the classification capabilities of the CS to Ethernet-specific header fields. Moreover, the network specifications also defined Ethernet-related operations such as packet forwarding, broadcast filtering, Proxy-ARP, and so on.

Therefore, we suggest using Ethernet as the technique to combine WiMAX and WLAN networks. The protocol stack of the data plane is shown in Figure 3.2. In this case, the WLAN interface of the station works just as that in other usual cases, while IEEE 802.3 MAC sits on the 802.16 MAC of the WiMAX interface. Within ASN, IEEE 802.3 PHY is deployed as the lower carrier in this approach. If other technologies are deployed, this Ethernet combination method can still work by encapsulating Ethernet frames into the generic routing encapsulation (GRE) tunnels. Under this Ethernet-based architecture, transparent link layer connectivity can be provided. For example, if the active interface is WLAN and an IP packet is generated at sometime, the packet will be encapsulated into an LLC frame first and then into an 802.11 MAC frame which will be transmitted via the WLAN air interface to AP. Here, we suggest the destination address in the 802.11 MAC frame to be the MAC address of ASN GW. This way, AP only works as a simple bridge between the station and ASN GW. When the frame is received by AP, the frame header is taken off with the destination and source addresses copied into the corresponding fields of the newly attached 802.3 MAC header. The generated 802.3 frame is further transferred by the interface of AP that corresponds to the destination MAC address (i.e., the address of ASN GW).

Figure 3.2 An example of the protocol stack of data plane.

3.3.2 Security issues

The security aspects of WLAN are defined by the IEEE 802.11i standard [113], which utilizes IEEE 802.1X [114] for its authentication and key management services. There are three main components in the 802.1X authentication system: supplicant, authenticator, and authentication server. In a typical WLAN network, the supplicant is the station

and the authenticator is AP, while the AAA server works as an authentication server. The 802.1X standard leverages extensible authentication protocol (EAP) [115] to provide a number of authentication schemes. It also defines an encapsulation method to carry EAP messages between the supplicant and authenticator, that is, EAP over LAN (EAPOL). If the supplicant and authentication server authenticate each other successfully, both independently generate a pairwise master key (PMK). The authentication server then transmits the PMK to the authenticator. The four-way handshake then uses the PMK to derive and verify the keys between the supplicant and authenticator.

Besides, the security architecture in WiMAX networks combines IEEE 802.16e security services and the EAP-AAA framework. ASN GW typically works as the authenticator in the 3-party AAA. A privacy key management version 2 (PKMv2) protocol is defined by IEEE 802.16e to protect EAP messages between MS and BS. BS acts as an EAP proxy that relays EAP messages to the authenticator via an authentication relay protocol. After the success of user authentication, a master session key (MSK) will be established at both the station and AAA server and then transferred to the authenticator to produce other authentication or encryption keys.

Taking into account the similarities of WiMAX and WLAN security architectures such as the 3-party EAP-AAA framework and the hierarchical key distribution, a combined security architecture is presented in Figure 3.3, where the three parties involved in the user authentication is user, ASN GW, and AAA server. We make a slight modification to the normal WLAN network by moving the authenticator function from AP to ASN GW. Both AP and BS only work as an EAP proxy, and thus their operations are transparent to the AAA server.

Based on this architecture, both WiMAX and its overlapped WLANs deploy the

Figure 3.3 Combined authentication architecture.

same AAA components, and the authentication operations can be performed only once as a result. For instance, a user switches on the device and selects WiMAX interface to serve first. The authentication operations between the user and AAA server will be performed via the WiMAX network. If the WLAN interface is switched on after a while, then the user authentication can be omitted during the WLAN initialization process because the user has been authenticated by the AAA server. Moreover, the WLAN initialization may be further simplified by reusing the keys distributed in the WiMAX entry process, if the policy permits. Due to this simplified access operation, the vertical handoff process under the proposed tightly coupled architecture can be expedited.

3.4 Summary

In this chapter, we introduced a tightly coupled and a loosely coupled interworking architecture, respectively, for WiMAX/WLAN overlay networks. We recommended the tightly coupled approach as the preferred one due to its better performance on implementation cost, vertical handoff, and joint radio resource management. Further, we addressed key mechanisms in terms of data path function and security function, for realizing the tightly coupled architecture, based on which, a seamless vertical handoff scheme will be proposed in the next chapter.

Chapter 4

Vertical Handoff Scheme for Wi-MAX/WLAN

In the design of heterogeneous overlay systems, one of the most important issues is vertical handoff support. Different from traditional horizontal handoffs which are mainly triggered by mobility to maintain the connectivity, vertical handoff can be proactively initiated within overlapped areas to connect to the most suitable network based on the user's or network operator's requirements. As indicated in Section 2.3.4, the vertical handoff between WiMAX and WLAN networks has not been thoroughly investigated. In this chapter, a seamless and proactive vertical handoff scheme is designed based on the tightly coupled architecture introduced in Chapter 3. The objective is to obtain always the best service for both mobile and fixed users. It has been discussed in Section 2.3.2 that the *best* service or network could be evaluated from a number of aspects including monetary cost, network conditions, power consumption, user preference and security, etc. In this chapter, we choose the performance of applications and network conditions as the main metrics. To achieve proactive handoff, we further derive evaluation algorithms for stations to estimate the conditions of both WiMAX and WLAN networks in terms of available bandwidth and packet delay.

4.1 Proposed Vertical Handoff Solution

For vertical handoff schemes, “seamless” and “proactive” are two desirable features [45]. A seamless handoff denotes that the handoff execution is transparent to upper layer applications, which indeed depends on the interworking architecture of heteroge-

neous networks. Based on the tightly coupled architecture addressed in Chapter 3, a seamless handoff can be expected. Detailed handoff execution operations will be discussed in the following subsection. On the other hand, a proactive user-based vertical handoff requires the handoff process to be controlled by the stations. This way, if QoS metrics such as network available bandwidth and packet delay are considered in vertical handoffs, the stations should be able to detect network conditions for a handoff decision. Consequently, algorithms for network condition detection need to be tightly integrated into QoS-oriented vertical handoff schemes [35]. Therefore, in this study, we first design a vertical handoff manager (VHOM) for stations to control the whole handoff process and discuss details of the handoff operation. Network condition detection algorithms are then derived for stations to estimate the available bandwidth and packet delay of WiMAX and WLAN networks, respectively.

4.1.1 Vertical handoff manager (VHOM)

VHOM works on the MAC layers of both IEEE 802.11 and IEEE 802.16e interfaces of the station. Major functions of VHOM include traffic measurement, network status detection, handoff decision, and connection transition, which are performed by four modules as illustrated in Figure 4.1:

- Traffic measurement module (TMM) periodically measures the performance of active applications running at the station and reports to the handoff decision module (HDM). The measurement results will be used to determine whether to initiate a vertical handoff for better service.
- Network condition detection module (NCDM) periodically detects the availability of current connected network and reports to HDM. Also, NCDM may be initiated by HDM to estimate the conditions of networks during the handoff procedure.
- Handoff decision module (HDM) is the core component of VHOM. It gathers the information from other modules and manipulates their operations. It makes a decision to launch a handoff process and selects the target network based on the available handoff policies.
- Connection transition module (CTM) transfers all connections of the station to the target network selected by HDM once a handoff decision is made.

Figure 4.1 Components of VHOM.

4.1.2 Vertical handoff procedure

In the existing QoS-oriented vertical handoff approaches (e.g., [32, 35, 116]) for heterogeneous overlay networks, QoS metrics are considered in handoff decisions. The handoff procedures are normally initiated when the stations move across the border of WLANs. As a result, both the fixed stations and the mobile stations within overlapped areas cannot benefit from vertical handoffs. We highlight that a vertical handoff could be initiated by two factors: mobility when a station moves out of the coverage of its connected WLAN, and QoS when the connected network cannot satisfy the requirements. As the mobility-triggered handoff has been well investigated in literatures, we focus on the operations of QoS-triggered vertical handoffs in the following subsection.

1) Service evaluation and handoff initiation

Once an application is established at the station, the TMM module will detect each packet of the application. Based on the delay sensitivity, the applications are classified into real-time and non-real-time applications, respectively. Since a real-time application is sensitive to latency, both the throughput and packet delay of the traffic should be measured. For a non-real-time application, the amount of transmission data is more important, and thus only the traffic throughput is measured.

Here, the transmission direction of the application should be taken into account.

For an uplink application, TMM can record the moment that the packet arrives at the MAC layer buffer of the station and that the packet is successfully transmitted by the connected network. Therefore, the calculated uplink traffic throughput and packet delay can well reflect the performance of the local connected network (WiMAX or WLAN). If an uplink application continually violates the QoS requirements for a given period, the handoff decision process will be started.

For the downlink traffic, however, the station cannot obtain the time information of the packet arriving at BS or AP. In this case, an end-to-end packet delay will be calculated rather than the delay purely introduced by the locally connected network (various solutions have been recommended in literatures to measure the end-to-end packet delay, e.g., [117-119]). If downlink traffic throughput or packet delay violates the QoS requirements, the poor performance may be introduced by the locally connected network or by other networks on the path between the two end nodes of this application. To avoid performing an unnecessary vertical handoff within the local network, the HDM module will inform NCDM to evaluate the conditions of the locally connected network first in this case. Only when the local connected network is proved to be working in a bad condition, the following handoff decision process can be started.

2) *Network condition detection and handoff decision*

In this phase, a decision of whether to perform a vertical handoff will be made by the HDM module. The main work is to decide whether the conditions of the other network that is not serving the station can satisfy the QoS requirements. The flow chart is shown in Figure 4.2 with the used parameters listed in Table 4.1.

Table 4.1 Parameters used in the algorithm.

APP	current application
BW _d /BW _u /BW	estimated available bandwidth of downlink WiMAX/uplink WiMAX/WLAN
Thr _{d_T} / Thr _{u_T} / Thr _{_T}	threshold for the available bandwidth of downlink WiMAX/uplink WiMAX/WLAN
t _d /t _u /t	estimated average packet delay of downlink WiMAX/uplink WiMAX/WLAN
Pd _{d_T} / Pd _{u_T} / Pd _{_T}	threshold for the average packet delay of downlink WiMAX/uplink WiMAX/WLAN

Figure 4.2 The QoS-triggered vertical handoff decision algorithm.

To evaluate the conditions of a network, the first step is to find the utilization of the medium, and then calculate the available bandwidth. If a real-time application is running on the station, the average packet delay of the network needs to be further evaluated. For WLANs, the utilization information can be obtained by collecting the NAV information. For WiMAX network, this information can be obtained by aggregating the number of allocated slots in the DL-MAP and UL-MAP messages broadcast by BS. We have designed novel algorithms for estimating network conditions based on the collected utilization information, which will be described with technical details in the following section. If the network evaluated by NCDM is WLAN, the network conditions are estimated based on the total radio resource because the medium of WLAN is contended by all stations including AP. In the WiMAX network, however, the radio resource has been allocated into a downlink portion and an uplink portion, which allows the network conditions to be evaluated for the downlink and uplink separately. Therefore, only when both the downlink and uplink network conditions satisfy the requirements, a decision of handoff to WiMAX can be made.

To make an effective handoff, the conditions of the target network must be good enough. It is required by the proposed scheme that the thresholds used in making vertical handoff decisions (e.g., Thr_{d_T} , Thr_{u_T} , Thr_T , Pd_{d_T} , Pd_{u_T} , Pd_T) are hysteresis-based. That is, the threshold should be better than the requirement of application for a

hysteresis value. For example, Thr_T should be the sum of aggregated throughput of applications and a predefined hysteresis, and Pd_T should be the lowest packet delay required by the real-time applications deducted by a hysteresis.

3) *Handoff execution*

Our scheme is based on the tightly coupled architecture designed in Chapter 3, where AP connects to the central gateway called ASN GW of WiMAX, just as its overlay BS. AP belongs to the same subnetwork at the IP layer with its overlay BS. As a result, the IP address of the station needs not to be changed after a vertical handoff, which makes a MAC layer handoff possible.

Based on this consideration, we deploy an ARP [120] method to execute vertical handoffs. Once a decision of handoff is made by the HDM module, the CTM module will be informed to issue a gratuitous ARP message and then transmit it through the target interface at the station. The message conveys the IP address of the station and the MAC address of the target interface. AP or BS in the target network relays the message to ASN GW, which then updates its ARP cache by binding the IP address with the MAC address contained in the message. Then, ASN GW issues an ARP reply message to the station. Hereafter, the data packets destined to the station will be transferred by ASN GW via the newly selected network.

It can be found that the handoff in our scheme is executed by only a round message trip from the station to ASN GW. The first gratuitous ARP message is used to notify ASN GW the handoff, which is necessary because ASN GW needs to redirect the packets destined to the station through the target network. The second ARP reply message is an indispensable acknowledgement for the station to start sending its packets via the newly selected network. Therefore, we could say that such execution minimizes the handoff signaling overhead. It is much faster than the Mobile IP based methods which introduce signaling overhead to the home network. Moreover, the whole handoff operations by our approach are performed at the MAC layer without the awareness of upper layers; thus, a seamless handoff is achieved.

4.1.3 Performance analysis and implementation details

- Packet loss during handoff process

For traditional horizontal handoffs, the performance of handoff solutions are usually

evaluated by packet loss, handoff latency, and signaling overhead. As discussed in the above section, a fast handoff execution can be achieved by our ARP-based solution with a very low signaling overhead. Moreover, the QoS-oriented vertical handoffs happen within the overlapped areas, where the two interfaces of the station can connect to both networks simultaneously. This is the trend of future multimode devices as discussed in Section 2.3. During the vertical handoff process, the inactive interface can monitor the conditions of the other network while the active interface keeps the current communication [7]. Therefore, the packet loss and handoff latency in this overlay network are not as critical as those in traditional handoff cases.

- Power consumption

Keeping two interfaces active simultaneously, the power consumption cannot be ignored especially for the battery-powered stations. In our design, the WiMAX interface of the station will switch to the sleep mode if not in service. Meanwhile, for the WLAN interface, it typically performs scanning operations continuously to connect to an AP. Owing to the small coverage of WLANs, such continuous scanning causes an unnecessary use of energy. Therefore, keeping the 802.11 interface turned on should be avoided by vertical handoff schemes. In existing works, various solutions have been addressed. The most intuitive way is to periodically turning on this interface [35], while which is known to be not effective [145]. More intelligent solutions are provided by obtaining the location information of AP from a server, GPS system [151], or BS [61, 68], but usually at the expense of increasing the device cost or modifying the specifications.

To this end, it is required by our scheme that the WLAN interface should be turned off after performing a mobility- or QoS-triggered handoff to WiMAX. It will be turned on when the WiMAX network cannot provide satisfied service and then NCDM needs to detect the conditions of WLAN. In other words, the WLAN interface is turned on only when it is needed by the station. Once it is switched on, the station will transmit a Probe Request frame to AP proactively rather than listening to the beacons broadcast by AP which is known to take longer time. If Probe Response frames are not received, the WLAN interface will be turned off again. Basically, the time interval to another scan trial depends on the performance of current applications and their desirability on a vertical handoff. If the network conditions are degraded seriously and quickly, the time

interval may be decreased dynamically in order to leave the network as soon as possible. Also, a shorter time interval may be set for a real-time application than a non-real-time application because the former has a higher sensitivity on the network performance. Moreover, if conditions are permitted, BS could periodically broadcast the density of APs within its coverage to help stations to adjust scan intervals [68]. Once an AP is found, the WLAN initialization process can be speeded up based on the proposed tightly coupled architecture, because the authentication and even the key distribution processes are simplified.

- Ping-pong effect

To avoid the ping-pong effect, the typical method of dwell timer plus hysteresis is deployed in our scheme. As discussed in the previous subsection, the thresholds used in making vertical handoff decisions (e.g., Thr_{d_T} , Thr_{u_T} , Thr_T , Pd_{d_T} , Pd_{u_T} , Pd_T) are required to be hysteresis-based to guarantee the target network conditions to be “sufficiently good”. On the other hand, the dwell time (Γ) for detecting the conditions of the other network is randomly taken from $(\Gamma_{\min}, \Gamma_{\max})$ with an aim to avoid more than one station handing over to the target network simultaneously. Since the handoff delay is not a critical issue as that in traditional handoffs mentioned above, a longer dwell time could be used to guarantee the conditions of the other network to be “continuously good”.

- Implementation complexity

In our design, the handoff process is completely controlled and executed by the stations. Both BS and AP only work as “transparent pipes”. ASN GW only needs to send an ARP reply message when it receives a gratuitous ARP message from the station during the handoff execution procedure. Therefore, our vertical handoff solution introduces a very low implementation cost to the network components.

4.2 Proposed Network Condition Estimation Algorithms

For multimedia applications over the network, available bandwidth and packet delay are two most important parameters for evaluating the performance of the network. The

available bandwidth of a link equals the difference between the total capacity and the traffic load over the link [121]. Usually, when the modulation and coding methods are known, the total capacity can be calculated. Therefore, the key idea of estimation is to find the utilization information of the link. In this section, we develop our scheme to estimate the utilization of WiMAX and WLAN networks, and further derive the available bandwidth and the packet delay in the networks based on this information.

4.2.1 Estimation in WiMAX

Defined by the IEEE 802.16e standard, the OFDMA technique is used at the PHY layer. The bandwidth is allocated in the form of data bursts, where each burst constitutes an integer number of slots. By the OFDMA technique, a slot is the minimum data allocation unit and defined in both time (symbol) and frequency (subchannel) dimensions. Both TDD and FDD are supported by IEEE 802.16e. As TDD is more popular for data and video services [122], we select TDD as an example in the following algorithms. An OFDMA-TDD frame structure is shown in Figure 4.3. In the standard, an adaptive split between downlink and uplink subframes is allowed. However, it is usually fixed or remained unchanged for a long period in practical applications. Based on the statistical results on traffic distribution, IEEE 802.16 recommended 2:1 as the default downlink to uplink subframe time ratio while WiMAX Forum recommended the ratio to be 27:15 [123, 124]. Hence, we take the fixed split case as an example for analysis. However, our algorithms can be extended to be used in FDD as well as adaptive split TDD

Figure 4.3 A sample of OFDMA-TDD frame structure.

operations.

