UNIVERSITE D'EVRY VAL-D'ESSONNE

Laboratoire de Réseaux et Systèmes Multimédia

THÈSE

pour obtenir le grade

DOCTEUR de l'Université d'Evry Val-d'Essonne

Spécialité: Informatique

présentée et soutenue publiquement

par

Quoc-Thinh NGUYEN-VUONG

Le 2 Juillet 2008

Mobility Management in 4G Wireless Heterogeneous Networks

JURY

Directeur:	M. Nazim AGOULMINE	Professeur (Université d'Evry Val-d'Essonne)
Rapporteur:	M. Djamal ZEGHLACHE	Professeur (Telecom & Management SudParis)
Examinateur:	M. Liam MURPHY	Professeur (University College Dublin)
	M. Zwi ALTMAN	Ingénieur, chercheur (France Télécom)
	Mme. Sylvie MAYRARGUE	Ingénieur, chef de projet (CEA Léti)
	M. Linas MAKNAVICIUS	Ingénieur, chercheur (Alcatel-Lucent Bell Labs)

Abstract

In the forthcoming era of seamless mobility, people will have an easy, universal, uninterrupted access to information, entertainment and communication ... when, where and how they want it. The ability to provide a seamless transition across heterogeneous networks will enable a new level of customer experience. This thesis contributes to the evolution of technology convergence by improving different aspects of the vertical handover management to make seamless mobility a reality.

In the first part of the thesis, we address the role of user in the inter-system mobility management. We show how users with their smart mobile terminals can overcome some obstacles and improve the performance of vertical handovers. We propose a terminal-controlled handover management which is built on the top of a new utility-based access network selection. The terminal is shown to be able to control its radio interfaces to optimize the power consumption as well as to control the handover initiation and handover preparation to ensure seamless services. We propose a new handover prediction scheme to assist the handover preparation at the application level by the terminal itself.

In the second part of the thesis, we consider the role of network control in the inter-system mobility management. We study a UMTS-WiMAX interworking solution including integration architecture, handover procedure, inter-system measurement and required cell overlap for seamless handovers. We examine the interworking and roaming solution across independent access networks using intermediary entities. Last contributions include a new load definition and a new load balancing index in order to hide the heterogeneity of different access technologies from load balancing process.

Key words: Heterogeneous wireless networks, interworking architecture, roaming, inter-system mobility, always best connected, utility, load balancing, cell overlap planning, UMTS, LTE, WLAN, WiMAX.

Résumé

L'évolution des technologies réseaux sans fils et cellulaires associée au développement des terminaux mobiles est en train d'ouvrir de nouvelles perspectives pour offrir aux utilisateurs un accès universel et ininterrompu au réseau, aux informations et aux services. La capacité à fournir un accès ubiquitaire et une mobilité transparente à travers des réseaux hétérogènes permettra d'enrichir l'expérience des usagers. Néanmoins, la mise en place de cet environnement pose des challenges de recherche extrêmement importants que cette thèse a comme pour objectif d'aborder. Elle présente un ensemble de solutions qui contribuent à l'évolution de la convergence de technologies en améliorant différents aspects du handover vertical, de la synergie entre technologies cellulaires et sans fils afin que la mobilité sans couture devienne une réalité.

La première partie de la thèse examine le rôle de l'utilisateur dans la gestion de la mobilité inter-système. Nous montrons comment les utilisateurs avec leurs terminaux mobiles intelligents peuvent surmonter certains obstacles dans la gestion de handover vertical et améliorer ses performances. Une solution de la mobilité contrôlée par le terminal d'usager a été proposée. Elle consiste en un nouveau mécanisme de sélection de réseaux d'accès avec de nouvelles fonctions d'utilité. Le terminal est en mesure de contrôler ses interfaces radio pour optimiser la consommation d'énergie ainsi que de contrôler l'initiation et la préparation de handover pour assurer des services sans interruption. Une nouvelle méthode de prédiction du handover pour assister à la préparation d'un basculement sans couture est proposée.

Dans la deuxième partie de la thèse, il s'agit d'aborder le rôle du contrôle du réseau dans la gestion de la mobilité inter-système. Une solution d'interfonctionnement entre UMTS et WiMAX, qui comprend l'architecture d'intégration, les procédures exactes de handover, la mesure inter-système et le chevauchement nécessaire entre deux cellules avoisinantes pour assurer un handover sans couture est proposée. Pour faciliter l'interfonctionnement et l'itinérance (roaming) entre différents réseaux d'accès indépendamment et sans accords directes entre opérateurs, des plateformes intermédiaires ont été proposées. Une contribution finale qui consiste à introduire une nouvelle définition de la charge et un nouvel indice de l'équilibrage de charge qui permettent de définir une solution unifiée de répartition de charge dans un contexte de réseaux hétérogènes.

Mots clés: Réseaux sans fils hétérogènes, architecture d'interfonctionnement, itinérance, mobilité inter-système, always best connected, utilité, équilibrage de charge, chevauchement des cellules, UMTS, LTE, WLAN, WiMAX.

Acknowledgements

First of all, I would like to express my deepest sense of gratitude to my supervisor, Prof. Nazim Agoulmine, for his guidance, encouragement and excellent advice throughout this research work.

I am thankful to the members of my thesis committee, Djamal Zeghlache, Liam Murphy, Linas Maknavicius, Zwi Altman and Sylvie Mayrargue for the time and effort that they invested in judging the contents of my thesis.

I am grateful to Yacine Ghamri-Doudane for his helpful discussions which significantly contributed to the results presented in this thesis.

I also owe thanks to Laurent Ouvry and Sylvie Mayrargue, who arranged me a working place at CEA-Léti (Grenoble) during my second year of doctoral work. They provided me not only with the good working facilities but also the good opportunity to get involved in some interesting projects. My thanks also go to the colleagues at CEA-Léti with whom I shared many nice moments.

I wish to thank to Djamel Khadraoui for welcoming me in CRP Henri Tudor/CITI Labs (Luxembourg) during the six last months of my PhD process.

I am pleased to thank to all my colleagues at Networks and Multimedia Systems Research Group (LRSM) for their support and their comradeship; especially to Vamsi Krishma Gondi, Mehdi Nafa and Elyes Lehtihet, who worked closely with me in SEIMONET and SUMO projects.

Finally, I take this opportunity to express my profound gratitude to my beloved parents for their invaluable love and support throughout the years. I cannot thank you enough for your prayers, unwavering support, encouragement, and for always believing in me. Many thanks to my six brothers, my parents-in-law for their moral support.

Lastly, but actually most importantly, I would like to thank my wife Thanh-Ha, for her devotion, love, encouragement, and patience.

Contents

List of Figures

List of Tables

Abbreviations

"...writing about 4G technology is like making alphabet soup - lots of acronyms."
paperboy@riderresearch.com

3GPP Third Generation Partners Project

3GPP2 Third Generation Partnership Project 2

AAA Authentication, Authorization and Accounting

ABC Always Best Connected

AHP Analytic Hierarchy Process

AMPS Advanced Mobile Phone System

AP Access Point

APN Access Point Name

AR Access Router

ASN Access Service Network

ASN GW Access Service Network Gateway

BER Bit Error Rate

BPSK Binary Phase-Shift Keying

BS Base Station

BSC Base Station Controller

CBC Cell Broadcast Center

CDMA Code Division Multiple Access

CIP Cellular IP

CoA Care of Address

CN Corresponding Node

CPICH Common Pilot Channel

CRRM Common Radio Resource Management

CSN Connectivity Services Network

CUSUM Cumulative Sum

D-AMPS Digital Advanced Mobile Phone System

DHCP Dynamic Host Configuration Protocol

DNS Domain Name Server

DSL Digital Subscriber Line

DSSS Direct Sequence Spread Spectrum

DVB-H Digital Video Broadcasting - Handheld

EAP Extensible Authentication Protocol

EDGE Enhanced Data rates for GSM Evolution

FA Foreign Agent

FDD Frequency Division Duplex

FHSS Frequency Hopping Spread Spectrum

GAN Generic Access Network

GANC Generic Access Network Controller

GERAN GSM / EDGE Radio Access Network

GGSN Gateway GPRS Support Node

GSM Global System for Mobile communication

GPRS General Packet Radio Service

GMM GPRS Mobility Management

GSA Global mobile Suppliers Association

GTP GPRS Tunnelling Protocol

FIFO First In First Out

FTTH Fiber To The Home

HA Home Agent

HAWAII Handoff Aware Wireless Access Internet Infrastructure

HDTV High Definition TiVi

HIP Host Identity Protocol

HMIP Hierarchical MIP

HLR Home Location Register

HO Handover

HSDPA High Speed Downlink Packet Access

HSUPA High Speed Uplink Packet Access

HSS Home Subscriber Server

IDMP Intra-Domain Mobility Management Protocol

IETF Internet Engineering Task Force

IP Internet Protocol

IPSec IP Security

IMT-2000 International Mobile Telecommunications-2000

ITU International Telecommunication Union

ITU-R Radio Communication Sector of the International Telecommunication Union

IWU InterWorking Unit

JTACS Japanese Total Access Communication System

LMA Local Mobility Anchor

LMDS Local Multipoint Distribution System

LTE Long Term Evolution

MAC Media Access Control

MAG Mobile Access Gateway

MAP Mobility Anchor Point

MBMS Multimedia Broadcast Multicast Service

MIH Media Independent Handover

MIP Mobile Internet Protocol

MIMO Multiple-Input Multiple-Output

MMDS Multichannel Multipoint Distribution Services

MME Mobility Management Entity

MN Mobile Node

MNO Mobile Network Operator

MOBIKE IKEv2 Mobility and Multihoming

MS Mobile Subscriber

MSC Mobile Switching Center

MVNO Mobile Virtual Network Operator

NMT Nordic Mobile Telephone

OCS Online Charging System

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiple Access

PDG Packet Data Gateway

PDN Packet Data Network

PDP Packet Data Protocol

PLMN Public Land Mobile Network

QoS Quality of Service

QPSK Quadrature Phase-Shift Keying

RAB Radio Access Bearer

RAM Random Access Memory

RAN Radio Access Network

RAT Radio Access Technology

RII Roaming Interworking Intermediary

RLS Recursive Least Square

RNC Radio Network Controller

RRC Radio Resource Control

RRM Radio Resource Management

RSS Received Signal Strength

RTP Real-time Transport Protocol

RTT Radio Transmission Technology

SAE System Architecture Evolution

SC-FDMA Single Carrier Frequency Division Multiple Access

SCTP Stream Control Transmission Protocol

SDMA Space Division Multiple Access

SDR Software Defined Radio

SDU Service Data Unit

SEGW Security Gateway

SGSN Serving GPRS Support Node

SIM Subscriber Identity Module

SINR Signal to Inference plus Noise Ratio

SIP Session Initiation Protocol

SLA Service Level Agreement

SLF Subscription Locator Function

SM Session Management

SMLC Serving Mobile Location Center

SNR Signal-to-Noise Ratio

SOFDMA Scalable Orthogonal Frequency Division Multiple Access

SS Subscriber Station

TACS Total Access Communication System

TCP Transmission Control Protocol

TDD Time Division Duplex

TD-SCDMA Time Division - Synchronous Code Division Multiple Access

TDMA Time Division Multiple Access

TTG Tunnel Termination Gateway

UDP User Datagram Protocol

UE User Equipment

UMA Unlicensed Mobile Access

UMB Ultra-Mobile Broadband

UMTS Universal Mobile Telecommunications Service

UPE User Plane Entity

UTRA UMTS Terrestrial Radio Access

UTRAN UMTS Terrestrial Radio Access Network

VDSL Very high bit-rate Digital Subscriber Line

VoD Video on Demand

VoIP Voice over IP

VLC VideoLAN client

VPN Virtual Private Network

WAC Wireless Access Controller

WAG Wireless Access Gateway

WCDMA Wideband Code Division Multiple Access

WiMAX Worldwide Interoperability for Microwave Access

WISP Wireless Internet Service Provider

WLAN Wireless Local Area Network

WMAN Wireless Metropolitan Area Network

Introduction

Following the explosive growth of the Internet during the last two decades, the current unprecedented expansion of wireless technology promises an even greater effect on how people communicate, interact and enjoy their entertainment. The growing advances in research and development of wireless communication technologies and the increasing capabilities of electronic devices are driving an evolution towards ubiquitous services to mobile users. Wireless networks become increasingly interoperable with each other and with the high-speed wired networks. This reflects a paradigm shift towards new generations of mobile networks where *seamless mobility* across heterogeneous networks becomes fundamental. This generation is referred to as fourth generation (4G).

Future users will be *always best connected* through different available access networks when they move from one place to another (at home, in the office, on the bus, on the train, in the shopping mall, in the cafe...). For example, a video teleconference can transparently switch from an enterprise Wireless Local Area Network (WLAN) to the traditional cellular environment when driving home and to the fixed home network when arrived. In fact, users can access and maintain a seamless connectivity *anywhere*, *anytime* via *any access technology* owned by *any operator* to use *any available service*. Handovers between the technologies are transparent to users, allowing a simplified and seamless on-the-move experience. In summary, *'seamless mobility is predicated on enabling a user to accomplish his or her tasks without regard to technology, type of media, or device, facilitating freedom of movement while maintaining continuity of applications experience'* [2].

This thesis contributes to the evolution of technology convergence by improving different aspects of the inter-system handover management to make seamless mobility a reality. This thesis represents the main contributions of the author's research studies during the past three years.

Problem statement

The first seamless mobility application that has been commercial is the Unlicensed Mobile Access (UMA) solution, allowing seamless handover between cellular and WiFi hot-spot for voice call services. However, the UMA technology has some drawbacks. It does not ensure the Quality of Service (QoS) of multi-service bearer and the handover between Universal Mobile Telecommunications Service (UMTS) and WLAN has not been yet supported. This solution is only suitable for home or SOHO (Small Office and Home Office) subscribers due to the access capacity limitations. Though the inter-system mobility has attracted immense research and development effort from the research community and standardization bodies, the seamless handover does not really happen due to many issues.

Vertical handover, a term used to indicate the handover between two access nodes of two different technologies, is an issue in heterogeneous networks since each technology has its own mobility management solution. The mobile terminal must be capable of adapting the service content and the communication parameters each time it changes the access network. The two most considered performance criteria for the handover design are latency and packet loss. Generally, multimedia applications,

one of main services in 4G networks, require a short handover latency, low jitter and minimal packet loss. Handover is required to be achieved seamlessly. It means that handover is transparent to users experience: users do not recognize handover occurrences at the application perception. Technically, it means that the handover interruption delay should be very small (below 50*ms*) and the packet loss ratio should be minor to not affect the service quality (tolerant loss ratio differs from different application types). Users want to have a continuous and qualified service and they do not care about which access technology they have connected.

In order to support the seamless handover from one technology to another, the access networks involved should be integrated. The interworking between Third Generation Partners Project (3GPP) and Third Generation Partnership Project 2 (3GPP2) networks and WLAN networks has been the topic of much work within 3GPP/3GPP2 standardization bodies, a collaboration between groups of telecommunications associations, to make a globally applicable 3G mobile phone system specification within the scope of the International Telecommunication Union (ITU)'s International Mobile Telecommunications-2000 (IMT-2000) project. The UMA solution mentioned above is also recognized as a 3GPP Generic Access Network (GAN) standard. Despite some discussions about the competition between Worldwide Interoperability for Microwave Access (WiMAX) and cellular networks, WiMAX is keen to cooperate by interworking with 3GPP/3GPP2 networks to provide the maximum value to operators allowing them to reach as many customers as possible and to best serve their customers. Since WiMAX is different from WLAN in terms of radio coverage, QoS, capacity and security, the 3GPP-WiMAX interworking needs more research efforts. The interworking architecture of different technologies must minimize changes to the existing infrastructure. In addition, most of the existing interworking solutions deal with the integration between two access networks deployed by the same operator or by two coordinated operators. An approach to facilitate a secure and seamless interworking and roaming in multi-operator environment has not been addressed yet.

From the users' perspective, they all expect future networks to offer more customized services at higher quality levels. Users require an ease of use and also an ability to control their services according to their preferences. With the increasing capabilities in devices, an obvious question is whether the users' terminal can control and manage the handover across different access technologies. Delegating the handover control to the mobile terminal could be a solution for users to get what they really want. However, such a terminal-controlled handover requires an interworking architecture where the the mobility management becomes a service independent of access network operator domains. An intelligent access network selection in the terminal is required to aid users to select the best access network. The network selection relies on different access network characteristics, different constraints on terminal capabilities, the mobility management policy, the user's contexts and user preferences. An efficient and pragmatic solution is required to solve trade-offs within the multi-criteria selection problem in heterogeneous networks. The fact that the mobile terminal is equipped with one reconfigurable interface or multiple radio interfaces should be taken into account when designing the vertical handover procedure. The power consumption of multi-interface terminals becomes one of the critical issues due to the limited battery capacity of portable devices. Optimizing the power consumption in the frame of the handover management has not been sufficiently addressed in the literature.

The determination of the right handover triggering instant is difficult in heterogeneous networks since many criteria (not only the link quality) should be taken into account in the decision algorithm. If the handover is initiated too early, the ping-pong handover problem may occur. The ping-pong effect causes the instability and service performance degradation. If the handover is initiated too late, mobile terminals may not have enough time for making handover, which leads to connection drop. Despite using different optimization techniques to determine the appropriate handover triggering instant, the interruption may be still present. As the vertical handover of an Software Defined Radio (SDR)-enabled device is a hard handover (i.e., a new connection will be established after the old connection is terminated), the handover interruption is inevitable. Also, the interruption occurs when the cell overlap

area is too small for a multi-interface terminal to complete the handover on the target interface before moving out of the serving cell coverage. The handover interruption time is very critical for real-time applications, particularly for streaming services. The seamless video-on-demand streaming in wireless networks for mobile users is a challenging task.

Objectives

In this thesis, we aim to address the different facets of the inter-system mobility management. The interworking architecture between UMTS and WiMAX is one of the objectives of this thesis. We address the UMTS-WiMAX vertical handover under the assumption that the dual-mode mobile device is equipped with a single SDR interface which can switch from UMTS mode to WiMAX mode and vice versa. To support seamless handover, the mobile device should be able to perform the inter-system measurement without affecting the on-going communications and complete the handover decision before moving out of the serving cell coverage. The latter requires a sufficient overlap area between adjacent cells. If the overlap area is unnecessarily large, it increases the operators' building cost. If the cell overlap area is too small, the network's connection loss ratio is increased because mobile terminals at the edge of a cell cannot receive support from neighboring cells in time to prepare the handover.

In the vision of open access networks where users can connect to any available access network of any operator, a more flexible and open solution is required to interwork the networks to offer real global interworking and roaming facilities. To this end, we aim to design a Roaming Interworking Intermediary (RII) platform which support all combinations of different radio technologies in a multi-operator environment. The RII will support secured roaming and seamless mobility across two *independent* access networks.

We also aim to investigate the role of user terminals in the inter-system mobility management. The users are in a strong position to control some parts of the handover process because their terminal can access to information on device capabilities and user preferences, and, most importantly, to knowledge of both serving and neighboring access networks. To enable a complete terminal-controlled handover procedure, we analyze the following aspects under the control of the user terminal: access network selection, handover initiation, multiple radio interface management and handover preparation. The coordination of these steps within a mobility management architecture is needed to provide a seamless terminal-controlled vertical handover.

In the converged network trend, the complementary characteristics of different access technologies promote their interworking. The resources of the interworked networks can be viewed as a shared resource pool. Balancing the traffic load across the integrated networks is both a motivation and a challenge. An efficient load balancing will lead to best utilization of the pooled resources and thereby to improve the user satisfaction level. In fact, the load balancing is related to the mobility management since it involves the users' network selection and the network-controlled vertical handover enforcement. The load balancing problem will be also analyzed in this thesis.

In short, the objective of the thesis is to optimize the inter-system mobility management, mainly between 3GPP and WLAN/WiMAX networks, by addressing the following aspects: interworking and roaming architecture, access network selection, inter-system measurement, required cell overlap, handover initiation, handover prediction-based adaptive streaming application and load balancing. The ultimate goal is to explore different directions to achieve the seamless mobility in the future converged 4G networks and propose efficient solutions.

Methodology

The inter-system handover is a process involving the management of network entities and end-user terminals. In order to satisfy the objectives outlined above, two different visions, one from user terminal side and one from network side, are employed to approach the inter-system mobility management issue. In the user-controlled vision, end-users are not anymore passive to enjoy services offered by their network operators. They can use their smart terminals to take over the control of the inter-system mobility management across *'multi-homing'* networks (or open access networks) as well as to handle handover optimizations to ensure the seamless services. This part contributes to the migration trend from the network-centricity towards terminal-centricity. In the second approach, we address the important role of the network control in the inter-system mobility management and the handover optimizations. With a global view on user requests, its load values, available resources, roaming agreements with interworked networks..., the network can also control the handover to ensure the seamless mobility for mobile users.

Outline

The following figure highlights the structure of this thesis which is organized as follows:

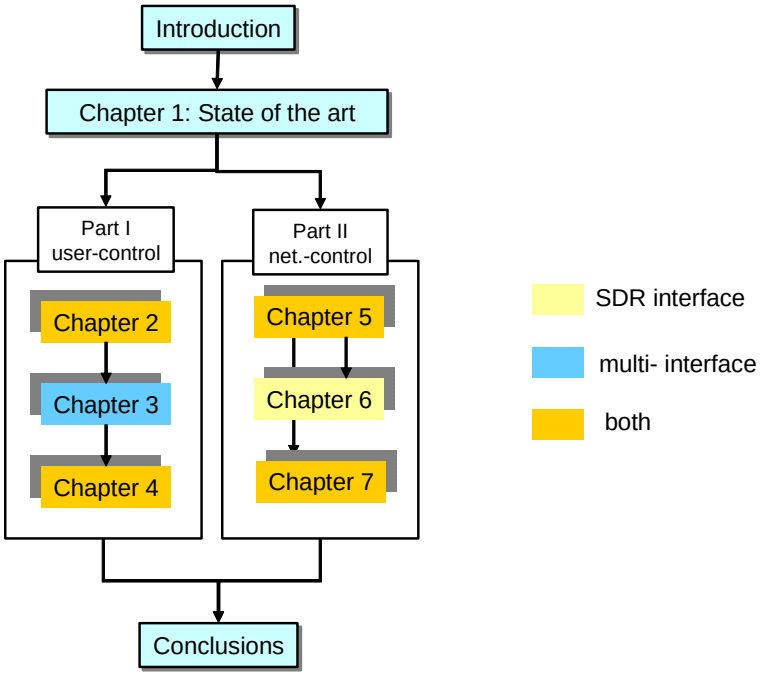

Figure 1: Organization of the thesis' chapters

- **Introduction**

  This chapter outlines the motivation and the scope of the work.

- **Chapter 1: Foundation of Mobility Management in 4G Wireless Heterogeneous Networks**

  In Chapter 1, we review the evolution of wireless mobile communication systems with a view to adopting the technology convergence as the core of the 4G concept. The motivations of the

4G networks are briefly identified. We address the main interworking architecture approaches proposed in the literature and in the standardization activities. The basic handover procedure and existing mobility management solutions are then reviewed and discussed. This chapter gives readers an overview of the active research initiatives in the area of mobility management in heterogeneous networks and identify the challenges behind the provisioning of seamless services during mobility.

Part I: User-controlled approach: In this part, we address the seamless handover management from the terminal perspective and investigate the role of the user in the mobility management. We show how users with their smart terminals can overcome some obstacles in the vertical handover management and how the performance can be improved under the user-centric terminal-controlled approach.

- **Chapter 2: Utility-based Access Network Selection**

  In Chapter 2, we first review the existing utility models including the single-criterion utility form and aggregate utility form used in access network selection and resource management problems. The limitations of the existing models are unveiled. We build up the utility theory and propose new single-criterion and multi-criteria utility forms to best capture user satisfaction and sensitivity in varying access network characteristics. This utility-based access network selection model is used in Chapter 3.

- **Chapter 3: Terminal-controlled Mobility Management Framework**

  In Chapter 3, we propose a terminal-controlled mobility management framework. The solution is devoted to mobile devices equipped with multiple radio interfaces. The proposed mobility management consists of a policy-based power-saving interface management coupled with a user-centric network selection solution, an adaptive handover initiation algorithm and a handover execution. It gives a complete solution from the architecture design to handover signaling exchanges and seamlessness optimization techniques.

- **Chapter 4: Handover Prediction-Assisted Seamless Media Streaming**

  In Chapter 4, we address seamless media streaming during horizontal and vertical handovers in heterogeneous networks. The solution is based on the terminal-controlled pre-buffering adjustment policy, running at the terminal side to maintain the appropriate amount of media content in the buffer. A practical handover prediction scheme is proposed to assist the right pre-buffering boosting decision.

Part II: Network-controlled approach: In this part, we address the seamless handover from the network perspective and highlight the crucial role of network control in the inter-system mobility management. We investigate how the network can assist and improve handover measurement, handover preparation, traffic load balancing, and security management.

- **Chapter 5: Interworking Architecture Design**

  In Chapter 5, we first address the proposed UMTS-WiMAX interworking architecture and the corresponding handover sequence charts. Second, we propose an RII functional entity to facilitate the interworking and roaming in a multi-operator environment. The solution allows a secured and seamless handover across two access networks of two independent operators.

- **Chapter 6: Inter-system Measurement and Required Cell Overlap**

  In Chapter 6, we address two important conditions for seamless handover between UMTS and WiMAX systems: inter-system measurement and required cell overlap. We investigate the possibility to make the UMTS-WiMAX inter-system measurement through a single reconfigurable radio interface terminal without affecting on-going sessions. We analyze the minimum cell overlap required to support seamless handovers between two adjacent cells within the same technology or different technologies in the UMTS-WiMAX networks.

- **Chapter 7: Load Balancing over Heterogeneous Wireless Packet Networks**

  In Chapter 7, we first highlight the limitations of the existing load balancing approaches and then address a new load balancing algorithm. The main contribution in this Chapter is to define a new load metric and a new balancing objective which makes it possible to reconsider the load balancing problem as a classic optimization problem. The proposed approach aims to hide the heterogeneity of different access technologies from the load balancing process.

- **Conclusions**

  This final chapter provides a summary of the thesis, discusses open issues and further research directions.

Chapter 1

Foundation of Mobility Management in 4G Wireless Heterogeneous Networks

In telecommunications, just like any other field of human endeavor, fashions come and fashions go. No sooner is one technology safely out of the laboratory than attention turns to the next new innovation. Over the last few years, 4G has been slowly taking shape as the next big development in wireless communications.

Alun Lewis, Independent telecommunications writer and consultant

This chapter presents a tracking of the evolution of mobile communication systems from the 1G analog communication networks to the 4G broadband converged networks. We address the 4G concept and provide an overview of the main approaches for interworking different technologies in the 4G heterogeneous environment. A basic handover procedure is presented and different mobility management approaches are discussed.

1.1 Evolution of mobile communication systems

1.1.1 Cellular technology evolution

Over the past 25 years, the evolution of the Internet and the advances of wireless technologies have made a tremendous impact on lifestyles around the world. Together, these two factors have changed the way people communicate, work, and get their entertainment.

With the introduction of cellular communications, we saw an increasing demand for wireless services. The growth was so rapid that by 2002, we witnessed a major shift in network usage: *for the first time in the history of telecommunications, the number of mobile subscribers exceeded the number of fixed lines*. And that trend continues. According to the ITU, by September 2005, the number of mobile subscribers exceeded 2 billion. According to Global mobile Suppliers Association (GSA)'s statistics [1], at the end of the first quarter of 2007, the number of mobile subscribers in the world exceeded 2.8 billion. Although the history of cellular networks has been rather brief, it has already seen three generations, and the fourth is emerging.

In mobile communication systems, there has been a paradigm shift every decade. The first generation (1G) systems in the 1980s were the original analog mobile voice networks. The second generation

[1] GSA is an organization dedicated to the promotion of the Global System for Mobile communication (GSM) mobile phone standard worldwide.

(2G) systems that emerged in the 1990s are based on digital technologies for mobile voice and data traffic. The third generation (3G) systems, firstly introduced in 2001 in Japan, are characterized by high-speed digital mobile voice, data and multimedia services. The pre-fourth generation (pre-4G) systems, a stepping-stone to 4G, will be commercialized around 2010. A full 4G is expected to be commercial around 2012. The evolution path of the mobile communication systems is depicted in Figure 1.1.

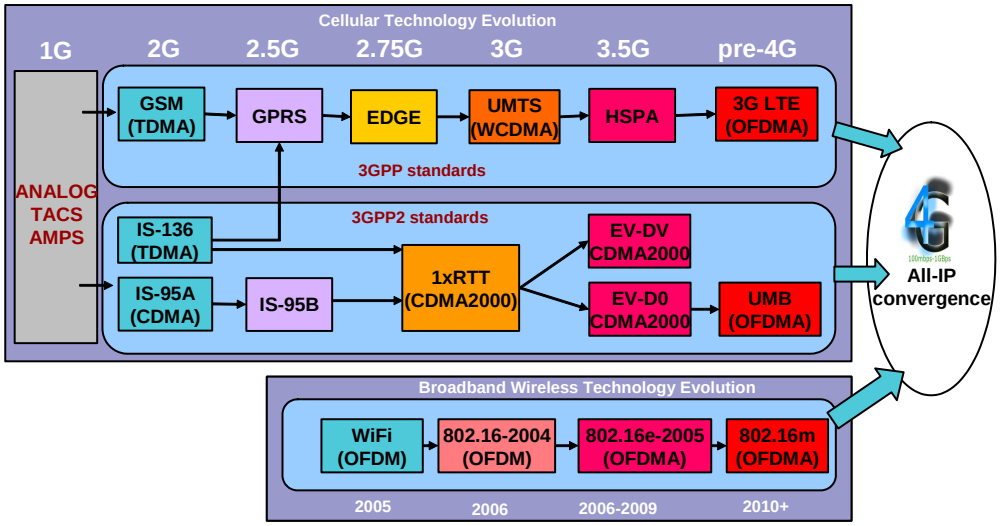

Figure 1.1: Wireless Technology Evolution Path

1G

1G systems, which initially debuted in Japan in 1979, are analog transmission systems. The key 1G standards included Advanced Mobile Phone System (AMPS), Total Access Communication System (TACS), Japanese Total Access Communication System (JTACS), and Nordic Mobile Telephone (NMT). 1G was indeed a major innovation in the telecommunication history. However, it was prone to the problems of quality of transmissions, security and inefficient utilization of the spectrum and capacity of available frequencies.

2G

2G networks introduced digital circuit-switched technology which uses the spectrum in a more efficient way. 2G networks are currently serving the vast majority of mobile subscribers and will remain in the market for a long time. It is likely that, in 2015, 2G will still be widely deployed [3]. The major 2G cellular standards are GSM, IS-136 and CdmaOne.

- GSM uses Time Division Multiple Access (TDMA) and Frequency Division Duplex (FDD). GSM has become the world's fastest-growing communications technology of all time and the leading global mobile standard.

- IS-136, known as Digital Advanced Mobile Phone System (D-AMPS), uses TDMA and Time Division Duplex (TDD). It is deployed throughout America, particularly in the United States and Canada. IS-136 is a digital overlay that is interoperable with the analog AMPS infrastructure (i.e., use AMPS for signaling to reserve resources). IS-136 allows data rates up to 30Kbps.

- CdmaOne refers to the original ITU IS-95 using Code Division Multiple Access (CDMA) that was first standardized in 1993. Today, there are two versions of IS-95, called IS-95A and IS-95B. IS-95A employs FDD with a channel bandwidth of 1.25-MHz for each direction, and supports

data speeds of up to 14.4Kbps. IS-95B can support data speeds of up to 115 Kbps by bundling up to eight channels. Due to its supportable data speeds, IS-95B is categorized as a 2.5G technology.

2.5G

2.5G is the realm of enhanced data services. The key 2.5G standards include General Packet Radio Service (GPRS), Enhanced Data rates for GSM Evolution (EDGE) and IS-95B. GPRS is an enhanced mobile data service for users of GSM and IS-136. In theory, GPRS supports transmission speeds of up to 172.2Kbps. GPRS is a packet-switched solution which is known as a migration strategy towards 3G. In the long path towards 3G, EDGE appeared as an enhancement for GPRS that can provide higher data rates (up to 384Kbps). By using EDGE, operators can handle three times more subscribers than with GPRS, triple their data rate per subscriber, or add extra capacity to their voice communications. EDGE is sometimes referred to as a 2.75G technology.

3G

3G networks are characterized by higher peak data rates, greater system capacity, and improved spectrum efficiency, among other capabilities. There is a range of technologies for 3G, all based around CDMA, including UMTS (with both FDD and TDD variants), CDMA2000 and Time Division - Synchronous Code Division Multiple Access (TD-SCDMA).

- UMTS, sometimes marketed as 3GSM, using Wideband Code Division Multiple Access (WCDMA) as the underlying air radio interface, has been standardized by 3GPP. UMTS is the 3G technology chosen by most GSM/GPRS mobile operators. The maximum user data rate is 1,920Kbps, but in real-world experience at the moment, it is only 384Kbps. To improve the performance of 3G UMTS, two standards High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA), jointly known as HSPA, have been developed. HSPA is usually referred to as a 3.5G technology.
  - HSDPA is a packet-based data service feature of the WCDMA standard that provides improved downlink data rates. The theoretical peak rate is 14.4Mbps, but the realistic end-user experience is initially likely to be 1.8Mbps or possibly up to 3.6Mbps. According to the latest report published by GSA in January 2008, 166 HSDPA networks have commercially launched in 75 countries.
  - HSUPA delivers substantial improvements in uplink data rates and QoS as well. The HSUPA standard enables users to transmit data upstream at a speed of 5.8Mbps. According to the same GSA's report, 51 operators have committed to deploy HSUPA, and 26 network operators have commercially launched HSUPA services in 22 countries.
- CDMA2000, direct successor to 2G CdmaOne, represents an entire family of technologies, including CDMA2000 1xRTT (Radio Transmission Technology), CDMA2000 EV-DO (Evolution-Data Optimized), and CDMA2000 EV-DV (Evolution- Data and Voice), standardized by 3GPP2. CDMA2000 is the 3G technology chosen by most CDMA mobile network operators.
  - CDMA2000 1xRTT officially qualifies as 3G technology, but it is considered by some to be a 2.75G technology. Although the peak data rate of 1xRTT can be up to 307Kbps, most deployments are limited to a peak of 144Kbps.
  - CDMA2000 EV-DO uses a separate 1.25MHz carrier for data, and supports up to 2.4Mbps downstream and up to 153Kbps upstream. 1xEV-DO Revision A supports Internet Protocol (IP) packets, increases the downlink peak rate to 3.1Mbps and substantially boosts the uplink rate to 1.2Mbps. 1xEV-DO Revision B enables operators to aggregate up to 15 1.25MHz channels to deliver up to 73.5Mbps. According to the report published in www.cdg.org site, 3G CDMA2000 EV-DO has surpassed 83 million subscribers in September 2007.

  - CDMA2000 EV-DV integrates voice and data over the same 1.25MHz carrier. CDMA2000 EV-DV offers a peak data rate of up to 4.8Mbps downstream and up to 307Kbps upstream. However, in 2005, Qualcomm put the development of 1xEV-DV on an indefinite halt, due to lack of carrier interest, mostly because both Verizon Wireless and Sprint chose EV-DO.

- TD-SCDMA was proposed by the China Communications Standards Association and approved by ITU in 1999. TD-SCDMA uses TDD mode and can operate in a minimum frequency band of 1.6MHz at 2Mbps or a 5MHz band at 6Mbps. Although the launch dates of TD-SCDMA have been pushed back time and time again, it is highly expected to be ready in time for Olympics 2008.

Though the roll-out of 3G networks was delayed in some countries by the enormous cost of additional spectrum licensing fees, GSM Association reported the crossing of the 200 million 3G subscribers in the second quarter of 2007. Both Japan and Korea, two first countries that launched 3G, continue to expand their 3G base with both reporting over 50% penetration. The 3G networks and their enhancements will continue to be deployed all around the world. 3G will probably have a lifetime similar to that of 2G - in the vicinity of 20 years.

pre-4G

While we are in the throes of finally seeing 3G networks deployed, there is already a buzz about their enhancement, going by names such as Long Term Evolution (LTE) and Ultra-Mobile Broadband (UMB).

- LTE - 3GPP Long-Term Evolution is the next version of the 3GPP-based radio standard. LTE is designed to provide higher data-rate (over 100 Mbps for downlink, and over 50 Mbps for uplink for every 20 MHz of spectrum), lower-latency and packet-optimized system compared to 3G. To this end, LTE uses Orthogonal Frequency Division Multiple Access (OFDMA) for the downlink and Single Carrier Frequency Division Multiple Access (SC-FDMA) for the uplink and employs Multiple-Input Multiple-Output (MIMO) with up to four antennas per station. 3GPP has recently reported LTE's peak theoretical downlink throughput rates of up to 326 Mbps in 2x20 MHz with 4x4 MIMO configuration. LTE is designed to be all-IP and to support mobility and service continuity between heterogeneous access networks. The LTE demonstrations and trials will be continued during this year 2008 and the first network deployment is expected around 2010.

- UMB - 3GPP2 Ultra Mobile Broadband, is the successor to CDMA2000 EV-DO, formerly known as EV-DO Revision C. UMB also incorporates OFDMA, MIMO and Space Division Multiple Access (SDMA) advanced antenna techniques to provide even greater capacity, coverage, and QoS. UMB can support peak download speeds as high as 280 Mbps in a mobile environment and over 75 Mbps for upstream transmission (with 4x4 MIMO configuration). UMB is expected to be commercially available in early 2009. However, up to now, Qualcomm is the only big name backer of UMB. The logical supporters of UMB will be the vendors who make equipment for CDMA since UMB is an upgrade path for that. The main ones are Alcatel-Lucent, Nortel, Motorola, ZTE and Samsung, but none of them has officially committed to releasing UMB products.

1.1.2 Mobile broadband wireless technology evolution

Wireless broadband communication is the confluence of the two most remarkable growth stories of the telecommunications industry in recent years: broadband communication and wireless mobile communication. The rapid mass-market growth of the mobile cellular systems with about 3 billion

subscribers have been previously summarized. During the same period, Internet has been evolving from a curious academic tool to having about a billion users. Parallel to the growth of Internet, the development of broadband technology has been accelerated to offer the high-speed Internet access services. The broadband access over the twisted-pair of telephone wires or over coaxial cable TV are the predominant mass-market technologies today. Recently, the advanced broadband access such as Fiber To The Home (FTTH) or Very high bit-rate Digital Subscriber Line (VDSL) has been being deployed to enable rich performance applications like High Definition TiVi (HDTV), video on demand at speed of Gigabits per second. In less than a decade, the broadband subscription worldwide has grown from virtually zero to 200 million [4].

The broadband wireless technology is about bringing the broadband experience over the air radio interface. The broadband wireless service can be distinguished into two types: *fixed broadband wireless* and *mobile broadband wireless*. The fixed broadband wireless technology such as Local Multipoint Distribution System (LMDS), Multichannel Multipoint Distribution Services (MMDS) and fixed WiMAX, are thought of as a competitive alternative to Digital Subscriber Line (DSL) or cable modem. Otherwise, the mobile broadband wireless technology like IEEE 802.11-based WiFi or Mobile WiMAX attempts to bring broadband applications to users on the move with the functionality of portability, nomadicity and mobility. In addition to the WiFi and mobile WiMAX, a few proprietary solutions, such as i-Burst technology from ArrayComm and Flash-OFDM from Flarion (acquired by QualComm), as well as the standard-based solutions like 3G cellular and beyond 3G systems also support the mobile broadband applications. Below, we introduce the WiFi and WiMAX technologies and their evolution.

WiFi

WiFi is based on the IEEE 802.11 family of standards. It is primarily a WLAN technology designed to provide in-building broadband coverage. WiFi becomes a defacto standard for broadband connection in homes, offices, and public hot-spot locations including hotels, airports, shopping centers, restaurants, cafes, and educational environments. In the past couple of years, a significant number of municipalities and local communities around the world have taken the initiative to get WiFi systems deployed in outdoor to provide broadband access to city centers as well as to rural and under-served areas.

The 802.11 specifications were initially introduced in 1997 for operation in the unlicensed 2.4GHz band, and they included two spread spectrum methods: 1Mbps Frequency Hopping Spread Spectrum (FHSS) and 1Mbps and 2Mbps Direct Sequence Spread Spectrum (DSSS). In 1999, IEEE 802.11b relying on DSSS transmission technology with support data of 11Mbps was published. Also in this year, IEEE 802.11a making use of Orthogonal Frequency Division Multiplexing (OFDM) transmission, increasing the speeds to a theoretical rate of 54Mbps was standardized. This standard operates in the 5GHz band. Published in 2003, IEEE 802.11g uses OFDM and operates in the 2.4GHz band. The standards 802.11b, 802.11a, and 802.11g are the most commonly used today.

The capabilities of WiFi are being enhanced to support even higher data rates and to provide better QoS support. In the large 802.11 family standards, three new emerging amendments are 802.11e, 802.11i and 802.11n. 802.11e addresses QoS extensions in WLAN. 802.11i enhances the security protection. However, the most eagerly awaited amendment is 802.11n. By using multiple-antenna spatial multiplexing technology, the IEEE 802.11n will support a peak layer 2 throughput of at least 100Mbps (maximal theoretical data rate is 540Mbps). The 802.11n is expected to provide significant range improvements through the use of transmission diversity and other advanced techniques. Though there are already many products on the market based on Draft 2.0 of this proposal (Draft 4.0 was approved in April 2008), the TGn working group is not expected to finalize the amendment until November 2008. 802.11n will be a main technology for WLAN networking in the future.

WiMAX

WiMAX is designed to accommodate both fixed and mobile broadband applications. The term 'WiMAX' was created by WiMAX Forum that was formed in June 2001 dedicated to promoting interoperability and compatibility of broadband wireless products based on the IEEE 802.16 standard. Originally, IEEE formed a group called 802.16 to develop a standard for Wireless Metropolitan Area Network (WMAN) in 1998.

- IEEE 802.16 first issued standards for the PHY and MAC layers of systems in the 10-66GHz bands, generally known as LMDS, in December 2001.
- In 2003, the 802.16a standard, using OFDM to mitigate the impairments fading and multipath, and operating in the 2GHz to 11GHz bands, was published. Further revisions to 802.16a were made and completed in 2004. This revised standard, IEEE 802.16-2004, replaces 802.16, 802.16a, and 802.16c with a single standard, called as fixed WiMAX.
- The IEEE 802.16e standard, also known as Mobile WiMAX, was initially designed to allow vehicular mobility applications. It was completed in December 2005 and was published formally as IEEE 802.16e-2005. It uses Scalable Orthogonal Frequency Division Multiple Access (SOFDMA), a multi-carrier modulation technique that uses sub-channelization, where channel bandwidths are selectable, ranging between 1.25MHz and 20MHz. The key attribute of IEEE 802.16e is the introduction of the handover capability for users moving between cells.

In October 2007, the Radio Communication Sector of the International Telecommunication Union (ITU-R) included WiMAX technology in the IMT-2000 set of standards, also known as 3G. Hundreds of WiMAX trials and deployments are now in progress around the globe. Even though the industry is still waiting for mobile WiMAX certified products and the first 802.16e roll-out, the IEEE keeps working on new 802.16 amendments. Two most relevant amendments in progress are 802.16j (Multihop Relay) and 802.16m (Advanced Air Interface). The goal of 802.16m is to achieve data rates up to 1Gbps for fixed users and 100Mbps for mobile users. It aims to improve the capacity and performance of Multimedia Broadcast Multicast Service (MBMS) and Voice over IP (VoIP). The driver behind 802.16m will be MIMO antenna technology on top of an OFDM-based radio system. The 802.16m is comparable with the LTE or UMB in terms of technology, capacity and services. It is expected that the WiMAX 2.0 based on 802.16m will be ready at the end of 2009.

1.1.3 Broadband wireless technologies: Comparative study

Feature	**HSPA**	**1x EV-DO**	**Mobile WiMAX**	**WiFi**
Standard	3GPP R6	3GPP2	IEEE 802.16e-2005	IEEE 802.11n
Peak DL data rate	14.4Mbps using all 15 codes	3.1Mbps (Rev A); 4.6Mbps (Rev B)	46Mbps @ 3:1, 2x2, 10Mhz	100Mbps
Peak UL data rate	5.8Mbps	1.8Mbps	7Mbps @3:1, 10Mhz	100Mbps
Bandwidth	5MHz	1.25MHz	3.5, 7,5,10,8.75MHz	20/40MHz
Duplexing	FDD	FDD	TDD initially	TDD
Multiplexing	TDM/CDMA	TDM/CDMA	TDM/OFDMA	CSMA-CA
Coverage	1-5km	1-5km	<3.5km	<300m
Mobility	High	High	Middle	Low

Table 1.1: Technology Features Comparison

The evolution path of cellular, WiFi and WiMAX technologies was summarized in the two above sections. Briefly, WiMAX occupies a somewhat middle ground between WiFi and 3G technologies when compared against the key dimensions of data rate, coverage, QoS and mobility support. Table

1.1 provides a summary comparison among 3G (3.5G HSPA, 1xEV-DO), Mobile WiMAX and WiFi 802.11n technologies. These three technologies are chosen since they are likely to dominate the mobile broadband telecommunication market today as well as in the next few years.

Two other standards-based technologies could also emerge in the future: IEEE 802.20 - Mobile Broadband Wireless Access (MBWA) and IEEE 802.22 - Wireless Regional Area Networks (WRAN). We address a brief introduction about these two under development standards:

- Much like 802.16e, the IEEE 802.20 hopes to define a broadband solution for vehicular mobility up to 250 km/h. The 802.20 standard was being positioned as an alternative to 3G cellular services since it can support high-speed handover and wireless network access. It is likely to be defined for operation below 3.5GHz to deliver peak user data rates in excess of 4Mbps and 1.2Mbps in the downlink and uplink respectively. At this point in time, the standard seems to be suspended owing to lack of consensus on technology and issues with the standardization process.

- The IEEE 802.22 standard aims to bring broadband access to rural and remote areas. Work on the 802.22 standard began in November 2004. The basic goal of 802.22 is to define a cognitive radio that can take advantage of unused TV channels in sparsely populated areas. Operating in the VHF and low UHF bands provides favorable propagation conditions that can lead to greater range (100km). The 802.22 standard is steadily processing and results are expected soon.

4G technology has not been officially defined yet. Most companies believe that the characteristics of IMT-Advanced, as defined by ITU, will represent a definition of 4G. Two expected requirements within IMT-Advanced are OFDMA-based technology and data rate support of 100Mbps for mobile applications. LTE, UMB, and WiMAX 802.16m all fulfill these requirements. Although several companies claimed that LTE/UMB/WiMAX 802.16m was 4G, we consider them, hereafter, as a pre-4G technology. Any claim that a particular technology is a 4G technology is, in reality, simply a market positioning statement by the respective technology advocate. A comparison among the three emerging technologies is presented in Table 1.2.

Feature	**LTE**	**UMB**	**WiMAX 802.16m**
Peak data rate (per sector @20MHz)	DL: 288Mbps (4x4) UL: 98Mbps (2x4)	DL:250Mbps (4x4) UL: 100Mbps (4x4)	DL: $\sim$ 350Mbps (4x4) UL: $\sim$ 200Mbps (2x4)
Latency	Link-Layer Access: <5 ms Handover: <50ms	LLA: <10ms Handover: <20ms	LLA: <10ms Handover: <20ms
MIMO [2] configuration	DL: 2x2, 2x4, 4x2, 4x4 UL: 1x2, 1x4, 2x2, 2x4	DL: 2x2, 2x4, 4x2, 4x4 UL: 1x2, 1x4, 2x2, 2x4	DL: 2x2, 2x4, 4x2, 4x4 UL: 1x2, 1x4, 2x2, 2x4
Bandwidth (MHz)	1.25, 1.6, 2.5, 5, 10, 15, 20	1.25 to 20	5, 10, 20, 40
Duplexing	TDD,FDD	TDD,FDD	TDD,FDD
Multiplexing	OFDMA and SC-FDMA	OFDMA	SOFDMA
Mobility	Up to 350 km/h	Up to 250 km/h	Up to 350 km/h

Table 1.2: Pre-4G technology requirements comparisons

Mobile broadband is the segment with the fastest growth in mobile communications. There is much hype in the wireless industry about WiMAX and LTE compared to UMB. Many dismiss UMB as being dead on arrival since it has so little traction with the operators. Whatever the part of the market share that each pre-4G technology will gain, one sure thing is that users in the near future will enjoy the mobile broadband services with rich media performance.

[2] MIMO configuration *x* x *y* means that the MIMO system uses *x* transmit antennas and *y* receive antennas.

1.2 4G wireless mobile heterogeneous networks

1.2.1 4G concept

Even though there are plenty of talks about 4G, there is not yet a universal agreed-upon definition of the 4G wireless mobile network up to now.

1.2.1.1 ITU's Vision: IMT-Advanced

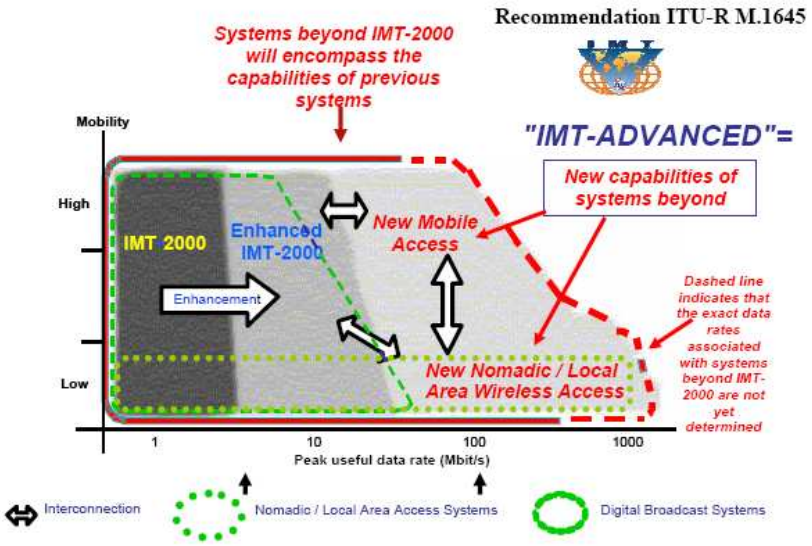

Figure 1.2: IMT-Advanced vision

According to [5], the ITU is currently establishing criteria for IMT-Advanced and will be screening various technologies for inclusion in the IMT-Advanced family. ITU-R has progressed from delivering a vision of 4G in 2002 (Figure 1.2) to establishing a name for 4G in 2005 (IMT-Advanced). The work of the ITU encompasses the important elements of business success in the wireless industry, especially the balance of a market and services view, a technology view, a spectrum view and regulatory aspects. A set of requirements by which technologies and systems can be determined as part of IMT-Advanced is being done by the ITU-R.