4.2.1.1 Available bandwidth estimation

At the MAC layer, a frame is divided into a downlink subframe and an uplink subframe. The BS determines the number of downlink and uplink slots that a station obtains in one frame, and then broadcasts the resource allocation results through the information elements (IEs) in DL-MAP and UL-MAP messages at the beginning of each downlink subframe. Therefore, the station can easily obtain the utilization of the WiMAX link by aggregating the number of allocated slots stated in DL-MAP/UL-MAP messages.

Let AAS_d and AAS_u denote the number of allocated downlink and uplink slots in one frame, respectively, which are averaged over f frames. T_f , T_{df} and T_{uf} denote the duration of a frame, a downlink subframe, and an uplink subframe, respectively. Then, the available bandwidth in downlink and uplink can be calculated by

$$\begin{cases} B_d = \left(1 - \frac{AAS_d}{s_d}\right) \frac{\delta_d s_d}{T_f} \\ B_u = \left(1 - \frac{AAS_u}{s_u}\right) \frac{\delta_u s_u}{T_f} \end{cases}, \quad (4.1)$$

where s_d (s_u) denotes the total slots in a downlink (uplink) subframe and δ_d (δ_u) is the number of bits transmitted in a downlink (uplink) slot which depends on the modulation and coding schemes deployed.

4.2.1.2 Packet delay estimation

The packet delay is defined as the latency for a packet to experience in the network, which can be calculated from the time that the packet arrives at the MAC layer of a station to the time that the packet is successfully transmitted. We divide the total delay into four components for analysis, namely, scheduling delay, queueing delay, mapping delay and transmission delay.

$$t = t_s + t_q + t_m + t_t. \quad (4.2)$$

t_s : the scheduling delay t_s is taken from the moment the packet arrives at the MAC layer buffer, to the moment this arrival information is obtained by BS.

- t_q : the queueing delay t_q is the time awaiting for the beginning of the frame allocated for transmitting the packet.
- t_m : the mapping delay t_m is taken from the beginning of the allocated frame to the first time slot appointed to the station.
- t_t : the transmission delay t_t is the time required to transmit the packet.

- Uplink packet delay estimation

IEEE 802.16e provides a number of ways for stations to request bandwidth from BS. Here, we take a typical example for analysis as shown in Figure 4.4, where each station can obtain an opportunity to send Bandwidth Request (BWR) message to BS for its newly arrived packets. When a packet arrives at the MAC layer, the station puts it into the buffer first. In the next frame, the station sends a BWR message in its opportunity to request transmission bandwidth for the packet. After that, the station waits for the UL-MAP message broadcast by BS in the following frames to obtain the transmission opportunity allocated to it. Once such an opportunity is found, the station transmits the packet in the appointed slots within the indicated frame. In the analysis, it is assumed that the packet arrival follows the Poisson distribution, which is a valid assumption because it has given sufficiently accurate results for the WiMAX networks [125-127].

Figure 4.4 An example of uplink packet transmission.

1) Calculation of t_s

The scheduling delay t_s in the uplink consists of two parts. The first part is the delay incurred by the arrival till the start of the next frame. Here, the header of the uplink subframe is seen as the beginning of the total frame. It is known that Poisson arrivals occur completely randomly in time [128], and then the first part of t_s can be approximated by $0.5T_f$. The second part is just the duration of the next frame, where the BWR message is transmitted. Then, we write the scheduling delay t_s as

$$t_s = 0.5T_f + T_f = 1.5T_f. \quad (4.3)$$

2) Calculation of t_q

After receiving the BWR message, BS has known the arrival of the packet at the station. The packet can be imagined to enter a common queue at BS, which is shared by uplink packets from all stations. If the next frame after sending the BWR message is granted to this packet, the queueing delay t_q will be zero. Otherwise, it will be increased by frame duration till one frame is granted.

Here, we model this common queue by the $D^X/D^m/1$ queueing system [129, 130]. The $D^X/D^m/1$ specifies a discrete-time batch arrival bulk service queueing model, where the time axis is divided into fixed-length time intervals. During one time interval, customers arrive in batches with a random size meanwhile multiple customers are served by a single server at a time. In the WiMAX system, the packets of users are the customers of this common queue, which are served by BS in bulk during one frame duration T_f .

The explicit formula for the equilibrium queue length of $D^X/D^m/1$ system is

$$E(N_q) = \sum_{k=1}^{m-1} \frac{1}{1-z_k} + \frac{A''(1) - m(m-1)}{2(m-\mu_A)}, \quad (4.4)$$

where $A(z)$ is the probability generating function of the newly arrival batch size, μ_A is the mean arrival batch size, m is the maximum number of customers that can be served during one serving period [129]. Since the packet arrival at a station follows the Poisson distribution, the packet arrival at the common queue of BS also follows the Poisson distribution based on the merging property. We use λ (packets/frame) to denote the mean packet arrival rate at the common queue. Then, we have $\mu_A = \lambda$, and $A(z)$ can be expressed as

$$A(z) = e^{\lambda(z-1)}, \quad (4.5)$$

where $z_1, \dots, z_{m-1}$ together with $z_0 = 1$ are the m roots of $z^m = A(z)$ in $|z| \leq 1$ [130]. An explicit expression for these roots is given by

$$z_k = \sum_{l=1}^{\infty} c_l w_k^l, \quad k = 0, 1, \dots, m-1, \quad (4.6)$$

where $w_k = e^{2\pi ki/m}$. For Poisson traffic, c_l can be expressed by the following expression

$$c_l = e^{-l\rho} \frac{(l\rho)^{l-1}}{l!}, \quad (4.7)$$

where $\rho = \lambda/m$ is the utilization of the queue server with $0 \leq \rho < 1$.

In our scheme, the number of packets that can be served in one uplink subframe is

$$m = \frac{\delta_u s_u}{L}, \quad (4.8)$$

where L is the average packet length at the station. After the station accesses the network, the radio resource will be utilized by it and other existing stations. Therefore, the utilization ρ will be expressed as

$$\rho = \frac{AAS_u}{s_u} + \frac{B}{\delta_u s_u / T_f}, \quad (4.9)$$

where B is the expected bandwidth of the desired station. Subsequently, we can get $\lambda = \rho m$. By combining (4.5) - (4.9), we can solve $E(N_q)$ for a given AAS_u . Finally, from the Little's theorem, we can obtain the queueing delay t_q .

$$t_q = \frac{E(N_q)T_f}{\lambda}. \quad (4.10)$$

It should be noted that (4.6) can be approximated in applications by a finite number of summations with result sufficiently close to the real value. For example, the iteration was suggested to stop when $|z_k^{l+1} - z_k^l| < 10^{-14}$ [130]. Assuming n to be the maximum number of summations for z_k ($0 \leq k \leq m-1$), the time complexity for calculating the total delay has the order of $O(n^2)$.

3) Calculation of t_m and t_t

At the uplink, all data bursts are allocated as horizontal stripes. The transmission starts at a particular slot and continues until the end of the uplink subframe. The resource allocation continues on the next subchannel horizontally as shown in Figure 4.3. Therefore, if the packet length L is larger than σ which is the number of bits that can be transmitted in one uplink subchannel, the transmission will definitely use the time slots at the head of the next allocated subchannel. In this case, the mapping delay t_m will be

zero and the transmission delay t_t will be the total uplink subframe duration T_{uf} . Then, the sum of t_m and t_t in this case is

$$t_m + t_t = 0 + T_{uf} = T_{uf}. \quad (4.11)$$

Conversely, if L is less than σ , the transmission will start from the end of the previous burst. Here, there may be two scenarios. In scenario S1, the residual slots of that subchannel are adequate, and then the data transmission will only use that subchannel. Otherwise, the data transmission will be continued to the next subchannel in scenario S2. Let P_{s1} and P_{s2} denote the probability of scenario S1 and S2, respectively. We can calculate the mean of the sum of t_m and t_t by the following expression

$$\begin{aligned} t_m + t_t &= E[t_m + t_t | S1] \cdot P_{s1} + E[t_m + t_t | S2] \cdot P_{s2} \\ &= \frac{1}{2} \left(\frac{L}{\sigma} + 1 \right) T_{uf} \cdot \left(1 - \frac{L}{\sigma} \right) + T_{uf} \cdot \frac{L}{\sigma} \\ &= \left[\frac{1}{2} + \frac{L}{\sigma} - \frac{1}{2} \left(\frac{L}{\sigma} \right)^2 \right] T_{uf}. \end{aligned} \quad (4.12)$$

Totally, we have

$$t_m + t_t = \begin{cases} T_{uf}, & L \geq \sigma \\ \left[\frac{1}{2} + \frac{L}{\sigma} - \frac{1}{2} \left(\frac{L}{\sigma} \right)^2 \right] T_{uf}, & \text{otherwise} \end{cases} \quad (4.13)$$

- Downlink packet delay estimation

In the downlink transmission, BS has complete knowledge about the traffic. There is no need to send BWR messages, and then the scheduling delay t_s will be $0.5T_f$. The calculation of the downlink queueing delay t_q is similar to that of the uplink. It only needs to change the uplink specific parameters including δ_u , s_u and AAS_u to the corresponding downlink parameters δ_d , s_d and AAS_d .

The mapping methods used in downlink are quite different from those in uplink. The standard requires all downlink data bursts to be rectangular. According to the scanning order, current mapping algorithms can be divided into two types. One type is scanning and assigning slots from symbol to subchannel and the other is from subchannel to symbol [131]. Here, we take the second type as an example. The average

mapping delay can be written as

$$t_m = T_p + \frac{1}{2} \rho (T_{df} - T_p), \quad (4.14)$$

where T_p denotes the time duration of the preamble, and ρ is utilization of the downlink resource which can be estimated just as (4.9) in the uplink case.

The transmission delay t_t in downlink depends on the burst size and the number of bits transmitted by all subchannels in one slot duration, $\mathfrak{G}$. It is supposed that one burst contains one packet. If the packet size L is larger than $\mathfrak{G}$, the packet will occupy the subsequent time slots and be packed to a rectangular according to the packet length. Otherwise, it will only occupy one slot duration. Then, the transmission delay can be approximated as

$$t_t = \begin{cases} T_s, & L \leq \mathfrak{G} \\ \left\lceil \frac{L}{\mathfrak{G}} \right\rceil T_s, & \text{otherwise} \end{cases}, \quad (4.15)$$

where T_s is the duration of one downlink time slot.

4.2.2 Estimation in WLAN

As described in Section 3.1.1, the MAC layer of WLAN deploys CSMA/CA which is connectionless and contention-based. NAV is the main scheme used to avoid collision by setting a busy duration on hearing frame transmissions from other stations. Hence, the utilization of the WLAN channel is well reflected by NAV.

Our work evolves from the work in [35], where the available bandwidth has been derived from NAV as

$$BW = B_0 - L \frac{NAV}{T_n + \frac{1}{2} T_{n,c} (N - 1)}, \quad (4.16)$$

where B_0 is the total system bandwidth, L is the mean frame size, T_n is the NAV duration for a successful frame transmission, $T_{n,c}$ is the NAV duration for a collision and N is the average number of trials of a transmission. In [35], a mean access delay has also

been derived as

$$\begin{aligned}
t_a = & \frac{b(p_c T_{col} + (1 - p_c) T_s)}{2} + b \frac{v(CW_{\min})}{2} \\
& + p_0 T_s \sum_{j=1}^{S-1} p_1^{j-1} + T_{col} \sum_{j=1}^{S-1} p_1^j + \sum_{j=1}^{S-1} p_1^j v(2^j CW_{\min}). \tag{4.17}
\end{aligned}$$

All the variables in (4.16) and (4.17) can be estimated by NAV. However, this access delay is defined as the time used from the moment when a packet becomes the head of the queue at the MAC layer to the moment when this frame is successfully transmitted. It is not the total delay since the time spent in the queue has not been included.

Here, we further extend the model based on the above expressions. Each station is regarded as a queueing system. The total delay experienced by a packet consists of two parts: the queueing delay and the service time. We notice that the access delay obtained in (4.17) is just the service time in this model, and thus we can estimate the queueing delay from this access delay with a precondition that the distribution of the service time is known. The authors in [132] has investigated the queueing characteristics of WLAN and concluded that the service time distribution can be well approximated by the exponential distribution, and M/M/1 provides a good approximation for the queueing system especially when the station works in non-saturated states, which are usually encountered in real implementations. Thus, we model this queueing system as M/M/1. With the mean packet arrival rate λ at the station, the queueing delay t_q can be obtained from the service time t_a by the following expression

$$t_q = \frac{\lambda t_a^2}{1 - \lambda t_a}. \tag{4.18}$$

Finally, the total delay can be expressed as

$$t = t_q + t_a = \frac{\lambda t_a^2}{1 - \lambda t_a} + t_a. \tag{4.19}$$

4.3 Performance Evaluation

To demonstrate the effectiveness and feasibility of our proposed approach, we have

built a simulation model by the simulation software QualNet 4.0. IEEE 802.11b and IEEE 802.16e are used as the WLAN and WiMAX standard, respectively. The channel bandwidth of WiMAX is 20 MHz. The default downlink to uplink time ratio in WiMAX is 2:1, with a default frame duration of 5 ms.

4.3.1 Evaluation of network condition estimation algorithms

- Available bandwidth estimation in WiMAX

We test the downlink and uplink available bandwidth in different data rate cases, which are determined by the modulation schemes and given by $\delta_d s_d / T_f$ and $\delta_u s_u / T_f$, respectively. The results estimated by the proposed algorithms have been compared with the cor-

(a)

(b)

Figure 4.5 Available bandwidth in WiMAX: (a) downlink; (b) uplink.

responding simulation results in Figure 4.5, for downlink and uplink, respectively. The AAS_d (AAS_u) occupation denotes the utilization of the medium, which equals AAS_d/s_d (AAS_u/s_u). It is shown in the figure that the simulated results and the estimated results match very well for both the downlink and uplink transmissions.

- Packet delay estimation in WiMAX

From the analysis in Section 4.2.1.2, we can see that the estimated packet delay is determined by several parameters such as the frame duration and the data rate. In the simulation, a downlink or uplink application runs on the desired station with a constant packet size of 500 bytes. The expected throughput is 800 kbps. For the downlink, the results are tested for different frame duration cases. 5 ms and 8 ms are selected as ex-

(a)

(b)

Figure 4.6 Mean packet delay in WiMAX: (a) downlink; (b) uplink.

amples of frame duration. The most robust modulation type is deployed that has a maximum available bandwidth of about 8.9 Mbps. There are 60 subchannels in the system and the number of bits transmitted by all subchannels of one time slot (Θ) is 2880 bits. The downlink time slot duration T_s is 0.2 ms.

For the uplink, the results are tested for different data rates. The default frame duration is 5 ms. The maximum available bandwidth for these two cases is 3.84 Mbps and 7.52 Mbps, respectively. The number of bits that could be transmitted by all time slots in one subchannel (σ) is 360 bits and 720 bits for two respective cases.

The results estimated by the proposed algorithm are compared with corresponding simulated results in Figure 4.6. It can be seen that the downlink delay increases progressively with the increase of AAS_d occupation. This is indicated in (4.14) that the mapping delay increases with the utilization. In addition, we can see that the downlink delay is much lower than the uplink delay of the same frame duration (5 ms). This is caused by the fact that the station needs to spend an extra frame to send the BWR message at uplink. Moreover, the resource allocation extends from the time domain to the frequency domain in uplink, while the case in downlink is reverse.

- Estimation in WLAN

The available bandwidth and packet delay of WLAN network estimated by our algorithms are compared with corresponding simulation results in Figure 4.7. The results are shown for different data rate type cases. The packet size of the desired station is 1000 bytes and the packet arrival rate λ is 50 (packets/second). In the packet delay estimation, we found that the slight difference between estimated and simulated results was mainly caused by t_a when the network was non-saturated, and then caused by both t_a and t_q when the network tended to be congested.

It is worth noting here that the computational complexity of the proposed estimation algorithms is very manageable. Our experiments were run on a 2.67 GHz processor with 2 GB memory. The calculation of available bandwidth in WiMAX is so simple that the time cost can even be ignored. Generally, the execution time taken for the calculation of packet delay in WiMAX and available bandwidth in WLAN was no more than 4 ms. The calculation of t_a was relatively complex which made the calculation of WLAN packet delay take about 10 ms. However, it was still small enough not to be concerned during a handoff process. Moreover, the execution time can be further

(a)

(b)

Figure 4.7 WLAN estimation: (a) available bandwidth; (b) mean packet delay.

decreased in the implementation by lower level programming languages.

4.3.2 Evaluation of the proposed vertical handoff scheme

On the proposed vertical handoff scheme, we have conducted simulations in an inter-working system which consists of one WiMAX network and two overlapped WLANs, as illustrated in Figure 4.8. We have compared our proposed VHOM scheme with two other reference schemes. The first one is a QoS-oriented vertical handoff scheme designed in [32] for the WLAN based overlay networks, where the QoS provided by a WLAN is considered in vertical handoff decisions when a mobile station moves across it. The second one is a WLAN-preferred scheme that is widely deployed in cellu-

Figure 4.8 Simulation topology of WiMAX/WLAN interworking system.

lar/WLAN systems as discussed in Section 2.3.1. As long as the station enters the coverage of a WLAN, a vertical handoff to the WLAN will be performed immediately. The performances of the schemes have been tested for both mobile and fixed stations.