According to the IMT-Advanced, *in systems beyond IMT-2000, there may be a need for a new wireless access technology to be developed around the year 2010, capable of supporting high data rates with high mobility*. As one can see in Figure 1.2, the main requirements for 4G will be likely to include two following conditions: peak data rate of 100Mbps for high mobility applications such as mobile access; and approximate 1Gbps for low mobility applications such as nomadic/local wireless access.

Following the paradigm of generation changes, it is expected that the 4G would follow sequentially after 3G as an ultra-high-speed broadband wireless network. This view is usually referred to as a *linear 4G vision* [6] [7]. Nevertheless, even if 4G is named as the successor of the previous generations, the future will not be limited to cellular systems and 4G will not be seen exclusively as a linear extension of 3G.

1.2.1.2 Convergence of heterogeneous networks

Unlike 1G, 2G and 3G, 4G is not a set of formally agreed end-to-end standards developed in the traditional top-down way that the telecommunications industry has used for years [8]. It is now widely accepted that 4G is a vision of an all-IP based, heterogeneous mobile broadband networks with multiple air interfaces, converged fixed-mobile networks, and multiple devices with multi-mode capabilities. 4G will provide end-users with an Always Best Connected (ABC) facility, low latency and high QoS broadband experience. The ABC means a seamless service provisioning across a multitude of wireless access systems and an optimum service delivery via the most appropriate available network.

4G will be a convergence platform providing clear advantages in terms of coverage, bandwidth and power consumption. 4G will ensure the seamless mobility and global roaming among various access technologies such as cellular networks, WiFi, WiMAX, satellite, Digital Video Broadcasting - Handheld (DVB-H). 4G services will be end-to-end QoS, high security, available at any time, anywhere with seamless mobility, affordable cost, one billing, and fully personalized. 4G is about convergence, convergence of networks, of technologies, of applications and of services, to offer a personalized and pervasive network to the users. Convergence is heading towards an advent of a really exciting and disruptive concept of 4G.

The 4G network will be an umbrella of multitude of technologies. The glue is likely to be **seamless mobility** over heterogeneous wireless networks. Inter-system mobility, mainly between 3G UMTS/LTE and WiMAX/WiFi, is the main aim of this thesis.

1.2.2 Motivations for 4G heterogeneous networks

As mentioned previously, the 4G mobile networks do not consist of only one access technology but multiple ones. It is needed to have a mechanism that enables seamless mobility among different systems. The motivation behind the heterogeneous networks comes from the fact that there is no technology that could offer ubiquitous coverage. No technology can provide simultaneously the high bandwidth, low latency, high mobility and wide-area data service to a large number of users. As all these systems have their own advantages and shortcomings, no single technology merits enough to replace all other existing technologies up to now, even pre-4G technologies. It is beneficial for mobile users to switch their connection among different access points of different technologies to maintain the connectivity all time, and to enjoy the best personalized services according to their own preferences.

During the evolution from the 1G to the pre-4G, a range of mobile wireless technologies have been developed. All these technologies were designed independently, targeting different types of services, data rates and users. The complementary characteristics of various technologies motivate the interest to integrate them together. An interworking approach can make the best use of advantages of all technologies and can eliminate their stand-alone defects. For example, an operator can deploy low-cost high-data rate WLAN/WiMAX that is either an extension of cellular network or inter-workable with cellular network so that the utilization of already deployed infrastructures can be maximized. The wireless broadband technologies like WiFi, WiMAX can be a good complement to cellular technologies in terms of geographical coverage and QoS.

One technology can be employed to extend the radio range of another technology. If one access technology is highly loaded, users can connect to another access technology to have the equivalent services. Also, if a user requires a higher QoS level which is not supported by its current access technology, he can automatically handover to another technology. These are the main motivations for research as well as business exploration of 4G convergence concept.

1.3 Interworking in 4G heterogeneous networks

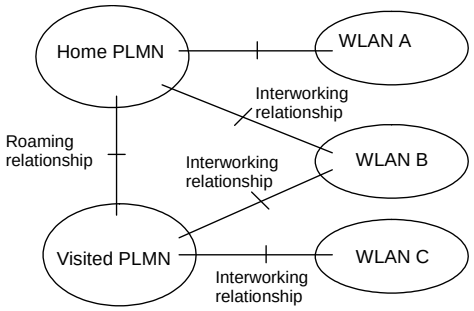

Figure 1.3: Interworking vs. roaming relationship (3GPP)

Interworking, integration and convergence are terms expressing the need for combining the advantages of diverse network technologies in order to get the best service for minimum investment from the network. The term *interworking* is much related to the *roaming* concept. However, according to 3GPP specifications, roaming is not specific to heterogeneous networks. 3GPP roaming relationship corresponds to a 3GPP subscriber using visited 3GPP network resources. The difference between interworking and roaming definition can be illustrated by Figure 1.3. The interworking relationship complements the well-known roaming definition. The 3GPP interworking relationship refers to a 3GPP subscriber using a non-3GPP radio interface to access 3GPP network resources. However, in a general meaning, interworking involves connecting two or more distinct access networks (not necessarily between 3GPP and non-3GPP networks) to allow end-users to access to these interworked networks and possibly to maintain the service continuity.

1.3.1 Interworking approaches

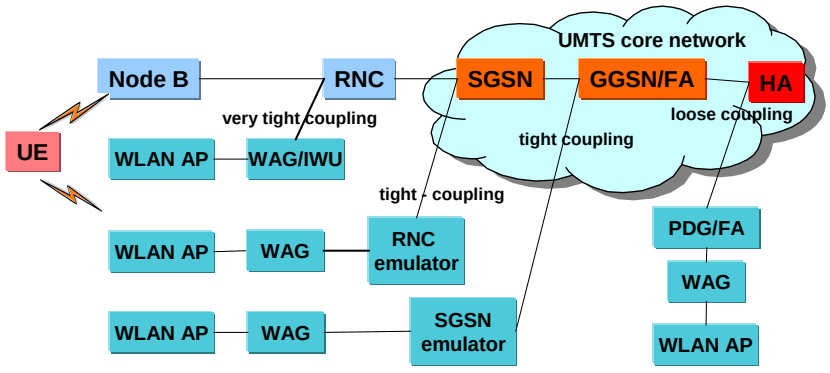

Figure 1.4: Different UMTS-WLAN interworking approaches

The interworking between different technologies, mainly between 3GPP and WLAN, has catched the attention of the research community and standardization bodies in the last few years. Broadly, the interworking can be classified into two approaches: *loose coupling and tight coupling* [9–11]. From a macro point of view the main difference is how and where a non-3GPP access network is coupled to the 3GPP/3GPP2 network. The distinction between tight-coupling and loose-coupling is based on the integration point of two networks involved as illustrated in Figure 1.4.

1.3.1.1 Loose-coupling architectures

Loose coupling offers a common interface for the exchange of information between the networks to guarantee service continuity. The two access networks have nothing in common, but the core networks are connected together. Loose coupling refers to the IP layer interconnection. The basic loose coupling interworking architecture between WLAN and UMTS is depicted in Figure 1.5. WLAN and UMTS are assumed to be in different IP address domains. The IP address is changed when the mobile terminal moves from one network to another. The heterogeneity of different access networks is managed and hidden by Mobile Internet Protocol (MIP). The integration point is the Home Agent (HA) of the MIP mechanism implemented in the Internet/external Packet Data Network (PDN). In general, the interworking point is placed after Gateway GPRS Support Node (GGSN).

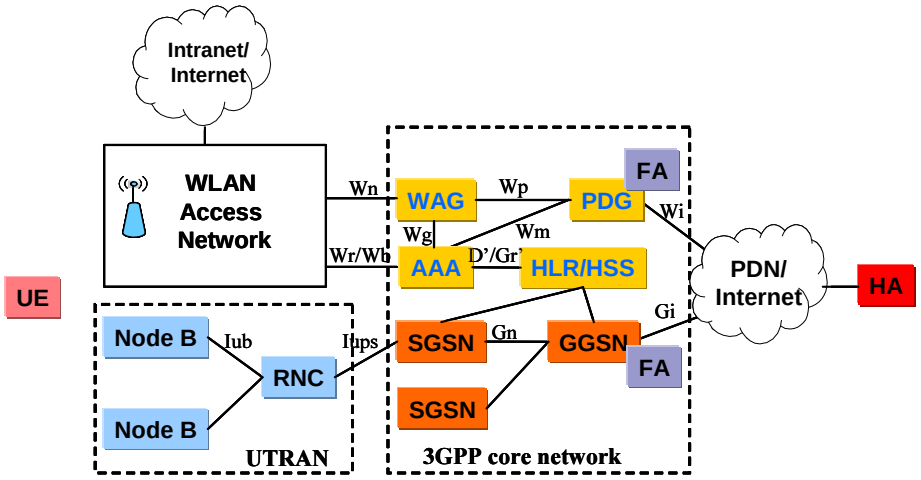

Figure 1.5: 3GPP -WLAN loose coupling interworking architecture

In the loose coupling interworking, the two networks communicate through the Internet. They operate independently of each other and roaming agreements between the corresponding operators are required to be established. To ensure service continuity with roaming capabilities, MIP and Authentication, Authorization and Accounting (AAA) functionalities are combined. The existing AAA solution in 3GPP network is used for managing and exchanging subscriber information and credentials between 3GPP and non-3GPP access networks. This approach separates completely the data path in WLAN and 3GPP networks. The WLAN data is never injected into the 3GPP core network. If two access technologies are deployed by a single operator, the Packet Data Gateway (PDG) may interface directly to the GGSN for signaling, otherwise, the signaling will be transported through the IP network.

The key of mobility management of this architecture is the MIP. The Foreign Agent (FA)s are located in the GGSN and the PDG while the HA is located in the PDN/Internet. When the mobile user moves across the networks, its home address is maintained. The major drawbacks are the handover latency and the packet loss due to MIP mechanism. In order to remedy this problem, pre-registration, pre-authentication, packet buffering and forwarding techniques have been studied. Many extensions of MIP have been proposed such as: Fast MIP, Hierarchical MIP, multiple Care of Address (CoA) registration MIP, layer-2 triggering based MIP, etc.

The loose coupling interworking architecture offers an easy and independent deployment. The loose-coupling interworking does not need drastic changes in existing infrastructures. There exists a variant of the loose coupling interworking that is sometimes referred to as an *open coupling*. In the latter scheme, no real integration between the two networks is present. WLAN and 3GPP are two independent systems that share a single billing scheme.

1.3.1.2 Tight-coupling architectures

In the tight coupling scheme, the non-3GPP access network is employed as a new radio access technology within the cellular one. The tight coupling makes two different Radio Access Technology (RAT)s working together with a single core network. The interworking point is at the 3GPP core network or at the UMTS Terrestrial Radio Access Network (UTRAN) as illustrated in Figure 1.4. When the integration point locates in the UTRAN, the interworking is known as a *very tight coupling* [10].

The 3GPP control protocols are reused in the WLAN and the data traffic is routed via the 3GPP core network to the outer entities. Two radio access networks are interconnected via layer 2. All the layer 3 protocols remain unchanged. The handover does not involve the change of remote IP address as well as the AAA policies. In the interworking reference model architecture depicted in Figure 1.4, the Radio Network Controller (RNC)/Serving GPRS Support Node (SGSN) emulator provides functionalities that are equivalent to those of an RNC/SGSN in an attempt to hide WLAN access network particularities from the UMTS. Its main function is to provide a standardized interface to the UMTS core network.

In the very tight-coupling solution, the WLAN is considered as part of the UTRAN. An important issue with the very tight-coupling scheme is the ownership of the WLAN. The most envisioned solution is that the 3GPP operator owns the WLAN part. Due to the scalability issue, it makes sense to introduce an InterWorking Unit (IWU) between the WLAN Access Point (AP)s and the RNC to share the control task of the RNC. The IWU will be implemented in the WLAN AP to either act as a pure traffic concentrator or be further responsible for control and supervision functionality [10, 12].

1.3.2 Interworking within 3GPP standards

1.3.2.1 Rel-6: 3GPP-WLAN interworking

The interworking between 3GPP and WLAN systems was considered by 3GPP TSG SA1 group. It has been specified in 6 different scenarios in 3GPP Release 6 [13].

1. **Scenario 1 - Common billing and customer care:** This is the simplest form of interworking, which provides only a common bill and customer care to subscribers. In fact, there is no real integration between the WLAN and 3GPP networks (open coupling).

2. **Scenario 2 - 3GPP system-based access control and charging:** This basically enables IP connectivity via WLAN for 3GPP subscribers. The scenario consists on introducing a new network element 3GPP AAA server to enable the WLAN authentication, authorization and charging mechanisms to converge towards 3GPP solution.

3. **Scenario 3 - Access to 3GPP Packet Switch (PS)-based services:** The goal of this scenario is to allow the cellular operator to extend access to its 3GPP based services to subscribers in a WLAN environment. Although the user is offered access to the same PS-based services over both the 3GPP and WLAN access networks, no service continuity across these access networks is required. The architecture reference model corresponding to non-roaming case of scenario 3 as specified in [14] is illustrated in Figure 1.6. This scenario introduces two main new network elements: Wireless Access Gateway (WAG) and PDG which enable secured accesses to different PDNs at Wi interface (Wi is similar to well-known Gi interface).

4. **Scenario 4 - Service continuity:** The goal of this scenario is to allow access to PS-based services as required by scenario 3, and additionally to maintain service continuity across the 3GPP and

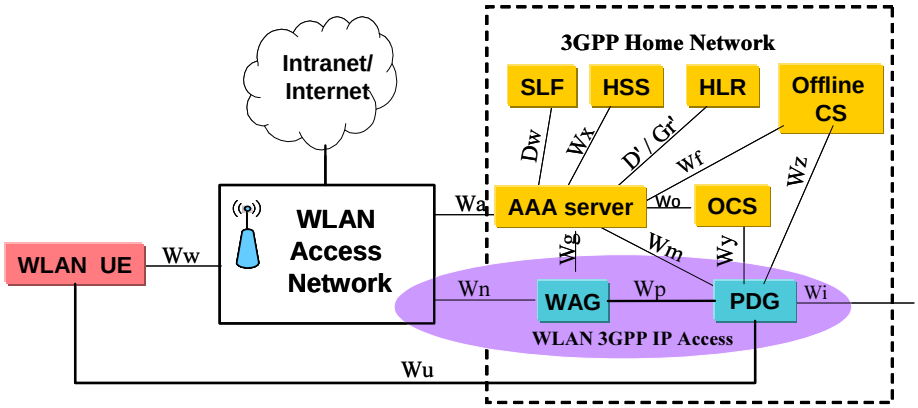

Figure 1.6: Scenario 3 non-roaming reference model (the shaded area refers to WLAN 3GPP IP access functionality)

WLAN systems. The change of access network may be noticeable to the user, but there will be no need for the user to re-establish the connection. Although service continuity is required by scenario 4, the service continuity requirements are not very stringent. This means that some services may not be able to continue after a handover to/from the WLAN.

5. **Scenario 5 - Seamless services:** This scenario is one step further than scenario 4. Its goal is to provide seamless service continuity between access technologies. The seamless service continuity is meant minimizing data loss and break time during the handover between access technologies.

6. **Scenario 6 - Access to 3GPP circuit-switched services:** The goal of this scenario is to allow the operator to offer access to circuit-switched services (normal voice calls) through WLAN access networks. Seamless mobility for these services should be provided.

In Release 6, 3GPP has specified scenarios 2 and 3, the other scenarios are going to be specified in Release 7 and in System Architecture Evolution (SAE) work item.

1.3.2.2 Rel-6: Generic Access Network

The GAN technology defines a new access network to the mobile core network that can be used to access the existing circuit-switched and packet-switched services [15] . The access network is based on use of unlicensed spectrum like WLAN and IP-based broadband access network. Specifications were developed in the UMA industry forum during year 2004 and were adopted in 3GPP Release 6 in April 2005 as the GAN standard. With GAN the end-user experience remains the same in the WLAN radio network as in GSM and WCDMA radio networks. The GAN solution is a scenario 6-like solution. Circuit-switched seamless handover is fulfilled but due to bad packet-switched handover, the GAN solution does not offer scenario 5 functionalities.

The interworking architecture between WLAN and 3GPP system according to GAN solution is illustrated in Figure 1.7. It introduces a Generic Access Network Controller (GANC) that acts as a Base Station Controller (BSC)/RNC emulator. When the mobile user establishes a communication in the zone covered by WiFi APs, it automatically attaches to them. There is software installed in the mobile terminal that can encapsulate the outgoing data to the cellular network in the IP packets. These packets are sent to the GANC that is responsible for transmitting the necessary information to the operator network. It redirects the IP packets to the target destination.

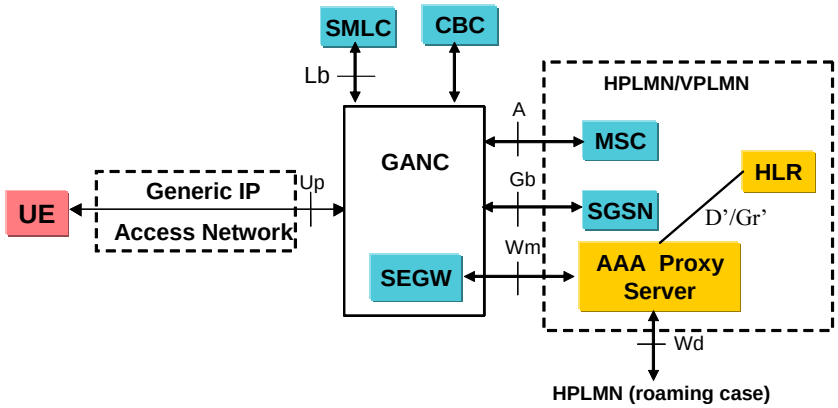

Figure 1.7: GAN architecture reference model

1.3.2.3 Rel-6: Tunnel Termination Gateway solution

The evolution of Tunnel Termination Gateway (TTG) to enable the interworking between 3GPP and WLAN systems was first mentioned in annex of [14] and fully described in [16]. A TTG is a separated equipment in charge of terminating the IP Security (IPSec) tunnels from the mobile stations and to map them to GPRS Tunnelling Protocol (GTP) tunnels established towards a GGSN as depicted in Figure 1.8. The idea of the present solution is to use the TTG and a subset of GGSN functions to implement the PDG functions.

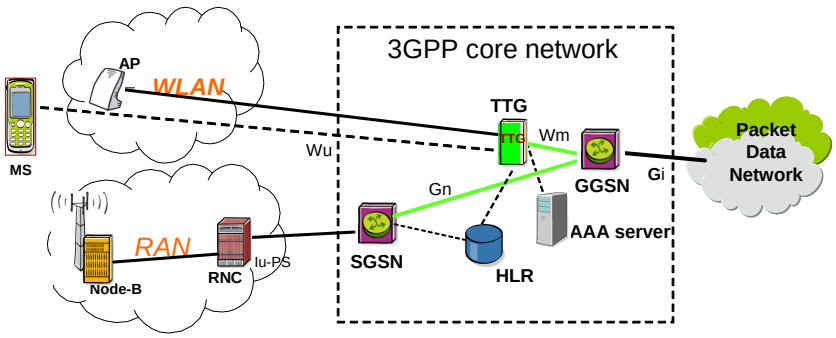

Figure 1.8: 3GPP-WLAN interworking architecture using TTG

The TTG can be considered as an SGSN emulator [17]. The functionality of TTG covers all aspects of PDG that are not covered by the GGSN. The TTG acts as the SGSN for the GTP tunnel establishment. It also acts as a WLAN user's proxy for the reason of transparency to the WLAN. In this scheme, the end-to-end tunnel from Mobile Subscriber (MS) to PDG is terminated at TTG and a GTP tunnel is established between the TTG and GGSN. The TTG supports location management and session management mechanisms such as SGSN context transfer, Packet Data Protocol (PDP) context update, and Home Location Register (HLR) update. This solution allows the re-use of existing GGSN and its existing capabilities such as charging frameworks. It offers a full interoperability with existing SGSN without upgrades. This solution satisfies service continuity requirements with little handover delay and no packet overhead. However, it requires a new network element TTG to be implemented.

1.3.2.4 Rel-7: SAE/LTE - non-3GPP interworking

Key benefits of the 3G LTE include significantly increased peak data rates, increased cell edge performance, reduced latency, scalable bandwidth, and co-existence with 2G/3G systems. Recently, the GSA has announced that 3GPP has approved the LTE technology specifications [1], leading to their inclusion in the forthcoming 3GPP Release 8.

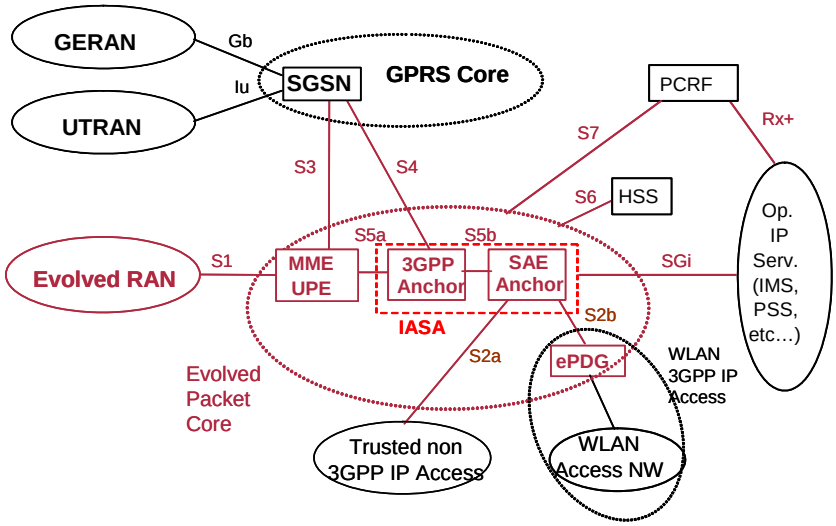

Figure 1.9: High level logical architecture for SAE/LTE system

The mobility between heterogeneous access networks, including service continuity within the LTE is also required. The 3GPP and non-3GPP access systems are integrated at the SAE Anchor entity as illustrated in Figure 1.9. The SAE Anchor represents functions grouped around the anchor point for handovers between 3GPP and non-3GPP access systems, whereas the 3GPP home anchor is the anchor point for handovers between 3GPP access systems. The SAE anchor allocates IP addresses for the MS as required by the used mobility protocol. This interworking architecture can be categorized as a loose-coupling approach. By providing a certain level of interaction between the SAE Anchor and the 3GPP Anchor within the evolved packet core, the MIP-based mobility signaling and tunneling only needs to be activated when the MS uses a non-3GPP access technology.

1.3.3 3GPP-WiMAX interworking

WiMAX-3GPP Interworking refers to the integration of WiMAX and 3GPP networks. For the moment, the WiMAX Forum considers the first three interworking scenarios mapping to the first three 3GPP-WLAN interworking scenarios [14]. Scenarios 4, 5 and 6 (addressing inter-system mobility) are currently out of the scope.

- Scenario 1 is the simplest case and does not have any impacts on either 3GPP or WiMAX architecture. A user will be charged on the same bill for his usage of both 3GPP and WiMAX services and custom care will be ensured without dependency on the access network he is connected to.

- In Scenario 2, a subscriber may use the WiMAX to access services but AAA operations are handled by the 3GPP system.

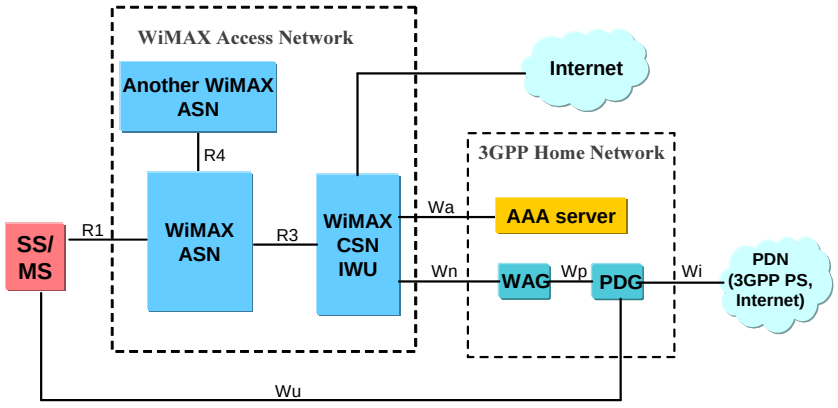

Figure 1.10: WIMAX - 3GPP interworking architecture reference model

- Scenario 3 allows operators to extend 3GPP PS-based services to the WiMAX network. In this scenario, an authenticated 3GPP subscriber can access to 3GPP PS services through a WiMAX access network interworking with its 3GPP Public Land Mobile Network (PLMN) (non-roaming case) or with a visited 3GPP PLMN (roaming case). The WiMAX access network is composed of WiMAX Access Service Network (ASN)s, connected to the 3GPP network through a WiMAX Connectivity Services Network (CSN) providing IP connectivity. The WiMAX-3GPP interworking according to scenario 3 is depicted in Figure 1.10.

1.4 Mobility management in heterogeneous networks

1.4.1 Handover terminologies

- *Horizontal vs. vertical handover:* During the handover process, the Mobile Node (MN)'s point of attachment changes from one access node to another one. In horizontal handover, the MN moves within a single access technology whereas in vertical handover, the access technology changes.
- *Make-before-break vs. break-before-make handover:* In make-before-break handover, the connection with the new target access node is established before releasing the connection with the old one. Conversely, in break-before-make handover, the old connection is terminated before the new one with the new target access node is established.
- *Hard vs. soft handover:* A hard handover is one in which the connection in the serving cell is released and only then the connection in the target cell is engaged. A hard handover is also known as a break-before-make handover. A soft handover is one in which the connection in the serving cell is retained and used for a while in parallel with the connection in the target cell. This handover is called a make-before-break handover.
- *Layer-2 vs. Layer-3 mobility:* The layer-2 (L2) mobility refers to the case where the MN roams among different access nodes while the point of attachment to IP network remains the same. Otherwise, the layer-3 (L3) mobility involves the change of IP addresses.
- *Global vs. local mobility:* The global mobility protocol handles mobility across access systems by associating the global IP address with the new local IP address at a fixed global mobility anchor. The mobility within one access system is managed by a local mobility management protocol.

1.4.2 Handover procedure

The handover procedure can be split into three phases: 1) neighboring cell discovery and measurement, 2) network selection and handover decision, and 3) handover execution.

1.4.2.1 Cell discovery & Measurement

The role of cell discovery and measurement is to identify the need for handover. This phase includes the following steps:

- *Neighboring cell discovery:* It is a preliminary step to be considered before carrying out the signal strength measurement. The MS can learn about its neighbors by scanning different channels or via the provisioning information from its current Base Station (BS).
- *Signal strength measurement:* The MS should synchronize in frequency and in time with its neighboring cells before it measures their radio link quality. The signal strength is averaged over time so that fluctuations due to radio propagation can be eliminated. Besides the measurement taken by the MS, the network makes itself the measurements such as the uplink quality, Bit Error Rate (BER) of the received data, etc.
- *Reporting of measurement result:* After the measurement, the MS sends measurement results to the network periodically or based on trigger events.
- *Information gathering:* Besides the physical link quality related parameters, in heterogeneous environments, the MS is required to collect other information like the terminal capabilities, service experiences status, context information, etc. to assist the vertical handover decision.

1.4.2.2 Network selection and Handover decision

This phase is responsible for determining when and how to perform the handover. We can divide this phase into different steps:

- *Network selection triggering (including handover initiation triggering):* Network selection is triggered taking as input the measurement results.
- *Network selection:* Network selection is the process of choosing the best access network among the multiple available ones. In heterogeneous environments, the MS must evaluate different criteria of each available network before selecting the best one. The selected access network must be commonly agreed between the user preferences and the network policy including the roaming agreement.
- *Handover initiation:* If the network selection results in change of access node, the handover initiation must follow right after. If the access technology of the selected access node is different from the serving access technology, a vertical handover is executed.
- *Pre-notification to all recommended target BSs:* The network selection gives a list of recommended BSs in the preferred networks order. In this case, the network may query the recommended BSs to check whether they can support the imminent handover from the MS. During this phase, certain pre-registration information of the MS will be relayed to the recommended target BSs for handover preparation purpose. At the end of this phase, the network can decide

which target access network to select and send its decision to the MS. Another option is that the network eliminates the undesirable BSs among the recommended ones and then sends back the list of desirable recommended BSs to the MS. Here, the target access network is selected by the MS. Such a pre-notification handover only exists if the MS and the network cooperate together during the network selection and handover decision phase. Otherwise, the MS or the network can decide solely the target access node.

1.4.2.3 Handover execution

The handover execution includes the connection establishment, the resources release and the invocation of proper security services.

- *(Re-)Authentication:* Once the target access network is selected and the handover decision is launched, the MS must use appropriate user credentials to authenticate with the target network and get valid encryption keys for communication sessions.
- *Execution:* Once the best access network is selected, and the re-authentication is successfully achieved, the communication session will be continued on the new radio interface through a new routing path. The change of routing path must be notified to the Corresponding Node (CN) or the content provider.

1.4.3 Mobility management classification

1.4.3.1 Link layer mobility management

The term link layer (L2) refers to everything that is below the IP layer. The L2 mobility management mechanism allows the MN to roam among different physical points of attachment while the point of attachment to IP network remains the same. The link layer mobility solutions are more or less link technology specific, for instance both IEEE 802.11f and GPRS networks can provide link layer mobility.

In this mobility management scheme, no IP subnetwork configuration is required upon movement. But some IP signaling may be required for the MN to confirm whether or not there was a change of the wireless access node. For example, an 802.11 network consists of several APs interconnected by means of a distribution system. When the Signal-to-Noise Ratio (SNR) drops below a certain threshold, the MN scans for the best available AP in the L2 network and re-associates with it. An L2 update frame is broadcasted in order to register the MN's current location with all bridges and switches in the distribution system. This L2 handover in WLAN was proposed in IEEE 802.11f amendment (withdrawn in Feb. 2006).

1.4.3.2 Network layer mobility management

1.4.3.2.1 Macro vs. Micro mobility

The mobility management at L3 can be broadly categorized into two types: *macro-mobility and micro-mobility* (see Figure 1.11). Such a distinction is based on the *domain* concept according to whether the movement of the mobile host is intra-domain or inter-domain. The mobility between different administrative domains is referred to as the *macro mobility* since it will be global and independent of underlying mechanisms such as routing protocols, link layer access techniques, and security

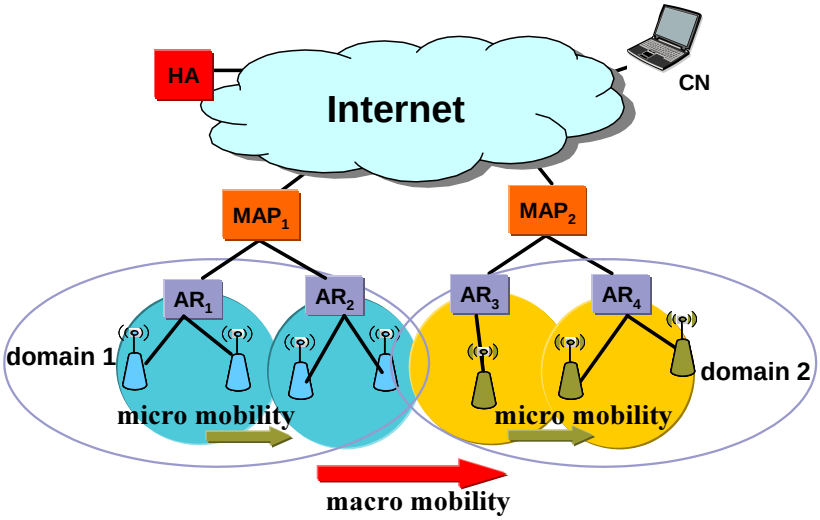

Figure 1.11: Micro vs. macro mobility management

architectures. On the other hand, the term *micro mobility* refers to the movement of the mobile host between different subnetworks belonging to a single domain.

As shown in Figure 1.11, a macro (global) mobility protocol is employed when an MN moves between two access domains. The mobility between two APs under the same Access Router (AR) constitutes an L2 mobility. Between these two lies micro (local) mobility. Micro mobility occurs when an MN moves between two APs connected to two different ARs.

i)Macro mobility

When an MN moves between the different subnetworks or different domains, its IP address will be changed. In order to maintain the connectivity, the MN should have a mechanism to inform quickly its CN about its new address or it should have a permanent IP address seen by its CN. MIPv4 [18] is proposed to solve the problem of node mobility. MIP is a standard that allows a user with a mobile device, whose IP address is associated with a particular network, to remain connected when moving to a network with a different IP subnetwork address. It is designed around two components namely, HA and FA. The HA maintains the database of current locations of all the mobile terminals under its control. When the MN moves away from the home network to a foreign network, the HA updates the current location of the MN with the address of the FA (called CoA). When a packet is addressed to the MN, it first reaches the HA, then HA encapsulates this packet with the FA address as a destination. Upon receiving this packet, the FA removes the IP header information inserted by the HA and sends the packet to the MN. An address translator implemented in the HA and a tunnel between the HA and the FA are required. Shortly, MIP keeps track of the location of the MN and delivers packets to its current location.

MIPv6 [19–21] includes many features for mobility support that are missing in MIPv4. MIPv6 can support the vertical handover in the network layer without the need of FA and address translator. MIPv6 uses two separate IPv6 addresses: the home address (HoA) as identifier and the CoA as locator. On its home link, the MN uses its HoA just like a stationary node. When the MN moves to a foreign link, it configures a CoA with the foreign prefix. It can provide a transparent movement to the transport and upper layers.

Besides the MIP, new Internet Engineering Task Force (IETF) work on the global mobility management protocols has proposed Host Identity Protocol (HIP) [22] and IKEv2 Mobility and Multihoming

(MOBIKE) [23, 24] solutions. The HIP provides a method of separating the end-point identifier and locator roles of IP addresses. In HIP, the Host Identifier (HI) is the public key of a public-private key pair. The HI is represented with a 128-bit long Host Identity Tag (HIT) that is created by taking a cryptographic hash over the corresponding HI. Since each HIT can be mapped dynamically to multiple IP addresses, HIP enables mobility and multi-homing. Handover management is carried out by means of direct peer notifications. Location management is provided by rendezvous servers. In fact, the HIP solution introduces a HIP layer between IP layer and transport layer. The upper layer sockets are bound to host identifiers instead of IP addresses. Thus, the HIP is sometimes referred to as a L3.5 mobility management.

The MOBIKE protocol facilitates Virtual Private Network (VPN) users to change from one address to another without re-establishing all security associations, or to use multiple interfaces simultaneously. The MOBIKE protocol is assumed to work on top of Internet Key Exchange version 2 protocol (IKEv2) [25]. It allows to change an IP address associated with IKEv2 and an IPsec tunnel to change without reusing IKEv2 from scratch.

ii)Micro mobility

To enhance the Mobile IP, the so-called micro mobility protocols have been developed to manage handovers within a single administrative domain. The micro mobility solutions aim to reduce the signaling overhead and the handover latency of the MIP mechanism. We can distinguish two kinds of approaches: tunnel-based and routing table-based.

The tunnel-based solutions consist on using the local and hierarchical registration. The tunnel-based solutions include Hierarchical MIP (HMIP) [26,27], Fast MIP (FMIP) [28] and Intra-Domain Mobility Management Protocol (IDMP) [29]. The HMIP is designed to minimize the signaling overhead to the CN and the HA. This is achieved by allowing the MN to locally register with a domain, named local mobility zone. These zones form the independent subnetwork domains which are connected to the Internet via a Mobility Anchor Point (MAP). The MAP acts as a local HA of the MN. When the MN moves inside the local MAP domain, it only needs to register the new location with the MAP. The handover is thus hidden at the HA and the CN.

The FMIPv6 [28] is an extension of MIPv6 that allows the MN to configure a new CoA before it moves to the new network and thus can use it immediately once connecting to the new network. The FMIPv6 can reduce the latency involved during the MN's binding update by providing a bi-directional tunnel between the old and the new AR. When the MN switches to the new link, the previous AR routes all packets to the new MN's CoA. Therefore, the MN updates its location and receives packets in parallel.

The routing-based solutions consist on maintaining host-specific routes in the routers to forward packets. The host-specific routes will be updated based on the host mobility information. Cellular IP (CIP) [30] and Handoff Aware Wireless Access Internet Infrastructure (HAWAII) [31] are the two examples of the routing-based micro-mobility protocols. A CIP system consists of a number of components to allow access, paging and mobility management. The concept behind CIP is similar to the mobility management of voice terminals in GSM. Its goal is to allow the idle MN to have discontinuous transmissions. As long as MNs can be traced within a paging area, they do not have to register every move during passivity. A visiting MN must register to the gateway and use its IP address as its care-of address. All the packets destined to the MN first reach the gateway from which they are routed to their respective IP address.

1.4.3.2.2 Host-based vs. Network-based mobility

Most of proposed micro mobility protocols like FMIPv6 [28] and HMIPv6 [27] are *host-based solutions* as they require host involvement at the IP layer. Recently, the success in the WLAN infrastructure market of WLAN switches, which perform mobility management without any host stack involvement, suggests a new way to manage the micro mobility. A new protocol that is network-based and that requires no software on the host is therefore desirable. The distinction between host-based and network-based mobility solutions is illustrated in Figure 1.12. The network-based mobility solution is a topic that has had a lot of attention within the IETF Network-based Localized Mobility Management (NETLMM) working group [32].

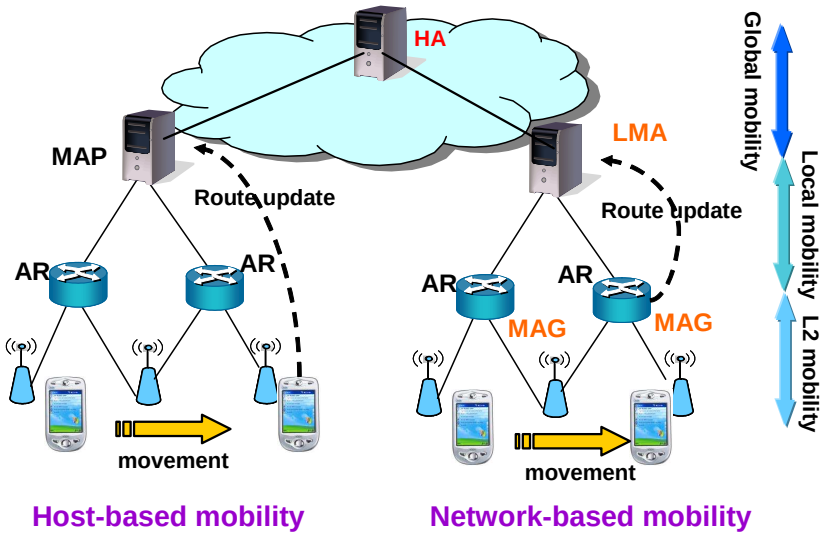

Figure 1.12: Host-based vs. Network-based mobility

Proxy MIPv6 (PMIPv6) [33] has been designated as the network-based localized mobility management protocol. The PMIPv6 protocol introduces a new entity, the Proxy Mobile Agent, which is a kind of MIPv4 FA sitting on the AR, also known as Mobile Access Gateway (MAG). Between AR and MN, a secure point-to-point link will be established. The MAG will handle mobility on behalf of the MN, using functionality similar to MIPv6's. Because of the use and extension of MIPv6 signaling and HA functionality, this protocol is referred to as Proxy MIPv6. The MAG is responsible for detecting the MN's movements to and from the access link and initiating binding registrations to the MN's Local Mobility Anchor (LMA). The MN may operate in an IPv4-only mode, IPv6-only mode or in dual IPv4/IPv6 mode.

When the MN changes its point of attachment, the MAG on the previous link will detect the MN's detachment from the link and will notify the LMA. The MAG will remove the binding and routing state for that MN. The MAG on the new access link upon detecting the MN on its access link will notify the LMA for updating the binding state. Once that signaling is complete, the MN continues to receive the Router Advertisements containing its home network prefix. Accordingly, the MN believes that it is still on the same link and it will use the same address configuration on the new access link.

1.4.3.3 Upper layer mobility management

Mobility management has also been considered at layers above IP (upper layers). For instance, the Transmission Control Protocol (TCP)-Migrate [34] adds mobility support to TCP sessions. Specifically, it implements extensions to the TCP protocol, so that TCP sessions can continue without interrupt when an endpoint changes its point of attachment. Similarly, mobile Stream Control Transmission

Protocol (SCTP) [35] builds upon the features of the SCTP transport protocol to offer transport layer mobility. Seamless mobility is inherent in the SCTP and is achieved through the multi-homing support feature of the SCTP and some of its extensions (e.g., Dynamic Address Reconfiguration) [36].

The mobility management at the application layer has also been studied. The most well-known solution is the Session Initiation Protocol (SIP)-based mobility management [37]. In this approach, the infrastructure of SIP is reused for mobility purposes. To make handover, the MN sends a re-INVITE message with its new address to its CN using the same call identifier as in the original setup [38]. The handover delay in a SIP-based mobility includes the L2 delay, the IP address configuration delay and the time required by the re-INVITE message to reach the CN. The common factor in these approaches is that they apply to specific protocols and applications and do not cover the full spectrum of Internet applications.

1.4.3.4 Cross-layer mobility management

The handover latency of MIP is due to the movement detection and registration. The proposed micro-mobility can only solve the latter one. By using the link layer information such as signal strength, the latency for handover detection can be reduced. The information from the link layer can also be used to notify the network to prepare the handover in the target network and to initiate the L3 handover procedure. Some algorithms use signal strength measurements directly to reduce handover latency [39], while others track the MNs via the received signal strength and use this tracking information to support low-latency MIP handover [40]. A seamless handover architecture for MIP, S-MIP [40] provides a way to combine a location tracking scheme and the hierarchical MIP handover to enhance the management process. The use of L2 hints for L3 handover has been widely explored in the literature. It refers to a L2/L3 cross-layer mobility management approach.

In the cross-layer perspective, the combination of MIP (network-layer protocol) and SIP (application-layer protocol) was investigated in [41, 42]. In this scheme, handover occurs at the network layer enabling connection oriented traffic to reach the destination using MIP. In parallel, the SIP peer is notified about the handover by using re-INVITE with the new CoA address indicated in the Contact field. Real-time traffic can be sent directly between peers avoiding suboptimal paths.

One can recognize that a single layer-specific mobility management protocol can hardly provide the advanced mobility support in heterogeneous networks. The intrinsic reason is that mobility brings about significant impacts on each layer, which in turn has its convenience to deal with different level mobility impacts. A multi-layer architecture that can make full use of each layer's contributions while still keeping the basic structure of the protocol stack is highly demanded [43].

1.5 Summary

Convergence stands as an unavoidable evolution of telecommunication networks. There is no doubt about this event because the ubiquity of broadband and IP is radically changing the business model of operators. A quadruple play service combining the triple play service of broadband Internet access, television and telephone with wireless mobile services is an example of convergence facility. A clear trend is emerging in the form of fixed and mobile telephony convergence, access technology convergence, service convergence, multi-standard device convergence, etc. Interworking of heterogeneous networks is inevitable for the sake of user service continuity taking advantages of each network. To this end, the mobility management is one important aspect that needs to be thoroughly studied. To provide seamless mobility over heterogeneous networks, it is very important to provide low latency

handover. The cross-layer mobility management approach and the hierarchical mobility management seem to be a good way to manage the mobility.

In this chapter, we have addressed a global view of the mobile network evolution from the first generation towards the next future generation which is likely to be characterized by the interworking of heterogeneous access networks. The mobility management, one of the key glues for the success of such a convergence, was introduced. In the following chapters, we will focus on our solution aiming at optimizing different aspects of the inter-system mobility management.

∼∼ △♡△ ∼∼

Part I
User-Controlled Approach

Mobile systems increasingly become an inseparable part of our daily lives in various branches of living (e.g., work, education, entertainment, health care, commerce...). And people are looking for a life that is more enriched and cultural, more flexible and diversified, more comfortable and safe, and more personal and convenient. Clearly, people expect that the next generation of mobile communication systems will provide something more than just "faster speed". Facing up to the ubiquitous access service, users will plan to take advantage of the providers' competition. Users will need easy-to-use multi-modal natural human interfaces like voice and gestures but also need to control and customize their perceived services. As mobile terminals are evolving towards being more intelligent and more powerful, they can aid users to handle the control without any expertise. It is time for network operators to delegate a part of the mobility management control to the mobile users (i.e., mobile terminal). In this part of the thesis, we explore how the mobility management can be achieved and improved under the user-centric terminal-controlled approach.

Chapter 2

Utility-based Access Network Selection

Customer experience is the defining success factor in business for the next twenty years. Learning from customers, creating the experience they want, measuring the success of what you do, continually fine-tuning the service and returning to customers to learn more. If you create a great customer experience, you'll almost certainly win.

Mark Hurst, Founder of customer experience consultancy Creative Good

ABC is a fundamental challenge for fourth generation heterogeneous wireless networks. The ABC concept refers to being not only always connected but also connected through the best available device and access technology at all times. The key to being "Always Best Connected" is the access network selection mechanism. Access network selection becomes an important step in the terminal-controlled handover management over heterogeneous wireless environments.

In this chapter, we analyze utility theory with a view to defining an appropriate decision metric for access network selection. Existing utility models are reviewed and their limitations are highlighted. We propose new single-criterion and multi-criteria utility forms to best capture user satisfaction and sensitivity in varying access network characteristics. We conduct simulations and analyses of the proposed model to show that it effectively allows end-users to select the best access network and helps operators to optimize the use of their resource.

2.1 Introduction

Advances in wireless communication technologies are driving the evolution towards ubiquitous and seamless service delivery across multiple wireless access systems. As described in Chapter 1, the integration of multiple access technologies deployed by different operators is fundamental for future 4G mobile heterogeneous networks. Future users will need this diversity and this interworking between access systems in order to maximize their profitability or improve the perceived QoS. New intelligent selection mechanisms are therefore needed to handle the complexity of the seamless handover and to select the best available access network that satisfies QoS requirements at the lowest cost and energy use.

In this vision, network selection becomes a key element of the handover procedure. Network selection includes the handover decision and drives the handover execution. In traditional homogeneous

networks, network selection is based only on factors of signal quality from serving and neighboring access nodes, like Received Signal Strength (RSS) or Signal to Inference plus Noise Ratio (SINR). But in heterogeneous networks with universal access facilities, the selection process becomes more complex because different access technologies usually provide different characteristics (QoS support, billing schemes, reliability degree...). Network selection becomes a multi-criteria decision-making problem that involves a number of parameters and complex trade-offs between conflicting criteria. A variety of access network characteristics have been considered and identified as potential network selection criteria in the literature [44–51]. Characteristics include link quality, availability, throughput, network load, file transfer delay, reliability, power consumption, bandwidth, cost of service, handover frequency, and terminal's velocity. Selection schemes consider subsets of these criteria in their decision-making strategy. In this chapter, we do not discuss a suitable subset of network selection criteria; rather, we focus on how criteria are used to make the right decision.

The complexity of access network selection is recognized as an NP-hard optimization problem [52]. There is no optimal solution since each user has his own preferences. Satisfying all criteria can prove difficult as some criteria may conflict. One user may prefer the cheapest access network while another may prefer the access network providing the highest performance. In fact, user preferences become a means to overcome the complexity of the making-decision process. They establish a rating relationship among a set of criteria and a degree of significance for each criterion. More precisely, each preference has a relative weight that users assign to each criterion depending on their requirements. Once the criteria are identified and the preferences are fixed, we need a method to compare candidate networks in order to identify the most suitable one. Usually, the decision will be based on perceived utility. This utility is also used to deduce the network operator's payoff in the radio resource allocation game. Utility is a key metric in network selection and resource allocation. Our goals are to adapt existing utility models to wireless system characteristics and user experiences, and to make it possible to analyze and quantify the QoS, and consequently the user satisfaction, offered by an access network node.

The remainder of this chapter is as follows. In Section 2.2, we summarize existing research on utility-based network selection metrics. Utility theory in wireless networks and the user acceptance probability concept are introduced and discussed in Section 2.3. Based on utility theory and user experiences, existing utility models and their limitations are thoroughly analyzed in Section 2.4 and Section 2.5. In these two sections, we also propose new single-criterion and multi-criteria utility forms to overcome the existing shortcomings. We demonstrate that our proposed models assist both users to select the best access network and operators to manage their resource allocation.

2.2 Related work and Motivation

Multi-criteria selection is a classic problem in economics and in many other fields. In network selection, one popular solution is a scoring method that quantifies the score (suitability level, value, worth) of a particular network [53] [54]. In general, the score of access network i is computed as $U_i = \sum_j w_j f_j(x_{ij})$ where x_{ij} is the value of criterion j in access network i, w_j is the preference weight of criterion j ($\sum_j w_j = 1$), and $f_j(.)$ is a normalized function. The normalized function is introduced to express different characteristics of different units with a comparable numerical representation. Normalized functions take various forms: a logarithm form was used in [44], an exponential form was proposed in [45] and a linear piecewise form was studied in [48, 51, 55].

In recent years, a utility-based microeconomics model has been applied to power control in wireless cellular systems [56], to radio resources management in wired and wireless networks [57–60], and to network selection strategy [47–50]. In this model, utility refers to the level of usability that a user derives from a given product, therefore reflecting customer decision experiences [61]. In access network

selection and radio resource management, it measures the users' satisfaction level corresponding to a set of characteristics of an access network, including the allocated resource parameters. Normalized function $f_j(.)$ is called a single-criterion utility function and total score U_i is a multi-criteria (aggregate) utility. In fact, the score (cost) of a particular access network is itself simply a utility. In addition to the logarithm, exponential and piecewise-linear utility forms mentioned previously, a sigmoid function has also been used to model single-criterion utility [49, 56–60].

Besides a weighted sum of all criterion utilities, an *acceptance probability* was also used as a decision metric. Acceptance has been defined as an outcome variable in the psychological process that users go through in making decisions. The concept of acceptance probability in radio resource management and access network selection was introduced in [62] and reused in [49, 58, 60, 63–67]. Acceptance probability means that a user may choose whether to accept (or select) an access network based on its intrinsic characteristics and on the amount of resources allocated to him. Generally, taking into account the user acceptance probability, a network operator plans his resource allocation in order to maximize revenue. As a microeconomics concept, acceptance probability is based on user utility and the price that the user is willing to pay for the connectivity service. Hence, the acceptance probability is approximately equivalent to an aggregate utility metric.

The Analytic Hierarchy Process (AHP) [68] is a technique developed by Saaty for multi-criteria decision analysis that has been recently applied to network selection [69] [70]. The AHP involves an importance-ratio assessment procedure and uses a hierarchy to establish preferences and order. It is also sometimes classified as a multi-criteria utility approach [71]. But AHP differs in the way in which it determines the weight assigned to a criterion and the score assigned to an alternative for each criterion. The AHP continues to be the subject of much debate, especially as it relates to utility theory. The choice between utility theory or the AHP is a matter for each user to decide [72]. In this chapter, we explore only the use of utility theory.

As described previously, many different single-criterion and aggregate utility functions exist. An obvious question is whether they are all suitable for modeling user behavior in the uncertain wireless radio environment. We will demonstrate that our proposed utility model not only allows users to select the best access network but also helps operators to optimize their resources allocation and enhance their revenues.

2.3 Utility theory

2.3.1 Utility theory for wireless network environments

Basic utility theory was developed by Von Neumann and Morgenstern [73]. Subsequently, the theory has been further explained and considerably developed. In microeconomics, utility means the ability of a product or a service to satisfy a human need. An associated term is utility function: the utility derived by a consumer from a product or a service. Different consumers with different user preferences (tastes) will have different utility values for the same product. This means that individual preferences should be taken into account when evaluating the utility. The concept of utility applies to both single-criterion (attribute, characteristic) and multi-criteria consequences. A utility function is defined mathematically as a function $U(\mathbf{w},\mathbf{x})$ from a set of observed product criteria $\mathbf{x}$ (by the user) and user preferences $\mathbf{w}$. As the user preferences associated to a set of criteria do not change when alternatives are considered, we can simply denote $U(\mathbf{x})$ as the utility function associated with criteria vector $\mathbf{x}$ for the product being considered. The mathematical properties of the function are described below.

Utility is an ordinal concept. It quantifies preferences among alternatives in the process of making a decision. The preference relationship can be represented by a continuous utility function. The the-

ory's fundamental assumption is that decision-makers are rational, that is, they will always choose the alternative with the highest utility value. However, just knowing that a user prefers item p to item q gives no indication of the extent of that preference. If $U(\mathbf{x}_p) = 3U(\mathbf{x}_q)$, p is preferred to q but p is not necessarily three times better than q.

When evaluating the utility of an access network, we distinguish between *upward* and *downward* criteria. A criterion is classified as *upward* if its utility is an increasing function of its value. Upward criteria include parameters such as allocated bandwidth, throughput, reliability degree, and RSS. Conversely, the utility of a downward criterion decreases in function of its value. Downward criteria include parameters such as network usage cost, energy consumption, bit error rate, transfer delay and handover frequency. Price is traditionally considered as a separate criterion that is completely different from others. However, price is not only the network usage cost but also, for example, the power dissipation at the terminal side. Accordingly, we prefer not to distinguish between price and other criteria. One access network is clearly preferred to another if it has higher values of upward criteria, lower values of downward criteria or both. In the following, we investigate the utility for upward criteria and then extend the results to downward criteria.

Let us consider a utility function $u(x_i)$ of an upward quality-related parameter x_i, $0 \leq x_i < \infty$. In general, we can consider x_i as an amount of resources that an access network can allocate to a user. Every parameter has an upper and a lower limit due to technological constraints and the user's requirements (i.e., $x_\alpha \leq x_i \leq x_\beta$). Utility therefore has an upper value. Consequently, we can normalize the utility by scaling to the interval [0,1], i.e., $u(x_i) \in [0,1]$. First, the utility function should be twice differentiable on interval $[x_\alpha, x_\beta]$. This reflects the fact that the utility level should not change drastically given a very small change in the value of a criterion (product's characteristic) and the marginal utility should be regular. Second, the utility function is a non-decreasing function of x_i. That is, the more resources allocated to the user, the higher the utility [60]. However, the improvement of the utility disappears when the allocated resources reach a certain threshold and the upper level of user satisfaction is obtained. In fact, it obeys the law of diminishing marginal utility, i.e., $\lim_{x_i \to x_\beta} u'(x_i) = 0$. The effect of diminishing marginal utility implies the concavity of $u(x_i)$ for x_i greater than a given value.

These requirements can be summarized in the following condition statements:

$$u'(x_i) \geq 0 \tag{2.1}$$

$$\exists x_c : u''(x_i) < 0, \quad \forall x_i \geq x_c \tag{2.2}$$

Similarly, when quality-related parameter x_i goes below a certain threshold and the utility comes close to zero, user behavior is indifferent to the decrease of x_i. In other words, the decrease in utility is negligible according to the decrease of the allocated resource if the latter is still less than a certain threshold x_v. This implies the convexity of $u(x_i)$ for x_i less than a given value:

$$\exists x_v : u''(x_i) > 0, \quad \forall x_i \leq x_v \tag{2.3}$$

Though the condition (2.3) is reasonable in wireless networks, it has not been considered in any of the existing utility-based network management solutions. The twice differentiability, and conditions (2.1), (2.2) and (2.3) stipulate the form of the utility function for an upward criterion. Conversely, the utility function of a downward criterion x_j, denoted as $v(x_j)$, should have the following properties: *(i)* twice differentiability, *(ii)* decreasing in function of x_j, *(iii)* concavity for x_j lower than a given value, and *(iv)* convexity for x_j greater than another given value. We can easily deduce the form of $v(x_j)$ as $v(x_j) = 1 - u(x_j)$.

In addition to the functional properties of a utility function, it is noteworthy that a utility function is unique up to a monotonic transformation [53] [73]. This means that applying a monotonic transforma-

tion to a utility function simply creates another utility function representing the same preference relationship. More generally, if $g(.)$ is an increasing function and $u(.)$ is a utility function, $g \circ u = g(u(.))$ is a utility function with the same preference relation as $u(.)$.

2.3.2 The concept of acceptance probability

Acceptance probability is a concept used in decision theory to measure the probability that a user accepts a given product or service. In radio resource management and access network selection, acceptance probability is an important indicator by which network operators measure or estimate the user behaviors in accordance with the operators' resource allocation strategies. If $\mathbf{x}$ is the characteristic vector of the access network under consideration and $a(\mathbf{x})$ is a random variable (whose value is 1 if the user selects this access network and 0 otherwise), then a($\mathbf{x}$) is a Bernoulli random variable with success probability $A(\mathbf{x})$. In fact, $A(\mathbf{x})$ is the user acceptance probability [64].

Acceptance probability has been defined as a tradeoff between the perceived QoS and the price to be paid [49, 58, 60, 62–67]. It is an increasing function of the utility and a decreasing function of the price. As we do not distinguish between price and other criteria, simply between downward and upward criteria, we can generalize acceptance probability as an increasing function of upward criteria and a decreasing function of downward criteria. That is,

$$\frac{\partial A(\mathbf{x})}{\partial x_i} \geq 0 \qquad \frac{\partial A(\mathbf{x})}{\partial x_j} \leq 0 \tag{2.4}$$

where x_i is an upward criterion and x_j is a downward criterion. If all the criteria satisfy the user (i.e., the utility of every criterion is equal to 1), he has no reason to refuse the service. We therefore have the following conditions:

$$\forall x_i \in \mathbf{x} : u_i(x_i) = 1 \Rightarrow A(\mathbf{x}) = 1 \tag{2.5}$$

Acceptance probability and aggregate utility are very similar concepts. Instead of selecting the access network with highest utility, the user can select the network with the highest acceptance probability. In fact, user acceptance probability is an indicator for operators and is logically evaluated by operators while aggregate utility is evaluated by users. Acceptance probability and aggregate utility are therefore similar only if both operators and users have access to the same information (that is, network selection criteria $\mathbf{x}$ and user preferences $\mathbf{w}$). Usually, a network operator does not precisely know the user preferences $\mathbf{w}$ or user-oriented network selection criteria (such as remaining battery in the terminal). To deal with these unobservable parameters, network operators can estimate their probability and use the expected values. The estimation of unobservable parameters is a classic problem in economics [74]. Hereafter, the acceptance probability is denoted as $A(\mathbf{x},\mathbf{w})$ to take into account user preferences $\mathbf{w}$ and access network characteristics $\mathbf{x}$.

In radio resource management, the network operator's decision metric (i.e., payoff) is mostly the revenue. The revenue, calculated over N users requesting a connection to the network, is: $R' = \sum_{k=1}^{N} p_k a_k(\mathbf{x}_k, \mathbf{w}_k)$ where p_k is the price that user k has to pay for the operator's connectivity service. Its expectation, also called potential revenue, is:

$$R = E(R') = \sum_{k=1}^{N} p_k A_k(\mathbf{x}_k, \mathbf{w}_k) \tag{2.6}$$

This is widely used to represent the operator's payoff in radio resource management. The user's payoff is, in fact, the aggregate utility.

2.4 Single-criterion utility function

Several works in the literature have addressed different forms of utility function, mainly step function, piecewise, logarithmic, exponential, and sigmoidal. Utility function has also been proposed to reflect the running application's degree of elasticity [75–77] or user tolerance (or irritation) in the face of service degradation [78]. Yet there is no consensus about a suitable form of utility function to model user satisfaction. Before proposing a new form of utility function, therefore, we present an overview of those that already exist.

2.4.1 Survey of single-criterion utility

2.4.1.1 Application's elasticity-based utility forms

Besides the access network characteristics criteria, the decision on network selection also depends on the applications that the user is running. The current applications represent the user's required QoS and preferences. There have been several initiatives to model the utility in accordance with the application's elasticity [75–77]. Elasticity was considered in the design of the utility function in [75] for admission control in the Internet and was used for load-balancing in beyond 3G networks in [77]. An application can be mapped to one of the following elasticity degrees based on its sensitivity to QoS parameters: inelastic, perfectly elastic and partially elastic. Different degrees of elasticity are identified by different utility functions.

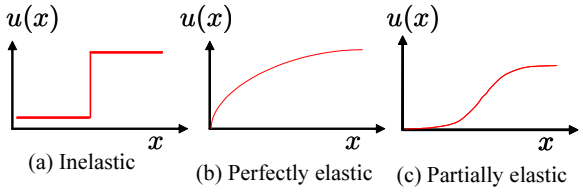

Figure 2.1: Utility function vs. application's elasticity

Real-time voice and video applications with constant bit rate are inelastic in their demand for bandwidth. Their utility is modeled as a step function with only two values, satisfied or unsatisfied, depending on whether the allocated bandwidth is above or below a given threshold (see Figure 2.1(a)). This kind of utility function is initially associated with the amount of bandwidth that the network promises to reserve for an application session in a wired network environment [75] [76]. We agree that inelastic applications are very sensitive to the variation of the resource allocation; utility modeling should take this sensitivity into account. However, the use of the step utility function as in [77], and particularly in network selection, is not appropriate. The QoS-related criteria of an access network are subject to variation due to the fluctuating nature of wireless networks. We cannot say that the user is happy if the allocated resource is B and not happy at all if it becomes $(B-\varepsilon)$. Furthermore, the QoS requirements for real-time voice and video applications are usually characterized by a minimum or maximum guaranteed amount [79] not by a fixed threshold. The step function therefore does not provide the regularity required of a utility function.

On the other hand, traditional data applications such as web browsing or email are *perfectly elastic* [75] [76] since they are assumed to tolerate variations in delay and to be able to use even the minimal amount of bandwidth. The utility for upward criteria in this case is modeled as an always concave function (see Figure 2.1(b)). We agree that non-real time applications are less sensitive to variations of network quality. Nevertheless, an always-concave function is not suitable for modeling user behavior.

First, the utility for a criterion like RSS cannot always be concave, especially for small values. A minimum level of RSS is required to establish and maintain the connectivity. Second, different users will assign different values to a given application and its QoS. A common, but we believe erroneous, assumption is that non real-time data applications have lower priority than real-time applications and that they are tolerant to delay. In fact, some users may assign a higher priority to their email traffic than to their video sessions. Some users may become very annoyed after waiting seconds for a web page to be refreshed or a whole day for a video clip to be downloaded. It is more accurate to model the utility based on user irritation, such as what amount of performance degradation (delay, for example) the user is willing to tolerate without complaint, rather than considering the application outages themselves. The unsuitability of the always-concave function parallels the argument in the convexity condition (2.3).

Lastly, the S-shaped utility function (see Figure 2.1(c)) was used for partially elastic applications like real-time applications with adaptive coding and a minimum of required bandwidth in [75] or non real-time data applications in [77]. This type of function satisfies all the conditions of a utility function identified in Section 2.3.1. We believe that it can be adapted to model the utility according to different degrees of user irritation or sensitivity for all network selection criteria, not just QoS-related criteria. As the application priority depends on the user choice, it is better to model it by user preferences. In other words, as discussed in [80], user preferences should be application-based and terminal capabilities-aware.

2.4.1.2 Evaluation of existing utility function forms

Utility form	Reference	Generalized mathematical formula	Increasing & Differentiability	Concavity	Convexity
Linear	[48] [55]	$u(x)=\begin{cases}0 & x<x_{min}\\ \frac{x-x_{min}}{x_{max}-x_{min}} & x_{min}\le x\le x_{max}\\ 1 & \text{otherwise}\end{cases}$	Yes	No	No
Logarithm	[44]	$u(x)=\ln(x)$ or $u(x)=\ln(1+ax)$ $(a>0)$	Yes	Yes	No
Exponential	[45]	$u_1(x)=e^{(x-M)}$ $(0\le x\le M)$	Yes	No	Yes
Exponential	[47] [81]	$u_2(x)=1-e^{-ax}$ $(a>0)$	Yes	Yes	No
Sigmoid	[56] [78]	$u_1(x)=\frac{1}{1+e^{\zeta(x_m-x)}}$ $(\zeta,x_m>0)$	Yes	Yes	Yes
Sigmoid	[49, 50, 60, 65, 77]	$u_2(x)=\frac{(x/x_m)^\zeta}{1+(x/x_m)^\zeta}$ $(x_m>0,\zeta\ge 2)$	Yes	Yes	Yes

Table 2.1: Utility theory-based comparative study of existing utility functions

We now investigate existing forms of utility function in the literature by examining the required properties: twice differentiability, increasing function, concavity and convexity. As stated earlier, if $u(x)$ is suitable for an upward criterion utility, $(1-u(x))$ is suitable for a downward criterion utility. We therefore examine only the utility forms of an upward criterion. The results are shown in Table 2.1. Figure 2.2 illustrates these different utility forms. We see that only the sigmoidal (S-shaped) functions can satisfy all necessary conditions of a utility function in our context of network selection. These functions are:

$$u_1(x)=\frac{1}{1+e^{\zeta(x_m-x)}} \quad (\zeta,x_m>0) \tag{2.7}$$

$$u_2(x)=\frac{(x/x_m)^\zeta}{1+(x/x_m)^\zeta} \quad (x_m>0,\zeta\ge 2) \tag{2.8}$$

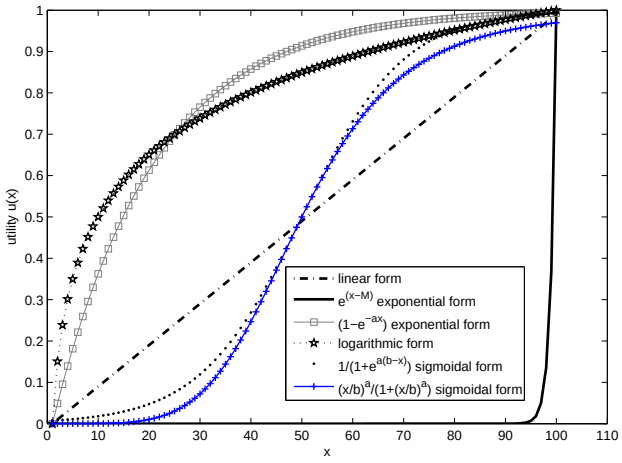

Figure 2.2: Illustration of different utility function forms

2.4.2 New single-criterion utility function

The sigmoidal form is suitable for modeling utility for each network selection criterion. However, tuning the parameters (e.g., ζ and x_m) to suit the technological and user constraints (i.e., lower limit x_α and upper limit x_β for each criterion), as well as user sensitivity, is challenging. Observing the sigmoid functions $u_1(x)$ and $u_2(x)$ given in (2.7) and (2.8), we see that $u_1(x_m) = u_2(x_m) = 0.5$. The value x_m corresponds to the threshold between the satisfied and unsatisfied areas of a specific parameter. The values of x_m and ζ determine the *center* and the *steepness* of the utility curve respectively. The parameter ζ makes it possible to model the user sensitivity to variation in access network characteristics. Note that x_m is user-specific and not necessarily the median of the interval $[x_\alpha, x_\beta]$. In addition to the four requirements identified in Section 2.3.1, we should redesign the sigmoidal utility function to satisfy the following conditions:

$$u(x) = 0 \quad \forall x \leq x_\alpha \tag{2.9}$$

$$u(x) = 1 \quad \forall x \geq x_\beta \tag{2.10}$$

$$u(x_m) = 0.5 \text{ for a given } x_m \tag{2.11}$$

Furthermore, the utility function should retain a steepness parameter so as to model user sensitivity.

Proposition 2.4.2.1. *Given a variation range of an upward criterion x, $x_\alpha \leq x \leq x_\beta < \infty$, and a middle point of the utility x_m, the suitable utility function for criterion x is:*

$$u(x) = \begin{cases} 0 & x < x_\alpha \\ \frac{(\frac{x-x_\alpha}{x_m-x_\alpha})^\zeta}{1+(\frac{x-x_\alpha}{x_m-x_\alpha})^\zeta} & x_\alpha \leq x \leq x_m \\ 1 - \frac{(\frac{x_\beta-x}{x_\beta-x_m})^\gamma}{1+(\frac{x_\beta-x}{x_\beta-x_m})^\gamma} & x_m < x \leq x_\beta \\ 1 & x > x_\beta \end{cases} \tag{2.12}$$

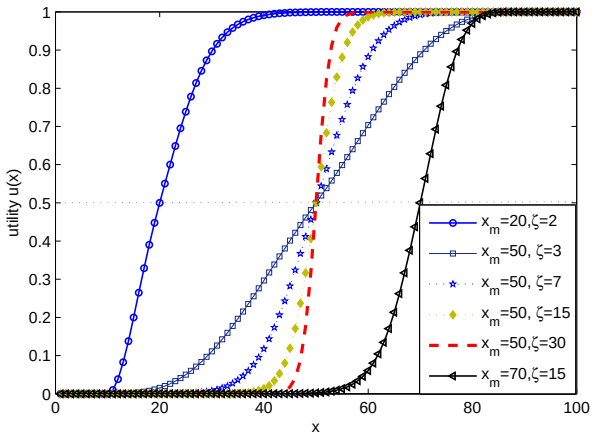

Figure 2.3: Single-criterion utility function forms for an upward criterion ($x_\alpha = 10$, $x_\beta = 90$)

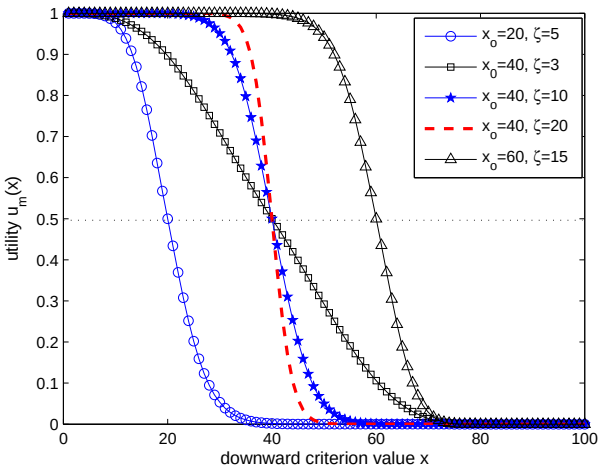

Figure 2.4: Single-criterion utility function forms for a downward criterion ($x_\alpha = 0, x_\beta = 80$)

where

$$\gamma = \frac{\zeta(x_\beta - x_m)}{x_m - x_\alpha} \tag{2.13}$$

$$\textit{and} \qquad \zeta \geq \max\{\frac{2(x_m - x_\alpha)}{x_\beta - x_m}, 2\} \tag{2.14}$$

ζ and γ are the tuned steepness parameters.

Proof. First, the proposed utility function clearly satisfies the conditions (2.9), (2.10) and (2.11). We see that the second and the third cases of (2.12) are similar to sigmoidal function $u_2(x)$. In order to show that (2.12) follows (2.1), (2.2), (2.3) and twice differentiability conditions, we only need to show that $\zeta \geq 2$, $\gamma \geq 2$ and the first derivative of $u(x)$ is continuous at x_m. From (2.14), we have $\zeta \geq 2$ and $\zeta \geq \frac{2(x_m - x_\alpha)}{x_\beta - x_m}$. Substituting the latter to (2.13), we have $\gamma \geq 2$. As (2.12) are differentiable, we have:

$$\lim_{x \to x_m^+} u'(x) = \frac{\zeta}{4(x_m - x_\alpha)} \tag{2.15}$$

$$\lim_{x \to x_m^-} u'(x) = \frac{\gamma}{4(x_\beta - x_m)} \tag{2.16}$$

By replacing (2.13) to (2.16), we have $\lim_{x \to x_m^+} u'(x) = \lim_{x \to x_m^-} u'(x)$. Hence, $u(x)$ is continuously differentiable and thus twice differentiable. $u(x)$ satisfies thus all requirements of a utility function. □

If a given criterion does not have an upper limit value (i.e., $x_\beta = \infty$), its utility will follow:

$$u(x) = \begin{cases} \frac{(\frac{x - x_\alpha}{x_m - x_\alpha})^\zeta}{1 + (\frac{x - x_\alpha}{x_m - x_\alpha})^\zeta} & x \geq x_\alpha \\ 0 & \textit{otherwise} \end{cases} \tag{2.17}$$

The form of the utility function for a downward criterion again is $(1 - u(x))$ where $u(x)$ follows (2.12) or (2.17) depending on whether the upper limit value x_β of the downward criterion exists or not.

Our proposed utility form offers a practical way to model user utility behavior with respect to the given user- and technology-related parameters x_m, x_α, and x_β. The steepness parameter ζ can therefore be tuned to capture user sensitivity. Some examples of our proposed utility forms are illustrated in Figure 2.3 (for upward criterion) and Figure 2.4 (for downward criterion) according to different values of x_m and ζ. Generally, to model a high user sensitivity to the variation of a criterion, the value of ζ is set to a high value and vice versa.

2.5 Multi-criteria utility function

2.5.1 Survey of multi-criteria utility

2.5.1.1 Additive aggregate utility

As previously mentioned, access network selection in heterogeneous networks is based on multiple criteria. A common approach to computing the aggregate multi-criteria utility of an access network is described as follows:

$$U(\mathbf{x}) = \sum_{i=1}^{n} w_i u_i(x_i) \quad \text{where} \sum_{i=1}^{n} w_i = 1 \tag{2.18}$$

where $\mathbf{x}$ is the vector of n considered criteria and w_i are the user preferences. This approach is referred to as an additive utility. The utility-based network selection schemes, addressed in [44–46, 54] and references therein, have used this additive utility approach.

Utility	w_i	Network A	Network B
$u(cost)$	1/3	0.5	0.8
$u(QoS)$	1/3	0.5	0.8
$u(load)$	1/3	0.5	0
Total Utility		**0.5**	**0.53**

Table 2.2: Case study: additive multi-criteria utility

Very similar to a classic scoring method, additive utility offers an easy and accessible way to aggregate different elementary utilities. It also allows users to introduce their preferences for different criteria. Although it is widely used and has some advantages, the additive utility also has serious limitations. A fundamental issue is whether the multi-criteria utility function can be separated into independent parts where u_i, the utility of criterion i, does not depend on the value of the other criteria. In this case, the elementary utilities $u_i(x_i)$ can be simply added to produce the aggregate utility. Unfortunately, the criteria are not always independent. An example is where an access network provides good utility for all selection criteria but one. The simple numerical study in Table 2.2 illustrates such a case. Access network B provides good utility for all selection criteria except network load (i.e., the access network is overloaded and its utility is zero). Under those circumstances, connecting to this network is not useful. However, the additive multi-criteria utility selects it. This limitation is due to:

$$\lim_{u_i(x_i)\to 0} U(\mathbf{x}) \neq 0 \quad \forall i = 1..n \tag{2.19}$$

2.5.1.2 Acceptance probability

As mentioned, acceptance probability has been widely used in radio resource management to mea-

sure the probability that a user is satisfied with the perceived utility u given the price p. Acceptance probability is modeled by:

$$A(u,p) = 1 - exp(-Cu^{\mu}p^{-\varepsilon}) \tag{2.20}$$

where $\mu > 0$ and $\varepsilon > 0$ control the user sensitivity to utility and price, and C is a positive constant representing the satisfaction reference value. The function is associated with the Cobb-Douglas demand curves [62]. This model was used like an aggregate utility-based network selection in [49]. As described in Section 2.3.2, acceptance probability is similar to aggregate utility if they are both computed with the same available information. Next, we will examine whether this Cobb-Douglas acceptance probability adequately models aggregate utility and user acceptance probability.

Acceptance probability can overcome the limitations of the additive utility (2.19). However, it has three limitations when measuring the user's satisfaction. The first visible limitation is the zero price effect, or $\lim_{p\to 0} A(u,p) = 1 \quad \forall u > 0$. An access network whose price is zero (e.g., free public WiFi) will always be selected even if it offers extremely poor connectivity. This is economically valid in general; but it should not be a factor in access network selection. The limitation is explained by the fact that acceptance probability does not take into account the future (next instant) service degradation penalty. A possible solution is to scale the price to interval [0, 1] via a downward utility function $u_p(p)$, i.e.,

$$A(u,p) = 1 - exp(-Cu^{\mu}u_p(p)^{\varepsilon}) \tag{2.21}$$

The second limitation is that utility u is computed for only one criterion (e.g., the allocated bandwidth). In multi-criteria network selection, the utility should include all characteristics except for price p. The solution is to define an overall utility as either the product over a set of elementary utilities or the weighted average over a set of elementary utilities [65]. The third limitation is that, even if the overall utility proposed in [65] is used and the zero price effect is removed, acceptance probability provides no means of introducing the user preference weights w_i as an additive multi-criteria utility approach would do.

In the use of (2.20) as a user acceptance probability, the three limitations still remain. This causes an error in the estimation of the user behavior and provides no way to consider the diversity of user preferences. Furthermore, from (2.20), we see that $A(u,p) = 1$ only if $u \to \infty$ or $p = 0$. The first condition ($u \to \infty$) never happens since utility is always assumed to be upper limited and normalized in $[0,1]$. Condition $u \in [0,1]$ was also used in all the papers [49, 58, 60, 62–67] that have adopted this acceptance probability. The second condition ($p = 0$) corresponds to the zero price effect mentioned above. If we use (2.21) to avoid the zero price effect, the acceptance probability is never equal to 1. The Cobb-Douglas acceptance probability is therefore not appropriate for modeling aggregate utility or acceptance probability.

2.5.2 New multi-criteria utility function

A multi-criteria utility should reflect the interdependence among the considered criteria. A basic question is whether a criterion can be completely compensated by another criterion or by a set of other criteria. In other words, the nullity of a specific elementary utility does not lead to an elimination of this access network in the selection process. Generally, when the user sets a non-zero preference weight for a criterion, it means that he considers this criterion in his evaluation. If its utility is zero (i.e., its value is below x_α for an upward criterion or above x_β for a downward criterion), the corresponding access network does not satisfy the technical or user constraints. Logically, this access network should not be selected: the non-zero preference criteria are not independent of each other. Because of these

limitations, we need to design a new multi-criteria utility form that satisfies the following requirements:

$$\frac{\partial U(\mathbf{x})}{\partial u_i} \geq 0 \tag{2.22}$$

$$\text{sign}(\frac{\partial U(\mathbf{x})}{\partial x_i}) = \text{sign}(u_i'(x_i)) \tag{2.23}$$

$$\lim_{u_i \to 0} U(\mathbf{x}) = 0 \quad \forall i = 1..n \tag{2.24}$$

$$\lim_{u_1,..,u_n \to 1} U(\mathbf{x}) = 1 \tag{2.25}$$

The aggregate utility should increase when the elementary utility increases (2.22). It should be an increasing function of upward criteria and a decreasing function of downward criteria (2.23). The condition (2.24) resolves the limitation (2.19) and the condition (2.25) reflects the fact that if all elementary utilities are equal to 1 (i.e., all criteria satisfy the user's expectation), the aggregate utility should be equal to 1. Finally, the user preference weights for different criteria are required to be considered in the aggregate utility form.

Proposition 2.5.2.1. *Given a network selection criteria vector* $\boldsymbol{x}$ *and the associated preference vector* $\boldsymbol{w}$*, a suitable multi-criteria utility function is formulated as:*

$$U(\boldsymbol{x}) = \prod_{i=1}^{n} [u_i(x_i)]^{w_i} \tag{2.26}$$

where n is the size of vector $\boldsymbol{x}$*,* w_i *is the preference weight for criterion i* ($\sum_{i=1}^{n} w_i = 1$)*, and* $u_i(x_i)$ *is the single-criterion utility of criterion i that follows the utility form proposed in Proposition 2.4.2.1.*

Proof. It is easy to verify that the proposed multi-criteria utility satisfies the requirements (2.24) and (2.25). Next, (2.22) is verified as $U(\mathbf{x})$ is an increasing function of each u_j. In fact,

$$\frac{\partial U(\mathbf{x})}{\partial u_j} = w_j [u_j(x_j)]^{(w_j - 1)} \prod_{i \neq j}^{n} [u_i(x_i)]^{w_i} \geq 0$$

Also, the partial derivative of $U(\mathbf{x})$ at an x_j is given as

$$\frac{\partial U(\mathbf{x})}{\partial x_j} = \left(w_j [u_j(x_j)]^{(w_j - 1)} \prod_{i \neq j}^{n} [u_i(x_i)]^{w_i} \right) u_j'(x_j) \tag{2.27}$$

$$= \frac{\partial U(\mathbf{x})}{\partial u_j} u_j'(x_j) \tag{2.28}$$

This proves that $U(\mathbf{x})$ is increasing for upward criteria and decreasing for downward criteria. So (2.23) is verified.

The last thing to show is that w_i represent the user preferences. As a monotonic transformation of a utility produces another utility with the same preference ranking, we can apply a logarithm transformation to $U(\mathbf{x})$:

$$V(\mathbf{x}) = \ln(U(\mathbf{x})) = \sum_{i=1}^{n} w_i \ln(u_i(x_i)) \tag{2.29}$$

If $v_i(x_i) = \ln(u_i(x_i))$, we see that $v_i(x_i)$ is an elementary utility function of criterion i (by the mono-

tonic transformation property). We have

$$V(\mathbf{x}) = \sum_{i=1}^{n} w_i v_i(x_i) \sim U(\mathbf{x})$$

Under this additive presentation, we see clearly that w_i are the user preferences. So (2.26) is a suitable multi-criteria utility form. □

The proposed multi-criteria utility satisfies all requirements of a utility function and avoids the limitations of existing models. Along with the limitations of the acceptance probability identified in Section 2.5.1.2, a suitable acceptance probability model should follow (2.4), (2.5) and (2.30).

$$\exists x_i \in \mathbf{x} : u_i(x_i) = 0 \Rightarrow A(\mathbf{x}, \mathbf{w}) = 0 \tag{2.30}$$

It can be seen that the proposed multiplicative utility form in Proposition 2.5.2.1 can be used to properly model the user acceptance probability. Hence, we have:

$$A(\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{n} [u_i(x_i)]^{w_i} \tag{2.31}$$

As demonstrated in the proof of Proposition 2.5.2.1, this multiplicative aggregate form fully satisfies the conditions (2.4), (2.5) and (2.30).

2.6 Performance evaluation

2.6.1 Validation of the proposed utility function

Criterion	Preference	x_m	x_α	x_β	ζ
Bandwidth (b)	$w_1 = 0.5$	40	5	90	2
Price (p)	$w_2 = 0.5$	30	0	80	3

Table 2.3: Parameters for utility computation

In this section, we validate the proposed single-criterion and multi-criteria utility forms through a network selection scenario. We assume that the selection decision is based on only two criteria, allocated bandwidth b and price p. The single-criterion utility is established based on the parameters given in Table 2.3. Recall that x_α and x_β are the lower and upper limits of each corresponding criterion, and x_m is the center of utility curve. The given values of price and bandwidth are relative ones so no units are necessary.

We first compare our multiplicative multi-criteria utility to the additive utility and the Cobb-Douglas acceptance probability. We use the original sigmoid form (i.e., $u(x) = \frac{(x/x_m)^\zeta}{1+(x/x_m)^\zeta}$) to compute the elementary utilities in the additive utility approach and the acceptance probability. We use the sigmoid form in Proposition 2.4.2.1 to compute the elementary utilities in the multiplicative multi-criteria utility. We use the original single-criterion utility forms to keep the existing solutions the same. The acceptance probability function used is $A = 1 - exp(-2 * u(b)^2 * p^{-0.2})$. In fact, we choose $C = 2$, $\mu = 2$ and $\varepsilon = 0.2$ in the acceptance probability form to scale its value in the interval [0, 1]. With another choice of these three parameters, the same functional form will be retrieved but the value of the acceptance probability may not be properly distributed in the interval [0,1].

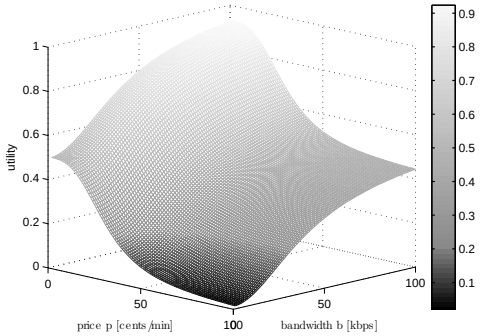

Figure 2.5: Variation of the additive multi-criteria utility

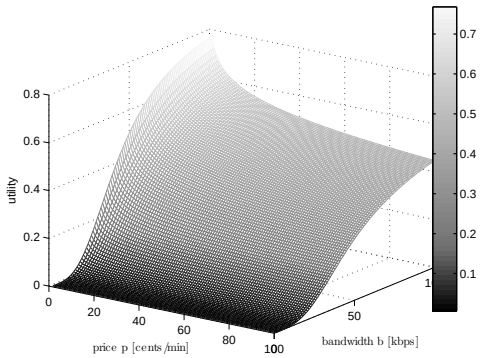

Figure 2.6: Variation of the Cobb-Douglas acceptance probability

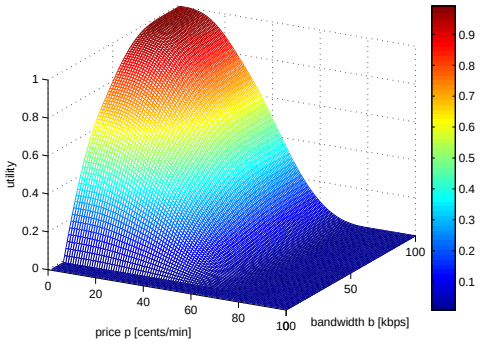

Figure 2.7: Variation of the proposed multiplicative multi-criteria utility

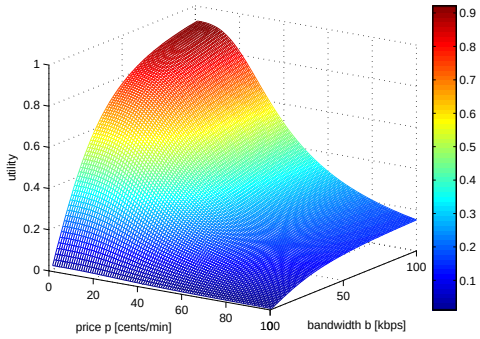

Figure 2.8: Multiplicative multi-criteria utility with original elementary utilities

As the two network selection criteria are changed, the utility metric (as calculated by the additive multi-criteria utility, the acceptance probability and the multiplicative multi-criteria utility) also changes as illustrated in Figure 2.5, Figure 2.6 and Figure 2.7. First, the additive multi-criteria utility in Figure 2.5 confirms the limitations identified in the previous section. An access network of $(p = 1, b = 1)$ is clearly assigned a higher utility than an access network of $(p = 30, b = 40)$ (denoted as $(1,1) \succ (30,40)$). This would lead to an unacceptable decision since the access network of $(1,1)$ could not satisfy the user's bandwidth requirements (minimum $b_\alpha = 5$ as shown in Table 2.3). But the access network of $(30,40)$ fully meets user expectations in both bandwidth and price. Similarly, $(100,100) \succ (80,90)$ is also unacceptable. In fact, the network of $(100,100)$ does not satisfy the user's constraint on the price (i.e., $p_\beta = 80$ as specified in Table 2.3). But the network of $(80,90)$ meets both bandwidth and price constraints. Until now, one may say that the identified limitations have been because the original single-criterion utilities do not capture the limits of each criterion. So we replace the original single-criterion utilities by the new proposed ones in the additive aggregate utility. In this case, the form of the utility variation is very similar to that shown in Figure 2.5. We see that the networks of $(1,1)$, $(30,40)$, $(80,90)$ and $(100,100)$ are assigned to the same utility level ($U(p,b) = 0.5$). It is still unreasonable that the network of $(30,40)$ has the same preference as the network of $(1,1)$.

The acceptance probability metric in Figure 2.6 can mitigate the *low* price effect of the additive utility. A zero price effect, which is not mitigated, is not presented here since the smallest value of

price is 1. The limitations of the approach when the price rises to the upper limit still remain. In fact, the unacceptability of $p > 80$ could not be captured. All the inconveniences of the additive utility and the acceptance probability are significantly improved by the use of our utility form as shown in Figure 2.7. Our multiplicative aggregate utility is a suitable form to model the utility in the access network selection problem.

One question would be whether it is possible to use the original single-criterion utility form instead of the proposed one to compute the multiplicative utility. The aggregate utility metric calculated by that approach is shown in Figure 2.8. In comparison with Figure 2.7, it can be seen that the utility metric does not take into account the limits of the network selection criteria. The utility metric cannot reach the maximal value ($U(p = 1, b = 100)$ is far from equal to 1) because the original elementary utility could not take into account the lower or upper limit of a criterion in its formula. This confirms the suitability of our single-criterion and multi-criteria utility forms.

2.6.2 Case study: the benefit to users

In this section, we investigate how the models help user terminals to select the best access network. We use a simulation scenario in which a mobile user moves across different available access networks. At each instant the user is able to choose from three available access networks: two WiFi and one UMTS. The selection is assumed to be based on price and achievable throughput criteria. In a real network environment, the achievable throughput can be estimated from the allocated bandwidth and the link quality (e.g., Bit Error Rate, modulation and coding scheme) [82, 83]. In this simulation, the values for price and throughput are randomly selected among a range of pre-defined values every 100 samples. The parameters used are shown in Table 2.4.

Parameter	Range	Preference	x_m	x_α	x_β	ζ
Throughput (tp)	$0 \div 900$	$w_1 = 0.7$	500	150	1200	3
Price (p)	$0 \div 50$	$w_2 = 0.3$	40	10	80	3

Table 2.4: Simulation parameters: User case

The user is assumed to be running a streaming application, so throughput is a more significant criterion than price (the user sets higher preference weight for the throughput criterion). In this simulation, we monitor the application's buffer evolution to evaluate the performance. The buffer size $b[t]$ is simulated by:

$$\begin{cases} b[0] = 600 \\ b[t] = \max(0, \min(600, (b[t-1] + tp[t] - 200))) \end{cases} \tag{2.32}$$

where 200 is the playback rate of the streaming application, 600 is the maximal memory size allocated to the buffer and $tp[t]$ is the achievable throughput at instant t. At the beginning of the simulation, the buffer is assumed to be filled. When the simulation starts, the user moves and selects the appropriate access network. The buffer is filled with media data at the current throughput rate every sample time. When the buffer runs out of content, the streaming application is interrupted.

We compare our model with the traditional additive aggregate utility model used in the network selection algorithm. We also use the proposed single-criterion utility in the additive aggregate utility to take into account the upper and lower limits of the considered criteria. In the simulation, the access network selection and the buffer size are updated each second (i.e., sample time). The simulation results in three scenarios as shown in Figure 2.9. The two top figures 2.9(a) and 2.9(b) give the same results for two access network selection schemes. In Figure 2.9(b), the outage of the buffer is explained by the fact that none of available access networks satisfies the requirements of the user's application.

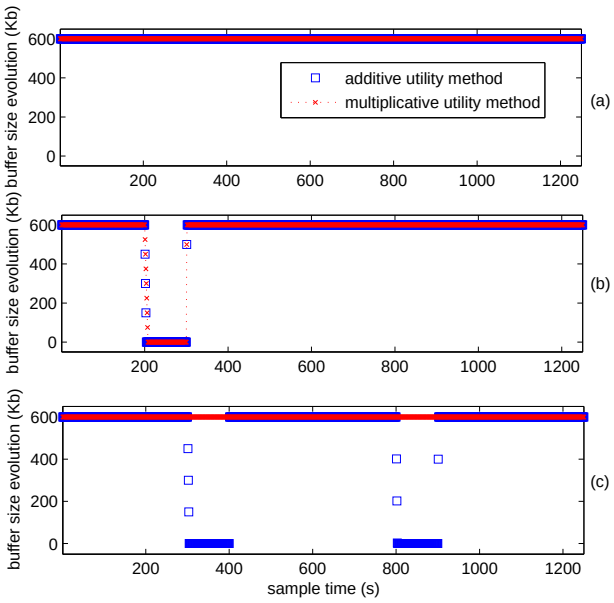

Figure 2.9: Streaming buffer evolution at the user side

In Figure 2.9(c), the buffer outage occurs twice with the additive utility-based network selection while our multiplicative utility-based solution delivers the video seamlessly. We feel that this is due to the limitations of the additive approach as described in the previous section whereby the high utility of lower price is considered to be more important than the low utility of lower throughput.

The simulation was repeated 1000 times. Each time, we recorded the length of time the streaming application was interrupted. With the additive solution, the interruption is about 7% of all running time whereas our multiplicative solution shows interruptions of less than 1% of the running time. We can conclude that the results confirm that our proposed utility forms serve users better than the existing one.

2.6.3 Case study: the benefit to network operators

After showing the advantages for users, we investigate our model's acceptance probability for the network operator's radio resource allocation. We compare it with the widely used Cobb-Douglas approach.

We consider the same radio resource allocation scenario (one access network operator and N users) that has been used in [58, 60, 62, 63, 66, 67]. We divide the N users into two classes of service: $N/2$ silver users and $N/2$ gold users. The access network selection algorithm is once more based on price and allocated bandwidth. The total bandwidth of an access network is considered to be limited. The network operator estimates the acceptance probability of each user and then determines the optimal resource allocation vector $b = [b_1, ..., b_N]$. The vector must satisfy $\sum_{k=1}^{N} b_k \leq T$ where T is the operator's total bandwidth. The network operator aims to maximize potential revenue, which is given by:

$$R = \sum_{k=1}^{N/2} p_s A_k(b_k, p_s) + \sum_{k=N/2+1}^{N} p_g A_k(b_k, p_g) \tag{2.33}$$

where p_s and p_g are the flat prices paid by all users in the same class of service (silver and gold). In this formula, A_k is the acceptance probability of user k. This is generally computed on bandwidth b_k

and price p_s or p_g. The original Cobb-Douglas acceptance probability has the following form (used in [58, 60, 62, 63, 66, 67] and references therein):

$$A_k(b_k, p_k) = 1 - e^{-K[u_{b_k}(b_k)]^{\mu}(p_k/\phi)^{-\varepsilon}} \tag{2.34}$$

where $K = -\log(0.9)$. The values of other parameters are detailed in Table 2.5. In our case, the multiplicative utility-based acceptance probability proposed in (2.31) is used:

$$A_k(b_k, p_k) = [u_{b_k}(b_k)]^{w_b}[u_{p_k}(p_k)]^{w_p} \tag{2.35}$$

The parameters for each class of service are also presented in Table 2.5. The values p_m and b_m are the same values x_m used in the single-criterion utility for price and bandwidth, respectively. The upper and lower limits of each criterion are set as $(p_\alpha = 0, p_\beta = 50)$ and $(b_\alpha = 0, b_\beta = 100)$. We assume that the operator has precise information about user preferences (w_s and w_g) as well as the user's sensitivity to bandwidth and price (μ and ε) in the Cobb-Douglas case.

Class	Price	$\mathbf{w}=(w_p, w_b)$	p_m	b_m	μ	ε	ϕ
Silver	$p_s = 20$	$w_s = (0.6, 0.4)$	20	30	2	6	10
Gold	$p_g = 40$	$w_g = (0.3, 0.7)$	40	60	3	4	40

Table 2.5: Simulation parameters: Operator case

We could not use potential revenue as a measure of comparable performance between the two approaches since acceptance probability in the two approaches is computed very differently. In order to compare the performance, we defined the resource efficiency metric as the ratio of the potential revenue to the potentially allocated resource. That is:

$$E = \frac{R}{B} \quad \text{where} \quad B = \sum_{k=1}^{N} b_k A_k(b_k, p_k) \tag{2.36}$$

The metric E can be seen as the amount of gained money per unit of allocated resource.

(N,T)	(10,400)	(10,500)	(20,800)	(20,900)
Cobb-Douglas case	0.44	0.441	0.421	0.443
Proposed case	0.547	0.525	0.553	0.608

Table 2.6: Resource efficiency metric

We conducted simulations for different values of N and T. For each simulation, we computed the resource efficiency index in the two approaches. The results are presented in Table 2.6. By using the multiplicative utility-based acceptance probability metric, the network operator can improve resource efficiency between 19% and 37% compared to the use of the original Cobb-Douglas metric. The results confirm that our proposed utility model also serves operators better.

2.7 Summary

In this chapter, we provided a complete utility theory for modeling single-criterion utility and multi-criteria utility in the context of wireless access network selection. The theory is based on a classic economic theory adapted to the behaviors of mobile end-users. The limitations of existing utility models were highlighted. Single-criterion and multi-criteria utility forms, with the ability to satisfy all the

utility properties and to address the limitations, were proposed. We showed that the proposed model can also be used as an acceptance probability metric for the network operators' radio resource management. The suitability and the effectiveness of the proposed model were validated by mathematical analysis and by simulations. We showed that our proposed utility model was not only useful for users' access network selection but also useful for operators' resource allocation management.

This chapter dealt mainly with utility-based access network selection decision. In the next chapter we describe a complete terminal-controlled handover management solution in which this access network selection model will play a crucial role. In addition to the selection decision-making, other aspects such as user preferences configuration, network selection triggering condition and handover threshold will be fully addressed in the next chapter. In the frame of the terminal-controlled handover management, this access network model will be again evaluated.

∼∼ △♡△ ∼∼

Chapter 3

Terminal-controlled Mobility Management Framework

If technology doesn't work for people...it doesn't work

http://www.usercentric.com

This chapter proposes a terminal-controlled mobility management framework in the 4G heterogeneous networks. The mobility management solution allows mobile users to freely handover between uncoordinated available access networks. The users are assumed to have access to both networks involved in the handover using a multi-interface terminal with help of a universal Subscriber Identity Module (SIM) card or a multi-homing contract. The terminal-controlled mobility management consists of a policy-based power-saving interface management scheme coupled with a user-centric network selection solution, a handover initiation algorithm and a handover execution. The network selection, an important step of the terminal-controlled mobility management scheme, is based on the utility model proposed in Chapter 2.

3.1 Motivation

In parallel to the evolution towards converged heterogeneous networks, the telecom market is facing a migration from network centricity towards user centricity. In the current network-centric approach, operators keep tight control over users so that their network is used to its greatest potential. End-users can only influence their preferences in a limited way. In the context of deregulated telecom market, the network-controlled handover management exhibits some serious limitations in the service continuity maintenance while handing over between two different network domains operated by different operators. Among these limitations, we can find complex issues such as security context transfer and data switching management. We believe that a user-centric vision will be a mandatory evolution trend in all-IP 4G networks as it represents the most efficient way to ensure an ABC service. In this vision, users will have greater control and will be able to select the access network with which they are most satisfied. The users are in this strong position because their terminal can access to information on device capabilities and user preferences, and, most importantly, to knowledge of both serving and neighboring access networks. The terminals will be thus able to trigger the handover at the right instant to achieve seamless handovers.

In heterogeneous environments, vertical handovers between different technologies can be performed either using one SDR interface [84, 85] or using multiple radio interfaces [11, 14, 15, 86, 87]. In this chapter, we consider the case where multimode terminals are equipped with multiple wireless access interfaces. Such devices are available in the market today. The vertical handover using multiple interfaces has already been commercially available through the UMA [15] solution and will be widely used in a near future. However, this solution requires a careful design to minimize side effects of additional electronic devices in the terminal. One of the most relevant issues is the high power consumption of multiple radio interfaces. The power consumption efficiency becomes *a must* because the battery capacity is limited in any portable device. Since only the terminal can be aware of its remaining battery capacity, it is clear that only the terminal can handle the power-saving mobility management. The handover should be initiated in an adaptive manner to optimize the power consumption and to guarantee uninterrupted services.

As discussed in Chapter 1, the mobility management for tight-coupling schemes is based on the existing cellular mobility solutions whereas the one for loose-coupling schemes is based on MIP mechanism. Most of the current solutions require agreements between the operators, who own different interworked access networks, to be established. The mobility management remains network-controlled. The user cannot maintain on-going sessions while handing over between two access networks belonging to two non-collaborating operators even if he has subscribed to these two networks. To overcome this issue, we propose a fully terminal-controlled mobility management scheme where different access networks may be completely independent. The terminal-controlled handover is built on the top of a very loose-coupling interworking architecture.