- For mobile stations

We have performed simulations for mobile stations in two scenarios. In the first scenario, a mobile station M1 moves towards an overlapped region as shown in Figure 4.8. There is an uplink connection running at M1 which is a non-real-time variable bit rate (VBR) application with an expected throughput of 800 kbps. Initially, VHOM found that the average throughput of the application fell below the acceptable 600 kbps. However, there was no other network available at that time. Since both the reference schemes trigger handoffs only by mobility, the performance of applications was not detected by them periodically. At the moment of T1, a WLAN network was found, and then a handoff was executed by the second reference scheme immediately. By our scheme, VHOM initiated the available bandwidth estimation process first. To guarantee the conditions of WLAN being sufficiently good, Thr_T is set to be 900 kbps. The estimated available bandwidth could not satisfy this requirement until T2, and then a handoff was performed at T2 as shown in Figure 4.9 (a). By the first reference scheme, the station did not execute handoff at T1 either due to the bad performance of WLAN.

In the second scenario, a mobile station M2 was initially served by the WLAN with the same application of M1. At the moment of T3, VHOM found that the throughput fell below the acceptable 600 kbps for three times, and then a handoff was executed because the estimated results showed that WiMAX was operating in a good condition

(a)

(b)

Figure 4.9 Throughput comparison in the movement scenario.

($\text{Thr}_{u,T} = 900$ kbps). It is shown in Figure 4.9 (b) that a handoff was executed at T4 by the second reference scheme when the station moved out of the border of WLAN. The first reference scheme executed the handoff a little earlier because a higher exit SNR threshold was set due to the bad conditions of WLAN. However, this scheme makes effects only around the border of WLANs. When the station stays inside the WLAN (before 80 s), the handoff cannot be initiated to proactively improve the QoS of applications. The simulation results show that the best service can always be achieved by the proposed VHOM scheme.

- For fixed stations

(a)

(b)

Figure 4.10. Simulation results for the fixed scenario:
(a) throughput; (b) end-to-end packet delay.

In the simulation, a station F located in an overlapped region without movements, which worked with a downlink real-time constant bit rate (CBR) application. The expected throughput is 500 kbps with a threshold of 450 kbps. The accepted end-to-end packet delay is 20 ms. Initially, the station was served by the WLAN. At around 27 s, the average end-to-end packet delay detected by VHOM could not fulfill the requirement for three times although the throughput was fairly well. By the VHOM scheme, the condition of the WLAN was evaluated first. The result showed that the latency introduced by the WLAN operations exceeded the 8 ms, which is the threshold. Then, VHOM began to estimate the status of the WiMAX network. At about 30 s, a decision of vertical handoff was made based on the fact that the downlink conditions of the

WiMAX network could satisfy the requirements ($Pd_{d_T} = 8$ ms). The throughput and end-to-end packet delay obtained by the schemes have been compared and illustrated in Figure 4.10. Both the reference schemes could not initiate handoffs for fixed stations. The results show that a much lower packet delay and more stable throughput can be obtained by our VHOM scheme.

4.4 Summary

In this chapter, we investigated the vertical handoff issue for the WiMAX/WLAN overlay networks. Based on the tightly coupled interworking architecture addressed in Chapter 3, we proposed a novel seamless and proactive VHOM scheme for stations to control the vertical handoff operations in the WiMAX/WLAN interworking networks. VHOM aims to provide always the best service for both mobile and fixed users. The performance of applications and network conditions are considered in the handoff process. In order to make the stations to be able to proactively evaluate network conditions for making handoff decisions, we develop algorithms to estimate the available bandwidth and packet delay in WiMAX and WLAN networks, respectively. These algorithms can also be deployed by other approaches. By the simulation experiments, we have verified that our proposed scheme can provide better service than other reference schemes.

Chapter 5

Proactive Load Balancing for Heterogeneous Overlay Networks

In Chapter 4, we propose a vertical handoff scheme for WiMAX/WLAN overlay networks. Indeed, it is also a resource management scheme from the perspective of individual users. In this chapter, we study the radio resource management for general heterogeneous overlay networks from the viewpoint of network operators.

In wireless systems, efficient utilization of radio resources is of great importance due to the limited bandwidth, especially when some parts of the system tend to be *congested* or *hot*. If the call is simply rejected or blocked due to the lack of bandwidth, high call dropping or blocking probability is introduced in hot areas with cold areas left underutilized, which is an undesired condition for network operators. On the other hand, from the algorithms derived in Section 4.2, we can see that the utilization of the network directly reflects the QoS provided. Keeping the traffic load over the network within an effective region will help to provide an estimated QoS for users. Therefore, load balancing is very important for wireless systems to achieve a better utilization of the overall radio resources and provide a good service support.

As reviewed in Section 2.4.3, many works of homogeneous systems reduce the call dropping probability in hot areas by degrading the QoS of ongoing calls to free more bandwidth for incoming calls. Although the system resource utilization is improved, the QoS provisioning to individual users becomes a crucial problem. In [90-92], load balancing has been taken into account in the call admission control process for heterogeneous interworking systems. However, by these approaches, if none of available networks can provide enough radio resource, the request call will still be blocked or

dropped (similar to legacy approaches of homogeneous systems). In heterogeneous interworking systems, the radio resources of all networks within the system can be viewed as being shared in a resource pool. Only considering the radio resource in the available networks of the request call cannot fully take advantage of the pooled resources. In [71] and [20], the proposals of exploiting vertical handoffs to balance the traffic load between networks are introduced, but no technical details are provided in [71]. The solution in [20] takes the perspective of stand-alone WLANs instead of balancing the traffic load over the entire system.

Motivated by the above considerations, we address a proactive load balancing solution for heterogeneous overlay networks from the viewpoint of the entire system. Firstly, we propose a call admission control scheme which aims to admit as many as possible calls into the system while keeping the networks working in effective states to provide QoS supports for users. Secondly, we derive a load balancing algorithm, which is invoked by the call admission control scheme when the available networks cannot admit the call without violating the QoS of other existing users. Vertical handoffs are initiated by the load balancing algorithm to make more room for the request call.

5.1 System Model

In this study, we consider a general two-tier heterogeneous interworking system as shown in Figure 5.1. This is the most popular structure for both researches and implementations, such as the cellular/WLAN overlay system and the WiMAX/WLAN overlay system addressed in Chapter 3. The tier-1 network covers the entire service area of the system and provides ubiquitous service for users. At hotspot areas such as offices, hospitals, campuses, airports, etc., a set of tier-2 networks with limited coverage is deployed to offload traffic from the tier-1 network. It is supposed that there is a BS in the tier-1 network and an AP in each tier-2 network.

As discussed in Section 2.4.1, in order to effectively utilize the pooled radio resources, the joint or common radio resource management is desired in the heterogeneous interworking system. To this end, a central radio resource controller (RRC) is deployed in the overlay system as illustrated in Figure 5.1, which is similar to the CRRM server (CRMS) defined by 3GPP [72]. The radio resources of all the networks are managed by this central RRC targeting to maximize the overall resource usage from

Figure 5.1 The system model.

the entire system view. The central RRC interacts with the local RRCs which are usually installed at BS or APs in the access networks. The major functions of the central RRC include collecting real-time traffic information of each network, making admission decisions for each request call, further deciding the serving network for the request call if both overlay networks are available, initiating the load balancing algorithm based on the defined policy and then selecting calls to perform vertical handoffs. According to the decision made by the central RRC, local RRCs perform the necessary activities to admit or reject the request call, or initiate vertical handoffs at the ongoing calls designated by the central RRC. In this chapter, we focus on the admission and load balancing operations at the central RRC.

By our proposed solution, vertical handoffs at ongoing calls may be proactively initiated by RRC to balance the traffic load among networks. Therefore, seamless handoffs are expected to guarantee the continuity of ongoing calls. As indicated in Section 2.3.1, multimode terminals in heterogeneous overlay systems can connect to multiple networks simultaneously. During vertical handoff process, the original interface can maintain the ongoing calls while the target interface performs handoff operations. Consequently, packet loss can be avoided, which significantly increases the feasibility of our solution.

It is analyzed in Section 2.2.2 that a tightly coupled interworking architecture can provide a good support on common radio resource management and seamless vertical

handoff. Our vertical handoff scheme designed in Chapter 4 further proves that the signaling overhead in a tightly coupled system could be very low. Therefore, our load balancing solution is very suitable for the tightly coupled systems. On the other hand, it is also discussed in Section 2.3.3 that PMIP can avoid the signaling overhead at wireless links because the mobile station is not required to participate in any mobility-related signaling. Meanwhile, PMIP is a network-controlled approach, which is consistent with our solution. Consequently, if there is necessary cooperation required by PMIP between the heterogeneous networks in a loosely coupled system, our load balancing solution is also applicable.

5.2 Call Admission Control Scheme

In this section, we propose a call admission control scheme for the two-tier heterogeneous overlay system composed by $N = \{N_0, N_1, \dots, N_M\}$, where N_0 is the tier-1 network and N_1 to N_M are the M overlapped tier-2 networks. The area covered by a tier-2 network is also referred to as a *double-coverage area*, where both the tier-1 network N_0 and a tier-2 network are available for the stations. Similarly, the area where only N_0 is in service is referred to as *single-coverage area*. For any network N_m ($0 \leq m \leq M$), it is assumed that the total amount of radio resource units provided is R_m . A resource unit denotes the basic unit of radio resource allocated for data transmission, which could be a time slot in time-division multiple access (TDMA) systems, a frequency band in frequency-division multiple access (FDMA) systems, or a logical allocation unit such as the logical channel in WCDMA.

The objective of the proposed admission control scheme is to accept as many calls into the system as possible while keeping networks in effective states for protecting the QoS of ongoing calls. Here, a network utilization factor U_m is defined for network N_m

$$U_m = \frac{L_m}{R_m}, \quad m = 0, 1, \dots, M, \quad (5.1)$$

where L_m is the number of resource units occupied by the calls being served by N_m . Meanwhile, a network utilization threshold θ_m ($0 \leq m \leq M$) is configured to determine whether N_m is working in an effective state. If N_m is a centrally controlled network (e.g., cellular network or WiMAX), θ_m will be configured by the local RRC according to

local admission control and QoS provision policy deployed. While if N_m is a WLAN (the most common tier-2 network), there is no central control and the radio resource is distributed in a contention manner. But the algorithms we derived in Section 4.2.2 show that the QoS metrics (e.g., available bandwidth, packet delay) provided to WLAN users can be well reflected by the network utilization. Therefore, we set a threshold θ_m to N_m to demonstrate the highest QoS requirement of ongoing calls recorded by the central RRC, that is, satisfactory service is always provided to all users as long as $U_m \leq \theta_m$. For example, if the highest QoS requirement is to keep the mean packet delay of the call in Figure 4.7 less than 15 ms (data rate 5.5 Mbps), we could set θ_m to be 0.7.

The call admission procedure is illustrated in Figure 5.2, where the request calls are differentiated based on the location of their associated stations. If the request call is located in the single-coverage area, a decision will be made to accept it to the tier-1 network N_0 or reject it. If the call is located in a double-coverage area, the central RRC needs to further decide whether it should be admitted into N_0 or its covered tier-2 network, N_k ($1 \leq k \leq M$). After a request call arrives, an admission decision will be made by the following steps:

- 1) The utilization factor of the network available to the request call will be estimated, under the assumption that the call is accepted into it. For any call, the tier-1 network N_0 is always available, and thus, the utilization factor of N_0 will be estimated first, which is

$$U_0^e = U_0 + \frac{r}{R_0}, \quad (5.2)$$

where r is the number of resource units consumed by the request call in the tier-1 network N_0 . If the tier-2 network N_k is also available, its utilization factor will be estimated as

$$U_k^e = U_k + \frac{r'}{R_k}, \quad (5.3)$$

where r' is the number of resource units consumed by the request call in the network N_k .

- 2) If the available network(s) can admit the request call within effective states, that is

Figure 5.2 Call admission control algorithm.

$U_0^e \leq \theta_0$ ($U_k^e \leq \theta_k$), the call will be admitted into the available network with lower estimated utilization factor.

- 3) Otherwise, the central RRC will invoke our proposed load balancing algorithm. The purpose is to create more room for the request call. The proposed load balancing algorithm does not simply offload traffic from one network to another, but minimize the variance of network utilization. By doing so, the fairness between networks can be protected. Meanwhile, the load balancing operations will not degrade the QoS of ongoing calls. The technical details of the algorithm are provided in the next section.
- 4) If a feasible solution could be found by the load balancing algorithm, the request call will be admitted into the network indicated by the solution. At the same time, the central RRC will inform the local RRCs to balance the load by initiating vertical handoffs at the calls that are designated by the solution.

- 5) While if there is no feasible solution, the central RRC needs to make further decision. In this study, the policy is to admit as many calls as possible into the system. Therefore, as long as the available network(s) has enough resource for the request call ($U_0^e \leq 1$ or $U_k^e \leq 1$), the call will be admitted into the network with lower estimated utilization.

5.3 Load Balancing Algorithm

As described in the above section, when the available network(s) cannot admit the request call without degrading the QoS of ongoing calls, the load balancing algorithm will be invoked. In this section, we present the design details of this algorithm. It's worthy to note that the algorithm can also be extended to be used in other approaches where load balancing is needed.

In this study, the objective of invoking the load balancing algorithm is to offload some traffic from the tier-1 network N_0 (as well as the tier-2 network N_k if available) to other networks within the system, so that more room could be created in N_0 (or N_k) for the request call. Here, we classify the tier-2 networks other than N_k into two categories: balanceable networks and non-balanceable networks. A balanceable network satisfies two conditions: 1) there are some calls located in this network but served by the tier-1 network N_0 ; 2) the utilization factor of this network is less than its corresponding threshold for the QoS support. The network that cannot satisfy these conditions will be referred to as a non-balanceable network. Apparently, only when one or more balanceable networks exist, it is possible to offload traffic from N_0 to such network(s) without degrading the service obtained by their ongoing calls, and then the load balancing scheme can make effect. Otherwise, the algorithm will be seen as infeasible.

It is supposed that the networks from N_1 to N_b ($1 \leq b \leq M$) are just the b potential balanceable networks at the time when the algorithm is invoked. Let c_i denote the number of calls located in the balanceable network N_i ($1 \leq i \leq b$), while these calls may be obtaining service from N_i or the tier-1 network N_0 . For the j -th ($1 \leq j \leq c_i$) call located in N_i , the amount of resource units needed by this call is $r'_{(i,j)}$ if the call connects to that tier-2 network N_i . While if the call connects to the tier-1 network N_0 , the amount of required resource is $r_{(i,j)}$. Here, an association variable is defined for this call, i.e.,

$$\eta_{(i,j)} = \begin{cases} 0, & \text{the call connects to } N_0 \\ 1, & \text{the call connects to } N_i \end{cases} \quad (5.4)$$

The function of the proposed load balancing algorithm is to re-associate the calls located in the balanceable networks as well as N_k if available, so that the variance of network utilizations can be minimized. We will derive the solutions for different cases that the request call is located in the double-coverage area and in the single-coverage area, respectively.

5.3.1 The request call is located in the double-coverage area

In this case, both the tier-1 network N_0 and the tier-2 network N_k are available to the request call. For a clear expression, it is assumed that N_k is just the network N_{b+1} . First, the utilization variable of each balanceable network can be expressed as

$$\Phi_i = \frac{\sum_{j=1}^{c_i} r'_{(i,j)} \eta_{(i,j)}}{R_i} \quad (1 \leq i \leq b) \quad (5.5)$$

Let η denote the association variable of the request call. If η equals 1, the request call will consume the resource of its covering tier-2 network N_{b+1} . Then the utilization variable of the network N_{b+1} can be given by

$$\Phi_{b+1} = \Phi_k = \frac{r' \eta + \sum_{j=1}^{c_k} r'_{(k,j)} \eta_{(k,j)}}{R_k} \quad (5.6)$$

Otherwise, if η equals 0, the request call will consume the resource of the tier-1 network N_0 , and then the utilization variable of N_0 can be written as

$$\Phi_0 = u_s + u_b + \frac{r(1-\eta) + \sum_{i=1}^{b+1} \sum_{j=1}^{c_i} r_{(i,j)} (1-\eta_{(i,j)})}{R_0}, \quad (5.7)$$

where u_s is the utilization of the resource induced by the ongoing calls located in the single-coverage area, and u_b is the utilization induced by the calls located in the non-balanceable networks but served by N_0 . Then, we can calculate the mean of the utiliza-

tion of these networks, which is

$$\overline{\Phi} = \frac{1}{b+2} \sum_{i=0}^{b+1} \Phi_i \quad (5.8)$$

Based on the above expressions, we formulate the load balancing problem as

$$\text{Minimize} \quad \sum_{i=0}^{b+1} (\Phi_i - \overline{\Phi})^2 \quad (5.9a)$$

$$\text{Subject to} \quad \Phi_i \leq \theta_i \quad (1 \leq i \leq b) \quad (5.9b)$$

$$\Phi_{b+1} \leq 1; \quad \Phi_0 \leq 1 \quad (5.9c)$$

$$\eta_{(i,j)} \in \{0,1\}; \quad \eta \in \{0,1\} \quad (5.9d)$$

By minimizing the variance of network utilizations in the objective function, the congested networks are enabled to return the effective states to the greatest extent, so that the possibility of initiating load balancing operations can be minimized. For the balanceable networks, the load balancing operations should not violate the QoS of ongoing calls in these networks which are reflected by their corresponding threshold θ . Therefore, $\Phi_i \leq \theta_i$ ($1 \leq i \leq b$) should be guaranteed as an important constraint. Meanwhile, in order to admit as many calls as possible, we set $\Phi_{b+1} \leq 1$ and $\Phi_0 \leq 1$, meaning that the request call will be accepted as long as the network N_0 or N_{b+1} (i.e., N_k) has enough available bandwidth.

This optimization problem is a quadratic binary programming problem, which can be solved by various methods [133]. If a feasible solution could be found, the request call will be admitted into the network indicated by η obtained in the solution. At the same time, $\eta_{(i,j)}$ will also be determined by the obtained solution for the j -th call located in the network N_i . If the calculated $\eta_{(i,j)}$ does not correspond to the network that is serving the call, then a vertical handoff should be initiated at this call.

5.3.2 The request call is located in the single-coverage area

If the request call is located in the single-coverage area, only the tier-1 network N_0 is available to it. Therefore, the request call can only consume the resource of the network N_0 , which means the association variable (η) of the request call is definitely equal

to 0. Meanwhile, since the tier-2 network N_k is not involved in this case, the algorithm is to balance the load among b balanceable networks N_1 to N_b as well as the tier-1 network N_0 . In this case, the calculation of Φ_i ($1 \leq i \leq b$) for each balanceable network is the same as that in (5.5). The utilization variable of the tier-1 network N_0 needs to be rewritten as

$$\Phi_0 = u_s + u_b + \frac{r + \sum_{i=1}^b \sum_{j=1}^{c_i} r_{(i,j)} (1 - \eta_{(i,j)})}{R_0}, \quad (5.10)$$

where u_s is the utilization of the resource induced by the calls located in the single-coverage area, and u_b is the utilization induced by the calls located in the non-balanceable networks but served by N_0 . The mean of the utilization of involved networks is given by

$$\bar{\Phi} = \frac{1}{b+1} \sum_{i=0}^b \Phi_i. \quad (5.11)$$

Then, we can formulate the load balancing problem for this single-coverage case as

$$\text{Minimize} \quad \sum_{i=0}^b (\Phi_i - \bar{\Phi})^2 \quad (5.12a)$$

$$\text{Subject to} \quad \Phi_i \leq \theta_i \quad (1 \leq i \leq b) \quad (5.12b)$$

$$\Phi_0 \leq 1 \quad (5.12c)$$

$$\eta_{(i,j)} \in \{0, 1\} \quad (5.12d)$$

If a feasible solution can be obtained, the request call will be admitted into the tier-1 network N_0 . Also, vertical handoffs may be initiated at ongoing calls according to their association variables determined by the solution.

5.4 Performance Evaluation

To demonstrate the effectiveness and feasibility of our proposed scheme, we simulated a two-tier heterogeneous overlay system in MATLAB and compared the proposed scheme with other reference schemes from various metrics. In this section, the system

configurations are described first, and then the simulation results for the comparison among different schemes are presented and discussed.

5.4.1 System configurations

Figure 5.3 The simulation topology.

The network topology deployed in the simulation is illustrated in Figure 5.3, with key parameters listed in Table 5.1. Six networks compose the two-tier overlay system. The radius of the tier-1 network N_0 is 1000 m and the radius of each tier-2 network is 200 m. At the beginning of each simulation run, half of the stations are evenly placed into five tier-2 networks, and the other half of the stations are evenly placed within the entire coverage of the tier-1 network. All the stations move in random routes. If a station moves in the single-coverage area, the velocity is 1.5 m/s. If the station moves into a tier-2 network, the velocity will be decreased to 0.2 m/s, because in real applications the stations usually move slowly or keep fixed for a period in hot spots.

The total resource units provided by the tier-1 network N_0 is 300 and the total resource units provided by each tier-2 network is 120. A resource unit could be a time slot, a frequency or a logical channel as mentioned in Section 5.2. In the simulation, the bandwidth of calls is also expressed in units (e.g., 256 kbps per unit). When a call arrives at a station, the number of requested bandwidth units is uniformly selected from the set $\{1, 1.2, 1.5\}$. The duration of the call is uniformly distributed between 0 s and

Table 5.1 Key parameters used in the simulation.

Tier-1 network	
Radius	1000 m
Velocity of stations	1.5 m/s
Total resource units	300
γ	6 ($d_{BS} < 300$ m)
	8 ($300 \text{ m} < d_{BS} < 600$ m)
	10 ($600 \text{ m} < d_{BS} < 1000$ m)
Tier-2 network	
Radius	200 m
Velocity of stations	0.2 m/s
Total resource units	120
γ	10 ($d_{BS} < 100$ m)
	12 ($100 \text{ m} < d_{BS} < 200$ m)

1000 s. Each simulation runs for a duration of 40000 s.

It is known that for the same call, the amount of resource consumed depends upon the wireless channel quality, which is usually reflected by the distance from the station to the BS (d_{BS}) or AP (d_{AP}). Closer stations experience higher channel gains so that higher-order modulations and higher data rates can be assigned with fewer resource units consumed. For simplicity, we assume that the wireless channel quality and data rates obtained by the station are only determined by the distance to BS or AP [134, 135]. In the simulation, each tier-2 network is divided into two regions according to the value of d_{AP} . For the call with per bandwidth unit, the number of resource units consumed (γ) is 10 (12) if $d_{AP} < (>)$ 100 m. Similarly, the tier-1 network is divided into three regions, i.e., $\gamma = 6$ (10) if $d_{BS} < 300$ m (> 600 m) and $\gamma = 8$ for the region between.

In the simulation, the quadratic binary programming problem was solved by the optimization package CPLEX [136]. Our experiments were run on a 2.67 GHz processor with 2 GB memory. Generally, the run time to solve the problem was no more than 15 ms, which is short enough not to be noticed during a call admission control process. The execution time can be further decreased in implementations since the central server of the integrated system usually has a much higher computational ability.

5.4.2 Performance comparison

We evaluate and compare our proposed scheme with other three reasonable reference schemes. The first reference scheme is based on the traditional WLAN-first admission control scheme, where load balancing is not taken into account. As long as a tier-2 network is available, the newly originated call will be admitted into it, and the call that is obtaining service from the tier-1 network will also be handed over into it immediately. The second reference scheme is based on [35], where the network load is considered in the vertical handoff decision process. When a mobile station moves into a tier-2 network, the utilization factor of the network is estimated first. If the estimated result is less than the given threshold required by the QoS, a decision of handoff to the tier-2 network is made. Otherwise, the station will keep the connection with the tier-1 network. The third reference scheme is based on [20], where the system can proactively initiate vertical handoffs at ongoing calls in order to balance the network loads. If the utilization factor of a tier-2 network exceeds a predefined threshold ϵ_w^u , the system will hand over the calls served by this network to the tier-1 network. Meanwhile, if the utilization factor of a tier-2 network falls below a threshold ϵ_w^l , the system will transfer traffic load from the tier-1 network to this network via vertical handoffs.

First, we compare the performance of these schemes from the perspective of users. We set the utilization threshold θ_i ($0 \leq i \leq 5$) used in our scheme to be 0.8. For the second reference scheme, the utilization threshold required by QoS is also 0.8. The thresholds $(\epsilon_w^u, \epsilon_w^l)$ defined in the third reference scheme are set to be (0.8, 0.4). We compare the schemes using various metrics with the variation of network loads. In Figure 5.4, the average blocking probability of the newly originated calls within the entire system is shown. Figure 5.5 illustrates the average dropping probability of mobility-triggered handoff calls. To evaluate the QoS obtained by the calls, we define a QoS satisfaction ratio for each call served by the tier-2 networks. This ratio denotes the proportion of time that a call obtains satisfactory QoS with the utilization of the network lower than the corresponding threshold during its existence. It is assumed that all the calls have the same QoS requirement which corresponds to the network utilization threshold 0.8. We average the QoS satisfaction ratio over all the calls and plot the result in Figure 5.6.

In the simulation, it was found that by the first reference scheme without the consideration of network traffic load, some networks were overloaded while others underutilized with the highest probability. As a result, the highest call blocking and

Figure 5.4 New call blocking probability.

Figure 5.5 Mobility-triggered handoff call dropping probability.

dropping probability are introduced by this scheme as shown in Figure 5.4 and Figure 5.5. The second reference scheme works better in these two aspects since the network traffic load is taken into account in the vertical handoff decisions. The scheme becomes effective only at the time as the mobile stations are moving into the coverage of tier-2 networks. This limits the quick response on the network load variations. Consequently, the QoS obtained by this scheme is as bad as the first reference scheme. Moreover, in our simulation, all the stations kept moving that provides opportunities for this scheme to operate during vertical handoffs. But if there are more fixed stations or more mobile

Figure 5.6 Average QoS satisfaction ratio of calls.

stations with low mobility in the system, the advantage of the scheme will be degraded. In the third reference scheme, vertical handoffs are dynamically initiated by the system in order to keep the load of tier-2 networks within effective region. As a result, the calls served by tier-2 networks can obtain a relatively good QoS support as illustrated in Figure 5.6. But the scheme consider stand-alone tier-2 network instead of the entire overlay system. When a tier-2 network needs to offload traffic, the tier-1 network may not have enough resource. Thus, a relatively high call dropping/blocking probability and low throughput are introduced by this scheme.

By our proposed scheme, once the available network(s) cannot accept the call without violating the QoS of ongoing calls, the load balancing algorithm will be invoked to create more room for the request call. Furthermore, both the calls in the single-coverage area and the calls in the double-coverage areas can initiate this process, thus resulting in the lowest call dropping and blocking probability as expected. Meanwhile, since the QoS of calls is considered in the admission control process, it can be observed from Figure 5.6 that the best QoS support can be achieved by the proposed scheme especially when the system is lightly loaded. It implies that for the system with the same traffic load, the best QoS support can be achieved by the proposed scheme. As the system tends to be heavily loaded, this benefit diminishes because more calls are admitted into the system by the proposed scheme than by other schemes. But the achieved QoS is still competitive with the third reference scheme and better than the other two reference schemes.

Figure 5.7 Average throughput of the integrated system.

Figure 5.8 Standard deviation of network utilizations.

Second, we compare the four different schemes from the viewpoint of network operators. For the network operators, an important issue is to maximize the overall utilization of the system meanwhile keep all the networks working in stable states. The overall throughput of the overlay system is illustrated in Figure 5.7, which denotes the average number of bandwidth units served by the system per second. We observe that the highest throughput can be achieved by the proposed scheme, which corresponds to the fact of the lowest call blocking/dropping probability. In Figure 5.8, we plot the standard deviation for the utilizations of all the networks. The utilization factor of each

network is recorded per 100 seconds. At each time point, the standard deviation of the recorded network utilization factors is calculated, and then an average result is further calculated over all time points. It is observed in the figure that the lowest deviation can be obtained by the proposed scheme, implying that the traffic load is distributed into the networks in the most even manner. In another word, by our proposed scheme, the networks can work in the stable status for a longer time and a good fairness between networks can be achieved.

5.5 Summary

In this chapter, we investigated the issue of traffic load balancing among different network domains of the heterogeneous overlay system. A call admission control scheme was proposed, followed by a load balancing algorithm which is used in the admission procedure. During the general call admission process, balancing the traffic load between the tier-1 network and the possible tier-2 network available to the request call has been taken into account. Once both networks cannot admit the call without violating the QoS of ongoing calls, the proposed load balancing algorithm will be invoked to create room for the request call by initiating vertical handoffs at ongoing calls. It has been shown by the simulation results that our scheme can utilize the overall resources of the interworking system effectively. More calls can be admitted into the system. Meanwhile, the networks can work in effective states with good QoS supports.

Chapter 6

Network Selection & Resource Allocation for Multicasting/Broadcasting

Group communications, including multicasting and broadcasting, are efficient mechanisms for delivering bandwidth-intensive multimedia content over wireless links. The one-to-many transmission provided by the group communication offers a great opportunity for network operators to transmit one stream to multiple users simultaneously. With the increasing popularity of multimedia services (e.g., IPTV, video conferencing, on-demand video, etc.), group communications have attracted extensive research interests in the wireless networking community.

Within a heterogeneous interworking system, the member stations of a multicast/broadcast stream may be distributed in many networks, meanwhile each member station may have more than one network available simultaneously. Then, the problem of selecting suitable networks for transmitting the multicast/broadcast stream arises naturally. If the member stations located within the same region connect to different networks while only one network can serve all of them, the system radio resources will be wasted; this will be even worse with the increase of the number of stations. Therefore, it is expected that the member stations of a stream are converged together such that fewer networks are involved in transmission. However, this could not be the unique factor that is considered in the network selection strategy for the multicast/broadcast service in heterogeneous interworking systems.

It is known that adaptive modulation and coding (AMC) has been widely deployed by advanced wireless communication technologies (e.g., WiMAX, CDMA2000, UMTS, WLAN, etc) as a resource allocation technique to effectively utilize the radio

resources. The main idea is to dynamically change the modulation and coding scheme with the objective of adapting the overall spectral efficiency to the channel condition. Typically, for a station close to BS presenting good channel conditions, a high order modulation scheme with low coding redundancy could be deployed to increase the spectral efficiency. Nevertheless, a station located far away from BS usually experiences worse channel conditions, and thus needs a more robust modulation scheme to protect the signal against noise and interference. A well-known digital communication principle states that "a greater data symbol modulation is more spectrum efficient but also less robust" [146]. In other words, a more robust modulation scheme corresponds to a lower spectral efficiency (i.e. data rate), and then consumes more radio resources for the same amount of data.

For a multicast/broadcast stream, to guarantee all the member stations connected to the same network to successfully receive the stream, the most robust modulation scheme corresponding to the member station(s) of the worst channel conditions has to be deployed. Meanwhile, if a member station with even worse channel conditions is also converged into this network, more radio resources may be consumed due to the use of a more robust modulation scheme. However, in a heterogeneous interworking system, this member station exhibiting the worst channel condition in this network may present better channel conditions in other available network(s). From the entire system point of view, fewer radio resources may be consumed if this station switches to another network. This reveals that the resource allocation for the multicast/broadcast stream should also be taken into account by the network selection strategy.

A simple but direct comparison of potential network selection strategies is illustrated in Figure 6.1. Under strategy 1, the member stations randomly connect to their available networks and the stream is then transmitted by all the three networks with a bandwidth cost of 10, 30, and 5, respectively, leading to the total bandwidth cost of 45. Strategy 2 mainly converges the member stations together and WiMAX is responsible of serving all three member stations by the most robust modulation scheme with a bandwidth cost of 40. However, with the resource allocation under consideration for a more comprehensive network selection, stations A and B connect to WiMAX with a bandwidth cost of 10, and station C connects to WLAN with a bandwidth cost of 5, leading to the total bandwidth cost of only 15 by strategy 3.

Therefore, we can see that intelligent network selection schemes are desired by the

Figure 6.1 Comparison of different network selection strategies.

multicast/broadcast service, especially when a large group of member stations is involved. Stations within the same network region should be converged together meanwhile the channel condition of the stations at different regions, i.e., resource allocation, should also be taken into account. As discussed in Section 2.5, among very few relevant works, network selection is considered in [12, 104, 105] but without mentioning resource allocation. The same bandwidth cost is assigned to all member stations covered by a network, i.e., an assumption not suitable for advanced wireless technologies supporting AMC.

In this chapter, we propose a Network Selection and Resource Allocation scheme for the multicast/broadcast service in heterogeneous wireless networks, referred to as NSRA. The objective is to minimize the overall bandwidth cost consumed by the multicast/broadcast stream within the entire system. We formulate NSRA as a binary integer programming problem, and further derive a Lagrangian based heuristic algorithm LR to solve the problem.

6.1 Design of the NSRA Scheme

6.1.1 System model and notation

As suggested in [137, 138], we mainly focus on the resource allocation for the wireless link, because the wireless link bandwidth is much more scarce and valuable than the bandwidth of wired links and thus becomes a dominant factor in the overall system optimization. The wireless link is the last hop of multicast or broadcast transmission, where broadcast can be seen as a special case of multicast. In the following analysis, the stream may be a multicast stream or a broadcast stream.

It is assumed that adaptive modulation and coding is supported by all the networks within the system, as this is the trend of advanced wireless technologies. For instance, in the WiMAX networks, each station negotiates its burst profile with BS before the connection starts and can also exchange its burst profile with BS during the connection [139]. The burst profile is a set of transmission settings including modulation and coding schemes that can reflect the channel conditions. Different burst profiles lead to different levels of robustness and transmission rates. Under worse channel conditions, a more robust burst profile will be deployed which corresponds to lower order modulation scheme and higher coding redundancy; hence, a lower data rate is obtained and more bandwidth resources are consumed by the same stream. For description convenience, we assume that the transmission parameters in each network within the system are defined in burst profiles like WiMAX.

Letting M denote the group of member stations requesting the stream, for each member station $m \in M$, there are a set of networks N_m available. These available networks of all the member stations compose a network set N . In the network $n \in N$, the set of burst profiles supported by its covering member stations is $B_n = \{b_1, b_2, \dots, b_p\}$, where $b_1 < b_2 < \dots < b_p$ in terms of robustness. We use (n, b) to denote the burst profile b ($b \in B_n$) in the network n , which introduces a bandwidth cost of $c_{n,b}$ for transmitting the stream. Correspondingly, we also have $c_{n,b1} < c_{n,b2} < \dots < c_{n,bp}$ because more robust burst profile provides lower data rate and hence consumes more bandwidth resource. In order to reduce the radio resources consumed by the stream, burst profiles with lower robustness are preferred for transmission; in other words, the finally selected burst profile must be one of the least robust burst profile supported by the member stations. Therefore, within B_n , only the least robust burst profiles supported by each member station covered by the network n need to be contained.

In the network n available to the member station m ($n \in N_m$), usually only a subset of burst profiles can be used by m limited by the channel conditions, which are listed in

$B_{n,m}$. If the least robust burst profile supported by m is b_{pp} , then $B_{n,m} = \{b_{pp}, b_{pp+1}, \dots, b_p\}$. Within the system, there are usually some of member stations located in the same region (i.e., with the same available networks and supported burst profiles), especially when the stream has a large group of members. If one of these member stations can successfully receive the stream, others can also at the same time. Hence, we include only one of these member stations into the set M in order to reduce the problem size. In other words, any m in M represents not only a member station but also logically a specific region. The notations used in the NSRA scheme are summarized in Table 6.1.

Table 6.1 Notations used in the NSRA scheme.

Notation	Description
N	the set of networks covering the member stations.
B_n	the set of burst profiles available at the network n , $n \in N$.
$c_{n,b}$	the bandwidth cost introduced with (n, b) , $n \in N$, $b \in B_n$.
$M_{n,b}$	the set of member stations that can be served by (n, b) , $n \in N$, $b \in B_n$.
M	the set of member stations in the system.
N_m	the set of networks available to the member station m , $m \in M$.
$B_{n,m}$	the set of burst profiles in n available to the member station m , $n \in N_m$.