In the following, to simplify the explanation, the terminal is assumed to have three radio interfaces: WiFi, WiMAX and UMTS. One should note however that the solution described here remains valid for any number of interfaces and underlying access technologies.

3.2 Very loose coupling architecture

The telecommunications market is influenced by three main drivers: users, network operators and service providers. In the user-centric vision, service providers are independent of the network operators. The network infrastructure can be divided into two parts: the access network and the converged core network. The former represents the current and future network operator domains while the latter is an external core network independent of operator domains. Our mobility framework does not require major changes in the access network operator and the service provider parts.

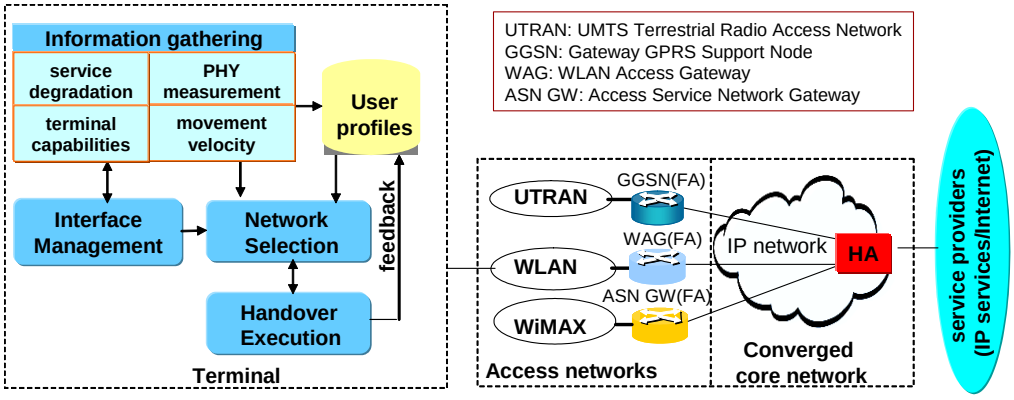

Figure 3.1: Very loose coupling interworking architecture

3.2.1 Converged core network

It is widely accepted that the 4G networks will be purely IP-based. Service continuity can be achieved using MIP with help of the HA deployed in the converged core network. The HA, which is not located in a particular access network, can provide the mobility management as a service for mobile users. The mobility management is thus seen as an independent third party service. Such integration of access networks can be referred to as a very loose-coupling interworking since the corresponding network operators may not have any Service Level Agreement (SLA) between them. The coupling point is located far from the radio interface. The proposed interworking architecture is depicted in Figure 3.1. Using MIP does not affect our terminal-controlled solution since the HA is located in the IP converged core network. In an IPv4-based network, we even do not need to implement FA entities in a particular access network since there exist solutions for MIPv4 management without it. The access networks thus remain unchanged in the very loose coupling architecture.

3.2.2 Mobile terminal

The handover management is based on the enhanced functionalities at the terminal side. The terminal must be equipped with multiple radio interfaces to access different access network technologies. Different elements involved in this handover management at the terminal side is illustrated in Figure 3.1. Functionalities of each of these elements are summarized in the following:

Information gathering: involves collecting all available information from the surrounding access networks. This information is used for managing radio interfaces, identifying the need for handover, and selecting the best access network among available ones. This is performed by the terminal, when it is powered on or when it moves across different radio coverage types, in order to find available access networks. This function allows to determine a generic set of parameters describing access networks and devices (type of access network technology, access network operator, QoS support, cost, remaining battery capacity...). In addition, the information related to the QoS status of the current running applications, the mobile terminal's velocity and the cell coverage radius is also collected. The gathering information can be regrouped into 4 categories:

- *PHY measurement:* The access network characteristics identify the available access networks and their radio link quality. Such information is measured by the physical radio interfaces periodically or when an event occurs. Different link indication parameters like RSS, SNR, SINR can be monitored.
- *Terminal capabilities:* The terminal capabilities information is related to multimode capacity, available radio interfaces and remaining battery. It can influence the network selection and handover management.
- *Service degradation:* The service degradation information is related to the trade-off between the application's QoS requirements and the quality of the current connection. It can be measured through Real-time Transport Control Protocol (RTCP) reports for instance. The combination of service degradation and radio link quality allows describing more accurately the quality of the current connection than the fluctuating physical-layer information only.
- *Terminal's velocity:* The terminal's movement velocity is also an important factor for mobility management in heterogeneous environments. The velocity estimation in mobile cellular systems can be achieved using the Doppler spread in the received signal envelope [88], the eigen-matrix pencil method [89], the time-frequency characteristics of the received signal [90] or the Global Positioning System (GPS)-assisted method [91].

User profiles: contain the user's identities for different access networks and subscribed services, user preferences, and mobility policy repository. In the context of the terminal-controlled mobility management, we are first interested in user preferences, which is a rating relationship among the parameters considered in network selection. Each preference has a relative weight that users assign to each criterion depending on their requirements. User preferences should be adequately configured for different contexts which are characterized by currently connected access node, terminal's velocity and running applications class. As user preferences influence the network selection and handover management process, future terminals will have to provide users with the facilities (e.g., Graphical User Interface dedicated to user preferences configuration) to specify and alter their preferences in an easy manner. Additionally, the terminal can maintain the mobility policy database that contains a black list of access network operators with whom the user has had a bad experience. The black list will be updated through feedback from handover execution failure and bad QoS as perceived by the application. The users can manually pre-specify the black list and remove a specific access network from this list.

Interface management: Based on the gathered information, the interface management will decide to turn on, stand by or turn off one or more radio interfaces to optimize the power consumption. Interface management becomes thus a constraint for network selection.

Network selection & handover decision: It refers to the process of deciding to which access network to connect at any point of time. This allows the best access network to be selected and handover to this access network to be initiated. In fact, based on the gathered information, the available interfaces, the user preferences and the black list, mobile terminal evaluates the neighboring access networks and select the most suitable one. Network selection is the heart of the handover procedure and a key decision enforcement of the terminal-controlled mobility management.

Handover execution: follows access network selection. Once the selected access network differs from the serving one, the handover execution is performed. The main goal of handover execution is to preserve session continuity while changing the point of attachment. If the handover fails, the network selection will attempt to select another access network. This experience will be registered in the experience repository.

3.3 Handover management

3.3.1 Information gathering

In order to effectively control the handover management, the terminal requires to have enough and reliable information to be able to make the right decision. On one hand, the terminal will scan and measure the physical signal strength from serving and neighboring access nodes. These measurement samples will not only serve as link quality indicator but also be used to estimate other information like terminal's velocity, achievable data rate or cell coverage size. The terminal will monitor its current IP connection status as well as the perceived quality of current applications. The remaining battery capacity and the power consumption rate of each integrated radio interface will be instantaneously monitored. In addition to information measured or estimated, the provisioning information from the network is also considered. In short, we only use information criteria that the terminal can measure or estimate without need of IP connectivity as well as those provisioned from the network side. The considered criteria include:

- ◇ *Access Network Identity (ANI):* identifies the access technology and the operator that has deployed this access network. It comes along with the cost information, supportable services as well as supportable data rates.

◇ *Cost* (c)*:* The cost of using a particular access network is a major criterion and potentially a decisive factor in a user-centric network selection. Different operators can propose different billing schemes. In fact, we can find two main options: billing based on duration for voice calls and billing based on the volume of downloaded data for data services. Operators can offer unlimited services (unlimited duration and unlimited downloaded data volume) for a flat price. In this work, we consider two kinds of cost model: cost per minute for voice services and cost per data volume for data and streaming services. In fact, the flat price for limited or unlimited services can be equivalently converted to the two above cost models. According to running applications, the terminal identifies the most suitable scheme.

◇ *Link quality (S):* The radio link quality like SNR or SINR reflects properly the wireless transmission channel quality. The RSS is also used to detect the presence of an access node and to initiate network selection and handover procedures.

◇ *Power consumption gain (ge):* The power consumption gain parameter is defined as the ratio of the power consumption rate e_i to the maximum achievable data rate r_i, $ge_i = \frac{e_i}{r_i}(J/Mb)$. The maximum achievable data rate of an access network can be estimated on the link quality and the corresponding modulation and coding scheme [92]. Users probably prefer to select a low power-consumption-gain access network to optimize the power consumption.

◇ *Battery lifetime* (L): The main source of power consumption of a portable device is related to Radio Frequency (RF) components of a radio interface [93, 94]. Though the energy consumed by radio interfaces is not the only source, it is the only variable part from the interface selection and handover management. The remaining battery lifetime is defined as:

$$L_i = \frac{E}{e_i^{ac} + \sum_j e_j^{sb}} \quad (h) \tag{3.1}$$

where E is the remaining battery capacity, e_i^{ac} is the power consumption rate of active interface i and e_j^{sb} is the power consumption rate of standby interface j. If the remaining battery lifetime L_i (corresponding to the case where interface i is selected) is low, the user who wants to elongate his device's lifetime will not use this interface to communicate.

◇ *Access network load* (ρ)*:* Even if the link quality criterion value from an access node is good, it may be heavily loaded (an important part of resources is occupied by other users). The load parameter, varying from 0 (unloaded) to 1 (heavily loaded) becomes also an important network selection criterion. The advantages of provisioning the load information in IEEE 802.21 Media Independent Handover (MIH) services have been investigated and demonstrated in [95]. Different from other criteria, the load cannot be measured from the terminal side. If an access network is heavily loaded, it cannot accommodate new connections without degrading the QoS of currently connected users. In fact, the network can also take into account different user classes. The premium class is the guaranteed one while the others are served only if resources are still available. From the user perspective, a premium user knows that he will always be served in priority. For regular users, the load parameter is much important because the loaded access network will not serve them well. At the network side, if it is heavily loaded, new connection requests of regular users will be refused or currently connected regular users will be enforced to handover to others access networks (see Chapter 7). Therefore, access networks have interest to inform users about their load. The degree of significance of the load parameter in the network selection algorithm is indeed a matter for each user to decide.

◇ *Velocity estimate (v):* The current mobile terminal's velocity can be used to discourage a terminal moving at a high speed to handover to small cell-size systems. The velocity parameter also serves

as a constraint in power-saving interface management.

- ◇ *Cell coverage estimate (R):* The radio coverage size of the current access node will be used to adaptively compute the handover initiation threshold. Based on the received signal strength measurements and the well-known values of the maximum power transmission of a base station and the receiver threshold (the minimum signal strength at the cell border), the cell radius (and also the path loss factor K_2) can be estimated as described in [96, 97] and references therein.

Some of the above criteria serve as criteria for network selection while others are used for interface management or network selection triggering. In the context of the terminal-controlled mobility management addressed in this chapter, the criteria considered for network selection are ANI, c, S, ge, r and ρ. Note however that the network selection criteria are not limited to those previously specified.

3.3.2 Power-saving interface management

As mentioned earlier, optimizing the power consumption of multi-interfaced devices is an important issue. According to [98], if a WiFi interface is added to a handset, two third of the battery lifetime is reduced. In order to optimize the power consumption for integrated WiFi and cellular devices, [99] proposed to activate the WiFi interface when the mobile receives a short message from the cellular network through a push mechanism on a call server. A power saving mechanism which activates WiFi interface by paging the mobile via its cellular interface, was proposed in [100]. However, these solutions require modifications to the network protocol stack to accommodate a push mechanism or a new paging scheme. They have addressed the power consumption of multiple radio interfaces only during the idle communication mode. In the terminal-controlled handover, we consider power-saving interface management as a step within the handover procedure. The power consumption efficiency is addressed in both idle and active communication modes.

Along with the emergence of multi-purpose terminals and many new multimedia data-intensive applications, the gap between the energy requirement and the terminal's battery capacity has progressively widened. Today's all-in-one portable devices integrated with additional peripheral devices like camera, MP3 player, FM radio...have lots of components that drain battery. Mobile devices are more and more equipped with multiple radio interfaces such as 2G/3G cellular radio transceivers, WiFi, WiMAX and Bluetooth. As each consumes energy while powered on, an efficient interface management is therefore required.

An interface can be in *active*, *standby*, or *turn-off* state. When an interface is in active state, it can receive and transmit signal. It can scan and measure neighboring cells and its battery is mainly consumed in this state. An interface in standby state can be periodically activated to scan new access networks and monitor the link signal quality of neighboring cells. Finally, if the radio interface is turned off, no communication or measurement occurs and the energy is not consumed.

Generally, a wireless radio interface always consumes energy unless it is powered off. The power consumed in active state is much more important than in standby state. To save energy, all interfaces are standby if there is no communication. However, mobile devices are most of the time in an idle communication mode and consequently standby interfaces still consume a significant portion of the terminal's battery. According to a large range of available products on the market and experimental results [94, 100], the power consumption of a standard WiFi interface in standby state varies between $40mW$ and $160mW$ or even more. The power consumption of a standby 3G cellular interface is about $10-20mW$ [93, 100]. The power consumption rate of mobile WiMAX interface is slightly smaller than that of WiFi interface but greater than that of cellular one [101].

In order to optimize the power consumption of multi-interface devices, we suggest turning off non-cellular interfaces during the idle communication mode. This is motivated by the fact that cellular interfaces consume less energy than non-cellular ones, and, more importantly, the cellular coverage is ubiquitous. The objective is to have only one interface active at a time for communication except during vertical handover periods. Also, if the remaining battery capacity becomes critical, turning off high-power-consumption rate interfaces will be a solution to elongate the device's lifetime. The proposed solution is expressed in terms of policy rules as described below:

- *Rule 1*: If no on-going communication occurs, turn non-cellular interfaces off and stand by the cellular interface. The user will be reached by a global identity like a global IP address. Incoming communications will be routed to the terminal via a default cellular interface.

- *Rule 2*: If an incoming communication is detected, the cellular interface will be activated to handle it, and the non-cellular interfaces will switch into standby state for possible imminent vertical handovers.

- *Rule 3*: If the user initiates an application session, the preferred interface for this application will be activated automatically. If the user manually turns on a non-cellular interface to search for available access networks, the selected interface will be activated. The other interfaces will switch to standby mode for possible handovers.

- *Rule 4*: If the terminal's velocity is greater than a predefined threshold (for example $5m/s$), the WiFi interface (in standby or active state) will be turned off. In fact, when the terminal moves at speed $v > 5m/s$, it will cross the WiFi cell in a short time (several hundreds of seconds). Connecting to WiFi cells in this case causes service degradation and energy consumption waste [102]. If the WiFi interface is a serving one, the network selection will be triggered to select another interface and another access network prior to turning it off. Similarly, if the terminal's velocity is greater than another threshold like $20m/s$, the WiMAX interface will be turned off. In fact, WiMAX systems are not designed to support a high QoS for users moving beyond this velocity.

- *Rule 5*: If the remaining battery lifetime using a non-cellular interface is less than a predefined threshold, this interface will be powered off independently of its current state. The network selection will be triggered to select another interface and access network if the turned-off interface is a serving one. Based on the user situation (at home or not at home) and the priority preference of power consumption criterion (high, medium, low or ignored), the appropriate threshold for turning off non-cellular interfaces is used (see Table 3.1 for some indicative predefined thresholds). This rule is essential for battery-limited portable devices and particularly for business men who want to be always reachable. Other users can set the battery lifetime threshold to 0 to take maximum benefit from other parameters. The threshold is set to 0 when the terminal is plugged into a power supply or when the terminal is at home situation.

	High	Medium	Low	Ignored
At home	0 min	0 min	0 min	0 min
Not at home	30 mins	20 mins	10 mins	0 min

Table 3.1: Example of battery lifetime thresholds configuration

3.3.3 Network selection & Handover decision

3.3.3.1 User preferences configuration

User preferences establish a rating relationship among criteria. Each weight is proportional to a degree of significance for each criterion in the network selection strategy. User preferences are thus represented as $w = \{w_j | w_j \in [0,1], \sum_{j=1}^{M} w_j = 1\}$, where M is the total number of considered criteria and w_j stands for the preference weight of criterion j.

As user preferences are user-specific, future devices need to integrate a network selection GUI to help users to configure their preferences. For ease of use, preferences are expressed under the priority meaning for each group of network selection criteria. We can gather different related parameters into the same group like cost (cost per minute, cost per data volume), or QoS (radio link quality, achievable data rate). Each criterion (or criteria group) has 4 levels of priority: high, medium, low, and ignored. Each priority level is associated with a discrete value p_j varying from 3 (high) to 0 (ignored). The preference weight for criterion j is equal to $w_j = p_j / \sum_{i=1}^{M} p_i$. The number of priority levels can also be divided into a larger scale like extremely high, very high, high, medium, low, very low and ignored. However, the large number of priority levels complicates the usage. Obviously, default values and guidelines are recommended to help non-expert users to get their user preferences well configured.

As there are different QoS requirements for different application classes, user preferences configuration should first take into account running applications. Secondly, with the same available access networks, the user decision may be different according to user situation. We propose therefore to configure the preference weight vector w_{kl} for each situation profile k and each running applications class l. Applications can be simply grouped into 2 classes: real-time (voice, streaming) and non real-time services (data downloading, web browsing). An application classification into conversational, streaming, interactive and background services [79] may be envisioned. If multiple applications belonging to different classes are simultaneously running, the preferences according to the highest priority class is used.

User situation profiles have already been implemented in some current mobile phone brands. They include home, in office, handset, indoor, outdoor, meeting, in car, and etc. Users can change the situation profile to better suit the context and environment. The change will come along with the modification to sound volume, ringing tone, warning tones, lights and call transfer. In our network selection solution, we go further by taking into account the situation profile in the network selection and the user preferences configuration. For example, if the user is at home, he could select home environment access networks like home WLAN or femtocell. If the user is in his office, he prefers to connect to his enterprise network. These policies can be specified in the experience repository. User preferences should be configured for different user situation profiles. For instance, when the user is at home, the network selection and interface management do not need to care about the power-saving criteria. The home situation profile is based on the location where the terminal's battery is charged. We believe that future mobile devices will have capabilities to identify automatically the situation profile based on the ANI-based location, mobile velocity, information collected by sensors integrated in devices [103].

The user preferences are therefore formulated under the form of a matrix of three dimensions $\{w_{klj}\}$: network selection criterion j, running application class l and situation profile k.

3.3.3.2 Network selection triggering conditions

After identifying the network selection criteria, the user preferences configuration and interface management, we address now network selection triggering conditions. Triggering network selection

depends on the gathered information and the interface management. If the selected access node is different from the serving one, the handover will occur. The handover initiation is therefore triggered by the network selection but the network selection does not always lead to a handover initiation.

We distinguish two types of network selection triggering conditions: *imperative* and *alternative*. All conditions that lead to an imperative handover to maintain the connectivity are called imperative conditions. Network selection is also triggered to find out a better access node in terms of user's satisfaction or QoS improvement. Conditions for such kind of network selection are known as alternative ones.

When the imperative triggering conditions are met, the serving access node is not appropriate for connectivity and it is not considered among candidate access networks. The imperative conditions are identified as follows:

- ◇ When an active interface is about to be powered off (due to the power-saving interface management), the network selection is initiated.
- ◇ When a handover execution to a selected target access node fails, the network selection is triggered to select another access node. Another possible option in this case is that the access network of the second highest preference level (during precedent selection evaluation) will be selected.
- ◇ If the RSS of the serving access node drops below a specific handover threshold θ_h, initiate the network selection. An adaptive handover threshold to ensure seamless handover is addressed and discussed in the next section.

The network selection is also initiated periodically or according to alternative conditions. The reason of having such alternative triggering conditions is that users may need an improved service or a better network. They include:

- ◇ A new access network is available. A signal strength level of an access node is good enough to be used in the network selection evaluation as well as to trigger the network selection if it is greater than a known threshold θ_o. In fact, θ_o is the received signal at the serving cell border, i.e., the minimum RSS value where the radio link can be hold.
- ◇ Users start an application requiring high QoS support that the current access network could not guarantee.
- ◇ Current applications suffer from service degradation.
- ◇ The network selection is triggered periodically to select the best access network according to user preferences.

A handover decision from a UMTS interface to a WiFi or WiMAX interface is depicted in Figure 3.2 (left diagram). Once a WLAN or WiMAX access network is discovered and its received signal is acceptable, the MN will trigger the network selection process. If the selected access network technology is WLAN or WiMAX, the vertical handover is then initiated. If the connection (including the authentication and QoS reservation) is successfully setup, the UMTS interface switches to standby state and the handover is complete. Otherwise, the MN remains connected to the UMTS network.

When the MN is communicating via a WiFi interface, the decision for handover from WLAN to UMTS/WiMAX is illustrated in Figure 3.2 (right diagram). When the WiFi interface is about to be powered off, the MN initiates the network selection algorithm to determine a suitable target UMTS or WiMAX access network. If the MN discovers new neighboring access nodes, the network selection

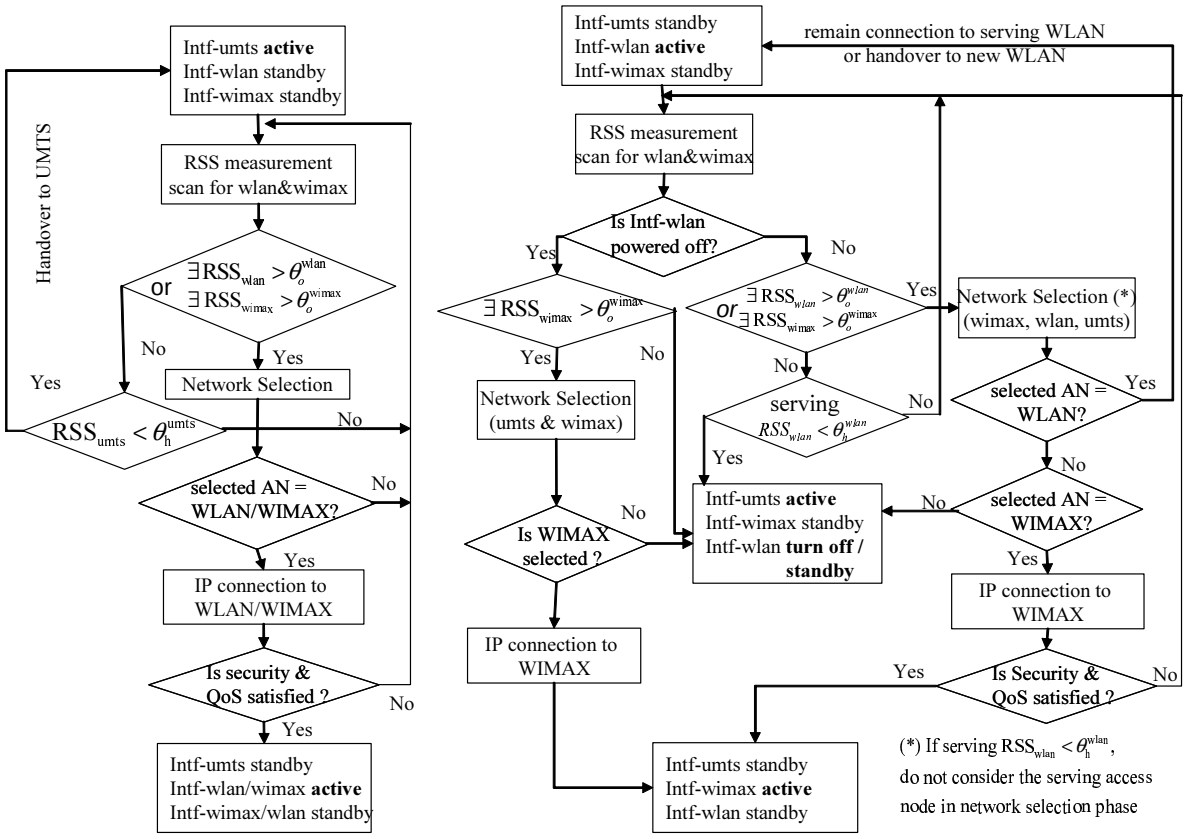

Figure 3.2: Network selection & handover initiation diagram: *Handover from UMTS to WLAN/WiMAX (left diagram), handover from WLAN to WiMAX/UMTS (right diagram).*

will be triggered. No handover occurs if the serving access network grants the highest utility. Otherwise, the horizontal/vertical handover to the selected access network will be performed. Finally, if the WLAN signal level drops below θ_h without detecting the presence of a WLAN or WiMAX access node, the MN will handover to the UMTS to maintain the connectivity. The vertical handover decision from WiMAX to UMTS/WLAN (not presented here) is very similar to that from WLAN to UMTS/WiMAX.

3.3.3.3 Adaptive handover threshold θ_h

When the received signal level of the serving access network drops below handover threshold θ_h, network selection and then handover execution will be initiated. An adaptive handover threshold θ_h is required to ensure seamless handovers. Without loss of generality, we present an approach to compute this threshold when handovers occur between UMTS, WLAN, and WiMAX systems. The WLAN radio coverage is very small compared to the UMTS or WiMAX one, whereas the radio coverage of UMTS and WiMAX is approximately the same. The UMTS and WiMAX are intended to be used as complementary technology and their cells are partially overlapped. In order to achieve seamless inter-system handover, the overlap between two adjacent cells should be large enough. However, this is just a necessary condition but not sufficient. The seamless handover also depends on the value of handover threshold θ_h since the communication interruption may occur due to an inappropriate threshold.

In the vertical handover with multiple interfaces terminals, packet loss occurs if the handover is completed after the MN moves outside the overlap region between the two cells involved. The in-flight

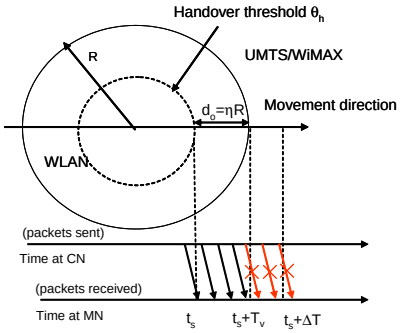

Figure 3.3: From WLAN to UMTS/WiMAX handover model

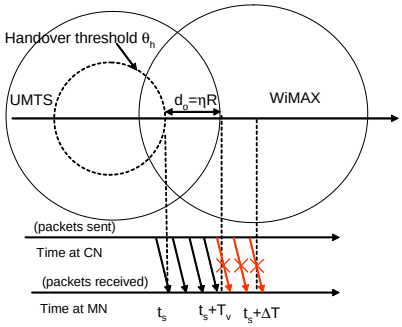

Figure 3.4: Handover between UMTS and WiMAX

packets following the old data path are not received by the MN via its old interface (e.g., three last packets in Figure 3.3 and Figure 3.4). The handover from UMTS/WiMAX to WLAN does not cause any loss of packets since the WLAN cell is fully covered by the UMTS or WiMAX cell. Hence, θ_h does not involve into the triggering conditions for this handover. Otherwise, the handover from WLAN to UMTS/WiMAX (Figure 3.3) or between UMTS and WiMAX (Figure 3.4) can cause packet losses if handover threshold θ_h is wrongly determined.

The parameters presented in Figure 3.3 and Figure 3.4 are described below:

- t_s : the instant when the received signal from the serving access node drops below handover threshold θ_h, $\theta(t_s) \leq \theta_h$. The network selection is triggered and the vertical handover to the selected access node is immediately executed.

- T_v : the time duration for the MN to cross the distance between the position where $\theta = \theta_h$ and the serving cell border, called crossing distance. The handover models illustrated in Figure 3.3 and Figure 3.4 correspond to the case where the MN moves on a straight line from the BS towards the cell border. The crossing distance is denoted as d_o. It is noteworthy that d_o is the shortest crossing distance (i.e., the worst case). Any other movement patterns (not going through the serving BS or not moving straightly) provides obviously a larger crossing distance compared to d_o. The adaptive handover threshold is thus based on this critical crossing distance d_o. We have

$$T_v = \frac{d_o}{v} \triangleq \frac{\eta R}{v} \tag{3.2}$$

  where $\eta \in [0,1]$, v is the terminal velocity and R is the serving cell radius estimate.

- ΔT : the handover delay computed from instant t_s to the instant when the CN receives a handover notification from the MN. Once the CN receives the routing update information from the MN, data packets will be routed to the MN via a new established path.

- δ : the packet delay computed from the instant when the CN sends a packet to the instant when this packet arrives at the MN's radio interface.

In order to achieve a seamless data delivery (without loss of packets), the last packet sent by the CN must arrive at the MN before the MN moves out of the overlap region. The seamless handover condition is therefore:

$$T_v \geq \Delta T + \delta \tag{3.3}$$

By substituting $T_v = \frac{\eta R}{v}$, we have:

$$\eta R \geq v(\Delta T + \delta) \tag{3.4}$$

It means that the handover should be initiated at distance $(1-\eta)R$ from the serving BS. According to the signal propagation model in [104,105], the received signal level θ_h at $(1-\eta)R$ can be formulated as:

$$\theta_h[dB] = K_1 - K_2 log((1-\eta)R) + X_\sigma \tag{3.5}$$

where K_1 represents the antenna gains and the signal's wavelength, and K_2 represents the path loss factor. X_σ is a zero mean stationary Gaussian random process modeling shadowing fading. The shadowing part in the received signal strength can be suppressed by passing through a low-pass filter, thereby (3.5) can be rewritten as:

$$\theta_h[dB] = K_1 - K_2 log((1-\eta)R) \tag{3.6}$$

From (3.4) and (3.6), we have:

$$\theta_h[dB] = \theta_o - K_2 log(1 - \frac{v}{R}(\Delta T + \delta)) \tag{3.7}$$

where θ_o is the received signal at the serving cell border. We see that θ_h depends on the packet transmission delay via the serving network, the handover delay to the target access network, the terminal velocity estimate and the cell radius estimate. In fact, the path loss factor K_2 and the cell radius R can be estimated based on the received signal strength measurements [96, 97]. The terminal velocity is also an estimated value [88–91]. The handover delay ΔT to a specific target access network is a given well-known value. The terminal can also estimate the packet delay δ based on the Round Trip Time (RTT) value. Accordingly, the handover threshold θ_h can be adaptively determined by the terminal itself. The handover threshold formula (3.7) remains valid for any couple of access technologies.

Now, we consider a fix handover threshold θ_h^*. From (3.3), packet loss happens if $(\Delta T + \delta - T_v) > 0$ where T_v will be determined in function of θ_h^*. From (3.2) and (3.6), the relation between T_v and θ_h^* is as follows:

$$\theta_h^* = \theta_o - K_2 log(1 - \frac{T_v v}{R}) \tag{3.8}$$

If the CN sends packets at a constant rate r, the number of lost packets is

$$NL = (\Delta T + \delta - T_v) r \tag{3.9}$$

By substituting T_v computed from (3.8) and $(\Delta T + \delta)$ computed from (3.7) in (3.9), the number of lost packets due to a fixed handover threshold can be expressed as:

$$NL = \left(10^{\frac{\theta_o - \theta_h^*}{K_2}} - 10^{\frac{\theta_o - \theta_h^a}{K_2}}\right) \frac{Rr}{v} \tag{3.10}$$

where θ_h^a is the adaptive handover threshold. We observe that if $\theta_h^* < \theta_h^a$, packet losses will occur ($NL > 0$). If we set a great value for θ_h^*, the handover from WLAN to UMTS/WiMAX will be triggered too early that the user could not benefit from the advantages offered by WLAN. In the handover between UMTS and WiMAX, a great value of θ_h^* means a great overlap between two inter-system cells. This implies a significant reduction of the radio coverage and an increase of the network deployment cost for operators. Conversely, a too small value of θ_h^* will lead to important packet losses. Figure 3.5 illustrates the number of lost packets in function of terminal velocity v for different choices of θ_h^*. The results are obtained in the case of the handover from WLAN to UMTS/WiMAX with the following setting parameters: $R_{wlan} = 200m$, $r = 100 packets/s$, $(\Delta T + \delta_{wlan}) = 2s$ and $K_2 = 40$. We see that the number of lost packets grows very fast with the increase of the velocity. The result shows the advantage

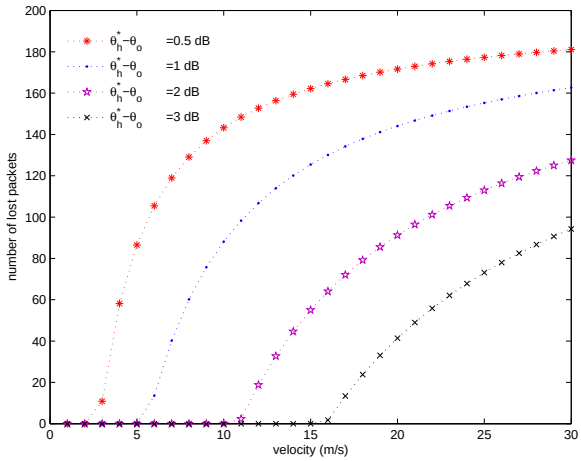

Figure 3.5: Packet loss due to a fixed handover threshold

of having an adaptive handover threshold to avoid packet losses during handover.

3.3.3.4 Network selection decision algorithm

After identifying the appropriate handover instant, the next step of the process is the network selection decision-making. We specify a network selection that performs in two phases: pre-selection and utility-based decision-making. In the pre-selection phase, undesirable access networks will be eliminated from the candidate list. First, the elimination involves access technologies whose corresponding radio interface is turned-off according to the interface management policies. These access networks are discovered right before taking the decision to turn off their corresponding interface. Secondly, the ANIs belonging to the black list established in the experience repository are filtered out. The pre-selection phase can also eliminate access networks that do not support the services required by the user.

After the pre-selection phase, our solution evaluates the utility for each remaining candidate access network, taking into account the user preferences configuration. The network selection criteria that are used in the utility-based evaluation are cost (c), power consumption gain (ge), maximum achievable data rate (r), access network load (ρ) and link quality (S). For each running application class and each situation profile, the preference weights for all considered criteria is deduced from the configured priority preferences. For example, if the priority preference vector is $p = [3,0,2,1,1]$ (that is, *c=high*, *ge=ignored*, *r =medium*, ρ=*low*, *S=low*), the associated preference weight vector becomes $\alpha = [0.43,0,0.29,0.14,0.14]$. Based on the utility value, the selected access network is the one that leads to the highest utility. The aggregate utility of an access network is in fact computed by:

$$U(\mathbf{x}_i) = \prod_{j=1}^{M} [u_j(x_j)]^{w_j} \tag{3.11}$$

where M is the number of considered criteria and x_j is the value of criterion j in vector $\mathbf{x}_i = [c, ge, r, \rho, S]$. w_j is the preference weight of criterion j corresponding to the current user situation and current running application class. u_j is the elementary utility of criterion j which takes form of a sigmoid shape proposed in 2.4.2.1. The details on the choice of this multiplicative aggregate utility function are explained in Chapter 2. In the above utility evaluation, if two access networks provides the same highest utility, each individual criterion, ordered from the highest to the lowest priority, is solely compared.

Two different access networks could not have all identical characteristics. Consequently, a selection decision is made as soon as a different utility is found for a particular criterion.

3.3.4 Handover execution

As the converged core network will be purely IP-based, we adopt the MIP mechanism [18,19,26–28] to maintain transparent network connectivity for mobile users on the move between different IP sub-networks. In fact, when a user roams into a foreign network, its terminal will acquire a new temporary address, called CoA. This address can be either obtained via an auto-configuration mechanism or be the address of the foreign network gateway, known as FA. The former is known as co-located CoA and the latter is known as FA-CoA. To accommodate the user roaming within IPv6 networks, a terminal can configure itself a new CoA without requiring an FA. The terminal then registers its CoA with its CN and its HA located in its home domain so that the packets destined for the user can be delivered to its current attached network.

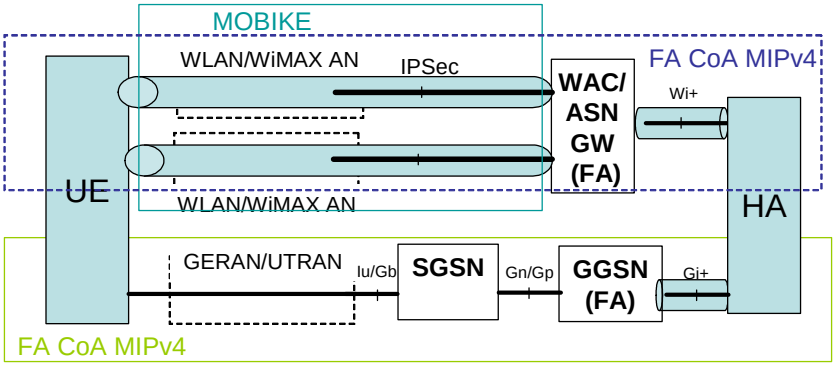

Figure 3.6: FA-CoA based mobility management solution

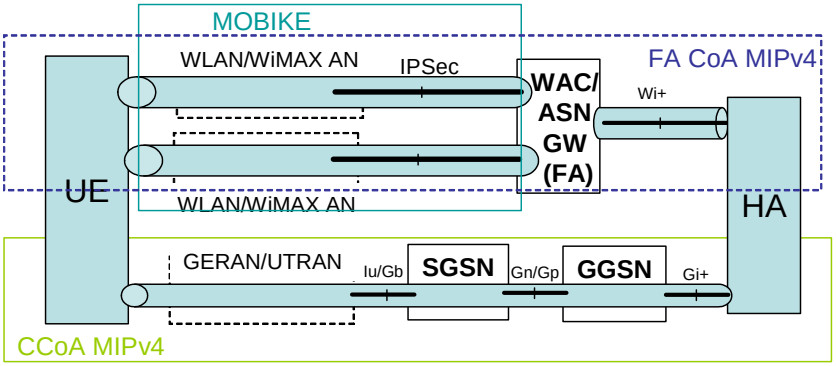

Figure 3.7: Co-located CoA based mobility management solution

In order to enhance the mobility management, we use, when possible, multiple CoA registration solutions for the handover execution [106–108]. Conceptually, the MIPv4/MIPv6 standard does not allow a terminal to register multiple CoAs bound to a single home address. A new identification called the Binding Unique Identification (BID) is used to associate with each binding cache entry to accommodate multiple binding registrations.

When the User Equipment (UE) is powered on, it searches for the available RAN and retrieves an associated CoA for each radio interface. The UE registers each of its CoAs with the HA. Conventionally

the CoA associated to the 3GPP radio interface is set to primary CoA. During the idle communication mode, only the primary 3GPP binding cache is updated when the UE changes its network domain. Therefore, the HA will use the primary 3GPP CoA to page the UE in case of an incoming call. When the access and core networks are IPv4-based and an FA is already deployed in the gateway of a particular access network, we suggest using the FA-CoA MIPv4 solution (as depicted in Figure 3.6). If a specific access gateway does not include the FA functionalities, the co-located CoA MIPv4 solution is employed as illustrated in Figure 3.7. Finally, the MIPv6 will be used if the access and the core networks are IPv6.

The main idea of the multiple interface handover execution is that a UE establishes a new connection with the target access network via its new target radio interface while maintaining the communication with its serving interface. The authentication with the target access network can be achieved through a universal pre-paid SIM card. After the handover is completed, the old interface remains active for a period of time to receive in-flight packets on the old data path, and finally switches to standby state. There exist two simultaneous communications via two different radio interfaces during a short handover execution period. One may note that another handover execution scheme like SIP-based handover [109] can be used in place of MIP-based handover without affecting our proposed interface management and network selection solution. In any case, we only need to know the handover execution delay to determine the appropriate handover threshold to achieve seamless handovers and reduce power consumption.

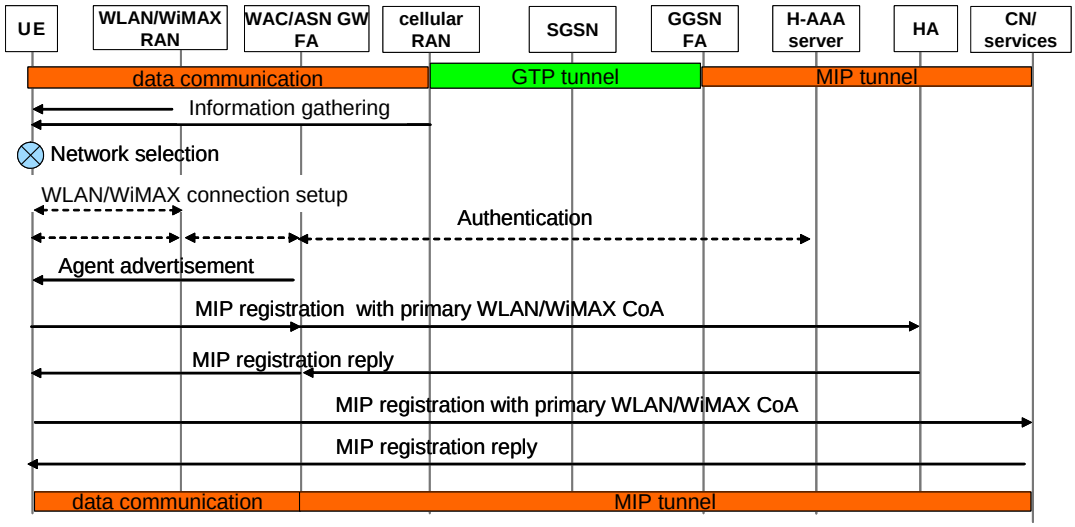

Figure 3.8: Handover procedure from 3GPP RAN to WLAN/WiMAX RAN

3.3.4.1 Handover from 3GPP RAN to WLAN/WiMAX

When the UE is communicating via 3GPP Radio Access Network (RAN) and the target access network, according to the network selection, is WLAN/WiMAX, the vertical handover from 3GPP to WLAN/WiMAX is triggered. First, the UE sets up the connection and authenticates with the target network. The foreign network authenticates the UE by exchanging the information with the H-AAA server. Afterwards, the UE acquires the corresponding CoA and sends a MIP registration with a primary WLAN/WiMAX CoA option to the HA and its CN. Once the MIP registration reaches the HA and its CN, these latter set the new WLAN/WiMAX CoA as primary CoA, return the MIP registration reply and use the new CoA path to forward or send data to the UE. After the handover is complete, the

cellular interface remains active for a period of time to receive the in-flight packets from the old path and finally switches to standby state.

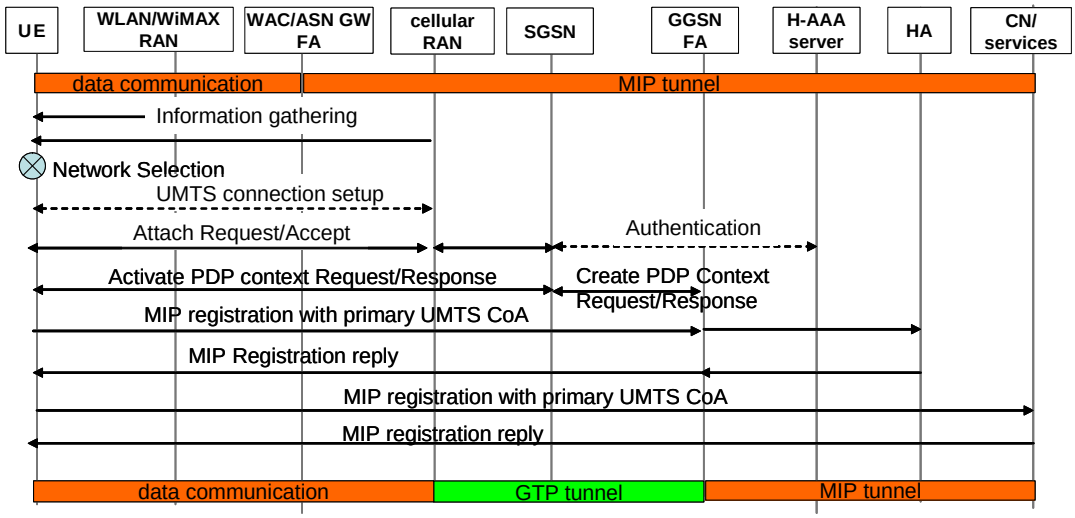

Figure 3.9: Handover procedure from WLAN/WiMAX RAN to 3GPP RAN

3.3.4.2 Handover from WLAN/WiMAX to 3GPP RAN

When the UE is communicating via WLAN/WiMAX, the cellular interface is in standby state and the 3GPP CoA in the HA and the CN are still maintained. If the association between the UE and the GGSN, known as PDP context, is not timer-expired, the UE only needs to send the MIP registration to the HA and its CN to execute the handover. If the PDP context is already released, the UE needs to reactivate the PDP context to establish the authentication with the 3GPP network. The in-flight packets destined for the UE via the old path are received as long as the UE is still covered by the old WLAN/WiMAX access network.

3.4 Performance evaluation

In this section we present the simulation results to highlight the benefits of the proposed network selection and power-saving interface management solutions. We show the advantages of the application-aware and situation-aware user preferences configuration. Finally, we evaluate the power consumption efficiency of the power-saving interface management and the power-aware network selection strategy.

We consider simulation scenarios in which a user holds a terminal equipped with 3 radio interfaces: UMTS/HSPA, WiFi and WiMAX. At each instant the user is able to choose from three available access networks: UMTS, WiFi and WiMAX and then to perform terminal-controlled vertical handover to the selected access network. The selection is based on five selection criteria mentioned previously. They are cost, power consumption gain, maximum achievable data rate, network load and signal quality (SINR for WiFi/WiMAX and energy per bit to noise power spectral density ratio E_b/N_o for UMTS). The values of signal quality, network load and achievable data rate for each available access network are randomly generated every several dozens of seconds (e.g., every 100 seconds for UMTS, every 75 seconds for WiMAX and every 50 seconds for WiFi). The achievable data rate is generated in such a way that it is correlated to the signal quality value. The cost (both cost per minute and cost per data volume) for UMTS access network is fixed during each simulation while the cost values of WiFi and

WiMAX are also randomly generated. The value range of selection criteria are given in Table 3.2. The setting parameters for each elementary utility are given in Table 3.3.

	UMTS	WiFi	WiMAX
cost	$20-70$	$0-35$	$5-50$
$e(W)$ in active state	1.2	4.5	3.5
$e(W)$ in standby state	0.005	0.06	0.03
Data rate $r(Kbps)$	$0-1000$	$0-1500$	$0-2000$
$S(dB)$	$1-12$	$1-25$	$1-25$

Table 3.2: Simulation parameters

Criterion	x_α	x_β	x_m	ζ
c	5	70	35	2
$ge(J/Mb)$	0	30	4	2
$r(Kbps)$	100	3000	600	3
ρ	0	1	0.5	2
$S = SINR$(dB)	5	25	15	3
$S = E_b/N_0$(dB)	2	12	6	3

Table 3.3: Setting parameters for elementary utility forms

The energy consumption rate values of UMTS and WiFi interfaces are the averaged values taken from the range presented in [93, 94] as well as in the off-the-shelf products' specification. The energy consumption rates of WiMAX interface are chosen following the fact that they are greater than those of UMTS interface and not much less than those of WiFi interface. The range of link quality value for UMTS (E_b/N_o) is taken from [104], and that of WiFi and WiMAX is extracted from [110, 111]. The cost parameter represents both cost per minute and cost per data volume. A relative cost value is used rather than a precise unit.

In the simulation, we assume that the probability of error in radio transmission is zero. The achievable throughput while connecting to a particular access network can be modeled as $r_i(1-\rho_i)$ where r_i is the maximum achievable data rate and ρ_i is the load of access network i before the user connects to it. The achievable throughput is an unknown parameter at the terminal side during the selection-making process.

3.4.1 Application-aware network selection

In the first simulation, we show the need and the advantage of having an application-aware user preferences configuration. We consider two different preference weight vectors: $p_1 = [1,1,2,3,3]$ and $p_2 = [3,1,3,1,1]$. Recall that the preference vector here is expressed in terms of priority (high, medium, low or ignored) and the vector elements correspond to five considered criteria $[c, ge, r, \rho, S]$ respectively. Configuration p_1 sets high preferences for QoS-related criteria while configuration p_2 emphasizes preferences on cost and achievable data rate criteria. Preferences configuration p_1 can be adopted for streaming users who want to achieve a seamless streaming by selecting high QoS access networks rather low-cost ones. On the contrary, p_2 is suitable for users who run non-real time data services (e.g., data downloading). These users prefer to select low cost or high speed access networks and they are willing to suffer from some moments of discontinuity.

We consider a mobile user who runs a streaming application on his mobile device. Assume that the streaming application duration is 1500 seconds and during this time, the user is crossing different access networks (UMTS/HSPA, WiFi or WiMAX). During the streaming session, we monitor the application's buffer size. The buffer evolution is simulated by:

$$\begin{cases} b[0] = 10R_{play} \\ b[t] = \max(0, \min(10R_{play}, (b[t-1] + tp[t] - R_{play}))) \end{cases} \tag{3.12}$$

where R_{play} is the playback rate of the streaming application and $10R_{play}$ is the maximum memory size allocated to the buffer (i.e., 10 seconds of video playing). At the beginning of the simulation, the buffer is assumed to be filled. When the simulation starts, the user moves and his terminal selects automatically the appropriate access network. The buffer is filled at the current throughput rate $tp[t]$ every sample time. When the buffer runs out of content, the streaming application is interrupted.

We observe the interruption time length during the streaming session for two network selection strategies. The percent of interruption time during the streaming session according to different playback rates are presented in Figure 3.10. We see that the interruption time of strategy p_1 is much shorter than that of strategy p_2. In fact, the strategy p_2 promotes low cost WiFi and WiMAX access networks which perhaps cannot guarantee the QoS required by streaming applications. In the same network coverage, strategy p_1 selects high cost cellular access networks if they assure QoS support and if there is no low cost WiFi/WiMAX access network assuring that quality. In the two cases, the buffer outage is explained by the fact that none of available access networks satisfies the requirements of the user's application. In spite of the trade-off between QoS and price, a rational user prefers to pay higher price for a good quality streaming service rather than to pay lower price for poor quality streaming. The configuration p_1 is more suitable than p_2 for users using streaming applications.