6.1.2 Problem specification

The function of our NSRA scheme is to select networks for transmitting the stream to all the member stations as well as to determine the number of resources allocated to the stream in each selected network (i.e., determine the burst profile deployed). The objective is to minimize the overall bandwidth cost consumed by the stream within the entire heterogeneous overlay system.

Here, we define a binary variable $x_{n,b}$ for the burst profile (n, b)

$$x_{n,b} = \begin{cases} 1, & (n,b) \text{ is selected for transmission} \\ 0, & \text{otherwise} \end{cases}.$$

The state “1” for $x_{n,b}$ denotes the network n to be selected to transmit the stream with the burst profile (n, b) . Therefore, the determination of each element in the vector $\mathbf{x} = \{x_{n,b} | n \in N, b \in B_n\}$ indicates the completeness of both the network selection and resource allocation operations. Meanwhile, “ $x_{n,b} = 1$ ” shows that $|M_{n,b}|$ member stations can successfully receive the stream with a paid bandwidth cost of $c_{n,b}$. Therefore, to

minimize the overall system bandwidth cost consumed by the stream, we need to select the burst profile (n, b) with larger $|M_{n,b}|$ (i.e., converging member stations) and lower $c_{n,b}$ (i.e., lower robustness).

Based on the above considerations, we formulate the NSRA scheme as the following binary integer programming problem (BIP)

$$\text{Minimize} \quad C = \sum_{n \in N} \sum_{b \in B_n} c_{n,b} x_{n,b} \quad (6.1)$$

$$\text{Subject to} \quad \sum_{n \in N_m} \sum_{b \in B_{n,m}} x_{n,b} \geq 1, \quad \forall m \in M, \quad (6.2)$$

$$\sum_{b \in B_n} x_{n,b} \leq 1, \quad \forall n \in N, \quad (6.3)$$

$$x_{n,b} \in (0, 1), \quad \forall n \in N \text{ and } \forall b \in B_n. \quad (6.4)$$

Without loss of generality, we assume an integer for the bandwidth cost $c_{n,b}$ of each burst profile. The constraint (6.2) ensures that any member station m ($m \in M$) can receive the stream from at least one available network with one of its available burst profile. Under the constraint of (6.3), in each selected network $n \in N$ the stream can be transmitted with only one burst profile, i.e., the stream can only be modulated and encoded once in a network. For the final optimal solution of the minimization problem, this constraint may be unnecessary. This is because, if two burst profiles b_1 and b_2 ($b_1 < b_2$) of a network are selected by a solution, b_1 will be redundant since all the member stations served by b_1 can be served by the more robust burst profile b_2 . Such cases can be removed during the minimization process. However, this constraint well reflects the overlay structure of the heterogeneous interworking system. It can then be used to expedite the minimization process and make the problem computationally cheaper as shown in the next section.

From the objective function (6.1) and the constraint (6.2), we can see that the BIP problem is an instance of set covering problem (SCP). In an SCP, each set covers some members with an assigned cost, and the problem calls for the minimum-cost group of sets such that each member is covered by at least one set. Each burst profile in our scheme can be seen as a set in SCP. Specially, our BIP problem presents an overlay structure where larger networks overlay smaller networks and more robust burst profiles overlay lower robust burst profiles from the viewpoint of member stations. It is known that the set covering problem is NP-hard in the strong sense [140], and an effec-

tive way to solve the problem is deploying heuristic methods to find optimal or near-optimal solutions. With the development of mathematical programming, a number of impressive heuristic algorithms have been derived for solving the real-world SCP instances (e.g., bus, railway and airline crew scheduling problems) [141]. The structural properties of real-world instances are utilized to improve the performance of the algorithms. In the following subsection, we derive a Lagrangian-based heuristic algorithm to solve the BIP problem which takes into account the overlay structure of our NSRA scheme. This is based on the consideration that Lagrangian approach is especially suitable for the integer programming problem with a relatively easy objective function complicated by a set of side constraints [142], just like our BIP problem. The Lagrangian approach has been increasingly used in large-scale mathematical programming applications since it usually can find good solutions at reasonable computational cost.

6.2 Lagrangian Algorithm

The Lagrangian approach relaxes the complicated constraints of the problem by introducing a penalty term into the objective function, and hence helps to find solutions easily. A lower bound to the optimum of the original minimization problem could be provided by the relaxed Lagrangian problem, based on which, reasonable feasible solutions of the original problem can be constructed. In this subsection, we first relax the BIP problem with a Lagrangian approach, and then derive a heuristic algorithm to find optimal or near-optimal solutions of the BIP problem based on the Lagrangian relaxation.

6.2.1 Lagrangian relaxation

By the algorithm, the constraint (6.2) is chosen to be relaxed. We define a Lagrangian multiplier vector $\mathbf{u} = \{u_m \geq 0 \mid m \in M\}$ for all the member stations in M , and attach it to the constraint (6.2). Penalty functions are then constructed and brought into the objective function in (6.1); we can then obtain the following Lagrangian relaxation (LR) problem

$$\text{Minimize} \quad C(\mathbf{u}) = \sum_{n \in N} \sum_{b \in B_n} c_{n,b} x_{n,b} + \sum_{m \in M} u_m (1 - \sum_{n \in N_m} \sum_{b \in B_{n,m}} x_{n,b})$$

$$= \sum_{n \in N} \sum_{b \in B_n} (c_{n,b} - \sum_{m: (n,b) \in B_{n,m}} u_m) x_{n,b} + \sum_{m \in M} u_m \quad (6.5)$$

$$= \sum_{n \in N} \sum_{b \in B_n} \varepsilon_{n,b} x_{n,b} + \sum_{m \in M} u_m$$

$$\text{Subject to} \quad \sum_{b \in B_n} x_{n,b} \leq 1, \quad \forall n \in N,$$

$$x_{n,b} \in (0, 1), \quad \forall n \in N \text{ and } \forall b \in B_n,$$

where $\varepsilon_{n,b} = c_{n,b} - \sum_{m: (n,b) \in B_{n,m}} u_m$ is called as the relative cost of the burst profile (n, b) . Here, u_m can be seen as the utility obtained when the member station m receives the stream, and then $\sum_{m: (n,b) \in B_{n,m}} u_m$ is the total utility obtained from the member stations that can be served by the burst profile (n, b) . In other words, if the burst profile (n, b) is selected for transmission, the system obtains a utility of $\sum_{m: (n,b) \in B_{n,m}} u_m$ with a bandwidth cost of $c_{n,b}$. Clearly, the burst profile with smaller relative cost $\varepsilon_{n,b}$ is preferred.

As the Lagrangian multiplier u_m of each member station $m \in M$ is nonnegative, meanwhile $(1 - \sum_{n \in N} \sum_{b \in B_{n,m}} x_{n,b}) \leq 0$ required by the constraint (6.2), the optimum of the LR problem is less than the optimum of the original BIP problem. Therefore, the optimal solution of LR problem under a given multiplier vector $\mathbf{u}$ provides a lower bound on the optimum of BIP. Because the complicated constraint (6.2) is relaxed, it is easy to solve LR problem and find lower bounds of the optimum of BIP.

6.2.2 Lagrangian heuristic algorithm

Based on the Lagrangian relaxation problem LR described above, we derive a heuristic algorithm in this subsection to find optimal or near-optimal solutions of the original BIP problem. The fundamental idea is to iteratively update the Lagrangian multiplier vector $\mathbf{u}$ to generate a series of optimums of the LR problem and feasible solutions of the original BIP problem. As described above, an optimal solution of the LR problem provides a lower bound on the optimum of BIP. Meanwhile, any feasible solution of BIP must correspond to an objective value greater than its final optimum and then constitutes an upper bound. The algorithm is to iteratively increase the lower bound and decrease the upper bound in order to make them as close to the optimum of BIP as possible.

Table 6.2 Lagrangian algorithm.

-
1. let $\underline{C} \leftarrow -\infty$, $\bar{C} \leftarrow -\infty$, $u_m \leftarrow 1, \forall m \in M$.
 2. **for** $\forall n \in N$
 - let $\varepsilon_{n,b} \leftarrow c_{n,b} - \sum_{m:(n,b) \in B_{n,m}} u_m$, $\forall b \in B_n$;
 - find $b' \leftarrow \arg\min_{b \in B_n} \{\varepsilon_{n,b}\}$; let $x_{n,b}(\mathbf{u}) \leftarrow 0$, $\forall b \in B_n \setminus \{b'\}$;
 - if** $\varepsilon_{n,b'} \leq 0$, **then** let $x_{n,b'}(\mathbf{u}) \leftarrow 1$, **else** let $x_{n,b'}(\mathbf{u}) \leftarrow 0$, **endif**;**endfor**;
 - let $C(\mathbf{u}) \leftarrow \sum_{n \in N} \sum_{b \in B_n} \varepsilon_{n,b} x_{n,b}(\mathbf{u}) + \sum_{m \in M} u_m$, $\underline{C} \leftarrow \max\{\underline{C}, C(\mathbf{u})\}$.
 3. let $\bar{\mathbf{x}}(\mathbf{u}) \leftarrow \mathbf{x}(\mathbf{u})$;
 - for** any m with $\sum_{n \in N_m} \sum_{b \in B_{n,m}} x_{n,b}(\mathbf{u}) = 0$
 - find $(n'', b'') \leftarrow \arg \min_{n \in N_m} \{\varepsilon_{n,b}\}$; let $\bar{x}_{n'',b''}(\mathbf{u}) \leftarrow 1$, $\bar{x}_{n'',b}(\mathbf{u}) \leftarrow 0$, $\forall b \in B_{n''} \setminus \{b''\}$;**endfor**;
 - $\bar{C} \leftarrow \min(\bar{C}, \sum_{n \in N} \sum_{b \in B_n} c_{n,b} \bar{x}_{n,b}(\mathbf{u}))$; **if** $\bar{C}$ changed, **then** let $\bar{\mathbf{x}} \leftarrow \bar{\mathbf{x}}(\mathbf{u})$, **endif**;
 - if** $\bar{C} \leq \lceil \underline{C} \rceil$, **then stop**, **endif**.
 4. **for** $\forall m \in M$
 - let $s_m \leftarrow 1 - \sum_{n \in N_m} \sum_{b \in B_{n,m}} x_{n,b}(\mathbf{u})$, $u_m \leftarrow \max(0, u_m + \lambda \frac{\bar{C} - C(\mathbf{u})}{\sum_{m \in M} s_m^2} s_m)$;**endfor**.
 5. **for** any (n, b) $n \in N$, $b \in B_n$
 - if** $x_{n,b}(\mathbf{u}) = 0$ and $C(\mathbf{u}) + \varepsilon_{n,b} > \bar{C}$, **then** $B_n \leftarrow B_n \setminus \{b\}$, **endif**;**endfor**;
 - if** $\lambda \geq \lambda_{thr}$ or $t \geq T$, **then stop**, **else** return to step 2, **endif**.
-

The details of the algorithm are illustrated in Table 6.2. $\underline{C}$ and $\bar{C}$ record the best known lower and upper bounds obtained by the algorithm, respectively. $\bar{\mathbf{x}}$ is the feasible solution of BIP corresponding to $\bar{C}$. Initially, the multiplier of each member station u_m is set to be 1. In step 2, the LR problem in (6.5) is solved under the multiplier vector $\mathbf{u}$ of current iteration. The relative cost $\varepsilon_{n,b}$ of each burst profile (n, b) ($n \in N$, $b \in B_n$) is calculated first. As above discussed, the burst profile with smaller relative cost is preferred by the system for the more utility obtained by it with lower bandwidth cost

paid. Meanwhile, it is required by the constraint (6.3) that at most one burst profile can be selected for transmission in each network. Therefore, for each network n , we find out its burst profile with the minimum relative cost, b' , because it presents the highest probability of being selected. If its relative cost $\varepsilon_{n,b'} \leq 0$, we set $x_{n,b'}(\mathbf{u}) \leftarrow 1$; otherwise, $x_{n,b'}(\mathbf{u}) \leftarrow 0$. Once the binary variables of all burst profiles are determined, the optimal solution of the LR problem under $\mathbf{u}$ is known as $\mathbf{x}(\mathbf{u}) = \{x_{n,b}(\mathbf{u}) \mid n \in N, b \in B_n\}$. At the same time, the optimum of the LR problem $C(\mathbf{u})$ can also be calculated, which provides a new lower bound on the optimum of BIP. The best known lower bound $\underline{C}$ will be updated if $C(\mathbf{u})$ is greater.

By step 2, the optimal solution of the LR problem $\mathbf{x}(\mathbf{u})$ is obtained. However, this solution may not be feasible to the original BIP problem. This is caused by the fact that the constraint (6.2) is relaxed in LR, and thus, it is not guaranteed that all the member stations can be served by the solution $\mathbf{x}(\mathbf{u})$. Even though $\mathbf{x}(\mathbf{u})$ can give us useful information on the structure of the optimal solution of BIP. We construct a feasible solution of BIP based on $\mathbf{x}(\mathbf{u})$ in step 3, which is referred to as $\bar{\mathbf{x}}(\mathbf{u})$. Firstly, we check each member station. If the member station m cannot be served by $\mathbf{x}(\mathbf{u})$ (i.e., $\sum_{n \in Nm} \sum_{b \in B_{n,m}} x_{n,b} = 0$), we find out its available network and burst profile with the minimum relative cost (n'', b'') , and set $\bar{x}_{n'',b''}(\mathbf{u}) \leftarrow 1$ to add this burst profile to the solution $\bar{\mathbf{x}}(\mathbf{u})$. If another burst profile b (definitely $b < b''$) has already been selected in the network n'' , we set $\bar{x}_{n'',b}(\mathbf{u}) \leftarrow 0$ because b is overlaid by b'' and thus unnecessary. Once all the member stations can be served by $\bar{\mathbf{x}}(\mathbf{u})$, it becomes a feasible solution of BIP. Then the corresponding objective value can be computed by (6.1), which provides a new upper bound of the optimum and may update $\bar{C}$. So far, both the upper bound $\bar{C}$ and the lower bound $\underline{C}$ at this iteration have been known. If these two bounds are identical, they are just the optimum of the BIP problem, and then the algorithm can be stopped here.

Otherwise, if the optimal solution is not found at this iteration, we need to update the multiplier vector $\mathbf{u}$ in order to improve the solution of the next iteration to make both the upper bound and the lower bound obtained as close to the optimum of BIP as possible. This is done in step 4 by a subgradient approach, where the subgradient of the current utility of each member station is defined as

$$s_m = \frac{\partial C(\mathbf{u})}{\partial u_m} = (1 - \sum_{n \in N_m} \sum_{b \in B_{n,m}} x_{n,b}(\mathbf{u})), \quad \forall m \in M. \quad (6.6)$$

If $s_m < 0$, more than one available network of m is selected for transmission, and then we reduce u_m for the next iteration. If $s_m = 0$, just one network serves m (the desired case), and then u_m is kept unchanged. If $s_m > 0$ (i.e., $s_m = 1$), none of the available networks of m is selected for transmission under current $\mathbf{u}$. Therefore, we increase the utility of such member station in order to increase the probability for its available networks to be selected at the next iteration. The parameter $\lambda > 0$ controls the step size along the subgradient direction. Following the classical Held-Karp approach [143], we set λ with a starting value, and halve its value when $\bar{C}$ has not been decreased for a specified number of iterations.

At the end of each iteration, the necessity of each burst profile is checked in order to continually reduce the problem size during the algorithm. If the burst profile (n, b) is not selected by the current solution of LR (i.e., $x_{n,b}(\mathbf{u}) = 0$) but with $C(\mathbf{u}) + \varepsilon_{n,b} > \bar{C}$, this burst profile will be removed from the burst profile set of the network n . This is because a burst profile leading to the lower bound greater than the upper bound cannot be in the optimal solution [144]. The algorithm stops when λ is less than a threshold λ_{thr} , or the number of iterations t exceeds a threshold T . In such cases, a near-optimal solution ($\bar{x}$ and $\bar{C}$) is achieved from the algorithm.

6.3 Implementation Details

As discussed in Section 2.4.1, joint or common radio resource management is desired in heterogeneous wireless networks. Usually, there is a central controller to manage all the radio resources within the interworking system. In such a centrally controlled case, the architecture introduced in Chapter 5 (see Figure 5.1) can be deployed for the NSRA scheme. The local RRCs installed in the access networks report the real-time information of its covering member stations (e.g., the joining or leaving of a member station, the burst profiles of a member station) to the central RRC, which performs the scheme from the entire system point of view. The algorithm could be invoked periodically or on demand. If the algorithm is invoked at some time and a new solution different from current solution deployed by the system is found, the central RRC will inform local RRCs to transmit the stream based on the new solution. If the serving network of a

member station is changed by the new solution, a vertical handoff is required. Therefore, similar to the scheme designed in Chapter 5, our NSRA scheme is also especially suitable for the tightly coupled and PMIP-based loosely coupled interworking systems because of their good supports on vertical handoff and common radio resource management.

On the other hand, the Lagrangian based algorithm also enables the scheme to be implemented in a distributed manner as suggested in [12]. But the Lagrangian algorithm and the protocol designed in [12] for the multicasting in heterogeneous interworking networks require the member stations to perform network selection operations and periodically exchange multiplier vectors with BS via the wireless links. As a result, considerable overhead is introduced into wireless links and the joining or leaving of any member station impacts the algorithm. By our algorithm, the variable $\mathbf{x}$ is defined for network burst profiles rather than for member stations. Therefore, our scheme can be performed completely at the network side. It is suggested that the local RRCs at access networks exchange necessary information (e.g., the bandwidth cost of their burst profiles, the multiplier information of their covering member stations) via the backbone network and perform the operations needed by the algorithm without the involvement of any member station.