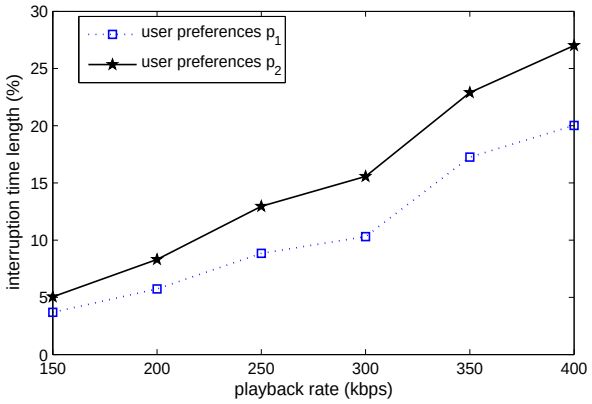

Figure 3.10: Application-aware user preferences for streaming services

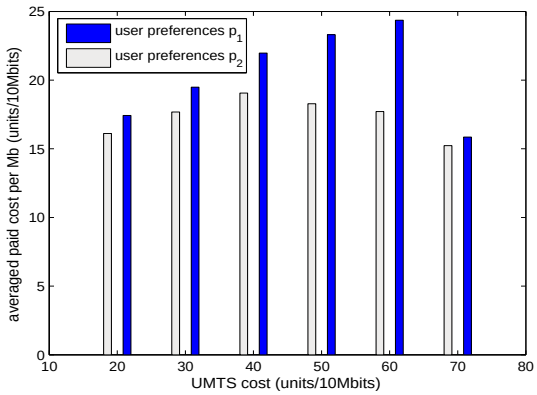

Figure 3.11: Application-aware user preferences for data downloading services

In the same network coverage scenario, we assume now that the user launches a data downloading application. We observe the cost that the user pays for each Mbit of downloaded data. The averaged cost per data volume for selection strategy p_1 and p_2 are presented in Figure 3.11. The results are obtained for different values of UMTS cost criterion (the UMTS cost is unchanged during each simulation). In this case, strategy p_2 is more cost-effective than strategy p_1. The gain of downloading speed is very small compared to the cost to pay. Consequently, it appears that preferences configuration p_2 is more suitable for non real time applications than p_1. The simulation results confirm effectively a need of having an application-aware network selection to best suit the diversity of QoS requirements of application classes.

3.4.2 Situation-aware network selection

Generally, if the user is at home, he does not need to care about power-saving. If he is in his office, he should select his secured enterprise networks for connectivity. Also, high-speed users should not select small-cell-size access networks such as micro-cells or WiFi cells since their terminal will quickly handover to another access network. Such strategies are indeed situation-aware network selections.

We consider two different situation profiles: at home and not at home. In the home environment, most of the time, the home WiFi networks provide a good QoS support and importantly a low-cost broadband access. Two user preferences configurations: $p_1 = [3,0,1,2,3]$ and $p_2 = [2,2,1,2,3]$ are considered. In the first configuration, the user does not care about power-saving issues and he sets high priority preferences for the cost and QoS-related criteria. Intuitively, this preferences configuration is suitable for the at-home situation. In the second configuration, the power consumption criteria are taken into account in network selection and the cost is set to a medium priority. As we aim to investigate the performance of streaming applications in this simulation, the QoS-related criteria in both configurations are intentionally set to high priority.

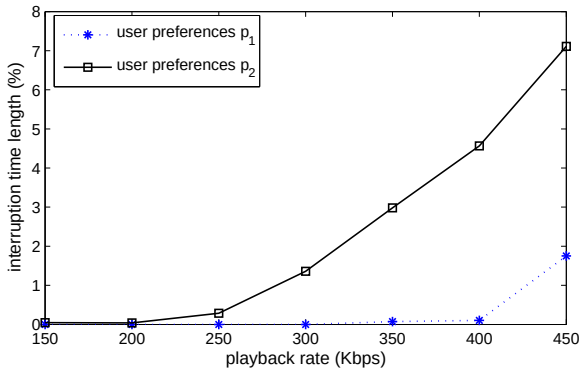

Figure 3.12: Streaming application performance in at-home network situation

First, we simulate a user who runs a streaming application in his home environment. The ratios of the interruption time length to the streaming session duration for the two considered user preferences configurations are presented in Figure 3.12. The interruption time of preferences configuration p_2 is much higher than that of p_1. The streaming quality of preferences p_1 is almost seamless. In fact, the network selection with preferences p_1 results in selecting most of time the home WiFi access network. Some small breaks occur when the playback rate is high since the simulated data rate of WiFi can be below the playback rate and no other access network offer better QoS. In the case of preferences strategy p_2, when the quality of the WiFi access is slightly degraded, the UMTS or WiMAX access networks will be selected instead. However, the quality of WiFi access network may be still better than UMTS or WiMAX access networks. The power consumption gain of UMTS and WiMAX interfaces contributes a significant factor in the aggregate utility evaluation, which leads to such a decision. That explains the difference between the outages of the streaming in the two network selection strategies. Consequently, the preference configuration should take into account the user situation to fulfill the user's requirement.

Second, we investigate the use of the preference configurations p_1 and p_2 when the user is not in his home environment (not at home situation). The user runs the same streaming application. The ratio of the interruption time length to the total simulation duration is monitored and presented in Figure 3.13. The preferences strategy p_2 yields better results in terms of streaming quality. In fact, the network selection using preferences p_1 promotes low cost access networks (like free public WiFi) which sometimes do not have good enough QoS support. As preferences strategy p_2 does not focus on

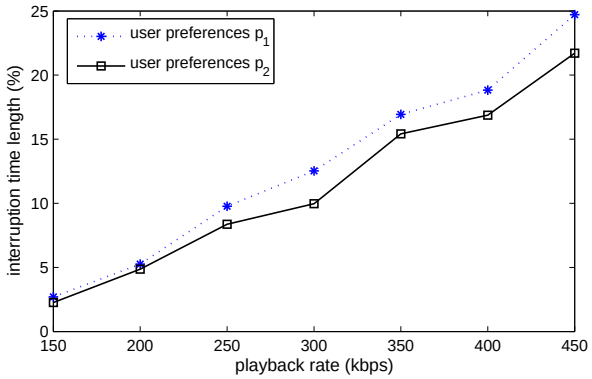

Figure 3.13: Streaming application performance in not-at-home network situation

the cost factor, such low cost access networks will not be selected. The result confirms that preferences strategy p_1 which is suitable for the home situation does not perform well in the not-at-home situation. Therefore, a preferences configuration specified for a particular situation will not be appropriate for use in another situation profile. The user preferences configuration should be situation profile-aware.

We have shown the need of an application-aware and situation-aware preferences configuration. The user preferences plays indeed an important weight in the selection decision. In the same network coverage, two different preferences configurations lead to two different selection results. An adequate user preferences configuration thus becomes an important step in the network selection design in heterogeneous network environments. However, the user preferences are user-specific as its name indicates. There is no optimal preference setting. The preference setting should provide different modes, one for novice users with default setting parameters and one for experts who can tune properly the parameters.

3.4.3 Power consumption efficiency

In this section, we evaluate the power-aware preferences configuration offered by the proposed network selection scheme and the efficiency of the proposed power-saving interface management policy. We consider a business man who takes a two-hour journey back home. During this long journey, he takes a train, gets on a bus, stops at a shopping mall and then walks to home. He runs a real time application (e.g., voice conversation and streaming video) on his terminal most of the time and wants to optimize the use of the terminal's battery in order to be always-on until arriving home. The user therefore sets the highest priority preference for the power consumption gain criterion from the network selection's GUI. The power-aware preferences configuration is $p = [1, \mathbf{3}, 1, 2, 1]$. We will compare this power-aware preferences configuration with a non-power-aware preferences configuration $p' = [2, \mathbf{0}, 1, 2, 1]$. The user preferences configuration p and p' are kept unchanged during the simulation.

We consider the three following strategies: non-power-aware network selection (NS) *without* Power-Saving Interface Management (PSIM), power-aware NS *without* PSIM, and power-aware NS *with* PSIM. The terminal's lifetime in function of the remaining battery capacity for these three cases is depicted in Figure 3.14. The device's lifetime is computed from the beginning of the simulation until the instant when the device runs out of battery. The results show that the power-aware preferences strategies (two upper curves of Figure 3.14) make it possible to elongate the device's lifetime significantly compared to the non-power-aware configuration (the bottom curve). In fact, between two alternatives with approximately the same quality, the power-aware strategy prefers the access network that has lower energy consumption gain.

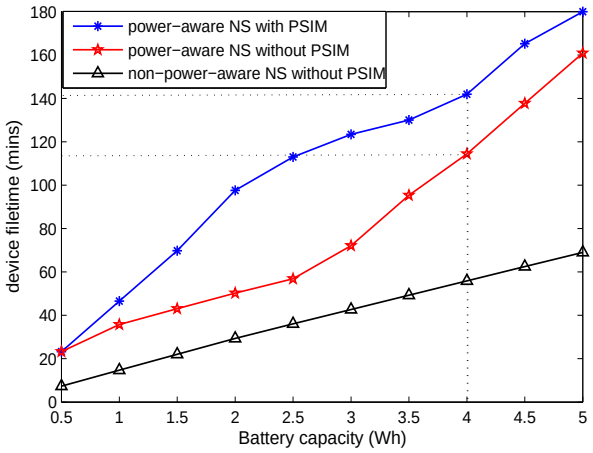

Figure 3.14: Portable device's lifetime vs. remaining battery capacity

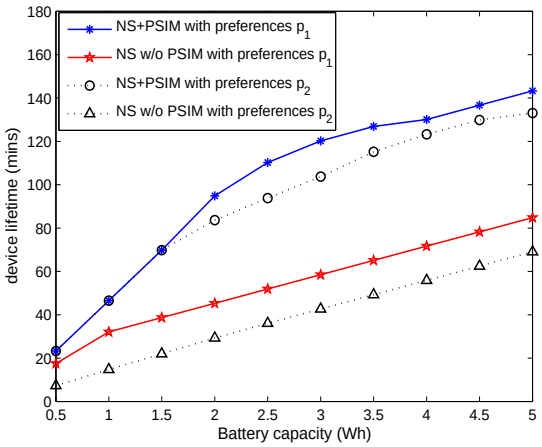

Figure 3.15: Portable device's lifetime for $p_1 = [3,2,1,1,1]$ and $p_2 = [3,1,1,1,1]$

When the power-saving interface management policy rules as described in Section 3.3.2 are implemented, the device's lifetime is clearly extended. From Figure 3.14, the network selection combined with PSIM (the asterisk curve) allows the user to remain connected until reaching his home with a smaller battery capacity usage $(3Wh)$ compared to the network selection without PSIM $(> 4Wh)$ (the stared curve). In fact, the proposed interface management prevents users from selecting high power-consumption interfaces if the remaining battery lifetime becomes low. These interfaces will be powered off to elongate the device's lifetime. Also, the WiFi and WiMAX interfaces are powered off if the mobile terminal's velocity matches the pre-defined turning-off thresholds (see Rule 4 in Section 3.3.2).

In order to confirm the advantage of the proposed PSIM, we conduct some simulations with random user preferences configurations. For each preferences configuration, we analyze the power efficiency of the NS combined with PSIM (NS+PSIM) and the NS without PSIM (NS w/o PSIM). The results are presented in Figure 3.15 for two preferences configurations $p_1 = [3,2,1,1,1]$ and $p_2 = [3,1,1,1,1]$. It is clear that the NS combined with PSIM can reduce the wasteful power consumption and thus elongate the device's lifetime.

Furthermore, the PSIM can also reduce significantly the energy consumed during the idle communication mode. If we keep all three radio interfaces in standby state, the total consumed power due to the RF communication is $P_{idle} = \tau(e^{sb}_{umts} + e^{sb}_{wlan} + e^{sb}_{wimax})$ where τ is the idle time duration and e^{sb}_i is the energy consumption rate of interface i in standby state. If our interface management is used, the WiFi and WiMAX interfaces are powered off. We can therefore minimize $\Delta P = \tau(e^{sb}_{wlan} + e^{sb}_{wimax})$. The amount of saved energy ΔP is noticeable since the idle communication duration is normally greater than the active one and the energy consumption rate in standby state of WiFi and WiMAX is considerable. For example, if $\tau = 5h$, $\Delta P = 0.45Wh$ ($e^{sb}_{wlan} = 0.06W$ and $e^{sb}_{wimax} = 0.03W$), the amount of saved energy ΔP is significant compared to the total battery capacity of normal portable devices (e.g., around 6.66Wh for Lithium-ion battery 1800mAh/3.7V). In short, the proposed power-saving interface management provides an efficient solution to optimize the power consumption and elongate the device's lifetime.

3.5 Summary

Existing mobility management solutions in heterogeneous networks require more improvements before they can be successfully deployed. The main challenges seem related to the business relationship

between access network operators and also the capital investment in network infrastructure modification. Accordingly, in this chapter we proposed a fully terminal-controlled mobility solution across heterogeneous networks without architectural changes in network operators' infrastructure. We described a user-centric approach for vertical handover management. We proposed in this work the user-centric network selection, the power-saving interface management and the adaptive handover initiation algorithm at the terminal to support seamless terminal-initiated and terminal-controlled vertical handover. The proposed access network selection is situation profile-based and application-aware (context-aware) to suit different communication contexts. It enables terminals to select the most suitable access network according to various access network characteristics. Multiple wireless interfaces of a terminal device are handled in both idle and active communication modes to optimize the power consumption. Interface management also serves as a complement to the network selection scheme by turning off inappropriate interfaces according to the terminal velocity and the remaining battery lifetime. The interface management solution optimizes the power consumption of battery-limited portable devices. We also address an adaptive handover initiation scheme to assist the service continuity. The solution is realistic and not much complex to implement in current mobile devices and networks.

In this chapter, we have shown that a seamless handover over a very loose coupling interworking can be possible thanks to the help of adaptive handover initiation threshold. However, the handover management is based on the fact that the terminal has multiple radio interfaces and the cell overlap is enough large. In the next Chapter 4, we address a terminal-controlled solution to seamless handover (for streaming users) in the case where the terminal is equipped with a single SDR interface as well as in the case where the cell overlap is not enough large to support the above multi-interface handover. Regarding the minimum required cell overlap condition for seamless handover of an SDR-enabled terminal, a thorough analysis is given in Chapter 6.

∼∼ △♡△ ∼∼

Chapter 4

Handover Prediction-Assisted Seamless Media Streaming

[...the wireless communications environment is not readily conducive to streaming video. On one hand, high-quality streaming video is no easy task. It involves downloading, decoding and playing video and audio simultaneously, and with no or very limited re-buffering taking place. On the other hand, the wireless environment is fraught with challenges like interference, multipath fading, bit stream errors and mobile terminal devices that are moving targets, darting in and out of areas that may have different transmission speeds as well as other characteristics....]

extracted from "Bringing Streaming Video to Wireless Handheld Devices" [112]

Among the challenges mentioned above, handover of mobile users is one of the important aspects that we address in this chapter. In one-way streaming media applications, the pre-buffering at the terminal-side is a well-known technique to overcome the shortage of media contents in the client's buffer. However, due to the limited memory at the terminal-side and the impairments of wireless channel, the media content in the buffer suffers from drastic fluctuations. In this chapter, we address a terminal-controlled adaptive pre-buffering policy based on handover prediction to support the high-quality seamless media streaming to mobile users. The handover prediction will be thoroughly analyzed.

4.1 Introduction

The increasing availability of wireless broadband networks has accelerated the widespread use of multimedia services. Wireless technologies have been evolving towards high data rates and high QoS support to ensure high-quality video streaming and video on demand services. The capabilities of portable devices continue to proliferate to accommodate future mobile multimedia services. However, ensuring seamless streaming over heterogeneous wireless access networks for mobile users remains a challenging task.

When one decides to bring video streaming applications to mobile devices, a set of challenges appears. One of them is the limited memory of portable devices. In a streaming application, the media content at the client is stored in a buffer. Basically, such a buffer is an area of memory used for temporary storage of data when a program or hardware device needs an uninterrupted flow of information. Buffers are typically created in Random Access Memory (RAM) rather than on the hard disk, as fetching data from RAM is faster than retrieving it from the platter technology used in conventional hard

drives [113, 114]. Hereinafter, the term "memory" is referred to as RAM. The good-quality seamless streaming service requires maintaining terminal-side buffer of proper size with media contents to ensure the continuity of the multimedia session. Without use of a buffer, one would either have to wait for the entire video to load to the local machine before playing, or endure playback with breaks and jumps in the data stream. In order to smooth the streaming playback, some may suggest configuring the buffer size to be very large. The problem is, the bigger the buffer, the less system RAM available for other tasks. More particularly, today's portable devices are increasingly multi-purpose and equipped with multiple peripheral devices like camera, music player...but their RAM memory is limited. Hence, the efficient utilization of the RAM memory in portable devices should be carefully considered, especially while managing the streaming buffer.

The buffer size affects the amount of startup latency because a client usually does not start playing a stream until the buffer is full. The conventional startup latency of 5 seconds is still frustrating for users so that several enhancements, called advanced fast start, have recently been proposed in new released media streaming clients to reduce such a delay [115]. The main idea is to start rendering content when media player receives a minimum amount of data. At the same time, the data are streamed at an accelerated rate, a rate that is faster than the encoded bit rate of the content, until the buffer is full. This provides an instant-on, always-on streaming experience by effectively eliminating buffering time.

Another well-known issue for streaming over wireless environment is the wireless network resource fluctuation which induces packet losses, delay, jitter and breaks during the video streaming. Besides variations of the wireless channel quality, the interruption during the handover process is also one of important sources of discontinuity. While the former issue can be resolved by intelligent streaming with the content adaptation [115], the later has not been sufficiently addressed in the literature. It represents the main target of this chapter.

The future 4G networks will provide users with facilities to move across different networks as mentioned in previous chapters. Basically, when an MN roams to a foreign network, it suffers a blackout period caused by handover. For the intra-system mobility between two WLAN APs for instance, the MN can communicate with only one AP at a time. Thus, it cannot communicate with a new target AP before stopping the communication with the old one. In fact, the MN must carry out the measurement, selection, authentication and association processes with the target AP. It also spends time on configuring a new CoA and then registering its CoA with its HA and its CN. The overall handover process induces a latency which causes packet losses and thereby streaming breaks. The same issue will happen if the MN uses an SDR-based reconfigurable interface [85] [84] for the vertical handover between two different wireless access systems.

Even if the MN is equipped with multiple radio interfaces to realize a *soft* inter-system handover (i.e., two communications at a time: one with the old access node and one with the new access node during the handover execution), handover interruption may still occur. If the cell overlap region is small, the MN may move out of the old radio coverage before the handover procedure is completed. The communication is therefore interrupted, which implies the shortage of the streaming buffer.

In this chapter, we propose a terminal-controlled pre-buffering adjustment policy, running at the terminal to maintain the appropriate amount of media content in the buffer. The main idea is to stream data at accelerated rates when necessary to fill the buffer to prevent the content shortage after taking through bad wireless channel experiences or prior to handover executions. To do so, we propose a practical handover prediction which assists the pre-buffering management.

4.2 Related work

In recent years, many techniques have been proposed to address the network resource fluctuation

for the multimedia streaming over fluctuating bandwidth and best-effort wireless networks. The source coding community has proposed a scalable video bit stream [116] in such a way to enable the server to adapt the video bit rate to the current available bandwidth. Other adaptive media streaming solutions that could adapt the stream quality with respect to variations of the transmission channel or the congestion state were studied in [117, 118]. In addition to the quality adaptation perspective, error-resilient coding and channel coding techniques [119, 120] were also used to mitigate packet losses and long delay in data transmission.

In an MIP-enabled wireless environment, different improvements were introduced to reduce the handover latency and packet loss caused by MIP handovers like the fast MIP [28], the hierarchical MIP [27] and the smooth handover scheme [121]. Yet, the packet loss is still present due to the small cell overlap region or due to the ineligible handover latency which may be up to few seconds. Although the pre-buffering techniques have been already used to overcome the possible shortage of media buffer, few studies address the seamless streaming considering the blackout of handover. As the media buffer size is limited, the pre-buffering operation should be adaptively handled. The inaccurate pre-buffering decision can lead to the shortage of the media buffer. As the pre-buffering at an accelerated speed implies the additional load on the server side as well as on the wireless link, it should be properly triggered when needed and when suitable. An accurate handover prediction combined with an accurate available bandwidth is thus crucial to determine the suitable pre-buffering initiation instant.

The available bandwidth estimation is an active area of research in its own right [122–125]. Usually, the available bandwidth estimation is based on the use of probing techniques, i.e., observing the behavior of the probe packets exchanged between the server and the client. The end-to-end available bandwidth corresponds to the available bandwidth of the wireless radio link since the radio link usually acts as a bottleneck for data transmission. If the server is cooperative for the available bandwidth measurement (i.e., sending probe packets to the client), technique *Pathload* [122] or *Pathchirp* [123] can be used. In fact, *Pathload* and *Pathchirp* use the time dispersion at the client of the received probe packets through a bottleneck radio link to deduce the available bandwidth. Otherwise, the technique *Spruce* [125], which consists of forcing an uncooperative server to send the probe packets to the client, can be employed. These techniques give satisfactory results; therefore we suppose that this information will be given by any of these methods. However, we will focus on the handover prediction aspect in the following.

In the literature, most of the handover prediction can be achieved by estimating the current position or matching the usual movement patterns of mobile users. A dynamic Gauss-Markov or hidden Markov model applied to historical and current movement information has been used to predict the user's location or user's speed [126] [127]. Such information is coupled with the knowledge of access network deployment maps to predict imminent handovers. Also, the handover prediction based on trajectories followed in the recent past and movement habits of mobile users was introduced in [128] and references therein. These solutions are not feasible in reality since user mobility behaviors as well as wireless network topologies change in an irregular manner. Accordingly, maintaining such a database of user movements and network topologies is very costly and difficult to implement in a large-scale real network environment.

A more pragmatic handover prediction based on the RSS was exploited in [129–132]. The expected amount of time before a connection to be lost has been calculated using a linear prediction from two consecutive averaged RSS values [129]. Such a linear prediction is not accurate as the RSS values are highly fluctuating due to fading and shadowing effects. The first-order Grey Model (GM) has been shown as an efficient solution to mitigate the noise in RSS values and to predict the future RSS values, which makes it possible to help the handover decision [130] [131]. Recently, the GM has been used to predict handover probability regions (low, medium or high) for an adaptive buffering to avoid the streaming interruption during handover in WLAN networks [132]. However, the given

buffering scheme does not take into account the achievable data rate of current streaming connections, the network resource availability and the handover latency. This solution does not show how the terminal can achieve the required pre-buffering data to overcome the handover interruption. It only informs about the necessity to pre-buffer. Besides, in [133] [134], an estimation of handover latency and transient packet losses was exploited to calculate the adaptive required buffer size. However, these solutions did not determine the pre-buffering initiation instant and required a cooperation between client and streaming server. Contrary to the existing solutions, based on the handover prediction (i.e., remaining time before handover or remaining time before moving out of the serving cell) and the available bandwidth, terminal devices will control the pre-buffering operation to ensure the seamless streaming.

4.3 Client-side adaptive pre-buffering management

We focus on providing an efficient terminal-controlled pre-buffering management scheme to prevent from streaming interruptions during handover. The proposed solution does not require any changes to the network infrastructure as well as to the handover procedure. We assume that the streaming buffer at the terminal is already set to an appropriate size to avoid network fluctuations as well as handover blackouts. The buffer size may be fixed but the media content in the buffer is variable. To achieve the seamless streaming, we focus on handling the amount of media content in the buffer by managing the pre-buffering operation. In fact, we force the media content to be streamed at an accelerated speed to fill the buffer at suitable instants. The pre-buffering scheme is managed and controlled by the terminal itself without any additional cooperation with the streaming server. No change is required at the server side, which makes our approach feasible.

When the User Datagram Protocol (UDP) is used to carry multimedia streams, the average data rate of UDP session should be equal to the playback rate R_r of the streaming application (i.e., encoded bit rate) in normal network conditions. In order to fill the buffer at an accelerated speed, the client can adjust its current connection data rate if the server supports data rate negotiation. Otherwise, the client needs to open $n = \lceil \Delta R / R_r \rceil$ (ΔR is the possible increased data rate which is constrained by the available bandwidth) supplementary connections of R_r to fill the buffer in time. Hereafter, we assume that the server streams the data to the client at a steady data rate of R_r for each UDP connection. As opening new UDP connection to increase the streaming speed will generate the additional load at the access network side and also at the server side, it is preferable to do it only when needed. We propose a pre-buffering policy at the terminal as follows:

- *Rule 1:* If the data in the buffer is less than a pre-defined threshold $b[t] \leq b_{min}$, open a supplementary UDP connection to fill the buffer during $\frac{b_{max}-b[t]}{R_r+\Delta R}$ where b_{max} is the maximum size of the buffer, $b[t]$ is the current amount of data in the buffer and ΔR is the streaming speed of the additional connection. Usually, it is equal to the encoded bit rate, i.e., $\Delta R = R_r$. In fact, the streaming speed of the additional connection is constrained by the maximum available bandwidth at the radio link R_{max}, that is $(R_r + \Delta R \leq R_{max})$. The threshold b_{min} can be determined as a minimum amount of data necessary for rendering media stream. The value of b_{min} can be set to $2-3$ seconds of playing-out video. This condition may happen due to the handover execution or the wireless network resource fluctuation.

- *Rule 2:* If the available bandwidth is high (for example $R_{max} > 2R_r$) and the buffer is not full, open a supplementary UDP connection to fill the buffer.

- *Rule 3:* If the handover is predicted to be imminent and the buffer is not full, trigger the pre-buffering operations. This is the main focus of this work to avoid streaming interruptions during handover. The details according to different handover scenarios are presented in the following.

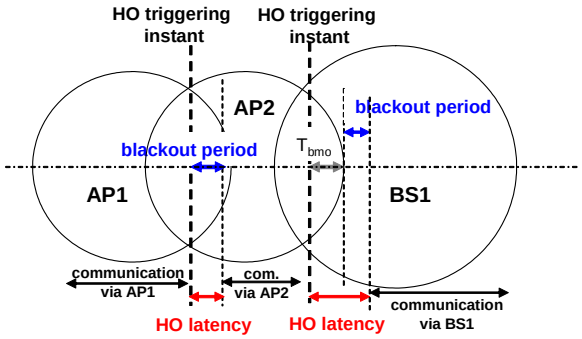

Figure 4.1: Horizontal and vertical handover model

To effectively manage the adaptive pre-buffering rule prior to the handover, the mobile terminal should predict the handover indicators like remaining time before handover and remaining time before moving out of the serving cell. Next, this information will be coupled with the knowledge of the handover blackout period and the available bandwidth on the current link to make the suitable decision about the pre-buffering. The blackout period will be the handover latency in the case of horizontal handover via the same radio interface (see Figure 4.1). For vertical handover using two different radio interfaces, the blackout period is the difference between handover latency τ_{ho} and time before moving out of the serving cell T_{bmo} (see Figure 4.1). The handover latency τ_{ho} depends on target access technologies. It corresponds to the delay for establishing the connection to a new access point and updating MIP bindings. The handover latency estimate was studied in [133]. For the sake of simplicity, a predefined well-known per-access technology handover latency value is assumed.

4.3.1 WLAN horizontal handover

The horizontal handover between two WLAN APs using one radio interface is difficult to achieve seamlessly. Similarly, the hard handover between two WiMAX cells or the vertical handover between two different technologies using an SDR-enabled device always suffers a blackout period during the switching from the serving BS to the target BS. Recall that the blackout period in this case corresponds to the handover latency, $\tau_{bo} = \tau_{ho}$. To achieve a seamless streaming over such a handover, we propose a pre-buffering scheme as follows:

Step 1 - Calculate the time before handover: When the terminal receives a good RSS from a neighboring AP, the remaining time before handover T_{bho} is predicted for each new pair of measured RSS values from serving and neighboring APs. The details of the prediction are presented in Section 4.4.2.

Step 2 -Calculate the required advance time: The required advance time constraint T_{ad} at instant τ to fill the buffer if the MN opens one supplementary connection of an allowable data rate ΔR is computed as:

$$T_{ad}[\tau] = T_{con} + \frac{(\min\{b_{max}, R_r\tau_{bo}\} - b[\tau])}{\Delta R} \tag{4.1}$$

where T_{con} is the time required to establish the additional UDP connection and $b[\tau]$ is the amount of data already in the buffer at instant τ. The increased data rate depends strongly on the available bandwidth constraint, i.e., $\Delta R = \min(R_{max}[\tau] - R_r, R_r)$ where $R_{max}[\tau]$ is the maximum available bandwidth

at instant τ. It should be noted that the advance time T_{ad} will not be changed every measurement sample due to the large time-scales fluctuation of the available bandwidth. Also, if $b_{max} < R_r\tau_{bo}$, the terminal can allocate more memory for the media buffer (set $b_{max} = R_r\tau_{bo}$) if possible to avoid streaming breaks during handover.

Step 3 - Pre-buffering triggering condition tuning: The advance time T_{ad} is a constraint for the pre-buffering initiation decision. On one hand, in order to successfully pre-buffer the required amount of streaming contents, the pre-buffering operation should be triggered under the following condition $T_{bho}[m] \geq T_{ad}[m]$. On the other hand, the pre-buffering should not be triggered too early since the user may change its movement direction and the handover may not occur. If the terminal keeps its movement pace (i.e., direction and velocity) and $T_{bho}[m]$ drops below $(T_{ad} + \Delta T)$ (where ΔT is the measurement interval), the next value $T_{bho}[m+1]$ is probably less than or equal to T_{ad}. A priori, the triggering condition should satisfy $T_{bho}[m] \leq T_{ad} + \Delta T$. To avoid the overestimation effect of the handover prediction, we suggest adding a hysteresis margin γ to the upper bound triggering condition. The choice of γ will be based on the confidence interval of the reliable predicted handover instant (i.e., those predicted when the MN is close to the handover instant). The pre-buffering is triggered at instant m if:

$$T_{bho}[m] \leq T_{ad}[m] + \Delta T + \gamma \tag{4.2}$$

The additional connection will endure until the buffer is fulfilled or until the handover occurs. One may note that if $T_{bho}[m] < T_{ad}[m]$, opening one supplementary connection is not enough to fill the buffer in time. In this case, we need a pre-buffering recovery step as described below.

Step 4: Pre-buffering recovery: During the pre-buffering operation, a severe drop of the time before handover estimation values at instant $j > m$ may occur. This may be due to an abrupt change of user movement pace. Therefore, during the pre-buffering procedure, we monitor the following condition:

$$\min\{b_{max}, R_r\tau_{bo}\} - b[j] > \Delta R T_{bho}[j] \tag{4.3}$$

If (4.3) occurs, it means that the time duration to fill the buffer is greater than T_{bho}. If $T_{bho}[j] > T_{con}$ (i.e., the remaining time before handover is large enough to establish a new connection), initiate new supplementary UDP connections of streaming speed $\Delta R'$. The value of $\Delta R'$ is:

$$\Delta R' = \frac{\min\{b_{max}, R_r\tau_{bo}\} - (b[j] + \Delta R T_{con})}{T_{bho}[j] - T_{con}} - \Delta R \tag{4.4}$$

In other words, the number of new supplementary UDP connections to compensate the abrupt change of T_{bho} during the pre-buffering procedure is $n' = \lceil \frac{\Delta R'}{R_r} \rceil$. Consequently, the total number of additional connections becomes $(1 + n')$.

4.3.2 Multi-interface vertical handover

In vertical handover using two different radio interfaces, the terminal can use a target radio interface to establish the connection with the target access network while its communication is still on-going via the serving radio interface. If the cell overlap is large enough and the handover instant is correctly decided, the terminal will finalize its handover procedure via the new interface before moving out of the old radio coverage. The session continuity is thus seamlessly maintained. When the cell overlap is too small, the mobile terminal's velocity is very high or both, even if the terminal initiates a handover as soon as it enters the overlap region, the blackout period will not be negligible. Note again that we do not aim at modifying the existing vertical handover decision algorithm. Once the handover is initiated,

we decide whether or not to trigger a pre-buffering. In fact, the pre-buffering decision is based on the estimate of the blackout period. That is:

$$\tau_{bo}[i] = \tau_{ho} - (i - t_h)\Delta T - T_{bmo}[i] \quad \text{for} \quad \forall i \geq t_h \tag{4.5}$$

where t_h is the handover initiation instant, T_{bmo} is the remaining time before moving-out of the serving cell and $\tau_{ho} - (i - t_h)\Delta T$ is the remaining time before the handover completion. The details of pre-buffering scheme are as follows:

Step 1 - Calculate the time before moving-out of the serving cell: When the vertical handover is initiated, MN starts predicting T_{bmo}. Based on the RSS values from the serving BS, the MN predicts the instant where the RSS drops below θ_{border}, the received signal strength at the cell border. The later is the minimum RSS value where the radio link can be hold. Accordingly, the value of T_{bmo} is predicted. The details of T_{bmo} prediction are presented in Section 4.4.3.

Step 2 - Pre-buffering triggering condition tuning: Basically, the pre-buffering operation is triggered if $\tau_{bo}[t_h] > \frac{b[t_h]}{R_r}$. It means that we need pre-buffering if the current buffered media is not enough to play out during the handover blackout. In order to take into account the prediction overestimation, we suggest adding a hysteresis margin γ in the triggering condition like in the case of the horizontal handover. That is to say, if $(\tau_{bo}[t_h] + \gamma) > \frac{b[t_h]}{R_r}$, initiate the pre-buffering process with the required accelerated data rate ΔR which is given by:

$$\Delta R = \frac{\min\{b_{max}, R_r\tau_{bo}[t_h]\} - b[t_h]}{T_{bmo}[t_h] - T_{con}} \tag{4.6}$$

The pre-buffering procedure can be successfully achieved if $\Delta R \leq (R_{max}[t_h] - R_r)$. The number of supplementary connections is:

$$n = \lceil \frac{\Delta R}{R_r} \rceil < \lceil \frac{R_{max}[t_h]}{R_r} - 1 \rceil \tag{4.7}$$

These connections will endure until the handover occurs or the buffer is fulfilled.

Step 3 - Pre-buffering recovery: To overcome the uncertainties of the user movement pace change, we suggest monitoring the following condition during the pre-buffering operation at each instant $j > t_h$:

$$\min\{b_{max}, R_r\tau_{bo}[j]\} - b[j] > \Delta R T_{bmo}[j] \tag{4.8}$$

If (4.8) occurs, new supplementary UDP connections of increased streaming speed $\Delta R'$ are required to be established if $T_{bmo}[j] > T_{con}$. In addition to n supplementary connections, we need to initiate $n' = \lceil \frac{\Delta R'}{R_r} \rceil$ new ones to boost the pre-buffering, where

$$\Delta R' = \frac{\min\{b_{max}, R_r\tau_{bo}[j]\} - (b[j] + nR_rT_{con})}{T_{bmo}[j] - T_{con}} - nR_r \tag{4.9}$$

4.4 Handover prediction

The handover prediction plays a crucial role in the previously proposed pre-buffering scheme. Two important parameters are the time before handover and the time before moving-out of the serving cell. To predict these two parameters, we employ the first-order Grey Model GM(1,1) filter. In fact, the grey prediction establishes a grey model extending from the past information to the future one based on the past and present known information (the measured RSS values). The GM method is chosen because of its efficiency and its low computational power [135].

4.4.1 Overview of GM(1,1)

The main feature of GM(1,1) is to build up a first-order ordinary differential equation from discrete sequence of measured RSS values. The RSS value at a distance d from an AP is computed using the COST-231 model [136]: $P_r[dB] = K_1 - K_2 log(d) + X_\sigma$ where K_1 depends on transmission and reception antenna gains, the antenna height, the center frequency and the signal's wavelength, and K_2 represents the path loss factor. X_σ is an independently and identically distributed zero mean stationary Gaussian random process modeling shadow fading.

Given a set of RSS from one visible AP at the terminal $X^{(0)} = \{X^{(0)}[1], X^{(0)}[2], \ldots, X^{(0)}[n]\}$ where $X^{(0)}[i]$ is the RSS at discrete time i, the accumulated generating sequence data $X^{(1)} = \{X^{(1)}[1], X^{(1)}[2], \ldots, X^{(1)}[n]\}$ are computed as $X^{(1)}[k] = \sum_{i=1}^{k} X^{(0)}[i]$. The first-order ordinary differential grey model is formulated as [137]:

$$\frac{dX^{(1)}[k]}{dk} + aX^{(1)}[k] = b \tag{4.10}$$

where $[a, b]$ are the parameters to be determined. By using the Least Square Error Method, $[a \quad b]^T = (B^T B)^{-1} B^T Y_n$ where

$$B = \begin{bmatrix} -0.5(X^{(1)}[0] + X^{(1)}[1]) & 1 \\ -0.5(X^{(1)}[1] + X^{(1)}[2]) & 1 \\ \vdots & \vdots \\ -0.5(X^{(1)}[n-1] + X^{(1)}[n]) & 1 \end{bmatrix}$$

and $Y_n = [X^{(0)}[2], X^{(0)}[3], \ldots, X^{(0)}[n]]^T$. The filtered and predicted RSS $\hat{X}^{(0)}[k]$ values are therefore given by:

$$\hat{X}^{(0)}[k+1] = (X^{(0)}[1] - \frac{b}{a})(1 - e^a)e^{-ak} \quad \text{for} \quad k \geq 1 \tag{4.11}$$

4.4.2 Time before handover prediction

The time before handover is a key indicator for an adaptive buffering streaming to overcome the blackout period during the horizontal or vertical handover using one radio interface. For example, the horizontal handover between two WLAN APs is initiated based on the RSS values. During the communication with the serving cell, the MN measures the RSS from both serving and potential neighboring cells. The handover will be triggered if $\hat{P}_s + \delta_h \leq \hat{P}_n$ where $\hat{P}_s$, $\hat{P}_n$ are the filtered RSS values from serving and neighboring cells respectively, and δ_h is a hysteresis margin used to avoid handover ping-pong effects. After the handover initiation, the communication may be interrupted during the blackout period as illustrated in Figure 4.1. If the handover triggering instant (or the remaining time before handover) can be predicted, the pre-buffering mechanism can be effectively handled to achieve a seamless streaming.

When the mobile terminal receives the good signal from a neighboring access node, it employs the GM(1,1) filter to predict the future RSS from serving and potential neighboring access nodes. Let us denote $P_s[i]$ and $P_n[i]$ the RSS values of serving and neighboring access nodes ($i \leq m$ where m is the index of the last RSS measurement sample). While the handover does not occur, compute $\hat{P}_s$ and $\hat{P}_n$, the predicted RSS values of P_s and P_n. We predict the future handover instant $k > m$ where $\hat{P}_s[k] + \delta_h < \hat{P}_n[k]$. Consequently, the time before handover is:

$$T_{bho} = (k - m)\Delta T \tag{4.12}$$

where ΔT is the RSS measurement sampling interval. In the case of the WLAN, for instance, ΔT is equal to the beacon broadcast period of $100ms$.

The handover prediction accuracy depends on the number of actual RSS samples $\{P[m-N], P[m-N+1], \ldots, P[m]\}$ employed as inputs in the GM(1,1) filter. Usually, the input sequence data size N is rather small since only two parameters are needed to be identified in (4.10). However, due to the RSS fluctuation, the longer the finite sequence data size N, the more regular the RSS predicted values $\hat{P}$. As the predicted values $\hat{P}[k]$ $(k > m)$ are computed on the measured values until instant m, the accuracy of $\hat{P}[k]$, hence the accuracy of T_{bho}, is a decreasing function of $(k-m)$. The predicted values T_{bho} are obtained under the assumption that the future RSS evolution would follow the fitting-curve (4.11). In other words, the terminal movement direction and its velocity are assumed to be unchanged from instant $(m-N)$ to instant k. Consequently, we cannot predict the time before handover too much in advance.

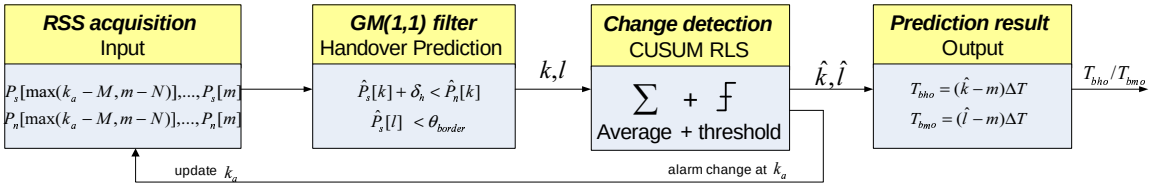

Figure 4.2: Handover prediction scheme

It is interesting to have a rough idea about the value of T_{bho} beyond which a pre-buffering will be triggered. From (4.2), the handover instant is required to be predicted at least T_{ad} seconds in advance. From (4.1), if we replace the increased data speed by $\Delta R = R_r$, we have $T_{ad} \leq T_{con} + \tau_{bo}$. Connection establishment duration T_{con} is indeed equal to several Round-Trip-Time (RTT) for exchanging signaling control messages between the client and the server. The value of T_{con} is around hundreds of milliseconds. Handover blackout τ_{bo} between two WLAN APs may be up to few seconds. Therefore, our prediction scheme is required to be able to forecast the handover instant a few seconds in advance.

As mentioned earlier, the handover prediction accuracy depends on the regularity of the RSS evolution. An abrupt change of the user's movement direction, the user's velocity, or both may cause errors in the handover prediction. The error due to a movement change will be completely removed after N measurement samples. In order to mitigate very soon the movement change effect, we suggest using a change detection mechanism in the handover prediction scheme (see Figure 4.2). The basic idea of the change detection mechanism is to detect as soon as possible a movement direction change, a brutal velocity change or both in order to adjust the GM(1,1) filter inputs.

A change in the movement direction or in the velocity does not induce an abrupt change in the RSS evolution. In fact, the latter is a function of the distance between the mobile user and the BS. Hence, it is challenging to detect immediately such a change. As the RSS sequence does not contain any explicit deterministic component, we cannot detect a change in the RSS evolution which is subsequently caused by a change in the movement pace. That explains why we employ a change detection after the GM(1,1) filter application. The change detection mechanism observes the evolution of the predicted handover instant $y_t = k$ which is a deterministic component. An abrupt fluctuation of y_t implies a change in the user movement pace.

The most successful algorithm of the sequential change detection is Page's Cumulative Sum (CUSUM) [138]. We employ the CUSUM Recursive Least Square (RLS) filter, which combines adaptive filters with the CUSUM test as a change detector, to detect changes in the sequence y_t. The CUSUM RLS

algorithm is described as follows:

$$\hat{\theta}_t = \lambda\hat{\theta}_{t-1} + (1-\lambda)y_t \quad (4.13)$$

$$\varepsilon_t = y_t - \hat{\theta}_{t-1} \quad (4.14)$$

$$g_t^1 = \max(g_{t-1}^1 + \varepsilon_t - \nu, 0) \quad (4.15)$$

$$g_t^2 = \max(g_{t-1}^2 - \varepsilon_t - \nu, 0) \quad (4.16)$$

The formula (4.13) yields the RLS estimate where λ is referred to as a forgetting factor. A large value of the forgetting factor ($\lambda = 0.9$) is intensionally selected to reduce the estimation noise. In (4.14), ε_t is the prediction error used as a distance measure for detecting a change. The test statistic g_t^1 (g_t^2) sums up its input ε_t, with the idea to trigger an alarm when the sum exceeds a threshold h. If the residual ε_t is a white noise sequence (no change occurs), the test statistic will drift away similar to a random walk. A subtraction of a small drift ν and a resetting to zero operation (once g_t becomes negative) are used to prevent false alarms and to prevent a long detection delay after a change. Since a change can result in an increase or a decrease of the handover instant y_t, we use two parallel tests g_t^1 and g_t^2 for detecting the increase or decrease in its mean value. When a change is detected at instant $k_a = t$ (i.e., $g_t^1 > h$ or $g_t^2 > h$), reset $g_t^1 = 0$, $g_t^2 = 0$, and $\hat{\theta}_t = y_t$.

The two design parameters of the CUSUM test are drift ν and threshold h. According to [138], drift ν is usually set to one half of the expected change $\nu = 0.5|\theta_{k_a} - \theta_{k_a-1}|$ whereas threshold h depends on the signal sequence characteristics and is usually determined by experience or by using a training set of data [139]. In the literature, different methods for choosing h based on a false alarm probability and a suitable delay for detecting a change (change detection delay) were proposed in [138] [140]. However, the formulation is asymptotic and difficult to apply in practice.

The outputs of the change detection algorithm are the estimator of the handover instant $\hat{k} = [\hat{\theta}_t]$ (with $[x]$ denoting the nearest integer of x) and the change instant k_a. Accordingly, we have $T_{bho} = (\hat{k} - m)\Delta T$. At the beginning of the prediction scheme, k_a is set to 0. Once a change is detected, a new value of k_a is thus updated and reported in the RSS input sequences. If no change is detected, the GM(1,1) filter uses the last N RSS values as inputs. Otherwise, recent RSS values associated with a new movement pace is used rather than all N last RSS values. Once a change is detected at instant k_a, such change probably occurs at instant $(k_a - M)$ where M is a change detection delay. In general, M is less than one half of input data sequence size $(M \leq N/2)$.

The choice of input size N is a trade-off between the prediction regularity and the change detection delay. If N is large, the change detection delay is too long. Otherwise, a small value of N implies significant fluctuations of T_{bho}. As the radio signal strength is spatially correlated [141], it is preferable to use the RSS sequence measured within a correlation distance. In urban environment, a correlation distance of $20m$ is widely adopted [142]. Hence, the value of N and M are determined from the measurement interval and the supportable movement velocity.

4.4.3 Time before moving out of the serving cell prediction

Contrary to the prediction of T_{bho}, the prediction of T_{bmo} is based only on the RSS of serving access node $P_s[i]$. The main idea is to forecast the future instant l when the user moves out of the serving cell (hereafter the moving-out instant). At this instant, $\hat{P}_s[l]$ becomes inferior to θ_{border} (i.e., $\hat{P}_s[l] < \theta_{border}$). It is the signal strength threshold below which the communication could not be maintained.

The prediction scheme for T_{bmo} is also illustrated in Fig. 4.2. The prediction scheme uses the available data sequence $\{P_s[m-N], P_s[m-N+1], \ldots, P[m]\}$. Similarly to the T_{bho} prediction scheme described above, in order to mitigate the effect due to movement change, we use a CUSUM RLS filter

to detect the movement change and to smooth the predicted moving-out instant. The change detection mechanism monitors the evolution of the predicted moving-out instant $y_t = l$. If a movement change is detected at k_a, the GM(1,1) filter input will be updated accordingly. The choice of data sequence size N and CUSUM design parameters (h, ν) are identical to those used in the T_{bho} prediction scheme. The predicted moving-out instant $\hat{l} = [\hat{\theta}_t]$ ($\hat{\theta}_t$ is the filtered value of predicted moving-out instant y_t) is used to compute T_{bmo}:

$$T_{bmo} = (\hat{l} - m)\Delta T \tag{4.17}$$

where m is the last measurement instant and ΔT is the time length between two consecutive RSS samples used in the prediction scheme.

4.5 Performance evaluation

The goal of this section is to validate and evaluate the effectiveness of the proposed handover prediction scheme by investigating different user movement scenarios. To this end, we focus only on the proposed pre-buffering policy involving the handover (rule 3 in Section 4.3) by assuming that the buffer is not full before the handover.

We simulate a user movement between two neighboring APs. For pedestrian users with the maximum velocity of $2m/s$ and measurement interval of $100ms$, we specify $N = 100$ (i.e., the RSS values measured within a correlation distance of $20m$). After having examined different movement change direction scenarios, we see that $\nu = 15$ and $h = 2000$ are a suitable choice for the change detection design parameters. By doing so, a change is detected after a maximum change detection delay of 40 samples. As a results, we set $M = 40$ throughout the simulations.

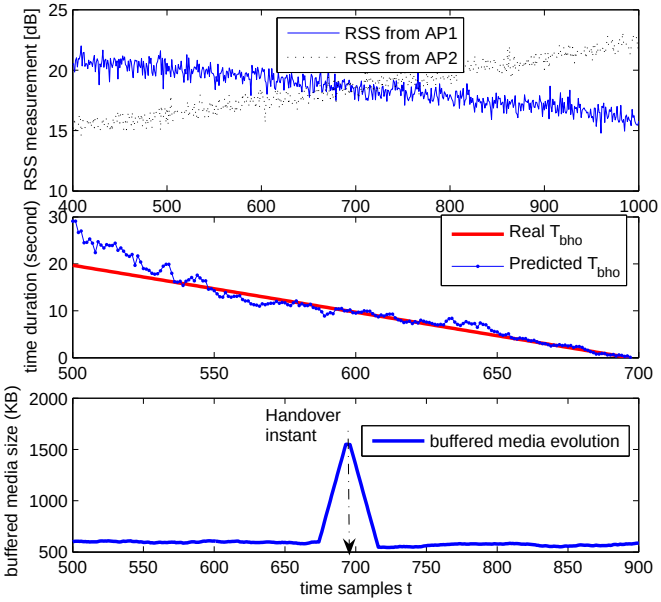

Figure 4.3: T_{bho} prediction and pre-buffering management

First, the user moves from AP1 to AP2 without change of direction and velocity. The RSS evolution from the two APs and the predicted T_{bho} result according to $\delta_h = 0$ are depicted in Figure 4.3. We can observe that the prediction result is more fluctuating when the user is far away from the handover instant and well coincided with real values (the real T_{bho} values are deduced from the handover instant)

when the user is close to the handover instant. The result shows that the proposed handover prediction makes it possible to predict the handover more than 10s in advance. We repeat this simulation many times and observe the fluctuation of T_{bho} at 10s and 5s before the handover instant. The confidence interval [143] of T_{bho} is computed as:

$$P(|T_{bho} - \overline{T}_{bho}| \leq \frac{t_{n,\alpha/2}\sigma_m}{\sqrt{S}}) = 1-\alpha \text{ where } n = S-1 \tag{4.18}$$

where S is the number of simulations, $\overline{T}_{bho}$ is the mean of T_{bho} over S simulations and σ_m is its standard deviation. In the above equation, $(1-\alpha)$ is the confidence level and $t_{n;\alpha/2}$ is the percentage point of Student's distribution such that $P(|t_n| > t_{n;\alpha/2}) = \alpha$. It means that we are $(1-\alpha)100\%$ sure that the predicted time before handover varies from $T_{lower} = (\overline{T}_{bho} - \frac{\sigma_m t_{n,\alpha/2}}{\sqrt{S}})$ to $T_{upper} = (\overline{T}_{bho} + \frac{\sigma_m t_{n,\alpha/2}}{\sqrt{S}})$. According to our simulation results, with the confidence level of 90%, the confidence interval of T_{bho} at 10s and 5s before handover (through 100 simulation repetitions) is $(10.57 \pm 0.33)s$ and $(5.24 \pm 0.19)s$ respectively.

The results confirm the reliability of our predicted T_{bho} and the overestimation tendency of the predicted values compared to real values. The later is due to the GM(1,1) filter nature and has been well-known recorded in the literature [131] [132]. The results show the necessity of adding a hysteresis γ in the pre-buffering triggering condition (4.2) to avoid the unexpected fluctuation and overestimation. In this case, γ of around 0.4s should be appropriate. We assume that $\tau_{bo} = \tau_{ho} = 2s$, $R_r = 500KB/s$, $T_{con} = 0.1s$, $\gamma = 0.4s$, the buffered media before the handover is maintained around $600KB$ and $b_{max} \geq R_r\tau_{bo}$. By applying the proposed buffering scheme, the evolution of the media in the buffer is illustrated in the bottom of Figure 4.3. The results show that the streaming delivery is seamlessly achieved throughout the handover interruption period.

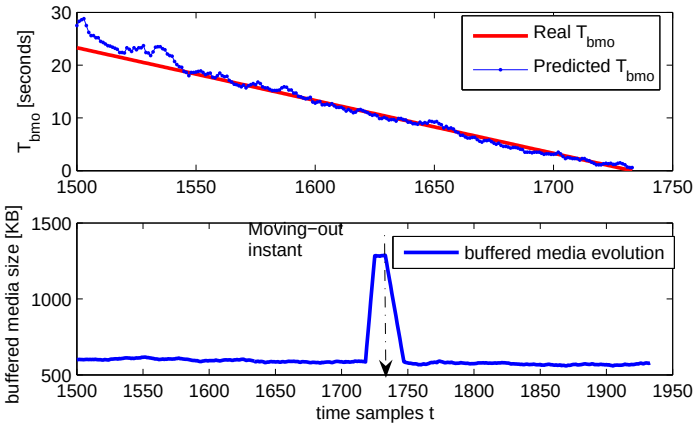

Figure 4.4: T_{bmo} prediction and pre-buffering management

In the same simulation, we use now the RSS sequence from AP1 to test the T_{bmo} prediction scheme. Assuming that $\theta_{border} = 10dB$, the predicted T_{bmo} is shown in Figure 4.4 (i.e., the moving-out instant is $t_l = 1733$). We observe that the predicted T_{bmo} is very close to the real T_{bmo}. With the confidence level of 90%, the confidence interval of T_{bmo} at 10s and 5s before the moving-out instant is $(10.88 \pm 0.37)s$ and $(5.33 \pm 0.17)s$ respectively. The results confirm the effectiveness and reliability of the T_{bmo} prediction scheme. They also stipulate the introduction of hysteresis margin γ in the pre-buffering triggering condition. We assume that $\gamma = 0.3s$, $\tau_{ho} = 3s$, handover instant $t_h = 1717$, $R_r = 500KB/s$, $T_{con} = 0.1s$ and the buffered media before the handover is maintained around $600KB$. By applying the proposed pre-buffering scheme, the evolution of the buffered media is represented in Figure 4.4. The result

confirms the effectiveness of the proposed handover prediction and the proposed re-buffering policy. In the following, we will focus on the T_{bho} prediction scheme since the same results can be found in the case of the T_{bmo} prediction scheme.

Secondly, the user follows the same trajectory described above but his velocity is varied. The change of velocity influences slightly the regular evolution of the RSS sequence and thereby the predicted T_{bho} as illustrated in Figure 4.5. We see a rapid decrease of T_{bho} around $t = 400$. It is due to an abrupt increase of the terminal's velocity. Otherwise, minor changes to the terminal's velocity do not affect the prediction of T_{bho}. The result shows that the prediction scheme performs well even in the case of the velocity fluctuation and that it provides a reliable T_{bho} for the pre-buffering policy. The evolution of the buffered media in Figure 4.5 is obtained with the same simulation parameters as described in the first simulation. We can conclude that our solution is robust against the velocity fluctuation.

Thirdly, we investigate the performance of our prediction scheme against the movement direction change. Two types of movement change can be distinguished: from very-soon-handover to no-handover situation and from no-handover to very-soon-handover one. The former is not a critical situation since the predicted handover, which does not occur due to the movement change, may induce a pre-buffering operation but does not affect the performance of the streaming application. On the contrary, the latter is critical since T_{bho} drops from an infinity value to a small value right after the movement change. We select this latter case to test our handover prediction scheme. Basically, the handover prediction should be able to detect such a change with an acceptable delay to adjust the T_{bho} and therefore to initiate correctly the pre-buffering operation. In this simulation, a movement direction change happens at instant $t = 1000$ and the handover occurs at instant $t = 1150$. The terminal's velocity $v = 1m/s$ is kept unchanged.

We compare our proposed prediction scheme with two others: one without change detection mechanism and one with capability to detect immediately the movement change. Such an immediate detec-

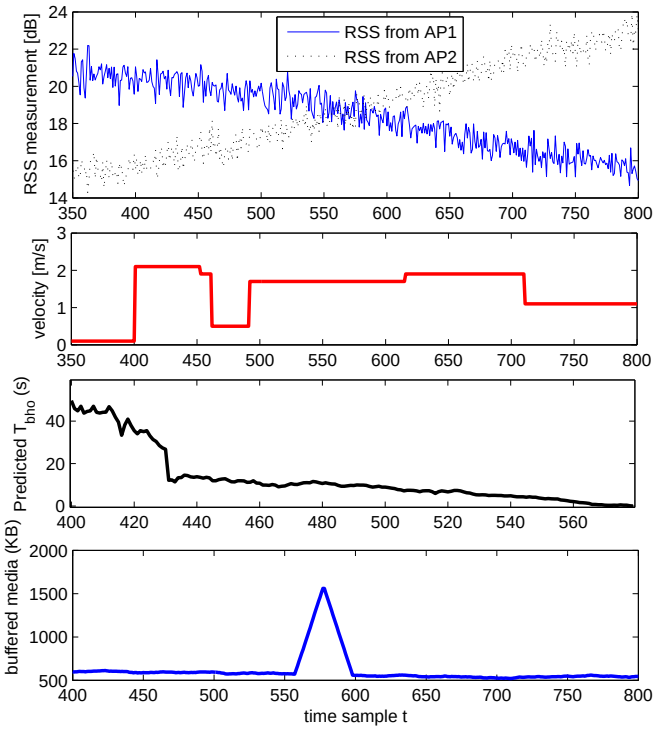

Figure 4.5: Performance evaluation for the variable movement velocity case

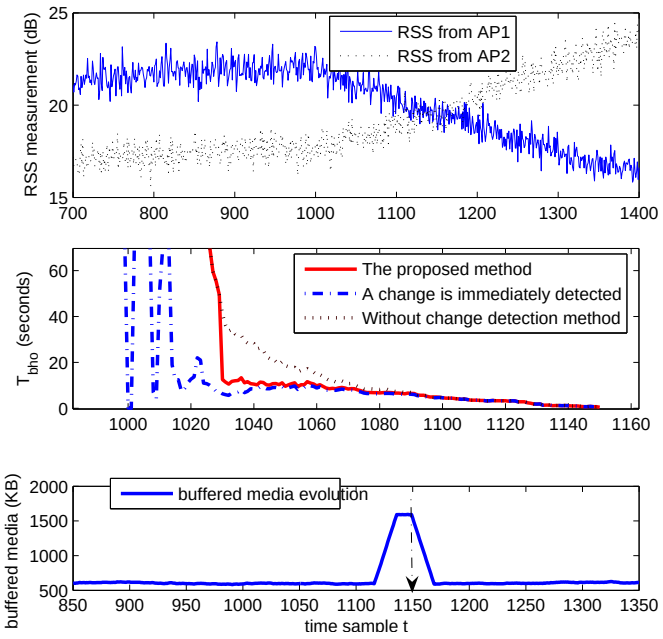

Figure 4.6: Performance evaluation for movement direction change

tion is impossible in reality. The time before handover prediction results from the three above schemes are presented in Figure 4.6. We see that the proposed prediction scheme yields a better result compared to the two others. If a change detection is not used, the error of T_{bho} is significant and such an error can be mitigated only after N time samples. If a movement change is immediately detected, the predicted values T_{bho} right after the change instant fluctuates due to the small size of the input RSS set used in the GM(1,1) filter. We can see that the suitability of the pre-buffering policy since the streaming is seamlessly achieved. The results prove the rationality of our proposed handover prediction scheme and the efficiency of the proposed pre-buffering scheme.

Real streaming experiments implementing our proposed handover prediction and streaming buffering policies have been tested in our lab[1]. The VideoLAN code has been modified to launch automatically required supplementary connections with the server to boost the download of the media content. The results show that the streaming is maintained seamlessly while a mobile terminal (laptop) is roaming among the WiFi APs. This result is a part of my colleague's work and is not presented here.

4.6 Summary

In this chapter, we presented a handover prediction scheme based on the received signal strength level using Grey-Model filter. We proposed to use the CUSUM statistics to detect the movement pace changes inside the handover prediction. More precisely, based only on the signal strength measurement, we estimate the remaining time before handover and the remaining time before moving-out of the serving cell. We have verified that our scheme is efficient and pragmatic to predict the handover. The predicted information is crucial for the handover preparation management. We have specified a terminal-controlled adaptive pre-buffering management using the handover prediction results at the terminal side. The solution aims at delivering seamlessly multimedia services to mobile users during both horizontal and vertical handovers. The solution is envisioned for handovers using reconfigurable

[1]The experiment and test-bed were set up and realized by my colleague Mehdi Nafa.

interface terminals and multi-interface terminals. It does not require modifications to current network infrastructures and the mobility management procedure. The proposed pre-buffering policies serve as a complement to the terminal-controlled mobility management proposed in Chapter 3 to offer mobile users the seamless streaming without any break.

One more time, we explore how the terminal can assist users to manage and to enhance the service quality and mobility performance. In the second part of this thesis, we will explore the control of the network side to improve the mobility management.

~~ △♡△ ~~

Part II
Network-Controlled Approach

In the first part of this thesis, we explored the user-controlled mobility management approach over heterogeneous wireless environments. Such approach provides users with facilities to personalize and customize their mobility services. However, the terminal might not have the possibility or capability to gather all information related to neighboring cells as well as available resources of the surrounding access networks to effectively handle the mobility management. In any case, the control from the network side is still primary from the operators' perspective. The concept of network-controlled handovers, commonly used in cellular technologies, becomes essential for distributing radio resources optimally and fairly, and for allowing each mobile user to take full advantage of the multi-access multi-technology capability.

Chapter 5

Interworking Architecture Design

Interworking: The act of working in together; interweaving
Webster's Revised Unabridged Dictionary

In addition to the very loose-coupling architecture described in Chapter 3, we propose in this chapter *i)* an interworking architecture between UMTS and WiMAX systems, and *ii)* an interworking and roaming architecture between 3GPP and non-3GPP systems using an RII functional entity. The integration of the RII into the on-going 3GPP LTE architecture to support interworking and roaming among different access systems, considering different service level agreements among operators, is presented. The signaling message sequence charts of different handover scenarios are detailed.

5.1 Introduction

A research trend which aims to integrate 3GPP/UMTS and WLAN to benefit the high data rate and low cost of WLAN has much attracted research community and standardization bodies for the last few years. Different interworking approaches have been summarized and discussed in Chapter 1. Initial study about interworking between 3GPP system and WiMAX has been achieved within WiMAX Forum [144] and for the moment it does not consider the real inter-system handover: it is based on reuse of the 3GPP-WLAN interworking model proposed by 3GPP. Accordingly, in Section 5.2 of this chapter, we propose a possible UMTS-WiMAX interworking architecture based on the 3GPP standards and present the corresponding handover procedures.

In heterogeneous networks, interworking and roaming can encompass a large number of possible scenarios and network configurations. In general, a roaming agreement is required to allow subscribers of one operator to gain access to networks of other operators. The agreement deals with technical and commercial aspects related to the roaming procedure, particularly how costs and earnings are divided. On the road to design the roaming between different networks, a third party roaming intermediary has been introduced [145–148]. The intermediary can enable the roaming between two networks without any direct agreements between operators of these networks. While the number of hot-spot operators has been rapidly increased, the roaming capability without direct agreement becomes crucial. Besides the roaming intermediaries among hot-spot operators, the roaming broker facilitating the roaming between mobile and hot-spot operators is also proposed [145]. It is responsible for providing the information of user's home services to the visited domain, taking care of the evolving relationship and determining signaling and accounting procedures. Unlike a broker, clearinghouse [147] does not resale the

WLAN access, instead provides a trusted intermediary for implementing roaming agreements. Most of the current solutions are proprietary ones. The mobility between the home and visited networks has not been addressed in any of the above solutions. In other words, they could not maintain on-going sessions while users roam between home and visited networks. In our work, we extend the roaming intermediary concept to deal with not only the roaming but also the interworking issue, called Roaming Interworking Intermediary (RII). Importantly, the proposed RII will enable the secure handover across different access systems and different operator domains without service interruption. Such a solution is presented in Section 5.3 of this chapter.

5.2 UMTS-WiMAX interworking architecture

One should take into account the differences between UMTS-WLAN interworking and UMTS-WiMAX interworking when designing the mobility management. The WLAN in hot-spot areas forms the micro-cells within the UMTS macro-cells. The mobility between UMTS and WLAN can be referred to fully overlapping handover. The long delay for switching from UMTS to WLAN connection does not much affect the performance. When the mobile terminal is connected to WLAN, it can maintain simultaneously the PDP context of UMTS so that it can reconnect immediately to UMTS without need of PDP context re-activation. On the contrary, the mobility between UMTS and WiMAX is referred to partially overlapping handover since the coverage of UMTS and WiMAX is approximately the same.

5.2.1 Proposed interworking architecture

5.2.1.1 Architecture description

The proposed architecture for UMTS-WiMAX interworking is depicted in Figure 5.1. The UE is an SDR-enabled mobile node that can communicate with both UMTS and WiMAX networks. SDR devices can reprogram to operate in different radio interface standards, allowing the efficient use of radio spectrum and power. They make it possible both to improve performance and customize radio devices to individual needs [149] [150]. The market for SDR-enabled handsets is expected to grow significantly over the next few years (i.e., 200M units by 2014) [151]. However, it can connect to only one access network at a time and the handover between UMTS-WiMAX is not easy to handle seamlessly. That represents the motivation for our work in this chapter.

The WiMAX Access Network provides the WiMAX access services. The mobility inside WiMAX network is managed by the WiMAX HA located between the Access Service Network Gateway (ASN GW) and the WAG. The WiMAX HA is not necessarily included in the 3GPP core network to keep its independence from the 3GPP system. The FA located in the ASN GW is the local FA in the interworking architecture. The WiMAX AN is connected to the UMTS network via a WAG and to the 3GPP AAA server for the WiMAX authentication process. The WAG is a gateway through which the data from/to the WiMAX AN is routed to provide the UE with 3GPP services. The WAG's functionalities include enforcing routing of packets through the PDG, performing accounting information and filtering out packets. The main functions of the PDG are to route the packets received from/sent to the PDN to/from the UE and to perform an FA. The mobility within the UMTS network is managed by its own mobility mechanism and the FA functions implemented in the GGSN. In order to enable the vertical handover between these two technologies, the HA is placed in the PDN and manages FAs of both WiMAX and UMTS networks.

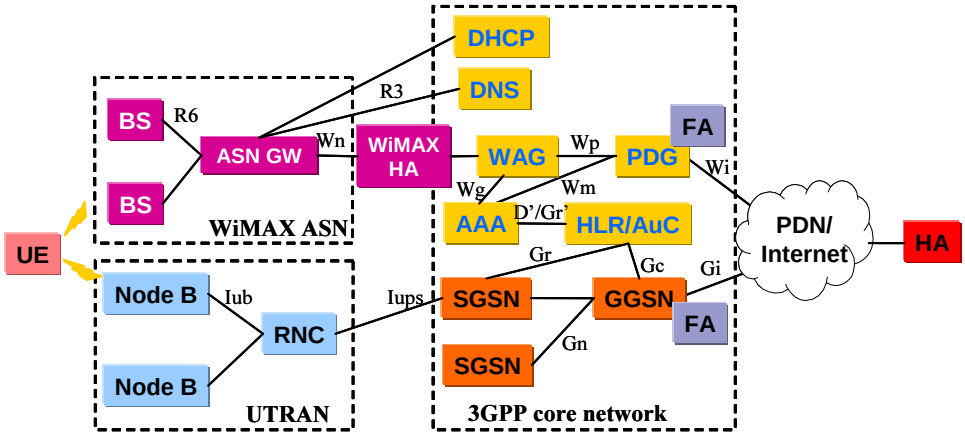

Figure 5.1: UMTS-WiMAX Mobile IP based interworking architecture

5.2.1.2 IP address management

In the WiMAX network, each time the UE changes its ASN GW, it will obtain a new local IP address through the Dynamic Host Configuration Protocol (DHCP) server. The ASN GW can learn this new local IP address and ask the DHCP server the WiMAX HA's address since it plays the role of the DHCP relay agent. The ASN GW informs the serving BS the UE's new local IP address and sends the MIP registration to the WiMAX HA. A generic IP-in-IP tunnel such as Generic Routing Encapsulation (GRE) [152] [153] may be used to transport IP packets between the WiMAX HA and the FA.

Each time the UE switches to a UMTS cell, it will initiate the PDP context activation procedure. No IP address is allocated to the UE at the PDP context activation. The remote address provided by the HA (or an external entity in the PDN) will be kept unchanged and will be informed to the GGSN via the PDP context activation. The remote IP address is a global home address. It may be a static address or a dynamic address attributed when the UE first time connects to the network, discovers and registers with the HA. The PDG/GGSN is responsible for relaying the allocated remote IP address to the UE.

5.2.2 Handover sequence chart

The mobility between two access networks is achieved by the MIP mechanism at the network layer. To reduce the interruption time during handover, before leaving the serving network, the UE prepares a new attachment in the target network. To reduce the packet loss, the old FA notifies the HA about the handover so that the HA can buffer inbound packets and forward them to the UE as soon as the HA receives the MIP registration update from the UE.

5.2.2.1 Handover from WiMAX access network to UTRAN

The handover scheme from a WiMAX cell to a UMTS cell is depicted in Figure 5.2.

1. During the communication, the WiMAX BS sends periodically a topology advertisement message to inform the UE of neighboring WiMAX BSs and Node Bs. Alternatively, the UE can scan different channels to discover the neighboring access networks.

2. The UE performs the synchronization and measurement. The WiMAX-UMTS inter-system measurement is fully addressed in the Chapter 6.

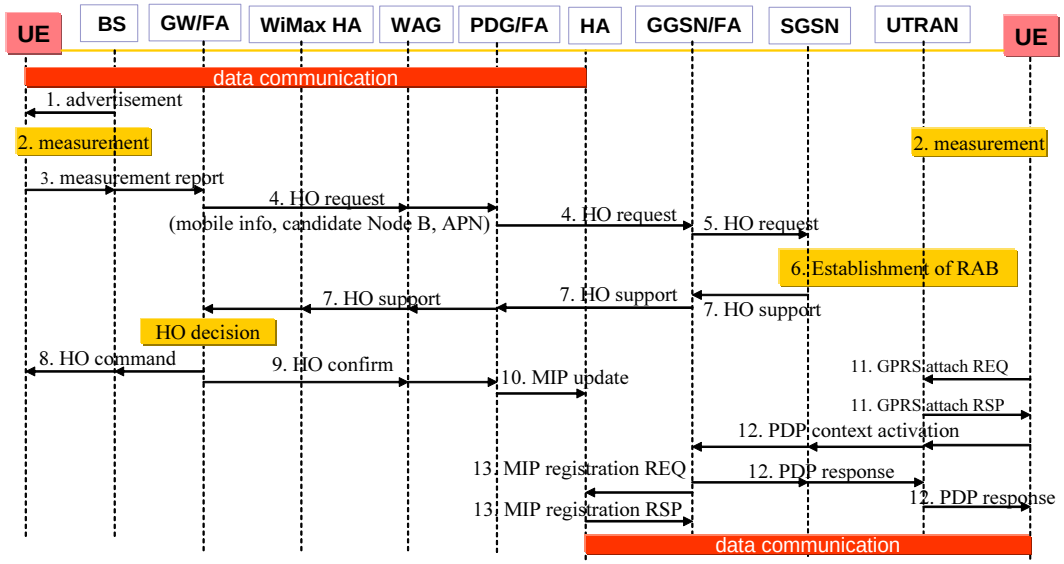

Figure 5.2: Handover scheme from WiMAX to UTRAN

3. After the measurement step, the UE sends the measurement report to the WiMAX BS. The report contains the signal quality level of each candidate UMTS cell.

4. If the triggering conditions for vertical handovers to the UMTS are met, the WiMAX BS initiates the handover procedure by notifying the potential target UMTS via handover (HO) request message routed through the core network. The message is sent to the PDG. The latter performs a DNS request to know the addresses of GGSNs which serve the current UE's Access Point Name (APN). The PDG selects one GGSN and sends the HO request to it. If the PDG does not receive any response from the GGSN for a certain time, it will select another GGSN in the list and resend the HO request message.

5. The GGSN sends the HO request message to the SGSNs who serve indicated Node Bs. In order to retrieve the address of the SGSN that serves a specific Node B, the Domain Name Server (DNS) server or the HLR is assumed to store this routing information.

6. The target RAN establishes a radio bearer resource for the UE.

7. The Node Bs which support the handover with the required QoS send a HO support message to the ASN GW.

8. Upon receiving the HO support messages, the ASN GW selects the best target UMTS cell and returns a HO command to the UE. This message includes the recommended target Node B and all the required information for setting up a new connection. The above exchange may require a large amount of information and add more latency to handover, it is therefore preferable to use a pre-configuration mechanism (a reference number to a pre-defined set of UMTS Terrestrial Radio Access (UTRA) parameters) [154]. This provides a temporary connection during which the UE can reconfigure the connection to the suitable one.

9. The ASN GW sends the handover confirmation including the target Node B identifier to the PDG/FA. The allocated resources in the WiMAX network will be released

10. Upon reception of the handover confirmation message, the PDG/FA sends a MIP update message to the HA to notify the UE's movement. The HA stops sending packets to the UE via this

PDG/FA and buffers inbound packets until it receives the MIP registration update from the target UMTS network.

11. The UE performs the GPRS attachment procedure to the UTRAN. The GPRS attachment procedure consists of accessing to the SGSN, authenticating with the AAA server and updating the location.

12. The UE starts the PDP context activation through which the UE informs its remote IP address to the GGSN.

13. After the connection is established between a new GGSN/FA and the UE, the GGSN/FA will perform the MIP registration with the HA including the UE's remote IP address and its CoA. The data will be transmitted to the UE via the new Node B and the handover procedure is completed.

5.2.2.2 Handover from UTRAN to WiMAX access network

The handover scheme from a UMTS cell to a WiMAX cell is depicted in Figure 5.3.

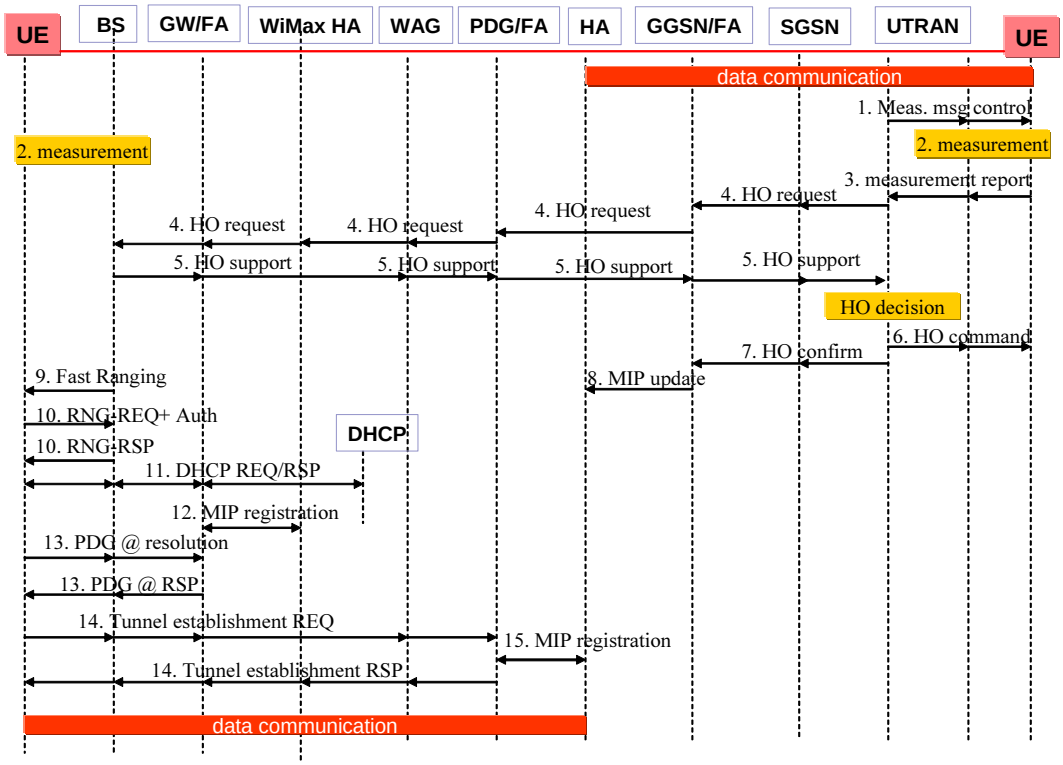

Figure 5.3: Handover scheme from UTRAN to WiMAX

1. During the communication in the UMTS FDD mode, the UTRAN sends the measurement control message to the UE including the neighboring WiMAX cell information, the compressed mode pattern...to help UE to discover neighboring cells.

2. The UE performs the signal strength measurement on the neighboring cells. The details of this inter-system measurement are fully addressed in Chapter 6.

3. After the measurement step, the UE sends the measurement report to the network. The report contains link quality indicators of the neighboring WiMAX BSs.

4. If the vertical handover to WiMAX conditions are met, the RNC initiates the handover procedure by notifying the potential target WiMAX BSs via a HO request message. This message includes the UE's APN, the candidate BS identifiers, the required QoS of UE's current applications. The message is sent to the GGSN. The latter performs the DNS request to learn the addresses of the PDGs which serve the UE's current APN. The GGSN selects one PDG in the result list and forwards the HO request message to the selected PDG. If the GGSN does not receive any response from the PDG after a certain time, it will send the HO request to another PDG in the list. The HO request message is transmitted to potential WiMAX BSs based on the routing information at the PDG.

5. The WiMAX BSs which support the handover with required QoS return a HO support to the RNC.

6. The RNC selects the best target WiMAX BS and sends a HO command to the UE. This message includes all the required information for setting up the connection to the selected target WiMAX BS.

7. Right after that the RNC sends a HO confirmation to the GGSN/FA. The UE disconnects from the UMTS network and starts the connection setup to the target WiMAX BS.

8. Upon receiving the handover confirmation, the GGSN/FA sends a MIP update message to the HA to notify the UE's movement. The HA stops sending packets to the UE via this GGSN/FA and buffers inbound packets until it receives a MIP registration update from the target WiMAX network.

9. Based on the information included in the HO request message, the WiMAX BS can provide a non-contention based initial-ranging opportunity to the UE by placing a fast ranging information element in the UL_MAP [155, 156]. This information will help the UE in the RAN connection setup. If not, the UE must perform the normal ranging procedure which takes more time.

10. The UE initiates the connection setup by exchanging Ranging Request (RNG-REQ)/ Ranging Response (RNG-RSP) with the target WiMAX BS. The details of network entry can be found in [156].

11. In the WiMAX access network, the UE performs a DHCP request to obtain a new local IP address. If IPv6 is used, the local address can be allocated by Stateless Address Autoconfiguration mechanism without DHCP server. During this step, the ASN GW learns the WiMAX HA's address for the MIP registration.

12. The UE performs a MIP registration to associate the UE's local address with its CoA.

13. The UE performs a DNS resolution to learn addresses of PDGs. The UE uses the APN to indicate the network service it wants to access. The DNS request is relayed to the ASN GW which relays the request to the DNS server. The UE selects one suitable PDG among the list of PDGs given in the DNS response. The selected PDG here may be different from the PDG selected by the GGSN during HO request/support step.

14. The UE establishes an end-to-end tunnel with the selected PDG using IKEv2 [25]. The UE informs the PDG about its local and remote IP address. Each time the UE changes its ANS network, it obtains a new local IP address and therefore a new tunnel should be correctly configured. For inter-WiMAX mobility, the time required for setting up a new IPSec tunnel may be too

long that the seamless handover cannot be achieved. To speed up this IPSec tunnel relocation, we suggest using the MOBIKE mechanism [23].

15. The PDG performs the MIP registration with the HA as soon as it is notified the UE's remote IP address. The data will be transmitted to the UE through the WiMAX access network. The handover is completed.

5.3 Interworking & Roaming architecture using RII

5.3.1 Overview of 3GPP LTE architecture

The main objective of 3G LTE architecture is to provide a higher data-rate, lower-latency and packet-optimized system that supports multiple RANs. In order to meet the requirements on high data rates with large transmission bandwidth and flexible spectrum allocation, OFDM and MIMO techniques have been chosen for Evolved-RAN [1]. The logical high level architecture of the evolved 3G is illustrated in Figure 5.4 (the left figure). A detailed architecture is presented in Figure 1.9 in Chapter 1 . The intra-LTE access system mobility is managed by two entities: Mobility Management Entity (MME) and User Plane Entity (UPE). The MME manages user contexts such as permanent and temporary identities, mobility states, location areas and user security parameters. The corresponding 2G/3G MME is SGSN. The UPE manages IP bearer service parameters and routing information. It is also responsible for triggering the paging when downlink data arrive for users. The corresponding 2G/3G UPE is the SGSN or SGSN+GGSN. The 3GPP Anchor is a functional entity that anchors the user plane to support mobility between 2G/3G and LTE access systems. The SAE Anchor manages the user plane to support mobility between 3GPP and non-3GPP access systems. The 3GPP anchor can be co-located with the MME/UPE or SAE Anchor or both.

The current 3G LTE architecture aims at developing an evolved 3GPP radio access and core networks to enhance the system performance. For the inter-system mobility management, the 3GPP cares much about the interworking between 2G/3G and 3G LTE while the interworking between 3GPP and non-3GPP access systems has not been adequately addressed.

5.3.2 Generic roaming & interworking architecture

We introduce a novel RII entity to facilitate the interworking & roaming in multi-operator environment among different access technologies and to enable the inter-system handover. The major difference between existing solutions and our proposed one is that **both roaming and handover aspects are addressed**. Our proposed architecture takes into consideration different contractual relationships between operators. If the operators have close SLAs, the roaming can be done directly between two involved networks. In this case, the RII will ease the roaming management and enhance the service continuity. On the other hand, if the operators have no agreement, the roaming will be handled with help of the RII entity. The interworking & roaming architecture among the 3GPP, WiMAX and WLAN networks is illustrated in Figure 5.4. The presented 3GPP network architecture is the 3GPP LTE [1]. Though our solution is in line with the 3GPP LTE architecture, this does not preclude the implementation of RIIs in the current cellular 2G/3G network systems.

Specifically, in order to support different interworking/roaming scenarios, we introduce three kinds of RII: local RII, core RII and global RII:

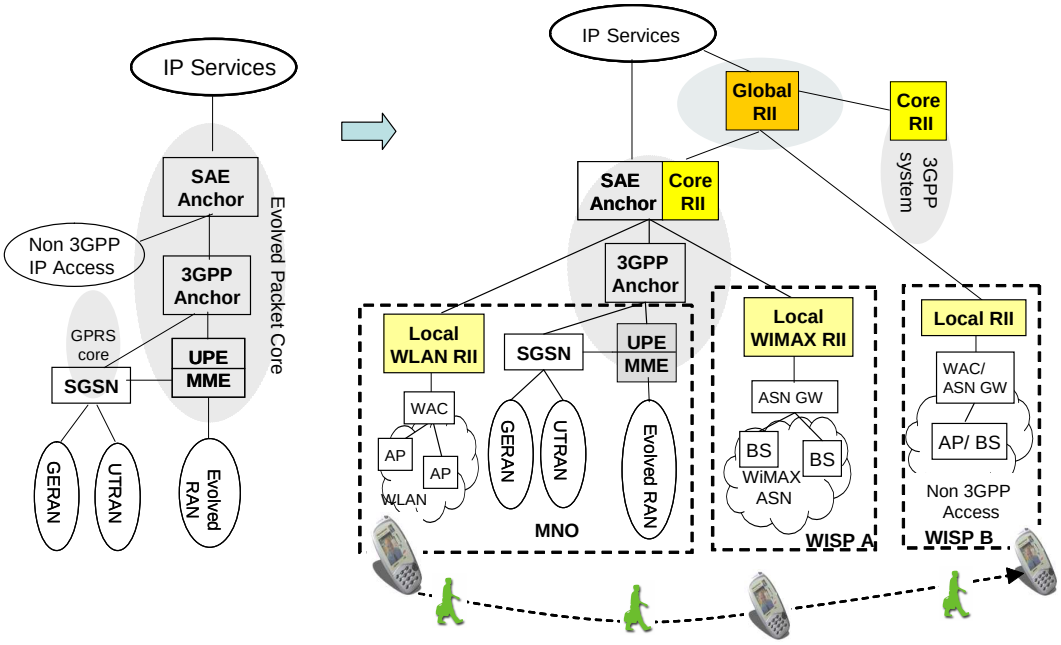

Figure 5.4: The current 3GPP LTE architecture [1] (left) and our proposed architecture enabling seamless and secure interworking/roaming with help of RII entities (right)

- **Local RII** is a control agent and a signaling gateway of a non-3GPP access system in the interworking/roaming architecture. The local RII can be implemented as a separate entity or integrated with an access gateway (e.g., ASN GW or WAC).

- **Core RII** is located in the 3GPP core network and served as a control agent and a signaling gateway between 3GPP and non-3GPP systems. The core RII performs as a local RII in respect of the global RII and as a global RII for different local RIIs under its control.

- **Global RII** is an intermediary for interconnecting access networks of different independent operators. It is an independent entity located outside the 3GPP core network and can be deployed by a Mobile Virtual Network Operator (MVNO). The global RII is a higher-tier RII that interconnects different core RIIs and local RIIs.

This architecture takes into account the different contractual relationships between operators while designing the interworking & roaming among different access systems. The WiMAX/WLAN access network can be owned by a Mobile Network Operator (MNO) or by a Wireless Internet Service Provider (WISP). In a multi-operator environment, different interworking/roaming scenarios can coexist:

- An MNO can deploy WLAN and WiMAX access technologies as an extension of its 3GPP network to best utilize its existing infrastructure and best serve its subscribers. In this case, in order to facilitate the inter-system mobility, the operator can implement a core RII in the 3GPP packet core network and a local RII for each non-3GPP access network. The local RII will be connected directly to the core RII (e.g., interworking between WLAN and 3G within the same MNO in Figure 5.4).

- An MNO can interwork with other WiMAX/WLAN operators if they have a close SLA (e.g., interworking between MNO and WiMAX WISP A in Figure 5.4).

- Importantly, our architecture allows the mobility between operators that have no direct SLA. The inter-system mobility is achieved through the global RII (e.g., interworking between MNO and WISP B or roaming between two 3GPP systems in Figure 5.4). By connecting the local/core RII to the global RII, the corresponding access system can benefit the roaming service to any other operators connected to the global RII.

5.3.3 Functionalities of RII

The RII consists of four different components: Mobility Management (MM), Security Management (SM), Network Selection (NS) and Presence Management (PM). Within an RII entity, the MM is a centralized component that interworks with three other components as illustrated in Figure 5.5a. In the global interworking and roaming architecture, the coordination between two interconnected RIIs is shown in Figure 5.5b. We can distinguish three kinds of information exchanged between RIIs: provisioning information between the NS components, security context between the SM components and all information related to handover and roaming between MM components. The details of such coordination will be presented through the handover procedure. Here we describe the functionalities of these four components.

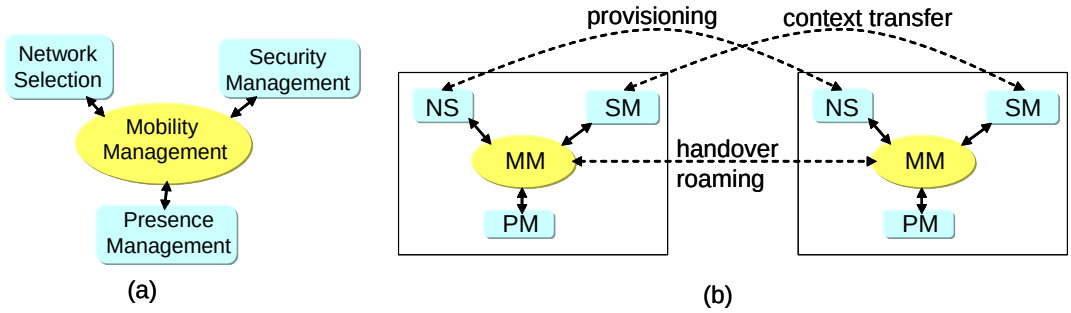

Figure 5.5: (a) Generic RII component interactions, (b) Information flows between two RIIs

- **Mobility Management:** The MM is responsible for preparing the handover by triggering the network selection (i.e., interaction between the MM and the NS), routing the handover preparation request based on the information from the PM component, checking the QoS support in candidate target access networks and assigning the connection setup information for an imminent handover terminal. It makes the handover decision and notifies the handover to the data plane anchor for handover execution preparation. The security context transfer between two involved systems is triggered by the MM (i.e., interaction between the MM and the SM). Once the handover is complete, the MM initiates the presence update (i.e., interaction between the MM and the PM) and the resource release in the old access network.

  The mobility within each access network is managed by its own mobility solution. The inter-system mobility is managed by MIP-based global mobility protocols. The MM in a local RII acts as an FA or a Proxy MIP client whereas the MM in a global RII can implement a HA for its own subscribers. There is no need to implement the FA/HA in the MM component if it has been already deployed in another entity within the same administrative domain. In addition to above functionalities, the MM of a global RII contains the operator database which stores the information like policy, SLA, accounting... of the operators that connect to it.

- **Security Management:** The SM is responsible for handling authentication, authorization and billing issues for roaming users. The SM encompasses the AAA functionalities defined in [14].

In addition, the SM can manage and communicate the user's security context (authentication identity, user identity, certificates, authorization and encryption keys) [157] for the roaming and inter-system handover preparation. It is in charge of authenticating and authorizing users based on subscriber profiles retrieved from the Home Subscriber Server (HSS) or from the security context transferred by the users' serving/home network. The SM in a core RII and in a local/global RII having its own subscribers acts as an AAA server. The SM in a global RII plays the role of a mediator for the roaming contract establishment and for the mobility context transfer to optimize the handover latency caused by the re-authentication procedure.

- **Network Selection:** The NS provides the provisioning information to serving users. Once the MM receives a list of candidate target networks from the UE during the handover preparation, the MM communicates with the NS to eliminate undesirable access networks.
- **Presence Management:** The PM stores and manages the presence information of users which describes how to reach them. The presence information specifies the serving access network, the serving RII and the location of users. Whenever a user roams to a different access network, at the end of the handover procedure, the user's presence information is updated in the RIIs involved. The paging mechanism is included in the PM to wake up standby users. The PM may also provide functionalities of a presence server.

5.3.4 Mobility management

5.3.4.1 Hierarchical mobility management

The proposed architecture allows users to roam among different access systems while maintaining on-going communication sessions. When a UE moves within WLAN/WiMAX systems, the handover will be managed by the WAC/ASN GW and the local RII. Similarly, the handover within 3GPP access systems is managed by the SGSN/MME and the core RII. When the handover occurs between two different access technologies or two different operator domains, the procedure will depend on their contractual relationship. Hence, the hierarchical mobility concept becomes a crucial tool for handover management.

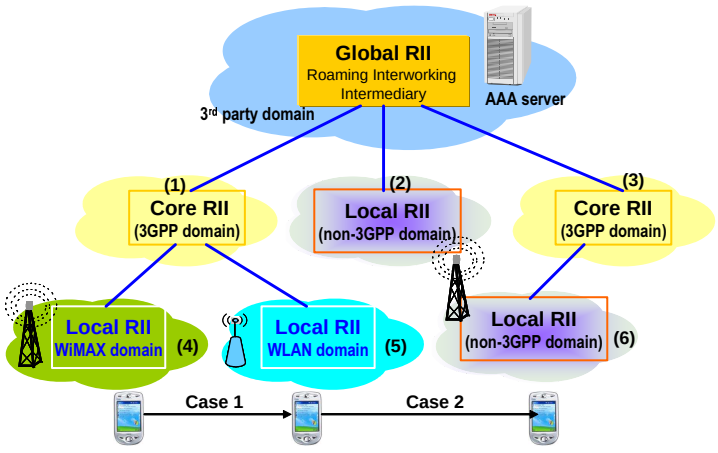

Figure 5.6: Hierarchical mobility management scheme

An example of hierarchical RII-based mobility management is represented in Figure 5.6. If the handover occurs between two indirectly interconnected access networks, handover signalings will go through intermediaries. For example, if the handover occurs between access network (4) and (5) (case

1), the core RII (1) will play the role of a mediator. If the handover occurs between two access systems that have no direct roaming agreement, case 2 in Figure 5.6 for instance, the handover is achieved with help of the global RII. The service continuity during roaming between two operators that have no existing agreement is one of the relevant advantages of our proposed solution.

5.3.4.2 Generic inter-system handover procedure

The main steps of a generic inter-system handover procedure is depicted in Figure 5.7. If we replace the global RII in Figure 5.7 by the core RII, the sequence chart becomes the handover procedure between two non-3GPP access networks that are tightly coupled with the same 3GPP access systems (case 1 of Figure 5.6). If we remove the global RII, the sequence chart corresponds to the handover procedure between two access systems of the same operator or two cooperative operators (handover between access network (1) and (4)). The details of each step is as follows:

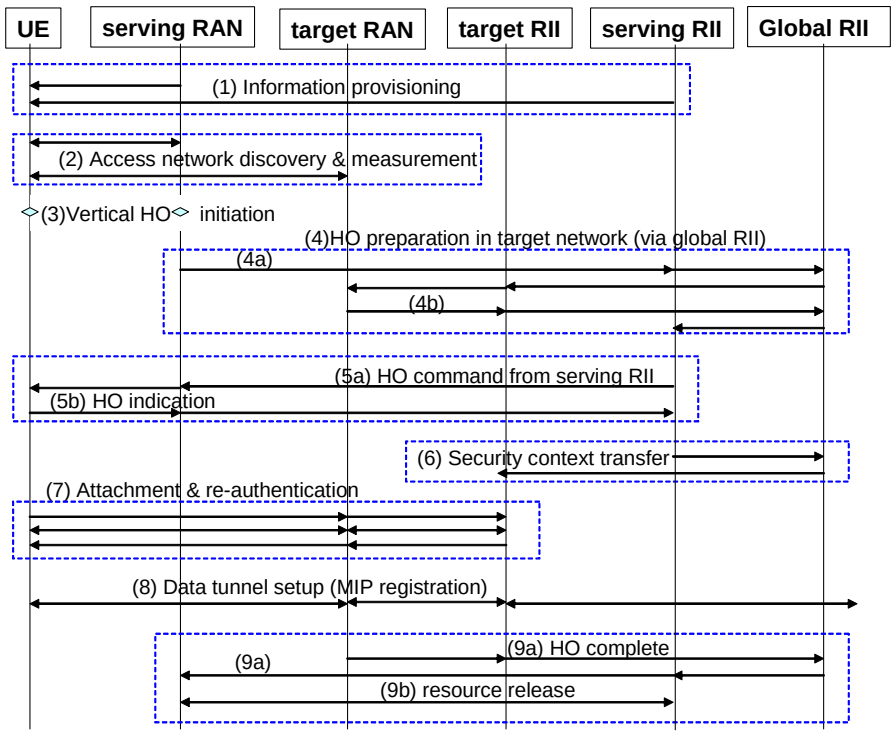

Figure 5.7: Generic Inter-system handover procedure

- **Step 1:** During the communication, the serving RAN sends the provisioning information to the UE to indicate the radio information of available neighboring cells. Such information helps the UE to synchronize with the neighboring cells and to monitor their signal strength. The NS component of the serving/home RII can provide the UE the information such as preferable or undesirable access networks and charging information of access networks for the UE's access network selection.

- **Step 2:** The UE selects preferred available access networks for the measurement purpose. The UE measures their link quality and send the measurement report to the network either periodically or event-based. The UE can also perform the scanning to discover new available access networks.

- **Step 3:** The handover can be initiated by the terminal or the network. An inter-system handover is triggered when *i)* the coverage of the same access system is not available, *ii)* the current system is heavily loaded or *iii)* the current system cannot satisfy user preferences or application requirements.

- **Step 4:** Once the vertical handover is initiated, the serving RAN will perform the handover preparation: checking whether the candidate target access network can support the imminent handover and performing the resource reservation in advance. The HO preparation request (message 4a) including potential target identities (ID) and the required QoS is sent to the serving RII which performs a network selection to eliminate undesired IDs. The request is then routed to the indicated target IDs via the global RII if needed. If the target RAN can allocate successfully the resource, the target RII will return a HO support (message 4b) to the serving RII. This message includes connection setup information.

- **Step 5:** The serving RII sends the UE a HO command (message 5a) including the recommended target IDs associated with their corresponding connection setup information. The serving RII may select one target cell, and send a strict HO command including only one selected target cell to the UE. If the UE receives a list of recommended IDs, it will select a suitable one and send a handover indication (message 5b) to notify its choice to serving RII for the handover execution preparation. The solution for access network selection is addressed in Chapter 2 & 3. Upon receiving the indication from the UE or after sending the strict handover command, the serving RII will send a HO notification to the user plane anchor point (e.g., SAE Anchor or HA) for the traffic redirection. Some techniques to minimize packet loss such as bi-casting and buffering can be used.

- **Step 6:** The serving RII sends the user's security context to the target RII to support the fast re-authentication. By transferring old security context from the serving network to the target network and reusing it with necessary adaptation, in this way a full security context creation is avoided. It reduces the handover latency which is significant in a roaming scenario between independent operators.

- **Step 7:** The UE performs attachment and re-authentication with the target access system. Based on the security context information, the target RII authenticates the UE without need to communicate with its home network.

- **Step 8:** Once the connection to the target RAN is successfully achieved, the UE sends the MIP registration to the HA to update the data plane path. The data tunnel is established to route the packets to the UE.

- **Step 9:** After the handover completion is notified (message 9a), resources in the old access network will be released (message 9b) and the presence information will be updated in the RIIs involved .

5.3.4.3 Detailed handover sequence charts

In this section, the vertical handover between UMTS and WLAN is presented in details. The message sequence chart for handover between UMTS and WiMAX basically follows the same principles except some following points. The ASN GW will replace the WAC in the handover solution and all handover decisions in the WiMAX side is done by this entity. The handover attachment in a target WiMAX access network will follow the procedure specified in the IEEE 802.16 standard [155, 156]. The specified handover sequences remain valid for both an SDR-enabled UE and a multi-interface UE

cases. However, the handover interruption time is different for the two cases: an SDR-enabled UE cannot maintain the serving communication while setting up the connection with the target access network whereas the multi-interface UE can.

5.3.4.3.1 Scenario 1: Tight-coupling using core RII

We consider a scenario where an MNO deploys the WiMAX or WLAN access network as an extension to their existing infrastructure. In this case, the ASN GW or the WAC emulates the role of the RNC. The scenario corresponds also to the case where an MNO interworks with a WLAN or WiMAX operators. This scenario refers to a tight-coupling interworking within the same administrative domain. Handover between UTRAN and WLAN/WiMAX is performed as a forward handover, i.e., the resources and address allocation are prepared in the target network before the UE is ordered to handover. The handover preparation is carried out by the core RII and the local WLAN/WiMAX RII. Some enhancements such as security context transfer and packet forwarding are used to reduce the interruption time during the handover.

Handover from UTRAN to WLAN access network (Figure 5.8):

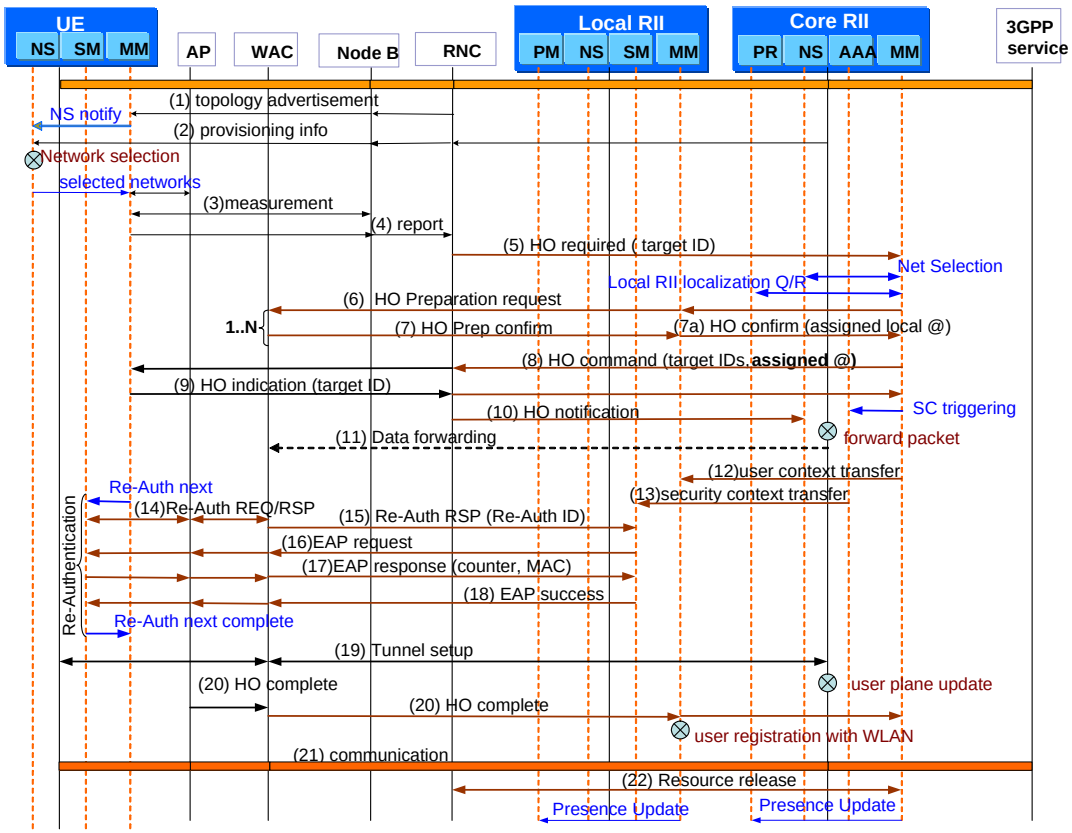

Figure 5.8: Handover from UTRAN to WLAN RAN

↪ 1) During the communication, the RNC sends the *topology advertisement* (or measurement control message) to the UE to indicate the radio information of its neighboring cells. The message includes the information which allows the UE to quickly synchronize with the neighboring cells.