On the other hand, it should be noted that the *member station* $m \in M$ in our algorithm indeed is a logical concept which represents a specific network region as discussed in Section 6.1.1. If multiple stations stay in the same region such that they have the same available networks and available burst profiles, only one of them will be included into M . Consequently, the joining or leaving of a station may not initiate the algorithm, if it does not change the burst profile sets B_n ($n \in N$) as which are used by other stations at the same region. In other words, only when the burst profile set B_n ($n \in N$) changes, our algorithm may be invoked. Therefore, the complexity of our algorithm will not always increase with the increasing number of member stations. It is bounded by the network size.

6.4 Performance Evaluation

In this section, we first describe the system configurations deployed in the simulation and then evaluate the performance of both the proposed NSRA scheme and the LR al-

gorithm with corresponding reference schemes or algorithms.

6.4.1 System configurations

We simulate three types of wireless networks (i.e., NX, NC, and NN networks), which can be seen as WiMAX, cellular, and WLAN networks, respectively. The key parameters of these networks are listed in Table 6.3. Taking the NX network as an example, there are 7 burst profiles available for the transmission of the stream, in the ascending order of robustness. As more robust burst profile corresponds to the lower data rate, the amount of radio resources consumed with different burst profiles are also listed correspondingly. For instance, if the most robust burst profile uses BPSK 1/2 and the least robust burst profile uses QAM64 3/4 as in WiMAX, their corresponding throughputs are 3.8768 Mbps and 34.904 Mbps; therefore, we assume the amount of resources consumed with these two burst profiles are 10 and 1, respectively. In the simulation, the bandwidth cost of the stream under a burst profile (n, b) is calculated by $c_{n,b} = c_n \times r_{n,b}$, where c_n is the bandwidth cost per unit resource in the network n and $r_{n,b}$ is the amount of resource needed under (n, b) .

In the simulation, the stations move along a random route. At the beginning of each

Table 6.3 Key parameters used in the simulation.

NX network	
Network Radius	3000 m
Number of burst profiles	7
Resource needed with each burst profile $r_{n,b}$	{ 1, 2, 3, 4, 7, 9, 10 }
Velocity of stations	100 m/min
NC network	
Network Radius	2000 m
Number of burst profiles	5
Resource needed with each burst profile $r_{n,b}$	{ 1, 2, 3, 4, 5 }
Velocity of stations	100 m/min
NN network	
Radius R_{NN}	200 m
Number of burst profiles	3
Resource needed with each burst profile $r_{n,b}$	{ 1, 2, 4 }
Velocity of stations	15 m/min

minute, the station randomly decides whether to move and the direction of movement. Within NN networks that are usually deployed at indoor areas such as home and office, the stations move with a relatively low speed. While in the NX and NC networks at outdoor areas, they move faster. The active duration of a member station is uniformly distributed between 0 and 60 minutes. Each simulation runs for a duration of 3000 minutes.

6.4.2 Evaluation of the NSRA scheme

To the best of our knowledge, our NSRA scheme is the first piece of work considering both the network selection and resource allocation for multicasting/broadcasting in heterogeneous interworking networks. To evaluate the performance of the scheme, we compare it with other three reasonable reference schemes. In the first reference scheme (Ref1), each member station randomly connects to an available network for service. In the second reference scheme (Ref2), smarter stations are used that always connect to the available network with the minimum bandwidth cost in order to reduce monetary cost for users. The third reference scheme (Ref3) is centrally controlled from the viewpoint of network operators, where an incoming member station is accepted to the network leading to the minimum increase on the overall system bandwidth cost. By all these approaches, the stream is transmitted in the selected networks with the least robust burst profile available to their served member stations to minimize the bandwidth cost consumed. In the simulation, our scheme is performed in a centrally controlled manner.

To thoroughly investigate the performance of different schemes, we compare the schemes in multiple scenarios by changing various factors including:

- Network size and structure
We use one NX network and 30 NN networks to compose a two-tier overlay network; one NX network, four NC networks and 40 NN networks to compose a three-tier overlay network. The NN networks are evenly distributed in each burst profile region of the NX network. There are at most 100 member stations in the two-tier overlay network and 200 member stations in the three-tier overlay network.
- Mobility of stream

(a)

(b)

Figure 6.2 Two-tier model with resident application.

(a) $c_{NX} = c_{NN} = 1$; (b) $c_{NX} = 5, c_{NN} = 1$.

For some multicasting and broadcasting services, such as IPTV and video telephony, users usually stay in indoor areas (e.g., home or office). We compare the schemes in such resident scenario and general mobile scenario, respectively. In a resident scenario, the member stations are uniformly distributed in the NN networks and moves only within the coverage of these networks. In a general mobile scenario, the stations are randomly distributed within the NX network at the beginning and then move in random routes.

- Bandwidth cost

We adjust the bandwidth cost $c_{n,b}$ of burst profile (n, b) by changing the bandwidth cost per unit resource c_n in the network n , based on the equation $c_{n,b} = c_n \times r_{n,b}$

(a)

(b)

Figure 6.3 Two-tier model with general mobile application.

(a) $c_{NX}=10$, $c_{NN}=1$; (b) $c_{NX}=20$, $c_{NN}=1$.

where $r_{n,b}$ is given in Table 6.3. The networks at the same tier are configured with the same bandwidth cost, and thus c_n can also be written as c_{NX} , c_{NC} and c_{NN} .

Figure 6.2 to Figure 6.5 compare the schemes in eight different scenarios. From these simulation results, we can see that Ref1 always introduces the highest system bandwidth cost. By Ref1, the member stations at the same network region cannot be converged together due to the loss of cooperation. The overlay networks covering the same network region are selected by different member stations at the same time. Especially when more member stations request the stream, almost all the networks within the system are selected for transmission with the most robust burst profile deployed. The system bandwidth is seriously wasted as a result.

(a)

(b)

Figure 6.4 Three-tier model with resident application.

(a) $c_{NX} = c_{NC} = c_{NN} = 1$; (b) $c_{NX} = 5, c_{NC} = 2, c_{NN} = 1$.

Compared to Ref1, Ref2 and Ref3 are more intelligent but they seem to work relatively well in limited cases. Ref2 achieves similar performance with the proposed NSRA scheme when there are few member stations, and there is a relatively large difference between the bandwidth costs of different tier networks since as where member stations tend to converge to the network with lower bandwidth cost. However, by Ref2, the network selection is performed by individual member stations without the entire system view. There exists the case where multiple member stations connect to several NN networks separately due to their lower bandwidth cost; however, the total bandwidth cost consumed in these NN networks is higher than the cost of serving all these

(c)

(d)

Figure 6.5 Three-tier model with general mobile application.

(c) $c_{NX}=10$, $c_{NC}=5$, $c_{NN}=1$; (d) $c_{NX}=20$, $c_{NC}=10$, $c_{NN}=1$.

member stations with one NC or NX network. With the increase of the number of member stations, this problem becomes more serious, thus making the performance of Ref2 even approaching that of Ref1.

By Ref3, the radio resources within the system are centrally controlled and the possibility of transmitting the stream over the same region via multiple networks can be decreased. However, the network selection for a newly joined member station is influenced by the previous state of networks. For example, the NX network with a burst profile of bandwidth cost 10 is selected at some time for the new member station A. Following A, member station B joins the group and also connects to the NX network in order to avoid additional bandwidth cost introduced to the system. If A leaves the

group after a while, the NX network will still be maintained by B, although the NN network covering B only produces a bandwidth cost of 1. Moreover, the NX network may be further kept by the later member stations. When more stations join the group, the network with a larger coverage has more opportunities to be continually utilized for transmission. That is why the bandwidth cost of Ref3 tends to be remained at a level in most scenarios. Another example is that 10 member stations join the group one by one, and each member station is connected to a NN network with a bandwidth cost of 1, leading to a total bandwidth cost of 10. Nevertheless, if all these stations connect to one NC network, the overall cost is only 2. That is why Ref3 works badly when there are few stations in the system, and there is a big difference between the bandwidth costs of different tier networks.

It is clearly shown in the figures that our NSRA scheme always introduces the minimal bandwidth cost into the system as compared to three reference schemes. This is because our scheme can select the networks for transmission from the entire system perspective, with excellent adaptability to the change of network status and link conditions. Member stations within the same network region can be effectively converged into one network by this scheme. Especially for the case that some networks are configured with a higher bandwidth cost (e.g., becoming congested), the minimal bandwidth cost obtained by our scheme means that our scheme can distribute the traffic from the hot area to cold areas most effectively. In this context, our scheme works well as a load balancing tool.

Another important observation obtained from these results is that, although the overall system bandwidth cost also progressively increases by our NSRA approach with the joining of member stations, the maximum cost is known, which indeed is the bandwidth cost of the most robust burst profile of the tier-1 NX network. The worst case by our approach is to select the tier-1 network to serve all the member stations with the most robust burst profile, which guarantees the bandwidth cost consumed by our scheme to be below a threshold. In Figure 6.2 (a), the NX network has a comparable bandwidth cost with NN network; therefore, it is selected for transmission by our NSRA scheme and by Ref3 most of the time. In the similar case of the three-tier network shown in Figure 6.3 (a), the bandwidth cost of the NSRA scheme is still 10, while the bandwidth cost of Ref3 is increased to be 20 because the NC network is also selected by Ref3 during some periods, and thus the performance is not guaranteed.

Table 6.4 Greedy algorithm for the NSRA scheme.

-
1. let $M' = \phi$, $x_{n,b} \leftarrow 0$, $\forall n \in N$, $\forall b \in B_n$;
 2. find $(n', b') \leftarrow \arg \min_{n \in N, b \in B_n} \{c_{n,b} / |M_{n,b} \setminus M'|\}$; let $x_{n',b'} \leftarrow 1$.
 3. let $M' \leftarrow M' \cup M_{n',b'}$;
if $|M'| \neq |M|$, **then** return to step 2, **endif**.
 4. **for** any (n, b) with $x_{n,b} = 1$, (in descending order of $c_{n,b}$)
if any $m \in M_{n,b}$ can find a $(n', b') \neq (n, b)$ with $x_{n',b'} = 1$, **then** let $x_{n,b} \leftarrow 0$, **endif**;
endfor;
-

Figure 6.6 Comparison with greedy algorithm and CPLEX.

6.4.3 Evaluation of the LR algorithm

In this subsection, we evaluate the performance of our LR algorithm from several aspects. Firstly, we compare the algorithm with a greedy-based algorithm, which is another representative heuristic algorithm for solving the set covering problem. We write a greedy algorithm for finding solutions for our NSRA scheme in Table 6.4, which iteratively selects the most cost effective burst profile until all the member stations can be served. A two-tier overlay network is simulated. Meanwhile, the NCRA scheme is also solved by the software CPLEX [136]. The results obtained by the three approaches are compared in Figure 6.6. We can see that our scheme can achieve equivalent performance with CPLEX and better performance than the greedy algorithm especially

when there are more member stations in the system. The reason lies in that the greedy algorithm is known to be an $H(n)$ -approximation algorithm [153], where $H(n) = \sum_{k=1}^n (1/k)$ is the approximation ratio and n is the size of the set to be covered. With the member stations joining, more sets are covered leading to a higher $H(n)$; consequently, the total cost through the greedy algorithm goes farther away from the optimum. Compared to the commercial software CPLEX, our free algorithm runs with a faster speed (the average running time is about 1/3 of that of CPLEX). Meanwhile, it is more easily to be embedded into the system and upgraded due to its small size.

Furthermore, we compare the LR algorithm with the Lagrangian algorithm derived in [12], where only the network selection is performed for the multicasting in heterogeneous wireless networks without the resource allocation taken into account, and all the member stations in a network are assumed to be with the same bandwidth cost. For comparison, we configure each network with only one burst profile. A three-tier overlay network is simulated and all the Lagrangian related parameters (e.g., λ) used in these two algorithms are identical. In the simulation, we found that if the network bandwidth costs are relatively low, the performances of the two algorithms are similar. Otherwise, more iterations are usually needed by the algorithm in [12] to obtain the optimal solution, because it reduces the upper bound slowly from a larger value to the optimum. Figure 6.7 compares the performance of the two algorithms, where c_{NX} is 20, c_{NC} is uniformly selected from 1 to 10, and c_{NN} is uniformly selected from 1 to 3. It is shown in Figure 6.7(b) that much less iterations are needed by our scheme, especially when there are not too many member stations in the system. When the member stations are less than 50, the number of iterations needed by the algorithm in [12] reaches the threshold $T = 2000$, and then only near-optimal solutions could be found at that time as can be seen from Figure 6.7(a). The possible reason is that the algorithm in [12] depends on the member stations to select network; thus, decreasing the member station number slows down the speed of the convergence of upper bound and lower bound. Conversely, our algorithm can always find the optimum with much less iterations.

6.5 Summary

In this chapter, a novel network selection and resource allocation scheme was proposed for the multicasting/broadcasting in heterogeneous wireless networks. The objective is

(a)

(b)

Figure 6.7 Comparison with the algorithm in [12].

to minimize the overall bandwidth cost consumed by the multicast/broadcast stream in the overlay system. We formulated the scheme as a binary integer programming problem, and then derive a Lagrangian-based algorithm to solve the problem. By the scheme, the stations at the same region can be effectively converged together transmitting the stream under possible highest modulation rates to reduce the resource consumption. The simulation results demonstrated that our scheme introduces the minimum bandwidth cost into the system; meanwhile our derived algorithm finds good solutions for the scheme with lower computational cost, as compared with other reference schemes and algorithms.

Chapter 7

Summary and Future Works

In this chapter, the main contributions of this thesis will be summarized and several future research topics will be discussed briefly.

7.1 Summary of Contributions

In this thesis, we investigate several important issues for the integration of heterogeneous wireless networks, including interworking architecture, vertical handoff and radio resource management. The objective is to extensively explore the complementary characteristics of different networks and effectively manage the overall radio resources within the system, so that both the users and network operators could profit from the interworking of heterogeneous networks. The major contributions are summarized as follows.

- The integration of WiMAX and WLAN networks are investigated, with loosely and tightly coupled interworking architectures addressed, respectively. With a loosely coupled approach, WiMAX and WLAN connect to Internet independently; while with a tightly coupled approach, WLAN AP connects to ASN GW of WiMAX directly just as an access network of WiMAX. Our investigation shows that the tightly coupled interworking architecture is preferential owing to its better performance on implementation cost, joint radio resource management, and seamless vertical handoff support provided by the link layer connectivity. For the implementation of such architecture, the key mechanisms are introduced such as establishing the data path between ASN GW and AP, and combining the security functions of WiMAX and WLAN networks. The proposed architecture provides a platform for the schemes based on the tightly coupled approach.

- A QoS-oriented vertical handoff scheme is designed for the WiMAX/WLAN overlay networks. The designed handoff protocol is based on the proposed tightly coupled interworking architecture, where a seamless handoff is achieved by executing the handoff at link layer. Both the performance of applications and network conditions are considered in vertical handoffs. When the service provided by the connected network cannot satisfy the QoS requirements, the station may switch to another network to seek better performance for current applications. The aim is to always obtain the best service for both mobile and fixed users. Simulation studies have demonstrated that the proposed scheme can keep the stations being best connected.

Furthermore, algorithms are derived to estimate the available bandwidth and packet delay provided by WiMAX and WLAN, respectively, which are deployed by the stations in the proposed vertical handoff scheme to evaluate the network conditions for making a handoff decision. They can also be used by other approaches.

- The radio resource management in general heterogeneous wireless overlay networks is studied from the perspective of network operators. A joint radio resource management model is addressed. To effectively utilize the overall radio resources pooled in the interworking system, an efficient load balancing scheme is proposed to be applied in the call admission control process. If a call arrives at the interworking system, where the available networks cannot provide it enough resource without degrading the QoS of the ongoing calls, the system will perform load balancing operations by proactively initiating vertical handoffs among networks in order to create more room for the request call. The load balancing algorithm is to minimize the variance between the network utilizations of the entire system. By our scheme, more calls can be admitted into the system compared to other reference schemes; therefore, the overall system throughput can be improved. The networks within the system can also work in more effective states and provide better QoS supports for users.
- Specially, the radio resource management for the multicasting/broadcasting service in heterogeneous interworking systems is investigated, because such service usually has a large group of members distributed in many networks within the sys-

tem, which significantly influences the system resource utilization. A novel scheme is designed to select suitable networks for transmitting the multicast/broadcast stream and allocate proper resources for the transmission in the selected networks. The scheme aims to minimize the overall bandwidth cost consumed by the stream within the entire system, which is formulated as an optimization problem and then solved by the derived Lagrangian-based heuristic algorithm. By our approach, the member stations located in the same network region could be converged together to ensure as few networks as possible to be involved in the transmission. Meanwhile, the stream can be transmitted with high enough modulation and coding rates in order to reduce the consumed radio resources. Simulation results demonstrate that our network selection and resource allocation scheme outperforms other reference schemes on saving system resources; moreover, the algorithm for solving the scheme can find good solutions with a reasonably low computational cost.

7.2 Future Works

The interworking of heterogeneous networks is expected to be a promising and feasible approach for the next-generation wireless networks to provide always the best service to users anytime and anywhere, based on the existing network architectures. Increasing interest is currently paid on the performance optimization of integrated networks, but still with many research topics to be investigated thoroughly in the future. Our forthcoming research works can be performed from various aspects. Here, we list several important topics to finish this thesis.

- In heterogeneous overlay networks, the stations with multiple interfaces (i.e., multimode terminals) have opportunities to connect to different networks simultaneously. This concurrent access to multiple networks is known as *multihoming*. In the thesis, this feature is deployed by the vertical handoff scheme to avoid packet loss during handoffs. With the development of relevant technologies (e.g., on the battery life), it is believed that the future multimode terminal could connect to multiple networks with its traffic distributed to different networks during a long period. Therefore, we could extend our vertical handoff scheme for the handoff of per-application running at the multimode station. When the serving network can-

not satisfy the QoS requirement of one application, a handoff decision could be made to transfer this application to another network while keeping other applications in the current network. Correspondingly, new mobility management protocols are needed to perform a *movement* of an application running at the station. Meanwhile, the admission control and resource allocation scheme could be further enhanced to select network for each call at a multimode station. The calls from the same station may be accepted into different networks from the network point of view, with the consideration of various factors such as the QoS requirement of calls, the traffic load of networks, and the preference on networks of different calls.