↪ 2) The NS component in the core RII sends the *provisioning information* to the UE which may include the preferable (or undesirable) access network list and the charging information of certain access networks. Based on the topology advertisement and the provisioning information, the UE

filters available access networks for the measurement purpose. The NS component in the UE notifies the MM the selected neighboring cells for measurement.

↪ 3-4) The UE performs the *measurement procedure* on neighboring cells and sends the *measurement report* to the RNC.

↪ 5) Based on the measurement report, the RNC decides whether to make horizontal or vertical handover. If the handover to a non-3GPP access network is made, the RNC will send the *HO required message* to the core RII, including potential target IDs and the required QoS of the current running application. Otherwise, the intra-3GPP handover signalings do not reach to the core RII. Upon receiving the HO required message, the core RII may make the *network selection* to eliminate undesirable target IDs based on its policy and its preferences. The core RII learns the address of the target RII by consulting its presence database (the *RII localization query/response* between the MM and the PM).

↪ 6) The core RII sends the *HO preparation request* to the local WLAN RII which in turn sends this preparation message to the WAC that controls the indicated target IDs.

↪ 7) The WAC sends back the *HO confirm* to the local WLAN RII if the WAC can reserve the required resource for the handing-over UE. The local WLAN RII adds the local assigned IP address within the HO confirm message and forwards it to the core RII. The local address IP is used by the UE to establish the IP connection in the WLAN access network. By doing so, the UE avoids the IP address allocation (by DHCP) while attaching to the target WLAN access network.

↪ 8) The core RII sends the *HO command message* back to the UE. It may send the UE a strict HO command including only one selected target cell.

↪ 9) The UE sends the *HO indication* including its choice of target cell to the RNC and the core RII. Next, the MM component in the core RII notifies the SM component to transfer the security context to the local WLAN RII (the *security triggering* message).

↪ 10) The RNC sends the *HO notification* to the SAE Anchor for traffic redirection.

↪ 11) The SAE Anchor stops sending packets to the UE via the 3GPP Anchor and starts to *forward the packets* to the WAC that controls the target cell. The WAC address can be retrieved by the DNS request from the target ID. The WAC buffers packets and waits for the UE attachment.

↪ 12-13) The core RII sends the *user mobility context* and the *user security context* to the local WLAN RII to assist the connection setup and the re-authentication process.

↪ 14-15-16-17-18) The *Extensible Authentication Protocol (EAP)-based re-authentication* procedure between the UE and the local WLAN RII is performed.

↪ 19) *Data tunnels* are established: one between the UE and the WAC and one between the WAC and the SAE Anchor using the IKEv2 protocol [25].

↪ 20) The *HO complete* is sent from the target AP to the WAC, from the WAC to the local WLAN RII and then from the local WLAN RII to the core RII.

↪ 21) The *data communication* is exchanged via the WLAN access network.

↪ 22) *Resources are released* in the UTRAN side. The local WLAN RII and the core RII *update the presence* information of the UE.

Handover from WLAN access network to UTRAN (Figure 5.9):

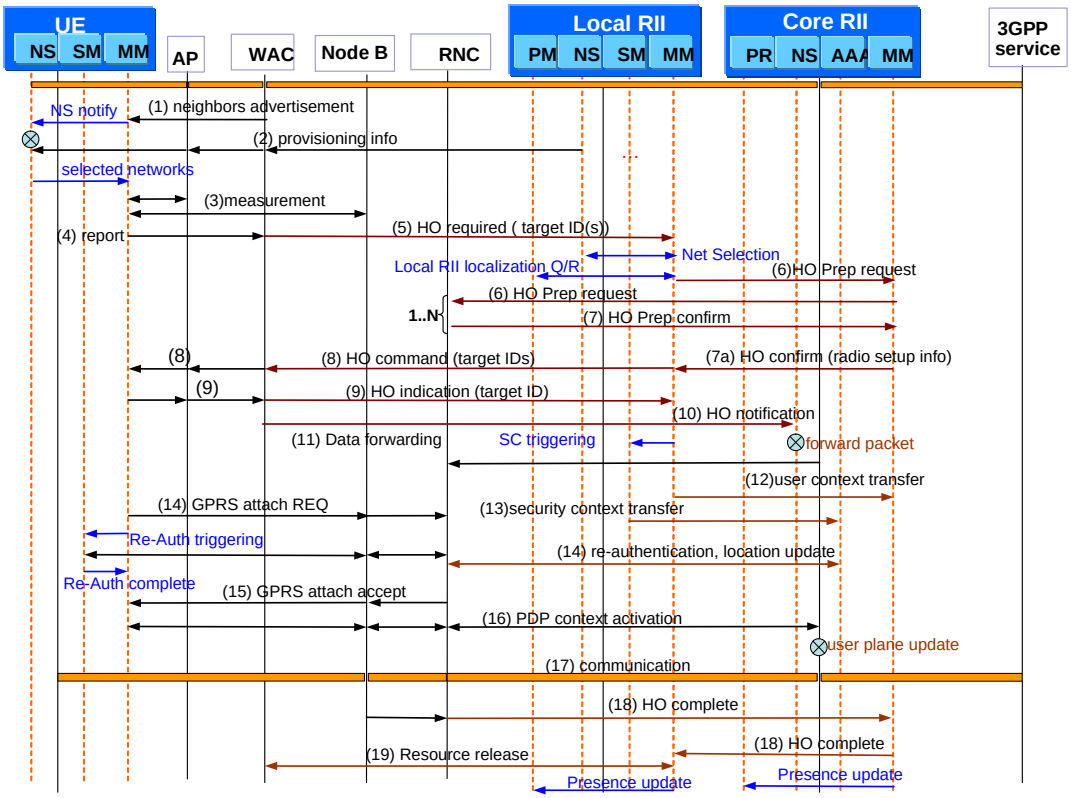

Figure 5.9: Handover from WLAN RAN to UTRAN

↩ 1) During the communication, the WAC sends the *advertisement* to the UE to indicate the radio information of its neighboring cells. This information helps the UE to quickly synchronize with the neighboring cells.

↩ 2) The NS in the local WLAN RII sends the *provisioning information* to the UE including the preferable (or undesirable) access network list and the charging information of certain access networks. The UE makes a preliminary network selection.

↩ 3-4) The UE performs the *measurement procedure* and sends the *measurement report* to the WAC. If the WAC does not have the control function, the UE will decide the handover initiation without sending the measurement report to the WAC.

↩ 5) If a vertical handover to a 3GPP access network is decided, the *HO required message* including the potential target IDs and the required QoS is sent to the local WLAN RII. The local WLAN RII may make the network selection and the RII localization query/response to learn the address of the target RII.

↩ 6) The local WLAN RII sends the *HO preparation request* to the core RII which in turn sends it to the RNC that controls the indicated target IDs.

↩ 7) The RNC sends back the *HO confirm* to the core RII if the RNC can reserve the required resource for the handing-over UE. 7a) the core RII inserts the reconfiguration information (a reference number to a pre-defined set of UTRA parameters) for radio setup in the HO confirm message.

↩ 8) The local WLAN RII sends the *HO command* message to the UE including the recommended target IDs and their corresponding pre-configuration reference.

↩ 9-10-11) The UE sends the *HO indication* including its choice of the target cell to the WAC and then to the local WLAN RII. The WAC sends HO notification to the SAE Anchor. The SAE stops sending packets to the UE via the WAC and forwards the packets to the 3GPP Anchor or even the SGSN/RNC that controls the target cell.

↩ 12-13) The local RII sends the user mobility and security context to the core RII.

↩ 14-15-16-17) The UE performs the GPRS attachment in the UTRAN to setup the connection with the target Node B and performs the PDP context activation.

↩ 18-19) The *HO complete* is sent from the Node B to the local WLAN RII and the resources allocated in the WLAN access network are released. The local WLAN RII and the core RII update the presence information of the UE.

5.3.4.3.2 Scenario 2: Roaming using global RII

This scenario describes the roaming between an MNO and a WiMAX/WLAN operator. The handover preparation is achieved through the global RII. The message sequence chart in Figure 5.10 and Figure 5.11 illustrates the handover procedure.

Handover from UTRAN to WLAN access network (Figure 5.10):

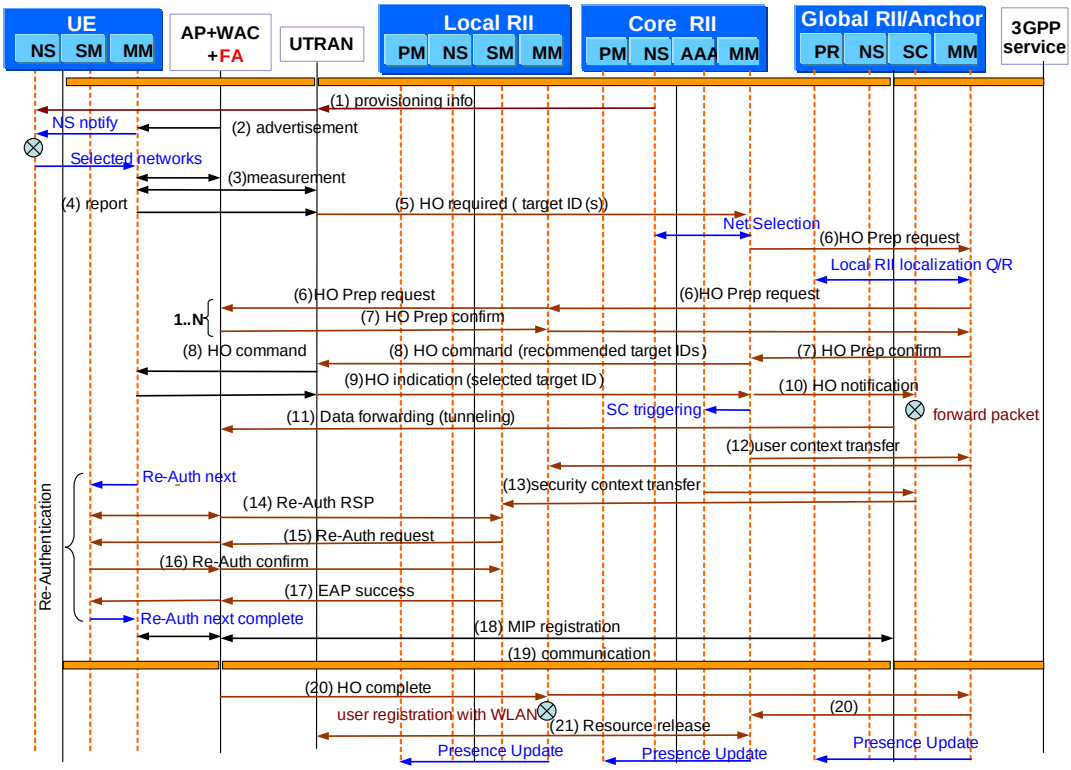

Figure 5.10: Handover from UTRAN to WLAN RAN via global RII

↪ 1-2-3-4-5) see the description of message 1-2-3-4-5 in Figure 5.8.

↪ 6) As the 3GPP network does not have a direct roaming agreement with the candidate access networks, the core RII sends the *HO preparation request* to the global RII. The global RII routes the HO preparation message to the WAC via the local WLAN RII. The global RII learns the address of the target local WLAN RII by consulting its presence database.

↪ 7-8-9-10) see the description of message 8-9-10 in Figure 5.8. Note that in the message 10, the *HO notification* will be sent to the Data Anchor (HA) instead of the SAE Anchor.

↪ 11) The HA stops sending packets to the UE via the SAE Anchor and forwards the packets to the WAC. The WAC address can be retrieved by the DNS request from the target ID. The WAC buffers the packets and waits for the UE attachment.

↪ 12-13-14-15-16-17) see description of message 12-13-14-15-16-17 in Figure 5.8.

↪ 18) The *MIP registration* is performed between the UE and the WAC/FA, between the WAC/FA and the local HA implemented in the local WLAN RII, and between the local HA and the Data Anchor Point.

↪ 19-20-21)The *data communication* is exchanged via the WLAN access network. The *HO complete* is sent from the target AP to the global RII and then from the global RII to the core RII. The resources are released in the UTRAN. The local WLAN RII, the core RII and the global RII update the presence information of the UE.

Handover from WLAN AN to UTRAN (Figure 5.11):

The messages sequence of the handover procedure from WLAN AN to UTRAN via the global RII is illustrated in Figure 5.11. Compared to the message sequence chart in Figure 5.9, the signaling between the local WLAN RII and the core RII will be exchanged via the global RII. The details of these messages are omitted.

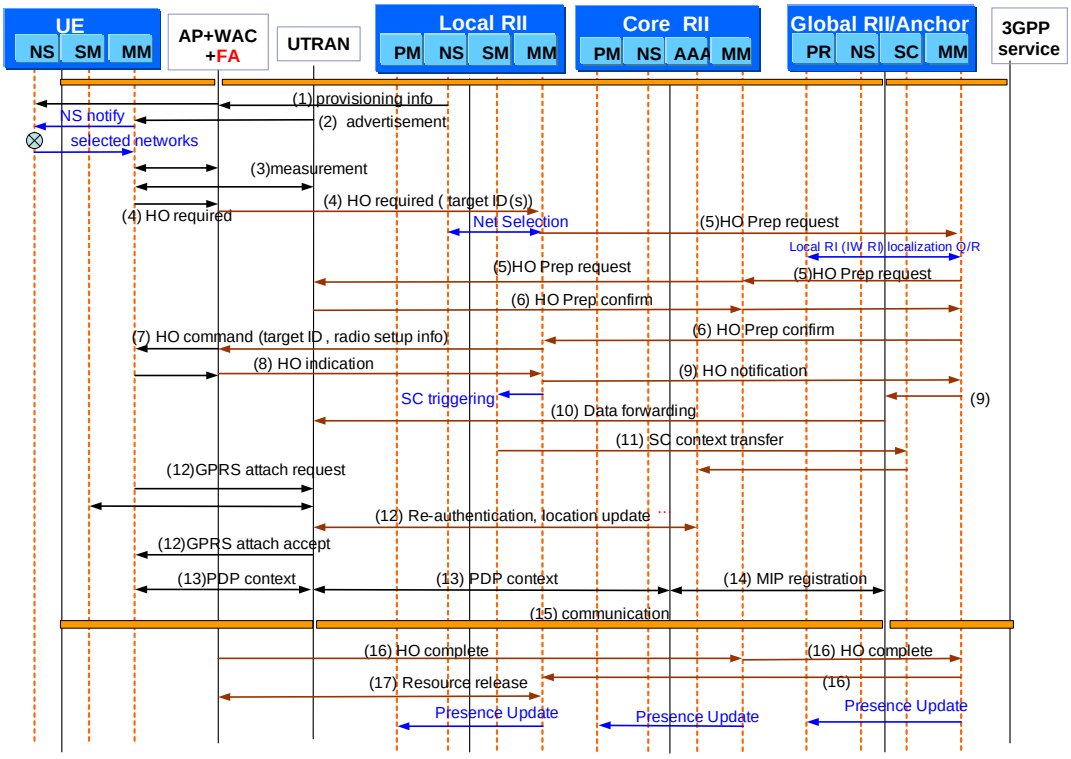

Figure 5.11: Handover from WLAN AN to UTRAN via the global RII

5.3.5 Advantages of RII solution

In this section, we qualitatively evaluate and compare the RII-based architecture with other existing solutions for interworking & roaming in multi-operator environments.

Seamless mobility: Most of third party intermediary solutions like broker, clearinghouse [145–148] have focused on the authentication, authorization and billing when users move outside their subscribed network and connect to a visited network. These solutions care about guaranteeing the access in the visited network rather than maintaining on-going communications. Despite much research effort has been done in designing the interworking solution between WLAN/WiMAX and 3GPP cellular networks, there is no solution for handover between two access systems of two independent operators. One of the disruptive advantages of the RII-based architecture is to enable the handover across different access systems without need of pre-existing agreement between operators. Compared to existing tight-coupling schemes, the RII-based interworking does not require many changes to access system architectures. Compared to existing loose-coupling solutions, the RII-based architecture supports make-before-break mobility even between access systems of two independent operators.

Secure and transparent mobility: The RII provides a secure mechanism for exchanging handover signalings among different access systems. We have extended the EAP protocol to carry the signalling exchanges between the RII entities in a secure manner. The details of such specification are given in [158]. The secure roaming agreement and billing control are established in an automatic manner based on user preferences and policy thank to the ontology method [159]. The roaming dealing is completely transparent to users. The security context is transferred from the serving access network to the target access network to enable the fast re-authentication.

Scalability: The RII can be used to integrate different access systems deployed by the same operator or different operators that may not necessarily have direct SLAs among them. The RII can be employed in a hierarchical structure to integrate access networks through first tier-RIIs, and then interconnect the first tier-RIIs through second tier-RIIs and so on to enable the global roaming. By establishing an SLA with the global RII, an operator can benefit the global roaming service to any other operators that connect to this global RII. Therefore, the wireless system interoperability using RII is scalable and flexible. Additionally, the load on the RII is limited since the latter is only involved in control signaling exchanges.

Feasibility: Though the solution is in line with the 3GPP LTE architecture, it does not preclude the implementation of RIIs in the current network environment for integration among 2G/3G and WLAN systems. The solution is feasible and economical since it does not require many changes in the existing network infrastructure. An operator that wants to benefit from such interworking and roaming facilities only needs to implement the local/core RII functionalities in its access gateway by software upgrading. Virtual operators have much interest to implement the global RII to provide third party roaming services. The feasibility of the proposed RII architecture and the handover sequence chart has been demonstrated through a test-bed which is described in details in [158].

Profitability: With the adoption of the proposed RII architecture, the network availability will be widely extended. First, users will have great interest since they can connect to any available access network. Second, the network infrastructure utilization will increase, which will give opportunities to operators to improve their profitability.

5.4 Summary

In this chapter, we first addressed a short-term solution for interworking between UMTS and WiMAX networks. The handover procedure with detailed message sequence exchanges was given. Secondly,

we presented a novel roaming and interworking intermediary (RII) which offers a flexible means for interworking and roaming among different access systems. The RII support all combinations of different radio technologies in a multi-operator environment with different contractual agreements among operators. Most importantly, the proposed RII solution allows users to freely and securely move across different access systems without need of a pre-existing subscription. Anyway, a direct or indirect SLA between two access networks' owners is always required. To enable the secure and seamless mobility, network-controlled operations such as on-line roaming agreement dealing, network provisioning, mobility and security context transfer...are specified. We specified in details the signaling control messages between all the network entities involved in the different vertical handover scenarios: between two tightly-coupled access networks and between two loosely-coupled access networks of independent operators.

On the top of this interworking architecture, we address in the next chapters the inter-system measurement, the required cell overlap to support seamless handovers and the radio resource management over heterogeneous networks.

~~ △♡△ ~~

Chapter 6

Inter-system Measurement and Required Cell Overlap

Now, these two (WiMAX and UMTS/HSDPA) are definitely complementary - there are lots of ways that they can feed of each other very synergistically to deliver services to make things better for operators and users.

Rupert Baines, Vice President of Marketing of PicoChip

In the precedent chapter, we proposed the interworking and roaming architecture between 3GPP cellular and non-3GPP systems. In this chapter, we focus on the interworking between UMTS/HSDPA and WiMAX systems by addressing the inter-system measurement and the required cell overlap analysis. They are two primary conditions necessary to achieve the inter-system handover. We investigate the feasibility of the UMTS-WiMAX inter-system measurement through reconfigurable radio-enabled terminals. We analyze the minimum cell overlap required for seamless handovers between two adjacent cells within the same technology and between different technologies in UMTS-WiMAX systems.

6.1 Introduction and Motivation

In heterogeneous environments, vertical handovers between different technologies can be performed either with multiple radio interfaces [11, 14, 17, 160] or with one unified Software-Defined Radio (SDR) interface [84, 85, 149]. Multiple radio interfaces often need application-specific integrated circuit (ASIC) devices and separate transceivers [150]. Hardware integration and power consumption are important issues since portable devices are limited by their batteries. But SDR devices can reprogram to operate in different radio interface standards, allowing the efficient use of radio spectrum and power. They make it possible both to improve performance and to customize devices to individual needs [149] [150]. In this chapter, we anticipate portable devices equipped with one reconfigurable radio interface. In order to perform inter-system measurement, an SDR-enabled device has to reconfigure its radio interface so that it can measure the signal strength from the non-serving technology. The key issues are (a) whether this inter-system measurement is possible without affecting on-going sessions and (b) whether the results are reliable enough for use in the handover decision. One of the goals of this chapter is therefore to explore UMTS-WiMAX inter-system measurement procedures.

The cell overlap area is also important for network planning and vertical handover management. The overlap area cannot be too large since this increases the number of Base Stations (BS) needed

and consequently the cost for operators. But the overlap distance should be large enough for seamless handover. This means that mobile terminals must have enough time to prepare handover (detecting neighboring cells, carrying out the measurement, initiating the handover and establishing the connection with the target cell) before losing the connectivity with the attached BS. If the cell overlap area is too small, the network's connection loss ratio is increased because mobile terminals at the edge of a cell cannot receive support from neighboring cells in time to prepare the handover. Cell overlap therefore depends greatly on signal strength and the measurement reporting period. The main focus of this chapter is to assist network planning by examining the cell overlap required to ensure seamless handover and optimize the deployment costs.

Radio cell planning consists of three phases: nominal planning, detailed planning, and network optimization [161]. The purpose of the nominal phase is to estimate the approximate number of cells and network elements. The detailed planning phase uses network planning tools to estimate the location of BS sites. This is done by considering real site locations and the propagation conditions calculated on digital maps and real user distributions. Once the network is deployed, field test measurements adjust and optimize the planning. In the detailed planning phase, the handover is considered in terms of signal threshold and network capacity, not the minimum cell overlap required for uninterrupted handovers. The handover operation is usually verified in the optimization phase. Our proposed method computes the minimum required cell overlap to be used as a constraint during the detailed planning phase. This gives network designers an additional tool in planning the location of cell sites. The computation can also be used to analyze the seamless handover capability of an existing or upgraded mobile network infrastructure.

6.2 Background knowledge

6.2.1 Overview of measurement and handover decision

When a mobile terminal moves from one BS in a network to another, it measures the signal strength of neighboring BSs and decides on the most suitable target for the handover. The measurement is imperfect because of a number of effects such as fast fading, short-term shadowing and multi-path distortion [162] [105]. At the mobile terminal, the signal is first averaged through a low-pass filter to eliminate high frequency components (Rayleigh fading) before measurement reports are sent to the network. Before the handover decision is made, the signal strength values are filtered through a linear averaging window or an exponential moving averaging window to smooth out shadowing effects on the network side [155, 163]. The measurement results and the reporting period are important in the handover decision. While the measurement mechanism within the 3GPP system [164] has been carefully studied, the UMTS-WiMAX inter-system measurement of an SDR-enabled terminal has not. We do so before moving to the analysis of cell overlap.

When the UE is at distances d_1 and d_2 respectively from BS A and BS B (two adjacent cells), the average received signal levels from BS A μ_A and from BS B μ_B are:

$$\mu_A(d_1)[dB] = K_1 - K_2 log(d_1) \tag{6.1}$$

$$\mu_B(d_2)[dB] = K_1 - K_2 log(d_2) \tag{6.2}$$

where K_1 represents the transmission and reception antenna gains and the signal's wavelength whereas K_2 represents the path loss factor. In the above formulae, the signal fluctuation parts due to fading are already smoothed out at the terminal side. A handover from BS A (serving) to BS B (target) occurs if the average received signal level from the serving BS A drops below that of the neighboring BS B. Due

to the uncertainty arising from fluctuations of the physical channel, the handover is initiated if and only if $\Delta\mu = \mu_B - \mu_A \geq h - \sigma$, where h is a hysteresis margin and σ is the standard deviation of the received signal.

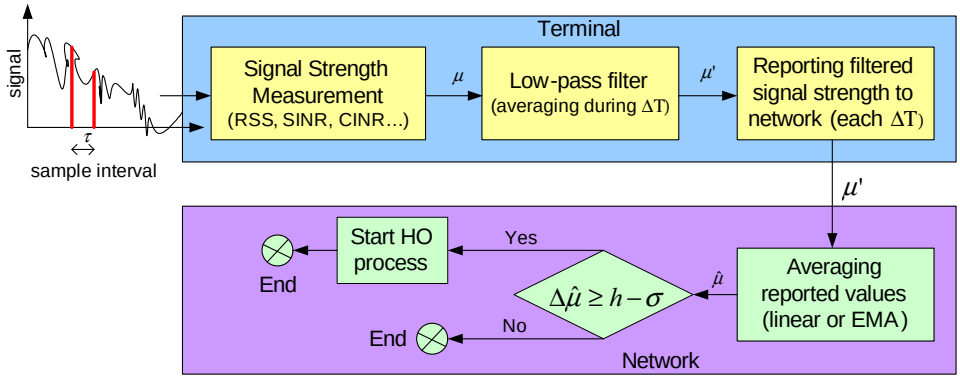

Figure 6.1: Measurement and signal strength-based handover decision

In the literature, most research has focused on handover decision algorithms. While these algorithms in homogeneous networks are mainly based on the strength of the signal received, a number of metrics are involved in heterogeneous networks. They include radio link quality, service cost, network load and power consumption. Yet the received signal strength is still vital in identifying the radio link state and the required cell overlap. The effects of channel impairments on handover decision algorithms have been addressed in [165–168] and references therein. However, the overlap area between adjacent cells has not been adequately addressed. Recently, in [168], cell overlap has been briefly discussed but only in the simple case of a user moving in a straight line between two WiFi access points. The proposed overlap formula cannot be used in other, less specific, cases. In this work, we discuss an accurate method of estimating the minimum cell overlap required to handle all possible movement directions.

6.2.2 Handover measurement in UMTS

The UMTS system encompasses two duplex transmission modes: FDD and TDD. In general, FDD is used to provide wide area coverage whereas the TDD usage is limited to complement FDD in hot spots or inside buildings [169]. We consider the UMTS FDD mode since it is widely commercially deployed. In UMTS-FDD networks, we can distinguish three types of measurement: intra-frequency, inter-frequency and inter-system measurement. The inter-system measurement is the most difficult procedure. Major challenges of inter-frequency/inter-system measurement are (*i*) to synchronize in time and in frequency with neighboring cells and then (*ii*) to measure the signal strength on the Common Pilot Channel (CPICH) on the frequency that is different from that of the on-going communication.

The UE in UMTS-FDD mode is continuously receiving on the downlink carrier and transmitting on the uplink carrier and thus there are no idle time slots for measuring on another frequency. In order to perform the inter-frequency measurements without having a dual receiver terminal, the compressed mode technique [170–172] is needed. The compressed mode means that the transmission and reception are halted for a short time to perform the measurement on other frequencies. The intention is not to lose any data but rather to compress the data transmission in the time domain. By using the compressed mode, the transmission gap length for measurement can vary from 3 to 14 slots located on one or two consecutive frames as illustrated in Figure 6.2. The instantaneous transmission power is increased in the compressed frame in order to keep the quality unaffected.

In the UMTS FDD measurement, in order to satisfy the required measurement accuracy, the intra-frequency measurement period (i.e., periodic reporting interval) is specified to 200 *ms* whereas the

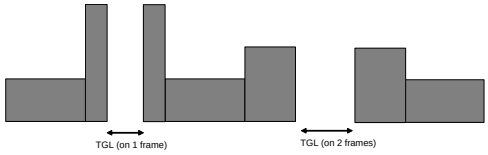

Figure 6.2: Compressed mode transmission

inter-frequency measurement period with the help of the compressed mode is given by [164]:

$$\Delta T_{inter\,freq} = Max\{480, \frac{T_{basic\_inter} \times 480}{T_{inter}} N_f\} \tag{6.3}$$

where T_{inter} is the minimum available time for inter-frequency measurement during $480ms$, $T_{basic\_inter} = 50ms$ is the basic inter-frequency measurement time period and N_f is the number of FDD frequencies to be measured. The value of T_{inter} depends highly on the compressed mode pattern configuration [170]. In short, the time for inter-frequency measurement is at least equal to $480ms$.

6.2.3 Handover measurement in WiMAX

With the completion of IEEE 802.16e-2005 standard [156] and the recent inclusion of WiMAX into the IMT-2000 family of technologies, the mobile WiMAX system has been attracting a very high level of interest. In this paper, we consider the mobile WiMAX operating in the frequency band of 3.4-3.8Ghz with the channel bandwidth of 5Mhz as specified in [4]. Until now, all mobile WiMAX profiles are based on TDD mode and use scalable OFDMA as the physical layer. As illustrated in Figure 6.3, each OFDMA frame is divided into downlink (DL) and uplink (UL) subframes separated by Transmit/Receive and Receive/Transmit Transition Gaps (TTG and RTG respectively). Each frame starts with a preamble, transmitted by one symbol over all subchannels with a specific Pseudo Noise code. The synchronization, equalization and signal strength measurement are performed on the preamble.

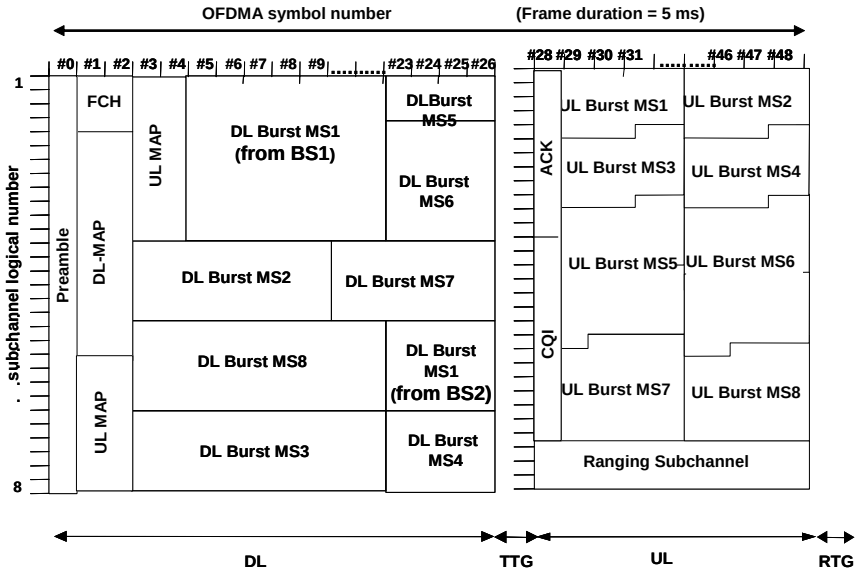

Figure 6.3: OFDMA frame structure for the channel bandwidth of 5MHz

During the communication the UE can ask the serving BS for a suitable scanning interval, where no

data transfer is scheduled between the UE and the BS, for seeking and measuring the signal strength of the neighboring BSs. For real time applications such as voice over WiMAX or video streaming which require a low latency, the allocation of a scanning interval may degrade the quality. Another solution for measurement in OFDMA systems is to use the idle slots in the downlink subframe. Indeed, data transmission from the serving BS to the UE is carried out only in some slots, known as data burst composed of some OFDMA symbols and some subchannels. For example, the MS1 in Figure 6.3 can communicate with the BS1 and then measure the signal strength from the neighboring BS2 without need of the scanning interval allocation.

6.3 WiMAX measurement period analysis

In this section we endeavor to quantify the required measurement report period in a mobile WiMAX system. Since a radio signal in wireless environments suffers from different effects such as multiple reflections, obstructions and noise, the received signal at the mobile terminal is the sum of a large number of scattered waves [105, 162]. The signal measured on each preamble fluctuates due to the fast fading effects. The terminal thus has to average the signal strength over a number of measurement samples before reporting the result to the upper layer or the serving BS. Assume that $(X_i)_{i=1..S}$ is a set of S samples of the measured RSS. Under the fast fading effects and the non-line of sight (NLOS) propagation environment, $(X_i)_i$ follows the Chi-Square distribution [162] [105] given by (6.4):

$$p(X) = \frac{1}{2\sigma_r^2} exp(-\frac{X}{2\sigma_r^2}) \tag{6.4}$$

where $2\sigma_r^2 = E(X)$ is the average signal power. The value of the measured signal strength is computed as:

$$\bar{X} = \frac{1}{S}\sum_{i=1}^{S} X_i \tag{6.5}$$

And the standard deviation of the measured signal strength is

$$\sigma_m^2 = \frac{1}{S-1}\sum_{i=1}^{S} (\bar{X} - X_i)^2 \tag{6.6}$$

The confidence interval [143] for our estimator follows

$$\mathbf{P}(|\mu - \bar{X}| \leq \frac{\sigma_m t_{n;\alpha/2}}{\sqrt{S}}) = 1 - \alpha \text{ where } n = S - 1 \tag{6.7}$$

In the above equation, $\mu = E(X)$ is the theoretical power signal value, $(1-\alpha)$ is the confidence coefficient or confidence level, and $t_{n;\alpha/2}$ is the percentage point of the Student's distribution such that $P(|t_n| > t_{n;\alpha/2}) = \alpha$. In other words, we are $(1-\alpha)100\%$ sure that the measured value varies from $L_1 = (\bar{X} - \frac{\sigma_m t_{n;\alpha/2}}{\sqrt{S}})$ to $L_2 = (\bar{X} + \frac{\sigma_m t_{n;\alpha/2}}{\sqrt{S}})$.

Hence, the absolute accuracy of the RSS measurement is:

$$E_1[dB] = 10\log_{10} L_1 - 10\log_{10}\mu = 10\log_{10}\frac{L_1}{\mu} \tag{6.8}$$

and

$$E_2[dB] = 10\log_{10}\frac{L_2}{\mu} \tag{6.9}$$

According to the standard, the absolute accuracy should be $\pm 4dB$ [155]. Besides the propagation phenomena in the wireless channel, the measurement error can be caused by an additive noise in the receiver equipment which is generally assumed to be about $\pm 2dB$ by the RF system designers. Thus, the E_1 and E_2 must vary from $-2dB$ to $2dB$ to satisfy the previous conditions.

S	10	20	30	**40**	50	60	70
$E_1[dB]$	-6.2	-2.9	-2.1	**-1.75**	-1.51	-1.29	-1.15
$E_2[dB]$	2.21	1.58	1.28	**1.14**	1.03	1.00	0.96

Table 6.1: Absolute accuracy with confidence level of 95%

To estimate S, we generate $X_i = -2\sigma_r^2 \ln U(0,1)$ following the Chi-Square distribution, where $U(0,1)$ is a uniform random variable on the interval $[0,1]$. Following the process described above, we calculate the absolute accuracy corresponding to different values of S. When the confidence level is 95%, the absolute accuracy is obtained as Table 6.1. We see that given the confidence level of 95%, at least 40 samples are needed to satisfy the measurement accuracy requirements. If the confidence level is 98%, the number of samples is at least 50. However, in most cases a 90%-95% level of confidence is largely sufficient for signal strength measurement in wireless cellular networks [173] [174]. As each signal sample from the serving BS is measured every 5ms-WiMAX frame duration (all WiMAX equipment will initially support only 5ms frames [4]), a time length of $200ms$ is needed to collect 40 measurement samples. If the UE can measure one signal sample from each neighboring cell every 2 frames (i.e., 20 samples per $200ms$), the required absolute measurement accuracy on neighboring cells will be maintained with a confidence level of 90%. A WiMAX measurement reporting period of $200ms$ (i.e., 40 frames) is therefore considered.

6.4 UMTS-WiMAX inter-system measurement

A seamless vertical handover between UMTS and WiMAX systems is possible only if the inter-system measurement through an SDR-enabled device is feasible and the required cell overlap is large enough. As the analysis of cell overlap is mainly based on signal strength, this section examines the UMTS-WiMAX inter-system measurement.

In homogenous UMTS or WiMAX networks, the neighboring cell information (such as cell identity, carrier frequency and scrambling code...) is provided to the UE via Radio Resource Control (RRC) messages or neighboring advertisement (MOB_NBR-ADV) messages respectively. This information helps the UE to synchronize with neighboring cells and measure their signal strength. In inter-system measurement, where the carrier frequencies of adjacent UMTS and WiMAX BSs are different, without the provisioning information, it takes time for the UE to discover neighboring cells in the other technology. In this work, we assume that the UMTS and WiMAX access networks are deployed by the same operator or by two operators with specific roaming agreements. Each access node is upgraded so that it is able to broadcast the neighboring cell information of other technologies as well as its own. This facilitates the discovery of neighboring cells in the other system. If the networks are operated by two different operators without this kind of agreement, inter-system measurement remains an open issue.

When the same technology coverage is not available, inter-system handover is required to maintain connectivity. Inter-system handover is also needed to balance the load between UMTS and WiMAX networks. If the currently connected network does not support the application's QoS, the UE may be

forced to trigger a vertical handover to another complementary technology. Inter-system measurement is therefore initiated only when a vertical handover is needed.

6.4.1 WiMAX to UMTS inter-system measurement

When the UE has a consistent communication with a serving WiMAX BS, it can allocate a scanning interval from that BS for inter-system measurement [155]. However, if the UE is running delay-critical, real-time applications, the scanning interval cannot be employed. In these cases, the UE has to use remaining idle slots during the downlink communication for measurements. This is possible since not all OFDMA symbols are addressed to the UE during the downlink subframe.

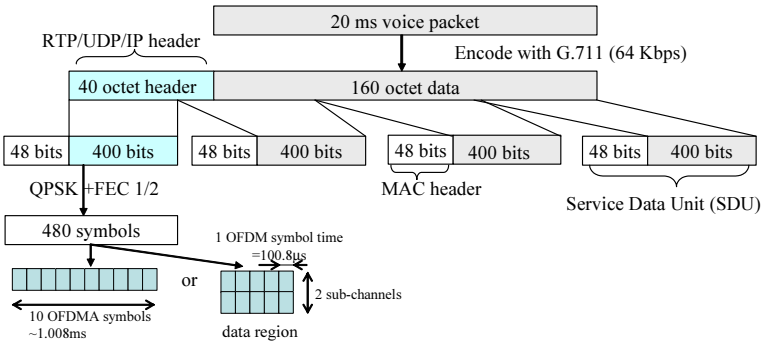

Figure 6.4: Voice over WiMAX transmission process

Let us now analyze the WiMAX to UMTS inter-system measurement in the case where the most delay-critical VoIP applications are used. We assume that the voice codec G.711 with the data rate of 64Kbps, used for high rate voice services, is employed here [4]. The data transformation and encapsulation from the application layer to the physical layer is illustrated in Figure 6.4. After having been encoded and then the RTP/UDP/IP header added, a 20ms packet of voice generates 200 octets of data to be transmitted at the MAC layer. The data must be transmitted during 20ms to maintain the voice quality (i.e., small delay and small jitter). Therefore, they will be divided into 4 Service Data Units (SDU) and each will be transmitted during one 5ms frame. As the handover usually occurs at the border of a cell, the most robust Quadrature Phase-Shift Keying (QPSK) modulation combined with a coding rate of 1/2 is considered. At the end, we have 480 modulation symbols to be transmitted by each 5ms frame.

In the OFDMA structure, one slot is composed of one subchannel by one OFDMA symbol which can carry 48 modulation symbols (i.e., each subchannel contains 48 data subcarriers). Accordingly, 10 slots are required in the downlink frame from the BS to the UE to ensure the above VoIP applications. The data region of these 10 slots can be formed by 10 OFDMA symbols over 1 subchannel or 5 OFDMA symbols over 2 subchannels as depicted in Figure 6.4. For any data region configuration, the time duration for the data reception in the downlink is less than $1.008ms$ (i.e., the duration of each OFDMA symbol is $100.8\mu s$). Similarly, if the UE runs a video streaming application with a data rate of 512Kbps, the downlink data region in each frame must be composed of 56 slots. In this case, the downlink data region can be formed by 7 OFDMA symbols over 8 subchannels. The downlink reception time is thus only $7 * 0.1008 = 0.7056ms$. Given that the DL subframe duration is not less than $2.5ms$ (half of a frame duration), the UE that uses the most critical application in terms of delay like VoIP or video streaming still has enough idle time in the DL subframe to measure the neighboring UMTS cells.

The WiMAX to UMTS inter-system measurement is somewhat similar to the inter-frequency measurement in UMTS. The minimal measurement reporting period in UMTS is $\Delta T_{inter\,freq} = 480ms$ [164] when the UE can reserve $50ms$ every $480ms$ for inter-frequency measurement. As analyzed above, the UE in the WiMAX communication mode can also reserve around $50ms$ for UMTS measurements during $480ms$ (e.g., 5 OFDMA symbol duration ($0.504ms$) every $5ms$-WiMAX frame). To keep it compatible with the GSM to UMTS inter-system measurement [164], we set the minimal WiMAX to UMTS measurement report period to $\triangle T_{w\rightarrow u} = 480ms$. Similar to $\Delta T_{inter\,freq}$, $\triangle T_{w\rightarrow u}$ varies according to the allocated measurement time period and the number of UMTS carrier frequencies to be measured N_u:

$$\Delta T_{w\rightarrow u} = \max\{480, \frac{T^w_{basic\_inter} \times 480}{T^w_{inter}} N_u\} \quad (ms) \tag{6.10}$$

where T^w_{inter} is the available time scheduled for measurement during $480ms$ of the WiMAX communication mode and $T^w_{basic\_inter} = 50ms$. The value T^w_{inter} depends on the number of idle slots allocated for measurement in each downlink subframe.

6.4.2 UMTS to WiMAX inter-system measurement

When the UE has an ongoing communication with a serving UMTS Node B, it has to enter the compressed mode [171, 172] in order to perform inter-system measurement on neighboring WiMAX cells. Since the compressed mode affects the capacity performance as well as the coverage of the UMTS system, it is activated only when necessary. The measurement period is sufficient if every 3^{rd} frame (one transmission gap every $30ms$) is compressed for inter-frequency measurement. In this case, the capacity degradation is around 19% [172]. If the UE can measure one signal sample from a WiMAX neighboring cell every scheduled transmission gap (in order to achieve 20 measurement samples as mentioned above), the measurement duration is $\Delta T_{u\rightarrow w} = 30 \times 20 = 600ms$. This can be considered a basic measurement reporting period.

Similarly, the UMTS to WiMAX measurement period $\triangle T_{u\rightarrow w}$ depends on the compressed mode pattern and the number of inter-frequency WiMAX cells to be measured N_w. That is,

$$\Delta T_{u\rightarrow w} = \max\{600, \frac{T^u_{basic\_inter} \times 600}{T^u_{inter}} N_w\} \quad (ms) \tag{6.11}$$

where $T^u_{basic\_inter} = 62.5ms$ is the basic inter-frequency measurement time period during each $600ms$ of the UMTS communication mode. This is because $T^u_{basic\_inter} = 50ms$ every $480ms$ [164]) and T^u_{inter} is the available time for WiMAX measurement during $600ms$.

6.5 Required cell overlap analysis

Cell overlap depends first on mobile terminal velocity. If the terminal velocity is too high, the time taken to cross the overlap area may be shorter than is required for handover preparation. For a given velocity, the required minimum cell overlap depends also on the handover delay (i.e. handover measurement).

6.5.1 Overlap in homogeneous UMTS or WiMAX networks

6.5.1.1 Cell overlap and crossing distance definitions

We consider three adjacent cells of the same radius R (as shown in Figure 6.5) and a UE moving from BS A to BS B with a constant velocity v. We assume constant velocity because, if cell overlap can support the movement of constant velocity v, it can support all non-constant movements whose velocity is less than or equal to v. If the distance between two BSs is D and the overlap distance between two cells A and B is D_{co}, we have $D_{co} = 2R - D$. The main focus of this section is the estimation of the minimum required cell overlap distance $D_{co_{min}}$, or, in other words, the distance threshold below which the handover cannot be seamlessly achieved.

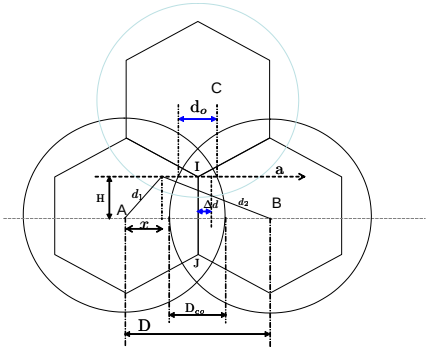

Figure 6.5: Overlapping model

We consider I the common vertices of three hexagons A, B and C and a straight line a going through the point I and parallel to AB. Based on the hexagon property, the distance H between the line a and AB is $H = \frac{D\sqrt{3}}{6}$. When the UE moves from BS A to BS B along a straight line, the distance between the point where the UE enters cell B and the point where it moves out of cell A is defined as a crossing distance. For example, if UE moves along the line a, the crossing distance is d_o. By the Pythagoras's theorem, we have:

$$d_o = 2\sqrt{R^2 - H^2} - D = 2\sqrt{R^2 - \frac{D^2}{12}} - D \tag{6.12}$$

The minimum crossing distance between two adjacent cells of radius R whose centers are separated by distance D is d_o. The proof is as follows: First, any move from BS A to BS B that does not follow a straight line results in a travelling distance that is longer than a straight line movement. Second, all straight line movements that do not cross the segment IJ (see Figure 6.5) are not part of the movement from BS A to BS B. For all the straight lines crossing K, a particular point in segment IJ, the line parallel to AB results in the smallest crossing distance. Of all lines parallel to AB, the line going through I or J gives the smallest crossing distance. Therefore, d_o, given by (6.12), is the minimum crossing distance between any two adjacent cells.

For seamless handover at link layer (L2) via a single radio interface, the UE generally has to be capable of triggering handover and starting a new connection with the target BS before it moves out of the radio coverage of its serving BS. The time from the instant when the UE arrives at the position I (or J) to the instant when the UE makes the handover decision is called handover delay, usually referred to as L2 handover delay. However, to achieve seamless handover at the application layer, other issues such as IP address assignment, security and session redirection should be considered. The minimum required cell overlap $D_{co_{min}}$ depends only on the L2 handover delay and should be computed when the UE follows the minimum crossing distance. In fact, half the minimum crossing distance $d_o/2$ has to be

at least as large as the distance travelled by the UE within the handover delay. The necessary condition is satisfied only if:

$$d_o/2 \geq \delta v \tag{6.13}$$

where v is the mobile terminal velocity and δ is the handover delay.

6.5.1.2 Handover delay

From the instant when the handover decision is made by the network to the instant when the UE stops communicating with the serving BS to launch the handover process, the delay is very small compared to the measurement delay. Handover delay is defined as the sum of the averaging delay and the hysteresis delay [165, 167, 168]. Below, handover delay is analyzed where the UE moves along the line a (see Figure 6.5).

At instant τ, we assume $\Delta\mu(\tau) = \mu_B(\tau) - \mu_A(\tau) = 0$. This corresponds to the instant when UE is at point I. Because of the network's averaging process, two filtered signals are equalized at instant $(\tau + \delta_{av})$, i.e. $\Delta\hat{\mu}(\tau + \delta_{av}) = 0$ where δ_{av} is the averaging delay. If a linear averaging window is used, the last reported signal is averaged over the N last reported values. The averaging window time is $T = N\triangle T$ where ΔT is the measurement reporting period. The averaging delay is $\delta_{av} = T/2 = \frac{N}{2}\Delta T$ since each average value has been made from the N past samples.

If an exponential moving averaging (EMA) window is used, the last reported signal $\mu[k]$ at instant k is smoothed by

$$\hat{\mu}[k] = (1-\alpha)\hat{\mu}[k-1] + \alpha\mu[k] \tag{6.14}$$

where $0 \leq \alpha \leq 1$ is a forgetting factor. The averaging delay can be estimated by finding the instant l where $\Delta\hat{\mu}[k+l] \geq 0$. From (6.1) and (6.2), we have

$$\Delta\mu[k] = \frac{K_2}{2}\log\frac{d_1^2}{d_2^2} \tag{6.15}$$

By the Pythagoras's theorem (see Figure 6.5) and replacing $H = \frac{D\sqrt{3}}{6}$, we have

$$\Delta\mu[k] = \frac{K_2}{2}\log\left(\frac{1+12(x/D)^2}{1+12[1-(x/D)]^2}\right) \tag{6.16}$$

For all values of separation distance D, x/D is a variable bounded between 0 and 1. $\Delta\mu[k]$ varies in function of $y = x/D$ rather then D itself. Hence, the averaging delay only depends on the forgetting factor α, i.e. $\delta_{av} = g(\alpha)\Delta T$ (for example $g(0.05) = 19$ or $g(0.01) = 84$). The equivalence between $N/2$ and $g(\alpha)$ in the averaging delay formula can be seen in the two averaging window cases. Where we do not distinguish between linear and exponential averaging processes, we use N_α as a common notation for both N and $g(\alpha)$.