- We have known that the radio resource management for the multicast/broadcast service is extremely important for the resource utilization of heterogeneous interworking systems. In the thesis, we investigate how to select networks and allocate resources for a multicast/broadcast stream in order to minimize the overall system bandwidth resource consumed by it. In the future, we will investigate how to allocate the limited radio resources pooled in the system to all the multicast/broadcast streams. The objective is to serve as many users as possible and then improve the revenue obtained by network operator. The factors to be taken into account include the number of member stations of each stream, the stream characteristics (e.g., throughput, importance), the channel conditions of member stations, etc.
- The multicasting/broadcasting of video streams over wireless networks has attracted considerable research attention due to the proliferation of advanced services like IPTV and video conferencing. To adapt to various conditions and requirements of group members, video encoding techniques has enabled encoding one video stream into layers, including the base layer and enhancement layer(s). The base layer is the most significant layer, which contains the data with the most important features of the program. The enhancement layers include the data that progressively refine the reconstructed video program. The more the number of layers a user has received, the better the video quality the user can experience.

Such layering approach provides much better flexibility in network selection and resource allocation for the multicasting/broadcasting of video streams over heterogeneous wireless networks. However, new challenges are also introduced

by the flexibility. For instance, a user may connect to the network with guaranteed service but higher monetary cost to receive the base layer, meanwhile connect to another network with degraded service but lower cost to receive enhancement layers. Meanwhile, the network operator may allocate more resources to the stream with a larger group to transmit more enhancement layers of this stream, such that more users could receive the video with a better quality. The objective is to maximize the utility obtained in the entire system. Also, our NSRA scheme can be extended to select networks and allocate radio resource for each layer of the stream in order to minimize the radio resource consumed by each layer of the stream.

List of Publications

- [1] **Dong Ma** and Maode Ma. Resource efficient multicasting/broadcasting in heterogeneous wireless networks. *Computer Networks* (in revision).
- [2] **Dong Ma** and Maode Ma. A QoS oriented vertical handoff scheme for WiMAX/WLAN overlay networks. *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 4, pp. 598-606, 2012.
- [3] **Dong Ma** and Maode Ma. Proactive load balancing with admission control for heterogeneous overlay networks. *Wireless Communication and Mobile Computing* (published online on 3 Nov 2011). Doi: 10.1002/wcm.1224.
- [4] Maode Ma and **Dong Ma**, “A remote healthcare system supported by heterogeneous wireless networks,” *Journal of Medical Imaging and Health Informatics*, vol. 1, no. 4, pp. 371-374, 2011.
- [5] **Dong Ma** and Maode Ma. A QoS-based vertical handoff scheme for interworking of WLAN and WiMAX. *IEEE Global Telecommunications Conference (GLOBECOM) 2009*, Honolulu, vol. 1–8, pp. 2260–2265, 2009.
- [6] Maode Ma and **Dong Ma**. Design of a heterogeneous wireless network-based remote healthcare system. *Proceedings of the First International Symposium on Bioengineering*, Singapore, pp. 263-270, 2011.

Bibliography

- [1] M. Ergen, *Mobile Broadband: Including WiMAX and LTE*, Springer, 2009.
- [2] P. Nicopolitidis, M. S. Obaidat, G. I. Papadimitriou, *et al.*, *Wireless Networks*, John Wiley & Sons, 2003.
- [3] B. G. Lee and S. Choi, *Broadband Wireless Access and Local Networks: Mobile WiMAX and WiFi*, Artech House, 2008.
- [4] E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless Communications*, vol. 10, pp. 49-55, 2003.
- [5] W. Song, *Resource Allocation for Cellular/WLAN Integrated Networks*, PhD thesis, Waterloo, 2007.
- [6] C. Makaya and S. Pierre, "An architecture for seamless mobility support in IP-based next-generation wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 1209-1225, 2008.
- [7] J. G. Atallah and M. Ismail, "Future 4G front-ends enabling smooth vertical handovers," *IEEE Circuits and Devices Magazine*, vol. 22, pp. 6-15, 2006.
- [8] D. Ma and M. Ma, "A QoS-based vertical handoff scheme for interworking of WLAN and WiMAX," in *IEEE Global Telecommunications Conference (GLOBECOM) 2009*, vol. 1-8, pp. 2260-2265, 2009.
- [9] D. Ma and M. Ma, "A QoS oriented vertical handoff scheme for WiMAX/WLAN overlay networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 4, pp. 598-606, 2012.
- [10] D. Ma and M. Ma, "Proactive load balancing with admission control for heterogeneous overlay networks," *Wireless Communications and Mobile Computing*, published online on 3 Nov 2011, Doi: 10.1002/wcm.1224, 2011.
- [11] D. Ma and M. Ma, "Resource efficient multicasting/broadcasting in heterogeneous wireless networks," *Computer Networks*, in revision.
- [12] D. N. Yang and M. S. Chen, "Efficient resource allocation for wireless multicast," *IEEE Transactions on Mobile Computing*, vol. 7, pp. 387-400, 2008.
- [13] H. Honkasalo, K. Pehkonen, M. T. Niemä, *et al.*, "WCDMA and WLAN for 3G and beyond," *IEEE Wireless Communications*, vol. 9, pp. 14-18, 2002.
- [14] M. M. Buddhikot, G. Chandranmenon, S. Han, *et al.*, "Design and implementation of a WLAN/CDMA2000 interworking architecture," *IEEE Communications Magazine*, vol. 41, pp. 90-100, 2003.
- [15] K. Ahmavaara, H. Haverinen, and R. Pichna, "Interworking architecture between 3GPP and WLAN systems," *IEEE Communications Magazine*, vol. 41, pp. 74-81, 2003.
- [16] ETSI TR 101 957, Requirements and Architecture for Internetworking between HIPERLAN/2 and 3rd Generation Cellular Systems, 2001.
- [17] C. Perkins, "IP mobility support for IPv4," IETF RFC 3344, 2002.
- [18] C. de Laat, C. Gross, L. Gommans, *et al.*, "Generic AAA architecture," IETF RFC 2903, 2000.
- [19] A. K. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS integration for next-generation mobile data networks," *IEEE Wireless Communications*, vol. 9, pp. 112-124, 2002.
- [20] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/WLAN integrated

- networks," *IEEE Network*, vol. 21, pp. 27-33, 2007.
- [21] V. S. Azhari, M. Smadi, and T. D. Todd, "Fast client-based connection recovery for soft WLAN-to-cellular vertical handoff," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 1089-1102, 2008.
 - [22] J. Y. Song, H. J. Lee, S. H. Lee, *et al.*, "Hybrid coupling scheme for UMTS and wireless LAN interworking," *AEU-International Journal of Electronics and Communications*, vol. 61, pp. 329-336, 2007.
 - [23] X. Gelabert, J. Pérez-Romero, O. Sallent, *et al.*, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *IEEE Transactions on Mobile Computing*, vol. 7, pp. 1257-1270, 2008.
 - [24] G. Lampropoulos, N. Passas, L. Merakos, *et al.*, "Handover management architectures in integrated WLAN/cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 7, pp. 30-44, 2005.
 - [25] K. Sabnani, "Converged networks of the future," in *NSF Wireless/Mobile Planning Group Wksp.*, 2005.
 - [26] 3GPP TS 43.318, Radio Access Network; Generic Access Network; Stage 2 (Release 10), v10.1.0, 2011.
 - [27] F. Xu, L. Zhang, and Z. Zhou, "Interworking of WiMAX and 3GPP networks based on IMS [IP Multimedia Systems (IMS) infrastructure and services]," *IEEE Communications Magazine*, vol. 45, pp. 144-150, 2007.
 - [28] S. Rizvi, A. Aziz, N. M. Saad, *et al.*, "Tight coupling interworking between UMTS and WLAN: challenges, design architectures and simulation analysis," *International Journal of Computer Networks*, vol. 3, pp. 116-134, 2011.
 - [29] B. Liu, P. Martins, A. E. Samhat, *et al.*, "A layer 2 scheme for Inter-RAT handover between UMTS and WiMAX in tight coupling architecture," in *Proceedings of IEEE PIMRC*, 2008, pp. 1-5.
 - [30] B. Cha, G. H. Kim, Y. Choi, *et al.*, "A vertical handoff algorithm using MN velocity for WiBro and WLAN," in *Proceedings of IEEE ICCIT*, 2008, pp. 226-232.
 - [31] N. Vulic, S. M. Heemstra de Groot, and I. G. M. M. Niemegeers, "Vertical handovers among different wireless technologies in a UMTS radio access-based integrated architecture," *Computer Networks*, 2011.
 - [32] A. V. Garmonov, S. H. Cheon, K. T. Han, *et al.*, "QoS-oriented intersystem handover between IEEE 802.11b and overlay networks," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 1142-1154, 2008.
 - [33] S. K. Lee, K. Sriram, K. Kim, *et al.*, "Vertical handoff decision algorithms for providing optimized performance in heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 58, pp. 865-881, 2009.
 - [34] W. Song, H. Jiang, and W. Zhuang, "Performance analysis of the WLAN-first scheme in cellular/WLAN interworking," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 1932-1952, 2007.
 - [35] C. Guo, Z. Guo, Q. Zhang, *et al.*, "A seamless and proactive end-to-end mobility solution for roaming across heterogeneous wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 834-848, 2004.
 - [36] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 1243-1254, 2008.
 - [37] Q. Zhang, C. Guo, Z. Guo, *et al.*, "Efficient mobility management for vertical handoff between WWAN and WLAN," *IEEE Communications Magazine*, vol. 41, pp. 102-108, 2003.
 - [38] B. J. Chang and J. F. Chen, "Cross-layer-based adaptive vertical handoff with predictive RSS in heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 3679-3692, 2008.

- [39] W. Lee, E. Kim, J. Kim, *et al.*, "Movement-aware vertical handoff of WLAN and mobile WiMAX for seamless ubiquitous access," *IEEE Transactions on Consumer Electronics*, vol. 53, pp. 1268-1275, 2007.
- [40] M. Nam, N. Choi, Y. Seok, *et al.*, "WISE: energy-efficient interface selection on vertical handoff between 3G networks and WLANs," in *15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2004, pp. 692-698.
- [41] I. Kanno, K. Yamazaki, Y. Ikeda, *et al.*, "Adaptive energy centric radio access selection for vertical handover in heterogeneous networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2010, pp. 1-6.
- [42] O. Ormond, J. Murphy, and G. M. Muntean, "Utility-based intelligent network selection in beyond 3G systems," in *IEEE International Conference on Communications (ICC)* 2006, pp. 1831-1836.
- [43] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multinetwork environments," *IEEE Wireless Communications*, vol. 11, pp. 8-15, 2004.
- [44] X. Guan, R. Tang, S. Bai, *et al.*, "Enhanced application-driven vertical handoff decision scheme for 4G wireless networks," in *International Conference on Wireless Communications, Networking and Mobile Computing (WiCom)*, 2007, pp. 1771-1774.
- [45] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in fourth generation heterogeneous networks," *IEEE Communications Magazine*, vol. 44, pp. 96-103, 2006.
- [46] S. Mitra, "Policy based admission control and handoff decision algorithm for next generation all-IP wireless network," in *International Conference on Electrical and Computer Engineering (ICECE)*, 2008, pp. 594-599.
- [47] D. Lee, Y. Han, and J. Hwang, "QoS-based vertical handoff decision algorithm in heterogeneous systems," in *IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*, 2006, pp. 1-5.
- [48] W. T. Chen and Y. Y. Shu, "Active application oriented vertical handoff in next-generation wireless networks," in *IEEE Wireless Communications and Networking Conference*, 2005, pp. 1383-1388.
- [49] W. Shen and Q. A. Zeng, "Cost-function-based network selection strategy in integrated wireless and mobile networks," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 3778-3788, 2008.
- [50] N. Nasser, S. Guizani, and E. Al-Masri, "Middleware vertical handoff manager: A neural network-based solution," in *IEEE International Conference on Communications (ICC)*, 2007, pp. 5671-5676.
- [51] W. Zhang, "Handover decision using fuzzy MADM in heterogeneous networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2004, pp. 653-658.
- [52] J. Zhang, H. C. B. Chan, and V. C. M. Leung, "WLC14-6: a location-based vertical handoff decision algorithm for heterogeneous mobile networks," in *IEEE Global Telecommunications Conference (GLOBECOM)* 2006, pp. 1-5.
- [53] J. Hou and D. C. O'Brien, "Vertical handover-decision-making algorithm using fuzzy logic for the integrated Radio-and-OW system," *IEEE Transactions on Wireless Communications*, vol. 5, pp. 176-185, 2006.
- [54] M. Liu, Z. Li, X. Guo, *et al.*, "Performance analysis and optimization of handoff algorithms in heterogeneous wireless networks," *IEEE Transactions on Mobile Computing*, vol. 7, pp. 846-857, 2008.
- [55] X. Gao, G. Wu, and T. Miki, "End-to-end QoS provisioning in mobile heterogeneous networks," *IEEE Wireless Communications*, vol. 11, pp. 24-34, 2004.
- [56] S. Y. Hui and K. H. Yeung, "Challenges in the migration to 4G mobile systems," *IEEE Communications Magazine*, vol. 41, pp. 54-59, 2003.
- [57] C. W. Lee, L. M. Chen, M. C. Chen, *et al.*, "A framework of handoffs in wireless overlay networks based on Mobile IPv6," *IEEE Journal on Selected Areas in*

- Communications*, vol. 23, pp. 2118-2128, 2005.
- [58] M. Abbassi, S. Khan, and M. Rahman, "A distributed signaling fast Mobile IPv6 scheme for next generation heterogeneous IP networks," *Lecture Notes in Computer Science*, vol. 7291/2012, pp. 43-51, 2012.
 - [59] X. Hu, Y. Bi, J. Huang, *et al.*, "Study on unified under layer trigger mechanism for handover optimization in mobile IP networks," in *2005 2nd International Conference on Mobile Technology, Applications and Systems*, 2005, pp. 1-6.
 - [60] K. S. Kong, W. Lee, Y. H. Han, *et al.*, "Mobility management for all-IP mobile networks: Mobile IPv6 vs. Proxy Mobile IPv6," *IEEE Wireless Communications*, vol. 15, pp. 36-45, 2008.
 - [61] W. S. Lim, D. W. Kim, Y. J. Suh, *et al.*, "Implementation and performance study of IEEE 802.21 in integrated IEEE 802.11/802.16e networks," *Computer Communications*, vol. 32, pp. 134-143, 2009.
 - [62] A. Pontes, D. dos Passos Silva, J. Jailton, *et al.*, "Handover management in integrated WLAN and mobile WiMAX networks," *IEEE Wireless Communications*, vol. 15, pp. 86-95, 2008.
 - [63] IEEE P802.21/D10.0, "Draft standard for local and metropolitan area networks: media independent handover services," 2008.
 - [64] K. Jang, B. Shin, and D. Hong, "Fast vertical handover utilizing sleep mode in WLAN and WiMAX heterogeneous networks," in *14th Asia-Pacific Conference on Communications (APCC)* 2008, pp. 1-5.
 - [65] J. Nie, X. He, Z. Zhou, *et al.*, "Benefit-driven handoffs between WMAN and WLAN," in *IEEE Military Communications Conference (MILCOM)*, 2005, pp. 2223-2229.
 - [66] J. Roy, V. Vaidehi, and S. Srikanth, "Always best-connected QoS integration model for the WLAN, WiMAX heterogeneous network," in *First International Conference on Industrial and Information Systems*, 2006, pp. 361-366.
 - [67] L. C. Wang, A. Chen, and H. H. Chen, "Network selection with joint vertical and horizontal handoff in heterogeneous WLAN and mobile WiMAX systems," in *IEEE 65th Vehicular Technology Conference (VTC)*, 2007, pp. 794-798.
 - [68] Y. Choi and S. Choi, "Service charge and energy-aware vertical handoff in integrated IEEE 802.16e/802.11 networks," in *26th IEEE International Conference on Computer Communications (INFOCOM)*, 2007, pp. 589-597.
 - [69] S. F. Yang and J. S. Wu, "Handoff management schemes across hybrid WiMAX and Wi-Fi networks," in *IEEE Region 10 Conference (TENCON)*, 2007, pp. 1-4.
 - [70] L. Giupponi, R. Agust í J. Pérez-Romero, *et al.*, "A framework for JRRM with resource reservation and multiservice provisioning in heterogeneous networks," *Mobile Networks and Applications*, vol. 11, pp. 825-846, 2006.
 - [71] R. Skehill, M. Barry, W. Kent, *et al.*, "The common RRM approach to admission control for converged heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 14, pp. 48-56, 2007.
 - [72] 3GPP TR 25.881. Improvements of RRM across RNS and RNC/BSS V5.0.0. 2001.
 - [73] F. Yu and V. Krishnamurthy, "Optimal joint session admission control in integrated WLAN and CDMA cellular networks with vertical handoff," *IEEE Transactions on Mobile Computing*, vol. 6, pp. 126-139, 2007.
 - [74] J. Luo, R. Mukerjee, M. Dillinger, *et al.*, "Investigation of radio resource scheduling in WLANs coupled with 3G cellular network," *IEEE Communications Magazine*, vol. 41, pp. 108-115, 2003.
 - [75] A. Hasib and A. Fapojuwo, "Joint radio resource management over very tightly coupled heterogeneous networks for multimode reconfigurable terminals," in *IEEE 63rd Vehicular Technology Conference (VTC)*, 2006, pp. 727-731.
 - [76] A. Hasib and A. O. Fapojuwo, "Analysis of common radio resource management scheme for end-to-end QoS support in multiservice heterogeneous wireless networks," *IEEE*

- Transactions on Vehicular Technology*, vol. 57, pp. 2426-2439, 2008.
- [77] H. Chen, S. Kumar, and C. C. J. Kuo, "Dynamic call admission control scheme for QoS priority handoff in multimedia cellular systems," in *IEEE Wireless Communications and Networking Conference (WCNC) 2002*, pp. 114-118.
 - [78] H. Chen, S. Kumar, and C. C. J. Kuo, "Differentiated QoS-aware priority handoff in cell-based multimedia wireless network," in *Proceedings of SPIE*, 2000, p. 940.
 - [79] 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, pp. 77-92, 1986.
 - [80] S. K. Lee, K. Kim, K. Hong, *et al.*, "A probabilistic call admission control algorithm for WLAN in heterogeneous wireless environment," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 1672-1676, 2009.
 - [81] D. K. Kim, D. Griffith, and N. Golmie, "A new call admission control scheme for heterogeneous wireless networks," *IEEE Transactions on Wireless Communications*, vol. 9, pp. 3000-3005, 2010.
 - [82] W. Song, Y. Cheng, and W. Zhuang, "Improving voice and data services in cellular/WLAN integrated networks by admission control," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 4025-4037, 2007.
 - [83] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/WLAN integrated network," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 725-735, 2009.
 - [84] Y. R. Haung and J. M. Ho, "Distributed call admission control for a heterogeneous PCS network," *IEEE Transactions on Computers*, vol. 51, pp. 1400-1409, 2002.
 - [85] B. Tarek and N. Nidal, "Efficient call admission control scheme for 4G wireless networks," *Wireless Communications and Mobile Computing*, vol. 9, pp. 489-499, 2009.
 - [86] Y. H. Chen, C. J. Chang, and C. Y. Huang, "Fuzzy Q-learning admission control for WCDMA/WLAN heterogeneous networks with multimedia traffic," *IEEE Transactions on Mobile Computing*, vol. 8, pp. 1469-1479, 2009.
 - [87] W. Li and X. Chao, "Call admission control for an adaptive heterogeneous multimedia mobile network," *IEEE Transactions on Wireless Communications*, vol. 6, pp. 515-525, 2007.
 - [88] S. Sen, J. Jawanda, K. Basu, *et al.*, "Quality-of-Service degradation strategies in multimedia wireless networks," in *IEEE 48th Vehicular Technology Conference (VTC) 1998*, pp. 1884-1888.
 - [89] 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, pp. 5-17, 2004.
 - [90] O. E. Falowo, S. Zeadally, and H. A. Chan, "Dynamic pricing for load balancing in user centric joint call admission control of next generation wireless networks," *International Journal of Communication Systems*, vol. 23, pp. 335-368, 2010.
 - [91] X. Gelabert, J. Pérez-Romero, O. Sallent, *et al.*, "On the suitability of load balancing principles in heterogeneous wireless access networks," in *Proceedings of WPMC*, 2005.
 - [92] A. Pillekeit, F. Derakhshan, E. Jugl, *et al.*, "Force-based load balancing in co-located UMTS/GSM networks," in *IEEE 60th Vehicular Technology Conference (VTC)*, 2004, pp. 4402-4406.
 - [93] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: an evolutionary game approach," *IEEE Transactions on Vehicular Technology*, vol. 58, pp. 2008-2017, 2009.
 - [94] Y. H. Chen, N. Y. Yang, C. J. Chang, *et al.*, "A utility function-based access selection method for heterogeneous WCDMA and WLAN networks," in *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1-5, 2007.

- [95] Y. Zhou, Y. Rong, H. A. Choi, *et al.*, "Utility-based load balancing in WLAN/UMTS internetworking systems," in *IEEE Radio and Wireless Symposium*, 2008, pp. 587-590.
- [96] A. Toskala and H. Holma, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*: Wiley, 2002.
- [97] Y. C. Lai, P. Lin, Y. Fang, *et al.*, "Channel allocation for UMTS multimedia broadcasting and multicasting," *IEEE Transactions on Wireless Communications*, vol. 7, pp. 4375-4383, 2008.
- [98] S. Choi, Y. Choi, and I. Lee, "IEEE 802.11 MAC-level FEC scheme with retransmission combining," *IEEE Transactions on Wireless Communications*, vol. 5, pp. 203-211, 2006.
- [99] M. Sun, L. Huang, S. Wang, *et al.*, "Reliable MAC layer multicast in IEEE 802.11 wireless networks," *Wireless Communications and Mobile Computing*, vol. 3, pp. 439-453, 2003.
- [100] B. S. Kim, S. W. Kim, and R. L. Ekl, "OFDMA-based reliable multicasting MAC protocol for WLANs," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 3136-3145, 2008.
- [101] T. Jiang, W. Xiang, H. H. Chen, *et al.*, "Multicast broadcast services support in OFDMA-based WiMAX systems [Advances in mobile multimedia]," *IEEE Communications Magazine*, vol. 45, pp. 78-86, 2007.
- [102] J. Lee, S. Pack, T. Kwon, *et al.*, "Reducing handover delay by location management in Mobile WiMAX multicast and broadcast services," *IEEE Transactions on Vehicular Technology*, pp. 1-1, 2011.
- [103] F. Hou, L. X. Cai, P. H. Ho, *et al.*, "A cooperative multicast scheduling scheme for multimedia services in IEEE 802.16 networks," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 1508-1519, 2009.
- [104] D. N. Yang and M. S. Chen, "Bandwidth efficient video multicasting in multiradio multicellular wireless networks," *IEEE Transactions on Mobile Computing*, vol. 7, pp. 275-288, 2008.
- [105] D. N. Yang and M. S. Chen, "Data broadcast with adaptive network coding in heterogeneous wireless networks," *IEEE Transactions on Mobile Computing*, vol. 8, pp. 109-125, 2009.
- [106] ANSI/IEEE Std 802.11-1999 Edition (R2003), "Information technology-telecommunications and information exchange between systems- local and metropolitan area networks- specific requirements- Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications," 2003.
- [107] IEEE 802.16-2005, "IEEE standard for local and metropolitan area networks - Part 16: air interface for fixed broadband wireless access systems for mobile users", 2005.
- [108] WiMAX Forum, "Network architecture - stage 2: architecture tenets, reference model and reference points base specification" 2012.
- [109] K. Etemad, "Overview of mobile WiMAX technology and evolution," *IEEE Communications Magazine*, vol. 46, pp. 31-40, 2008.
- [110] D. Niyato and E. Hossain, "Wireless broadband access: WiMAX and beyond-integration of WiMAX and WiFi: optimal pricing for bandwidth sharing," *IEEE Communications Magazine*, vol. 45, pp. 140-146, 2007.
- [111] M. Ibrahim, K. Khawam, A. E. Samhat, *et al.*, "Analytical framework for dimensioning hierarchical WiMAX-WiFi networks," *Computer Networks*, vol. 53, pp. 299-309, 2009.
- [112] M. Riegel, "Ethernet services over mobile WiMAX," *IEEE Communications Magazine*, vol. 46, pp. 86-93, 2008.
- [113] IEEE Std 802.11i/D4.1, "Wireless medium access control (MAC) and physical layer (PHY) specifications: medium access control (MAC) security enhancements," July 2003.
- [114] IEEE Std 802.1X-2001, "Port-based network access control," 2001.
- [115] B. Aboba, L. Blunk, J. Vollbrecht, *et al.*, "Extensible authentication protocol (EAP)," IETF RFC 3748, 2004.

- [116] M. J. Kim, S. W. Son, and B. H. Rhee, "QoS based provisioning vertical handover between IEEE 802.11 and 802.16," in *IEEE 11th International Conference on Advanced Communication Technology (ICACT)* 2009, pp. 1415-1418.
- [117] L. De Vito, S. Rapuano, and L. Tomaciello, "One-way delay measurement: state of the art," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, pp. 2742-2750, 2008.
- [118] B. Ngamwongwattana and R. Thompson, "Sync & sense: VoIP measurement methodology for assessing one-way delay without clock synchronization," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, pp. 1318-1326, 2010.
- [119] D. Constantinescu, P. Carlsson, A. Popescu, *et al.*, "Measurement of one-way Internet packet delay," in *17th Nordic Teletraffic Seminar (NTS17)*, 2004.
- [120] D. C. Plummer, "An Ethernet address resolution protocol (ARP)," IETF RFC 826, 1982.
- [121] R. Prasad, C. Dovrolis, M. Murray, *et al.*, "Bandwidth estimation: metrics, measurement techniques, and tools," *IEEE Network*, vol. 17, pp. 27-35, 2003.
- [122] C. So-In, R. Jain, and A. K. Tamimi, "Scheduling in IEEE 802.16e mobile WiMAX networks: key issues and a survey," *IEEE Journal on Selected Areas in Communications*, vol. 27, pp. 156-171, 2009.
- [123] R. Jain, C. So-In, and A. Al Tamimi, "System-level modeling of IEEE 802.16e mobile WiMAX networks: key issues," *IEEE Wireless Communications*, vol. 15, pp. 73-79, 2008.
- [124] B. H. Kim, J. Yun, Y. Hur, *et al.*, "Capacity estimation and TCP performance enhancement over mobile WiMAX networks," *IEEE Communications Magazine*, vol. 47, pp. 132-141, 2009.
- [125] J. H. Yeh, J. C. Chen, and P. Agrawal, "Fast intra-network and cross-layer handover (FINCH) for WiMAX and mobile internet," *IEEE Transactions on Mobile Computing*, vol. 8, pp. 558-574, 2009.
- [126] D. Chuck and J. M. Chang, "Bandwidth recycling in IEEE 802.16 networks," *IEEE Transactions on Mobile Computing*, vol. 9, pp. 1451-1464, 2010.
- [127] Y. Zhang and M. Fujise, "Energy management in the IEEE 802.16 e MAC," *IEEE Communications Letters*, vol. 10, pp. 311-313, 2006.
- [128] H. C. Tijms and J. Wiley, *A First Course in Stochastic Models* vol. 2: Wiley Online Library, 2003.
- [129] Y. Q. Zhao and L. L. Campbell, "Equilibrium probability calculations for a discrete-time bulk queue model," *Queueing Systems*, vol. 22, pp. 189-198, 1996.
- [130] A. Janssen, J. Leeuwaarden, "Analytic computation schemes for the discrete-time bulk service queue," *Queueing Systems*, vol. 50, pp. 141-163, 2005.
- [131] T. Ohseki, M. Morita, and T. Inoue, "Burst construction and packet mapping scheme for OFDMA downlinks in IEEE 802.16 systems," in *IEEE Global Telecommunications Conference (GLOBECOM)* 2007, pp. 4307-4311.
- [132] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of IEEE 802.11 MAC protocols in wireless LANs," *Wireless Communications and Mobile Computing*, vol. 4, pp. 917-931, 2004.
- [133] A. Chinchuluun, P. M. Pardalos, R. Enkhbat, *et al.*: *Optimization and Optimal Control: Theory and Applications*, Springer, pp: 198-226, 2010.
- [134] A. Biagioni, R. Fantacci, D. Marabissi, *et al.*, "Adaptive subcarrier allocation schemes for wireless OFDMA systems in WiMAX networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, pp. 217-225, 2009.
- [135] W. Chu and Y. C. Tseng, "Performance analysis of IEEE 802.11 DCF in a multi-rate WLAN," *IEICE transactions on communications*, vol. 90, pp. 2836-2844, 2007.
- [136] CPLEX: *Mathematical Programming Optimizer*, <http://www-01.ibm.com/software/integ-ration/optimization/cplex-optimizer/>, 2010.
- [137] J. Liu, B. Li, Y. T. Hou, and I. Chlamtac, "On optimal layering and bandwidth allocation

- for multisession video broadcasting," *IEEE Transactions on Wireless Communications*, vol. 3, pp. 656-667, 2004.
- [138] T. Kwon, Y. Choi, and S. K. Das, "Bandwidth adaptation algorithms for adaptive multimedia services in mobile cellular networks," *Wireless Personal Communications*, vol. 22, pp. 337-357, 2002.
 - [139] W. H. Kuo, T. Liu, and W. Liao, "Utility-based resource allocation for layer-encoded IPTV multicast in IEEE 802.16 (WiMAX) wireless networks," in *IEEE International Conference on Communications (ICC)*, 2007, pp. 1754-1759.
 - [140] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, pp. 221-222, 1979.
 - [141] S. Umetani and M. Yagiura, "Relaxation heuristics for the set covering problem," *Journal of the Operations Research Society of Japan*, vol. 50, pp. 350-375, 2007.
 - [142] M. L. Fisher, "An applications oriented guide to Lagrangian relaxation," *Interfaces*, pp. 10-21, 1985.
 - [143] M. Held and R. M. Karp, "The traveling-salesman problem and minimum spanning trees: Part II," *Mathematical Programming*, vol. 1, pp. 6-25, 1971.
 - [144] J. E. Beasley, "Lagrangian relaxation", in C. R. Reeves (ed.): *Modern Heuristics Techniques for Combinatorial Problems*, Blackwell Scientific Publications, pp: 243-303, 1993.
 - [145] S. Lee, S. Seo, and N. Golmie, "An efficient power-saving mechanism for integration of WLAN and cellular networks," *IEEE Communications Letters*, vol. 9, no. 12, pp. 1-3, 2005.
 - [146] L. Nuaymi, *WiMAX Technology for Broadband Wireless Access*, John Wiley & Sons, pp. 45-46, 2007.
 - [147] G. R. Hiertz, D. Denteneer, L. Stibor, *et al.*, "The IEEE 802.11 universe, " *IEEE Communications Magazine*, vol. 48, no. 1, pp. 62-70, 2010.
 - [148] C. Cox, *An Introduction to LTE: LTE, LTE-Advanced, SAE, and 4G Mobile Communications*, John Wiley & Sons, 2012.
 - [149] H. Wu, C. Qiao, S. De, *et al.*, "Integrated cellular and ad hoc relaying systems: iCAR," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2105-2115, 2001.
 - [150] G. Ruggeri, A. Iera, and S. Polito, "802.11-based wireless-LAN and UMTS interworking: requirements, proposed solutions and open issues," *Computer Networks*, vol. 47, no. 2, pp. 151-165, 2005.
 - [151] M. Ylianttila, J. Makela, and K. Pahlavan, "Analysis of handoff in a location-aware vertical multi-access network," *Computer Networks*, vol. 47, no. 2, pp. 185-201, 2005.
 - [152] J. C. Chen and H. W. Lin, "A gateway approach to mobility integration of GPRS and wireless LANs," *IEEE Wireless Communications*, vol. 12, no. 2, pp. 86-95, 2005.
 - [153] V. Chvatal, "A greedy heuristic for the set covering problem," *Mathematics of Operations Research*, vol. 4, pp. 233-235, 1979.
 - [154] H. Soliman, C. Castelluccia, K. E. Malki, *et al.*, "Hierarchical Mobile IPv6 mobility management (HMIPv6)," IETF RFC 4140, 2005.
 - [155] G. Koodli, "Fast handovers for Mobile IPv6," IETF RFC 4068, 2005.

List of Abbreviation

3GPP	Third generation partnership project
AAA	Authentication, authorization and accounting
ABC	Always best connected
AC	Admission control
ACK	Acknowledgement
AMC	Adaptive modulation and coding
AMPS	Advanced mobile phone system
ANNs	Artificial neural networks
AP	Access point
ARP	Address resolution protocol
ASN	Access serving network
ASN-GW	ASN-gateway
BER	Bit error rate
BIP	Binary integer programming problem
BPSK	Binary phase shift keying
BS	Base station
BSS	Basic service set
BWR	Bandwidth request
CBR	Constant bit rate
CDMA	Code-division multiple access
CRMS	CRRM server
CRRM	Common RRM
CSMA/CA	Carrier-sense multiple access with collision avoidance
CSN	Connectivity serving network
CTM	Connection transition module
CTS	Clear-to-send
DCF	Distributed coordination function
DHCP	Dynamic host configuration protocol

DL	Downlink
EAP	Extensible authentication protocol
EAPOL	EAP over LAN
EDGE	Enhanced data rates for GSM evolution
FDD	Frequency division duplexing
FDMA	Frequency-division multiple access
FFT	Fast Fourier transform
FMIP	Fast handovers for Mobile IP
FQAC	Fuzzy Q-learning admission control
GAN	Generic access network
GPRS	General packet radio system
GRE	Generic routing encapsulation
GSM	Global system for mobile communications
HDM	Handoff decision module
HHO	Horizontal handoff
HMIP	Hierarchical Mobile IP
IE	Information element
IETF	Internet engineering task force
IS	Interim standard
JRRM	Joint RRM
JSAC	Joint session admission control
LLC	Logical link control
MAC	Medium access control
MADM	Multiple attribute decision-making
MBMS	Multimedia broadcast multicast service
MBS	Multicast broadcast service
MIH	Media-independent handoff
MIMO	Multi-input multi-output
MSK	Master session key
NAV	Network allocation vector
NCDM	Network condition detection module
NLOS	Non-line-of-sight
OFDM	Orthogonal frequency-division multiplexing

OFDMA	Orthogonal frequency division multiple access
PHY layer	Physical layer
PMIP	Proxy Mobile IP
PMK	Pairwise master key
PKMv2	Privacy key management version 2
PMP	Point-to-multipoint
QoS	Quality of service
RRC	Radio resource controller
RRM	Radio resource management
RSS	Received signal strength
RTS	Request-to-send
SIM	Subscriber identity module
SIR	Signal-to-interference ratio
SMDP	Semi-Markov decision
SNR	Signal-to-noise ratio
TDD	Time-division duplexing
TDMA	Time-division multiple access
TMM	Traffic measurement module
UL	Uplink
UMA	Unlicensed mobile access
UMTS	Universal mobile telecommunications system
URM	Unreserved resource for multicasting
USIM	User service identity module
VBR	Variable bit rate
VHO	Vertical handoff
VHOM	Vertical handoff manager
WCDMA	Wideband CDMA
WiMAX	Worldwide interoperability for microwave access
WLAN	Wireless local area network