Due to the uncertainty arising from fluctuations of the physical channel, the handover is initiated only if $\mu_B - \mu_A \geq h - \sigma$, where h is a hysteresis margin and σ is the standard deviation of the received signal. A hysteresis of $(h-\sigma)$ dB shifts the handover decision point from the instant when $\mu_B(d_2) - \mu_A(d_1) = 0$ (i.e. $d_1 = d_2$) to the instant when $\mu_B(d_2) - \mu_A(d_1) = h - \sigma$. This kind of delay is referred to as hysteresis delay δ_{hyst} and corresponds to the travelling distance Δd as illustrated in Figure 6.5. From

(6.1) and (6.2), the hysteresis margin is obtained as:

$$h - \sigma = K_2 \log \frac{d_1}{d_2} \tag{6.17}$$

Substituting $d_1^2 = H^2 + (D/2 + \Delta d)^2$, $d_2^2 = H^2 + (D/2 - \Delta d)^2$ and $H = \frac{D\sqrt{3}}{6}$ in (6.17), we have

$$\frac{D^2/3 + (D/2 + \Delta d)^2}{D^2/3 + (D/2 - \Delta d)^2} = 10^{\frac{2(h-\sigma)}{K_2}} \tag{6.18}$$

The value of Δd is obtained by solving this equation and eliminating the solution greater than D. The hysteresis delay δ_{hyst} is therefore given by:

$$\delta_{hyst} = \frac{\Delta d}{v} = \frac{D}{v} \frac{3(1+A) - \sqrt{3(-1+14A-A^2)}}{6(A-1)} \tag{6.19}$$

where $A = 10^{\frac{2(h-\sigma)}{K_2}}$. If we denote $C = \frac{3(1+A)-\sqrt{3(-1+14A-A^2)}}{6(A-1)}$, the total handover delay becomes:

$$\delta = \delta_{av} + \frac{D}{v} C \tag{6.20}$$

6.5.1.3 Overlap distance

The minimum cell overlap required for seamless handover corresponds to the maximal allowable value of separation distance D. From (6.12), (6.13) and (6.20), we have the following inequality:

$$2\sqrt{R^2 - \frac{D^2}{12}} \geq 2\delta_{av} v + (2C+1)D \tag{6.21}$$

Solving the inequality, we obtain

$$0 \leq D \leq \frac{\sqrt{48R^2(3C^2+3C+1) - 12\delta_{av}^2 v^2} - 6\delta_{av} v(2C+1)}{4(3C^2+3C+1)} \tag{6.22}$$

Therefore, the minimum required cell overlap is:

$$\begin{aligned} D_{co_{min}} &= 2R - \frac{\sqrt{48R^2(3C^2+3C+1) - 12\delta_{av}^2 v^2}}{4(3C^2+3C+1)} \\ &\quad + \frac{6\delta_{av} v(2C+1)}{4(3C^2+3C+1)} \qquad (6.23) \\ &\triangleq f(R, v, K_2, N_\alpha, \triangle T, \sigma, h) \qquad (6.24) \end{aligned}$$

where C is a function of h, σ and K_2. The value of $D_{co_{min}}$ given in (6.23) is the minimum required cell overlap corresponding to a specific set of environment parameters (K_2, σ), a specific handover measurement configuration (h, ΔT, N_α) and a terminal velocity v.

Network planning already considers a number of aspects like supported services, user density, network capacity and mobility support. This theoretical analysis gives network designers another means of ensuring seamless mobility. In all cases, minimum required cell overlap is a necessary condition

(not a sufficient condition). Seamless mobility cannot be achieved if the cell overlap is smaller than its minimum required value.

6.5.2 Overlap in heterogeneous UMTS-WiMAX networks

Handover between WLAN and 2G/3G is described as fully overlapping since the WLAN cell size is small compared to the cellular cell size [11, 160]. Cell overlap between UMTS and WiMAX has to be addressed since their cell sizes are similar. An overlapping model between UMTS and WiMAX cells is shown in Figure 6.6. The radio coverage of UMTS and WiMAX cells is assumed to be different. As the cell organization between heterogeneous and homogeneous cells is not the same, it is impossible to determine the minimum crossing distance as previously described. We therefore compute the required overlap distance when a UE moves along the straight line AB with constant velocity v. The crossing distance is PQ and the minimum required overlap distance between UMTS and WiMAX cells is $D_{co_{min}}(PQ)$. If the crossing distance is less than PQ (when the UE moves along a straight line other than AB), the required cell overlap becomes larger than $D_{co_{min}}(PQ)$. As a result, $D_{co_{min}}(PQ)$ is a lower bound of the minimum required cell overlap.

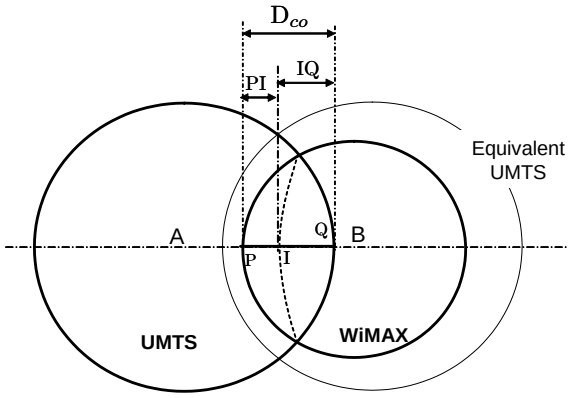

Figure 6.6: Required overlap between UMTS and WiMAX cells

When the UE moves along AB, $D_{co}(PQ) = PQ = PI + IQ$, where I is the point at which the signal quality level from the UMTS cell is equal to the level received from the WiMAX cell. If g_u is a function that can translate WiMAX signal strength to UMTS signal strength with equivalent quality, we have $g_u(\mu_B(I)) = \mu_A(I)$ ($\mu(I)$ is the received signal strength at the point I). Assuming that the UE moves from the UMTS cell to the WiMAX cell, the handover decision should be made before the UE goes beyond the point Q. Therefore, $g_u(\hat{\mu}_B(Q)) - \hat{\mu}_A(Q) \geq h_u - \sigma_u$ where $h_u - \sigma_u$ is the hysteresis margin in the UMTS handover decision. In fact, the travelling distance is IQ due to the decision delay caused by the UMTS to WiMAX inter-system measurement. If the WiMAX cell is replaced by an equivalent UMTS cell (see Figure 6.6), IQ can be considered to be half of the required overlap between two inter-frequency UMTS cells. This computation uses the UMTS to WiMAX inter-system measurement period instead of the inter-frequency measurement period: $IQ \geq \frac{1}{2} f(R_u, v, K_2^u, N_\alpha^u, \triangle T_{u\to w}, \sigma_u, h_u)$ (the subscript or superscript u means UMTS-related parameters). Similarly, PI should be larger than or equal to half of the minimum required overlap between two WiMAX cells, computed using a WiMAX to UMTS inter-system measurement period instead of a WiMAX measurement period. The required overlap between UMTS and WiMAX cells is:

$$\begin{aligned} D_{co}(PQ) &\geq \frac{1}{2} f(R_w, v, K_2^w, N_\alpha^w, \triangle T_{w\to u}, \sigma_w, h_w) \\ &+ \frac{1}{2} f(R_u, v, K_2^u, N_\alpha^u, \triangle T_{u\to w}, \sigma_u, h_u) \end{aligned} \tag{6.25}$$

where f is (6.24) and the subscript (superscript) u and w denote the UMTS-related and WiMAX-related parameters respectively. The lower bound of the minimum required overlap between UMTS and WiMAX cells is thus given by:

$$\begin{aligned} D_{co_{min}} &= \frac{1}{2} f(R_w, v, K_2^w, N_\alpha^w, \triangle T_{w\to u}, \sigma_w, h_w) \\ &\quad + \frac{1}{2} f(R_u, v, K_2^u, N_\alpha^u, \triangle T_{u\to w}, \sigma_u, h_u) \end{aligned} \tag{6.26}$$

In the numerical analysis in Section 6, the designated minimum required overlap between UMTS and WiMAX cells in fact means its lower bound value. This is as important in network planning as the minimum required cell overlap. If the overlap is less than this lower bound, the handover cannot be seamlessly achieved.

Table 6.2 summarizes the computation of minimum required cell overlap in four different cases: intra-frequency UMTS cells, inter-frequency UMTS cells, homogeneous WiMAX cells and heterogeneous UMTS-WiMAX cells. The second column indicates the formula used to compute the required minimum cell overlap value while the last column gives the corresponding measurement reporting period.

Case	$D_{co_{min}}$	Measurement report period
Intra-freq. UMTS	(6.23)	$\Delta T_{intra} = 200ms$
Inter-freq. UMTS	(6.23)	$\Delta T_{inter} \geq 480ms$
WiMAX-WiMAX	(6.23)	$\Delta T = 200ms$ (Section 6.4.2)
WiMAX-UMTS	(6.26)	$\Delta T_{w\to u} \geq 480ms$ (cf. (6.10))
		$\Delta T_{u\to w} \geq 600ms$ (cf. (6.11))

Table 6.2: Summary on the required cell overlap computation

6.6 Numerical analysis

This section considers the effects of the main parameters: (a) averaging window size N, (b) mobile terminal velocity v and (c) cell size R. The analyses are conducted for two cells of the same technology as well as two cells of different technologies. For simplicity, we discuss only the use of the linear averaging window. From (6.23), we see that the required cell overlap can be explicitly computed for the given values $R, v, K_2, \Delta T, N, h$ and σ. With a long measurement averaging window time, the standard deviation of the received signal σ is small and vice versa. This relationship between σ and $N, \Delta T$ is as follows [165]:

$$\sigma(N, \triangle T) = \sqrt{\frac{\sigma_s^2}{N} + \frac{\sigma_0^2}{N}[1 + 2\sum_{k=1}^{N}(1 - \frac{k}{N})\frac{sin(2\pi k f_m \triangle T)}{2\pi k f_m \triangle T}]} \tag{6.27}$$

where σ_0 is the standard deviation of the log-normal shadowing, σ_s is the standard deviation of the Rayleigh fading on each measurement report and f_m is the shadowing frequency. Since the signal strength in each measurement report has already been averaged over a number of measurement samples[1], the standard deviation of this averaged value is σ_s. We therefore prefer to replace σ by (6.27)

[1]The sample at instant k is modeled as $s[k] = K_1 - K_2 \log d[k] + u[k] + 20\log(e[k])$ where u is a zero-mean Gaussian

in order to highlight the effect of the averaging window of size N. In order to avoid unnecessary handovers, we select $h = 1.82\sigma$ [167]. Some of these parameters are taken from standards documents [4, 155, 172] and some from other research papers [163, 164, 175]. They are shown in Table 6.3.

Parameters	UMTS intra-frequency	UMTS inter-frequency	WiMAX
Frequency band [172] [4]	2GHz	2GHz	3.5GHz
Frame duration [172] [4]	10ms	10ms	5ms
σ_0 [163] [175]	10dB	10dB	9dB
f_m [167]	0.4Hz	0.4HzdB	0.4Hz
K_2 [163] [175] [164]	37.6dB	37.6dB	47dB
ΔT(cf. Table 6.2)	200ms	480ms	200ms

Table 6.3: Simulation parameters

6.6.1 Influence of the averaging window size on intra-system cell overlap

To investigate the effects of the averaging window size on the required cell overlap, we considered three scenarios: (1) two intra-frequency UMTS cells, (2) two inter-frequency UMTS cells and (3) two WiMAX cells. We assume that the UE moves at the speed of $v = 100km/h$ and the UMTS and WiMAX cell radius are $R_u = R_w = 4km$. For each scenario, the variation of the required cell overlap by averaging window size is shown in Figure 6.7. The required cell overlap is expressed as a ratio between the overlap distance and the cell diameter (i.e. $D_{co_{min}}/2R$).

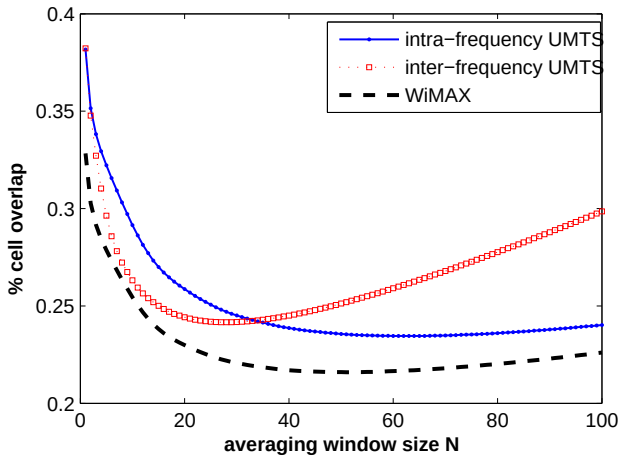

Figure 6.7: Required cell overlap area vs. averaging window size

Figure 6.7 shows that the minimum value of the required cell overlap between two adjacent intra-frequency UMTS cells is 23.45% of the cell diameter ($D_{co_{min}} = 1876m$). This corresponds to an averaging window size of $N = 63$ measurement reports (linear averaging window size $T = 12.6s$). However,

random variable with the standard deviation σ_0, e is a Rayleigh random variable and $d[k]$ is the distance between the UE and the BS.

N is specified in the handover algorithm of the UE in order to support the mobility management and optimize the handover procedure. For ease of notation, we use $D_{co_{min}}$ as its minimum value according to the variation of N. That is:

$$D_{co_{min}} = \min_{N} f(R, v, K_2, N, \triangle T) \tag{6.28}$$

With two adjacent UMTS cells operating in different frequencies, the minimum required cell overlap is 24.17% of the cell diameter when other parameters are unchanged. This is larger than between two intra-frequency cells because the inter-frequency scenario needs a longer measurement report period and therefore a higher averaging delay and larger cell overlap. Between two adjacent WiMAX cells, the required overlap area is around 21.6% of the cell diameter, smaller than for UMTS systems.

When the averaging window size N is small, the cell overlap needs to be large. When N is small, the standard deviation of the signal σ is significant, implying a large hysteresis delay. The hysteresis delay becomes an important factor in the overall handover delay and results in a large cell overlap distance. When the averaging window size increases and σ decreases, the overlap distance decreases until a minimum value is reached. If N still increases, the overlap distance then increases. In this case, the averaging delay is an important factor in the overall handover delay while the hysteresis delay remains stable. For all defined handover algorithms (the choice of the averaging window size and the hysteresis margin), network planners should therefore make sure that overlap distance between any two adjacent cells is not less than its $D_{co_{min}}$ if uninterrupted handovers are to occur.

6.6.2 Influence of averaging window size on inter-system cell overlap

The averaging window size used in the UMTS and WiMAX handover measurement process is shown respectively as N_u and N_w on the overlap area between cells. As each handover algorithm can be designed differently, the averaging window size can differ from one system to another. From Section 6.4, we assume that the WiMAX to UMTS measurement period is $\triangle T_{w \to u} = 480ms$ and the UMTS to WiMAX measurement period is $\triangle T_{u \to w} = 600ms$.

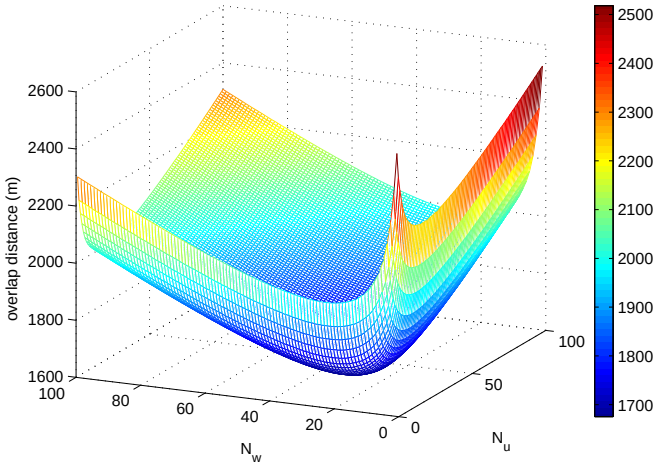

Figure 6.8: Required cell overlap area between UMTS and WiMAX cells vs. averaging window size

Considering that $R_u = 4km$, $R_w = 3km$ and $v = 100km/h$, the required cell overlap distance is illustrated in Figure 6.8. The minimum value is $D_{co_{min}} = 1770m$. This is 22.13% of the UMTS cell diameter or 29.50% of the WiMAX cell diameter. With $R_w = 4km$, the required cell overlap area is

equal to 23.40% of the UMTS/WiMAX cell diameter. For seamless handover, more overlap is required on the WiMAX side than on the UMTS side. The effect of the averaging window size parameter on cell overlap is similar with either homogeneous or heterogeneous cells.

The minimum required cell overlap depends significantly on the choice of the averaging window size. In practice, if the value of N (or α) is known, we only need to estimate the required cell overlap that corresponds to this value. In the following section, we consider that the minimum cell overlap is implicitly the minimum value with respect to the variation of the averaging window size.

6.6.3 Influence of velocity

The mobile terminal velocity has a direct impact on cell overlap. When the UE moves quickly, cell overlap must be large enough to allow the UE to perform the necessary measurements. Figure 6.9 shows simulation results that indicate the increase of the cell overlap with increasing mobile terminal velocity. Results are provided for three scenarios: inter-frequency UMTS, homogeneous WiMAX and UMTS-WiMAX interworking. The cell radius was fixed to $R_u = R_w = 4km$. In network planning, the required overlap area should be calculated from the maximum terminal speed that the system can support.

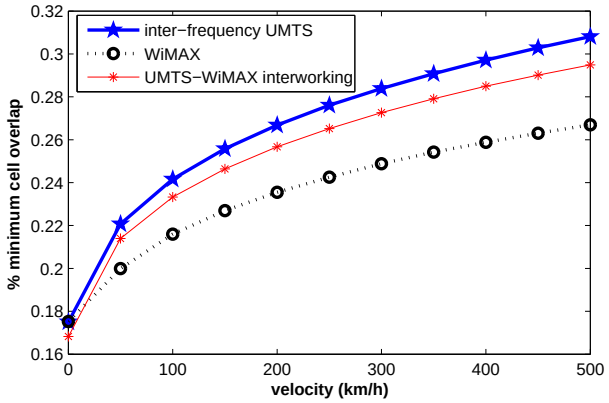

Figure 6.9: Required overlap area vs. mobile's velocity for $R_u = R_w = 4km$

From Figure 6.9, it can be seen that, even when the mobile terminal velocity is close to zero, the required cell overlap area is not less than 17% in all three scenarios. This is explained by the fact that the hysteresis delay still plays an important role in the handover decision. The hysteresis margin is used to avoid the ping-pong effect due to the log-normal shadowing fluctuation. This is directly related not to the movement but to the signal propagation. The cell overlap area is therefore still important for seamless handover even for low speed mobile users.

The same figure shows that a minimum overlap area of about 30% of the cell diameter is required for a maximal velocity of $500km/h$ in the inter-frequency UMTS network. Similarly, the overlap area in the WiMAX system is about 26.7%. In the cellular system, overlap of around 20%-30% is normal [2]. So the range value of cell overlap to support an extremely high terminal velocity in the UMTS and WiMAX system is still acceptable.

[2] See http://www.umtsworld.com/technology/coverage.htm

6.6.4 Influence of cell size

The radio coverage of a cell depends on radio conditions, the transmission power of its base station, the data rates of supportable applications, and the number of connected users. In the future, wireless mobile networks will support applications with very high data rates. As there will always be a tradeoff between cell range and transmission data rate, high-speed communication is possible only for the radio coverage around the BS. This means that the cell sizes will be reduced. In order to avoid service degradation and to maintain seamless service delivery, the required overlap between the adjacent cells must take into account the influence of reduced cell sizes.

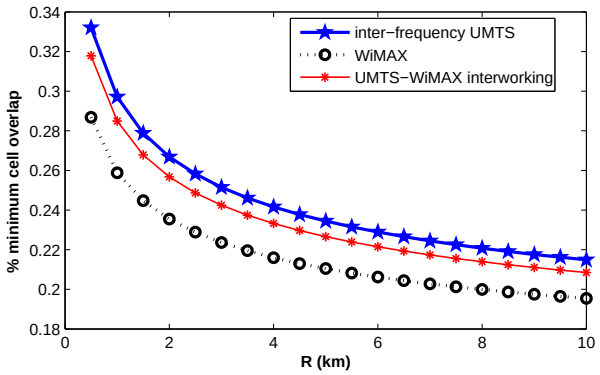

Figure 6.10: Required cell overlap area vs. cell radius for $v = 100km/h$

Figure 6.10 shows the variation of the required cell overlap according to cell radius for three cases: inter-frequency UMTS, homogeneous WiMAX and UMTS-WiMAX interworking. The required cell overlap area decreases with the cell size. This does not mean that the overlap distance decreases; the decrease is in the ratio between the overlap distance and the cell diameter. When the cell size is $R = 500m$, the cell overlap area is around 29% for the WiMAX system or 33% for the inter-frequency UMTS system. For a high speed of $500km/h$ and $R = 500m$, the required overlap can reach more than 40% (42% for inter-frequency UMTS case). In short, required cell overlap must be calculated for the smallest allowable reduced cell size. Not surprisingly, the handover decision based on the received signal measurement does not scale well for highly mobile terminals crossing very small cells. The handover algorithm for small-cell systems (such as picocells or microcells) must be different from the macro cell system. A dedicated handover management solution should also be designed for extremely high speed terminals.

6.6.5 Integration of parameters

All the above parameters influence the minimum required cell overlap area. The maximum allowable movement velocity, the minimum reduced cell size and the most suitable averaging window size (the size that which gives the minimum cell overlap value) are then considered together. The results are shown in Table 6.4.

We chose a maximum supportable velocity of 200km/h for the UMTS system and $100km/h$ for the WiMAX or UMTS-WiMAX interworking system. The results show that, for data-intensive and high speed mobile users, the minimum required overlap in intra-frequency UMTS systems must be at least 35.5% of the cell diameter in order to guarantee seamless handover. The cell overlap area in WiMAX systems (at least 28.7%) is less than in UMTS systems and in UMTS-WiMAX interworking systems. The inter-frequency UMTS organization requires the largest overlap (37.4%).

	UMTS intra-frequency	UMTS inter-frequency	WiMAX	UMTS-WiMAX
$v(km/h)$	200	200	100	100
$R(km)$	0.5	0.5	0.5	0.5
$D_{co_{min}}(\%)$	**35.5**	**37.4**	**28.7**	**31.8**

Table 6.4: Overlapping area ratio

Our analysis shows that the required cell overlap for usual values of the mobile terminal velocity and the cell radius lies mostly between 20%-30% of the cell diameter. For extreme cases, such as very high movement velocity or significantly reduced cell size, the cell overlap must be considerably larger, around 40% of the cell diameter. The results also show that the required cell overlap is never less than 17%. In fact, the required overlap between adjacent UMTS and WiMAX cells is about the same as between two UMTS or WiMAX cells.

Minimum required cell overlap is a useful tool in network planning. It can be used in the detailed planning phase for seamless handover between any two planned adjacent cell sites. It can also be used to evaluate the seamless mobility support of an existing network and to determine where additional cell sites are needed. For example, an operator willing to deploy WiMAX cells in addition to an existing UMTS network to offer higher data applications or to support higher speed users can use our solution to calculate the minimum overlap condition during the nominal and detailed planning phases.

6.6.6 Use cases

In this section we present several examples where our solution can facilitate and assist the network planning task. As a first example of use, the minimum required cell overlap condition can be used to check the seamless handover possibility between any two planned adjacent cell sites.

Second, we consider a UMTS network upgrade scenario. Assuming that the cell radius of the initial UMTS system is $1km$. Using the environment parameters in Table 6.3, the minimum required overlap between two intra-frequency UMTS cells is $550m$ (i.e., the maximal allowable separation distance is $D = 2R - D_{co_{min}} = 1450m$) to support the maximum movement speed of $80km/h$. In order to offer high-data applications, to support a large number of subscribers or both, the cell radius is thus reduced to, for example, $600m$. An operator has a choice between UMTS and WiMAX to upgrade the network. As the handover between UMTS and WiMAX can be achieved seamlessly; the choice of WiMAX is more interesting in terms of QoS and building cost.

Assume that the cell radius of WiMAX is the same magnitude order as that of UMTS ($R_w = 600m$). Given the maximum movement speed of $80km/h$, the maximal allowable separation distance between UMTS and WiMAX cells is about $845m$, and that between two WiMAX (or intra-frequency UMTS) cells is $875m$ (or $840m$ respectively). Taking into account this maximal allowable separation distance condition, network designers can determine the prior positions as well as the number of added WiMAX cells. Figure 6.11 illustrates the position of the added WiMAX cells, taking into account the minimum required overlap condition with respect to the user movement speed. More precisely, in high-speed movement regions (main streets), the separation distance between two adjacent cells should not be greater than the maximal allowable separation distance to ensure seamless mobility. For example, to achieve seamless mobility between U_1 and U_2, we added a WiMAX cell W_1. In the low or zero-speed movement region (buildings), a small overlap distance is sufficient (e.g., overlap between W_3 and U_6).

Last but not least, we consider the case where an existing UMTS network needs to be upgraded to ensure service continuity for users on newly deployed high-speed trains (movement speed of $300km/h$).

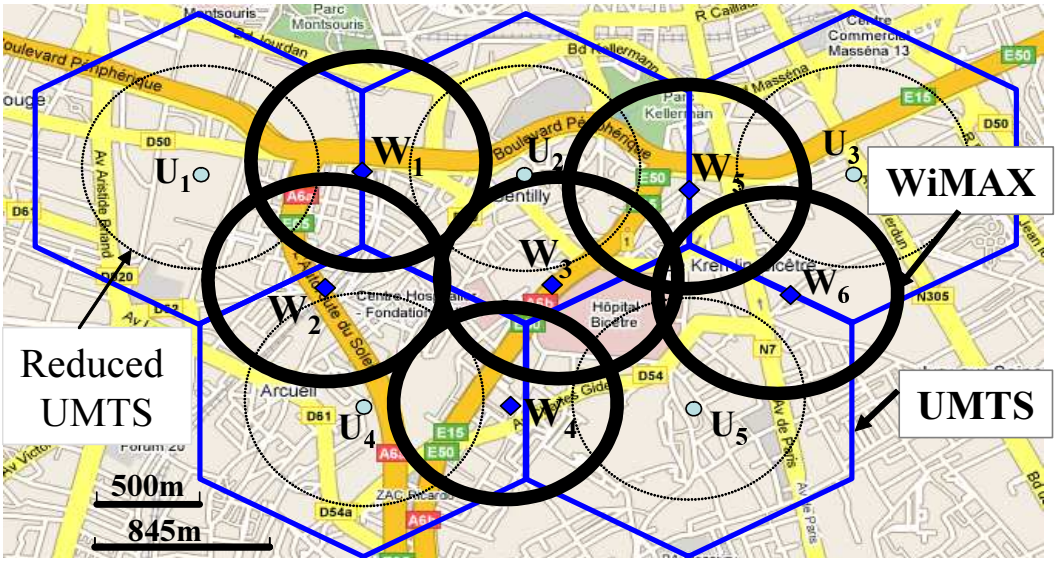

Figure 6.11: Network upgrade scenario 1

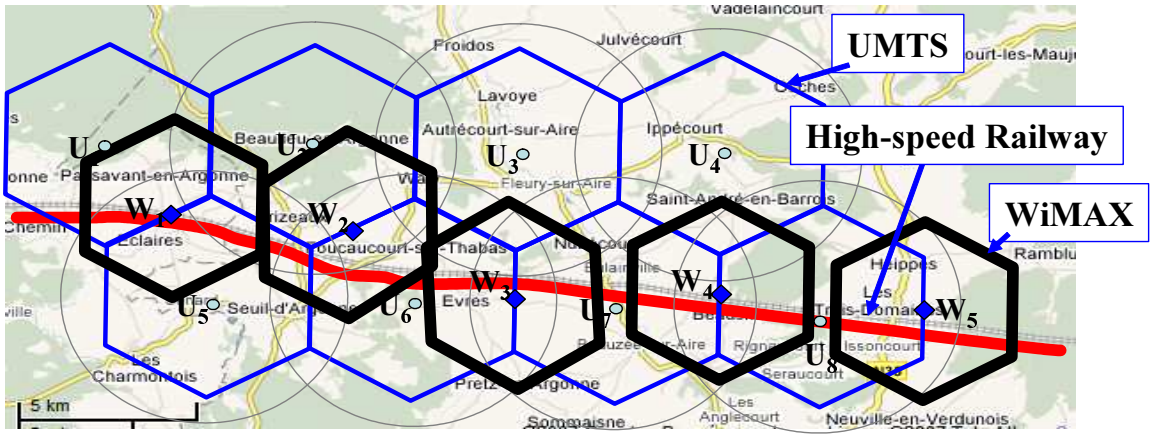

Figure 6.12: Network upgrade scenario 2

Assuming that the UMTS cell radius is 5km. Before the arrival of high-speed railways, the maximum movement speed is 100km/h and thus the maximal allowable separation distance is 7700m. To ensure seamless handover for a movement speed of 300km/h, the separation distance should not be greater than 7300m. Accordingly, the operator can deploy new WiMAX cells along the railway as illustrated in Figure 6.12. Given the WiMAX cell radius is 4km, a separation distance between UMTS and WiMAX should be less than 6600m to ensure seamless mobility. Considering this condition, a possible organization of new added WiMAX sites is determined as in Figure 6.12. For example, to enable the seamless handover between U_1 and U_5 cells, a WiMAX cell W_1 is added.

6.7 Summary

In order to achieve the seamless vertical handover between UMTS and WiMAX systems, we addressed in this chapter two important aspects: inter-system measurement and required cell overlap. First, we demonstrated that it is possible to make the UMTS-WiMAX inter-system measurement

through a single reconfigurable radio interface terminal without affecting on-going sessions. The inter-system measurement scheme was introduced and the measurement report period was discussed. We also studied the measurement reporting period for the signal strength measurement in the WiMAX system. Second, we analyzed the minimum cell overlap required to support seamless handovers between two adjacent cells within the same technology or different technologies. More precisely, we focused on the required overlap area between any two adjacent cells (UMTS - UMTS, WiMAX - WiMAX, and UMTS - WiMAX). An analytical method of estimating the minimum required overlap between two adjacent cells of the same or different technologies was proposed. The minimum required cell overlap is computed based on measurement and handover parameters, physical channel characteristics and mobile terminal velocity. The various parameters influencing cell overlap were analyzed to show how the required overlap can be computed and used to ensure seamless handover as the network is being planning. This method provides network designers with a useful tool with which to provide seamless handover as networks are planned or upgraded.

∼∼ △♡△ ∼∼

Chapter 7

Load Balancing over Heterogeneous Wireless Packet Networks

Load balancing involves the fine tuning of a computer system, network or disk subsystem in order to more evenly distribute the data and/or processing across available resources.

Computer Desktop Encyclopedia

The ability to balance the load between different access systems is one of the essential motivations to integrate different access technologies in a converged 4G network. But, it is also a big challenge to balance the load effectively. Load balancing is an aspect that requires a strong control from the network side. We address the load balancing in this thesis since it is greatly related to the mobility management. Firstly, the load of an access node depends on the access network selection. If many users select the same access node, this access node probably becomes overloaded. Secondly, to balance the load, the network has to force some users to carry out the handover. In this chapter, we propose a new approach to compute the load which can hide the heterogeneity of different access technologies from the load balancer. The objective of load balancing is also redefined to improve the overall network performance and to minimize unnecessary load balancing operations.

7.1 Introduction and motivation

Along with the rapid growth in demand for high data rate and high QoS multimedia communications as well as the scarcity of radio resources, an efficient Radio Resource Management (RRM) scheme is highly required. An operator can deploy different technologies or interwork with other technologies owned by other operators to enable the global roaming capability through a coordinated heterogeneous access environment as discussed in Chapter 5. An advanced Common Radio Resource Management (CRRM) is a motivation for interworking among these networks and also a challenge to overcome. The stronger the coupling between access networks, the more efficient the resources can be commonly utilized. We consider the tight coupling approach where different access technologies are deployed by a single operator or by cooperative operators. Available radio resources of coupled networks are jointly managed. We suggest adopting the CRRM architecture introduced by 3GPP [176] and further used in [177–182]. CRRM is defined as a platform to gather information from the BS of different access systems, and to control the resource allocation of all BSs to optimize the overall system performance.

In the joint RRM research area, most of previous work mainly focused on identifying the functionalities of the CRRM architectural components and designing protocols for control exchanges between these components [176–178, 183]. Also, the resource allocation scheme which aims at determining the amount of resources allocated to each user in such a way to maximize the operators' revenue (briefly addressed in Section 6.3 of Chapter 2) or the users' satisfaction has been increasingly studied [184–186]. However, the load balancing between different BSs and different access technologies has not been sufficiently considered. In fact, the load balancing plays an important role in the CRRM. The load balancing consists of accepting or denying a new user request and enforcing users connected to a heavily loaded BS to handover to a lightly loaded one. Although the load balancing is much related to the resource allocation, they are two separable aspects. The load balancing can be considered on the one hand as an objective of the resource allocation scheme and on the other hand as a constraint for the resource allocation optimization. In this work, we only focus on the load balancing issue.

An adaptive threshold for load balancing based handover enforcement initiation was introduced in [187]. Although this approach makes it possible to detect the need of initiating a handover, the suitable target access network is not addressed. Another solution for RRM algorithm based on fuzzy logic and reinforcement learning was presented in [180] [181]. However, the admission control is just an initial step in the load balancing process as it only deals with the arriving communications. Even if an efficient admission control algorithm [180, 181, 188] is used, overload situations might still occur due to the mobility of high-rate data users or the inherent fluctuation of the transmission channel.

All the load balancing solutions have been based on a fundamental resource unit notion, called *load*. The load metric represents the occupation ratio of a BS. The load of a cellular network is usually computed through the received power and the interference level [189] whereas the load of a WLAN is simply computed through the number of users connected to an AP [180] [181]. The load can be computed in different manners for different systems. As a result, the same load value for two different systems does not mean the same load situation. As such a comparison is the basis of any cross-system load balancing solution, having a same semantic of the load metric is mandatory. Furthermore, the existing load computation methods, which are based on the interference [189] or the throughput [190], do not allow the load variation anticipation prior to the situation where a user moves into/out of a cell. In fact, the estimation of future interference or throughput values is really challenging. Accordingly, we cannot be able to make the right decision to achieve an efficient resource balancing.

In this chapter, we introduce a new approach to quantify the load in wireless packet networks and a novel load balancing algorithm to improve the above limitations. A new load metric and new balancing index are proposed in Section 7.2. Section 7.3 is devoted to discuss the load balancing algorithm.

7.2 Load metric & balancing index

7.2.1 Load metric definition

Along with the increase of multimedia and data-intensive applications, future 4G networks will experience an extremely high load. We present here only the cross-system downlink load balancing. However, the solution is still valid for uplink load balancing. Traditionally, the load metric corresponding to the resource occupation ratio varies from 0 to 1. In wireless packet networks, the channel access is dynamically assigned to mobile users by a scheduler [191] running in the BS (see Figure 7.1). The scheduler decides which packets are transmitted to their corresponding destinations at an instant (depending on the required QoS of each user and radio link quality between the user and the BS). Contrary to a fixed resource allocation in circuit networks, the resource allocation in packet networks is much more dynamic. An overload situation will cause a delay or packet loss to some specific connections,

but not necessarily an outage of connections. It is thus interesting to be able to estimate the overload degree. The way to balance the load in packet networks is thus different from circuit networks.

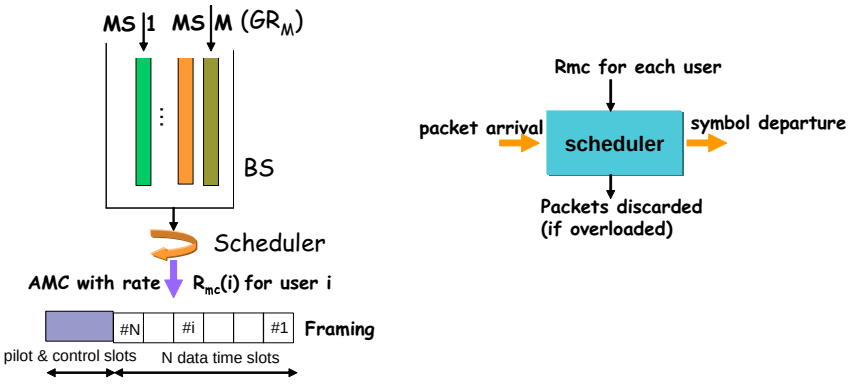

Figure 7.1: Scheduler in a base station

The packet scheduling is an active research area. Generally, based on the transmission channel estimation, the BS will adapt the modulation and coding scheme to transmit packets in such a way to maximize the throughput and minimize the packet error rate. Recently, the QoS priority has been also taken into account in the packet scheduling [191]. Compared to the load balancing, a global strategy involving all the BSs in the system, the packet scheduling is just a local strategy at each individual BS. We can see that if the total requesting resource (i.e., packet arrival rate mapped with the modulation and coding rate) is higher than the capacity of the BS (i.e., symbol departure rate at physical layer), some users will not get their required QoS. In other words, the BS is overloaded.

We define load ρ as the ratio of the required resources to the total resources. If the amount of the required resources of all users connected to a BS is greater than or equal to its total resources, this BS is considered as overloaded. In differentiated QoS wireless networks, the objective of the scheduler is to guarantee the QoS required by the non-best-effort users. Hence, the required resources information used for load computation is the guaranteed bit rate η required by running applications of non-best-effort users. Alternatively, the required resources of a communication is its data arrival rate at the BS. As a First In First Out (FIFO) buffer is implemented at the BS for each connection, the packet arrival rate can be easily retrieved. In the following, for simplicity, each communication is assumed to have a guaranteed bit rate $\eta(Kbps)$. At the physical layer, multiple transmission modes comprising of a pair of modulation scheme and Forward Error Control (FEC), as in IEEE802.11/15/16, 3GPP and 3GPP2 standards, are available to each user. Given the modulation and coding rate of $R_{mc}(bits/symbol)$, the packet of N_p bits is mapped to a block of N_p/R_{mc} symbols after modulated and coded. Hence, the required resources of a communication can be expressed as $\frac{\eta}{R_{mc}}$ $(Ksymbol/s)$.

The total resources of a BS can be referred to as the number of data symbols that the BS can transmit in downlink during one second, i.e., data symbol rate R_s. For example, in HSDPA system [192], the channel multiplexing is in time domain where each Transmission Time Interval (TTI) consisting of three slots (or 2ms) can carry 480 data symbols. Within each TTI, a maximum of 15 parallel codes can be assigned to one user or shared between several ones. Hence, the total resources become $15 \times 480 symbols/(2ms) = 3.6 Msymbols/s$. In an OFDM system like WiMAX, 3G LTE, or IEEE802.11a/g, the resources consist of OFDM symbols in time domain and sub-carriers in frequency domain. The downlink data symbol rate is equal to *(number of downlink OFDM symbols)*×*(number of data sub-carriers)/(frame duration*). Meanwhile, in the direct-sequence CDMA system like UMTS, CDMA2000 or IEEE 802.11b, the symbol rate depends on the spreading factor[1] $SF(chips/symbol)$

[1]Direct Sequence spreading process is done by directly combining the baseband information to high chip rate binary code.

of the used code. For instance, the symbol rate of 802.11b corresponding to the use of Baker code or Complementary Code Keying (CCK) is equal to $1Msymbol/s$ or $1.375Msymbol/s$ respectively [193]. As the chip rate $R_c = R_s \times SF$ ($chip/s$) is a fixed quantity, we choose it as the total resources parameter.

Now let M denote the number of currently connected users at a BS of total resource R_c. Each user i is characterized by a required guaranteed bit rate η_i, a modulation and coding rate R^i_{mc} and an associated spreading factor SF_i. If SF does not exist, we set $SF_i = 1$. The load of a BS is thus given by:

$$\rho = \frac{1}{R_c} \sum_{i=1}^{M} \frac{\eta_i SF_i}{R^i_{mc}} \tag{7.1}$$

This load metric definition takes into account not only the user's required resource but also the radio link quality between the user and the BS. If the link quality is so poor to guarantee the connection or the user is outside the corresponding BS's radio coverage, the corresponding modulation and coding rate R_{mc} will be set to 0. If the BS accepts this user request, its load becomes infinity. Thus, the load balancing algorithm should refuse the connection or force this user to handover to another neighboring access network. Using this definition, the resources heterogeneity among different access systems will be hidden from the load balancing problem. In other words, the balancing scheme is based only on the load values of different access nodes regardless of underlying technologies and underlying scheduling schemes. The load balancing over heterogeneous networks is somewhat similar to that over a homogeneous network.

7.2.2 Load balancing index

One of the key elements in the load balancing is the balance index used to measure the balance of resources in a system. Such an index was first introduced in [194] and recently used in [190]. It is defined as:

$$\xi_1 = \frac{(\sum_i \rho_i)^2}{K \sum_i \rho_i^2} \tag{7.2}$$

where K is the number of neighboring BSs over which the load can be distributed. In fact, ξ_1 is a correlation factor between the load vector $[\rho_1, \ldots, \rho_K]$ and the vector $[1, \ldots, 1]$. If all BSs have the same load level, then $\xi_1 = 1$. The load balancing target is to maximize ξ_1. However, this balance index has serious limitations. Consider a scenario where a new user at the overlapped zone of three BSs as depicted in Figure 7.2(a) wants to initiate a communication. Given that $\{\rho_A = 0.8, \rho_B = 0.4, \rho_C = 0.3\}$ are the current load of BS A, B and C respectively and $\{\Delta\rho_A = 0.1, \Delta\rho_B = 0.2, \Delta\rho_C = 0.7\}$ are the added load if the new user attaches to BS A, B and C, respectively. By using objective function ξ_1, the new user attaches to BS C as it results in the highest balance index $\xi_1 = 0.89$. Unfortunately, the BS C becomes overloaded ($\rho_C = 1$). As the same connection will generate different added loads when connecting to different access nodes, it becomes difficult to maintain all BSs at the same load value. Also, in a heavily loaded system, the balancing objective ξ_1 tries to evenly distribute the load to all BSs, which leads to a situation where all BSs will be overloaded. It may be better to degrade the QoS of only a set of users instead of all users. When the load between the BSs has not been balanced yet but all the BSs are not in the imminent overloaded situation, it is not necessary to maximize ξ_1 by forcing the users to attach to another BS. To resolve the overload situation in the exemplary scenario, one may suggest adding a constraint like $\rho_i < 1 \; \forall i$ while trying to maximize ξ_1. It seems to be a good solution in a lightly loaded system. But, this constraint is never satisfied in a heavily loaded system. In fact, the objective of load balancing algorithm is to minimize the effect of overload situation and not to avoid the overload situation (because it is not always guaranteed in a finite capacity system).

The Spreading Factor is the ratio of the chips (UMTS = 3.84Mchips/s) to baseband information rate.

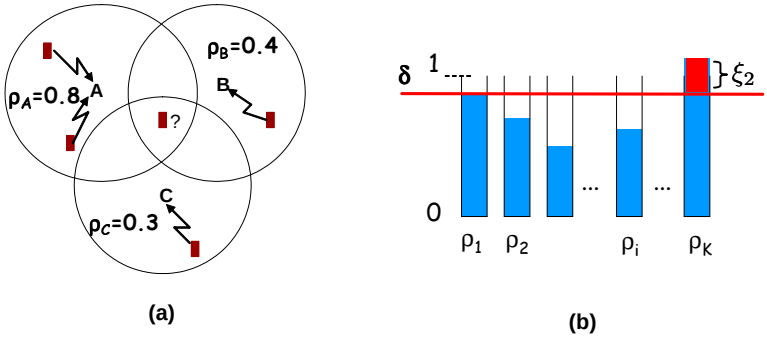

Figure 7.2: (a) Problem of using ξ_1; (b) Load balancing index ξ_2 computation

In order to improve the revealed limitations, the objective of our load balancing scheme is to reduce overloading situation in access networks. The idea is to detect imminent overload situations and start to redistribute the load from heavily loaded access nodes to lightly loaded ones. A system is considered as load-balanced if all BSs have a load below a specific threshold $0 < \delta < 1$. It is motivated by the avoidance of unnecessary load balancing operations which waste the resource and cause undesired handover overheads. Usually, in a load control strategy, operators reserve an amount of resources $(1-\delta)$, known also as *guard channel* ratio, for handing over users as well as for system redundancy. The choice of threshold δ can be inspired by the research on guard channel optimization in [195] and we do not address such a choice in this work. Accordingly, we propose a new balance index ξ_2 (see the illustration in Figure 7.2(b)):

$$\xi_2 = \sum_{i=1}^{K} (\rho_i - \delta)^+ \tag{7.3}$$

where $a^+ \triangleq \max(a, 0)$. If $\exists \rho_i > \delta$, then $\xi_2 > 0$. The greater index ξ_2, the closer to an overload situation the network is. Note however that $\xi_2 > 0$ does not mean an overload situation since ξ_2 may be greater than 0 but $\rho_i < 1 \forall i$. The objective of the load balancing is now to minimize ξ_2. In the previous scenario, the overload situation does not occur while using ξ_2 as an objective function since $\xi_2(C)$ (that is the value of ξ_2 while network C is selected) is clearly greater than $\max\{\xi_2(A), \xi_2(B)\}$ for any chosen δ. In fact, we have $\xi_2(C) = 1-\delta$, $\xi_2(B) = (0.6-\delta)^+$ and $\xi_2(A) = (0.9-\delta)^+$.

7.3 Load balancing algorithm

7.3.1 Optimal algorithm

7.3.1.1 Optimization formulation

We provide here a formulation of an optimal load balancing algorithm. Assuming that at a given instant our system consists of M currently connected users and K BSs of different access technologies. Let us denote $\mathcal{W} = (w_{ij})$, $i = 1..M$, $j = 1..K$ as a generated load matrix where w_{ij} is the load generated at BS j while user i attaches to it. If user i is not in the radio coverage of BS j, then $w_{ij} = \infty$. The balancing algorithm will be triggered upon the imminent overload situation. Results of the algorithm should come out with an assignment $\sigma = (\sigma_{ij})$, where $\sigma_{ij} = 1$ if user i is assigned to attach to BS j and $\sigma_{ij} = 0$ otherwise. The assignments $\sigma^\star$ are given as

$$\sigma^\star = \arg\min_{\sigma} \sum_{j=1}^{K} (\rho_j - \delta)^+ \text{ where } \rho_j = \sum_{i=1}^{M} w_{ij}\sigma_{ij} \tag{7.4}$$

subject to the following conditions: $\sigma_{ij} = 0$ if $w_{ij} = \infty$ and only one element σ_{ij} in each row i of matrix σ is non-zero. We assume that if user MS i is in coverage of a particular BS then MS i will be allocated the resource ($\exists j : \sigma_{ij} = 1$). In other words,

$$\exists j : w_{ij} \neq \infty \Rightarrow \sum_{j=1}^{K} \sigma_{ij} 1_{\{w_{ij} \neq \infty\}} = 1 \tag{7.5}$$

One may note that the constraint on binary integer variables σ_{ij} makes our optimization problem non-convex, and therefore far more difficult to solve. In the worst case where any user can connect to any BS, by using potentially exhaustive search, we need to compute the values of ξ_2 for K^M possibilities of σ to find out $\sigma^\star$. The optimal algorithm is impractical for implementation since it requires an exponential computation time, especially in a large wireless network with thousands of users and BSs. Also, such an assignment may lead to a reallocation of resources for all users which implies a significant amount of handovers and overheads.

In a lightly loaded system, there may exist many possible solutions $\sigma^\star$ for problem (7.4). In this case, even if the system is load-balanced, the resource utilization may not be optimized. We therefore define a second optimization objective based on the total user satisfaction index to enhance the load balancing operations.

Assume that each user MS i requires a guaranteed bandwidth η_i and the SNR of the radio link between MS i and BS j is γ_{ij}. Inspired by [82], the achievable throughput of MS i if it connects to BS j can be estimated as

$$T_{ij} = \frac{\eta_i}{\rho_j} g(\gamma_{ij}) = \frac{\eta_i}{\rho_j} \frac{L}{M} [1 - 0.5 \exp(-v\gamma_{ij})]^M \tag{7.6}$$

where M is the block size, L is the number of data bit within the block size M and v is the specified constant depending on the considered technology and $\rho_j = \sum_{i=1}^{M} w_{ij}\sigma_{ij}$ is the load of BS j. In fact, $g(\gamma_{ij})$ is the probability that the radio frame of size M is transmitted without errors. And $\frac{\eta_i}{\rho_j}$ represents the achievable data rate if user MS i connects to BS j. Shortly, for a given vector η, given SNR matrix (γ_{ij}), given elementary load matrix (w_{ij}) and a selected assignment σ, we can easily deduce the achievable throughput matrix (T_{ij}). According to a particular assignment σ, we can rewrite the achievable throughput of MS i as follows:

$$T_i(\sigma) = \sum_j \frac{\eta_i}{\rho_j} g(\gamma_{ij}) \mathbf{1}_{\{\sigma_{ij}=1\}} \tag{7.7}$$

Inspired by the utility function form proposed in Chapter 2, we use the modified Sigmoid form to model the user satisfaction degree based on its estimated achievable throughput:

$$u_i(T_i) = \begin{cases} \frac{(\frac{T_i - \eta_i^{min}}{\eta_i - \eta_i^{min}})^\zeta}{1 + (\frac{T_i - \eta_i^{min}}{\eta_i - \eta_i^{min}})^\zeta} & T_i \geq \eta_i^{min} \\ 0 & \text{otherwise} \end{cases} \tag{7.8}$$

where η_i^{min} is the minimum acceptable bandwidth threshold of MS i. The parameter ζ is the tuned steepness parameter that follows $\zeta \geq 2$. As a result, the second optimization objective is to find out an assignment $\sigma^\dagger$ that maximizes the total user satisfaction:

$$\sigma^{\dagger} = \arg\max_{\sigma^{\star}} \sum_{i=1}^{M} u_i(T_i(\sigma^{\star})) \tag{7.9}$$

where $\sigma^{\star}$ are the assignments that satisfy the load balancing condition (7.4). Obviously, the second optimization objective is used only when the load balancing index $\xi_2 = 0$ is obtained and more than one assignment $\sigma^{\star}$ are found.

7.3.1.2 Illustration example

Let us consider the following networks composed of 3 BSs and 6 MSs (as depicted in Figure 7.3).

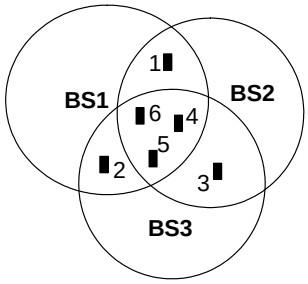

Figure 7.3: An illustration example

The generated load matrix of 6 MSs are assumed as follows:

$$W = \begin{pmatrix} 0.4 & 0.3 & \infty \\ 0.5 & \infty & 0.6 \\ \infty & 0.3 & 0.5 \\ 0.2 & 0.3 & 0.4 \\ 0.3 & 0.2 & 0.4 \\ 0.4 & 0.2 & 0.3 \end{pmatrix}$$

Recall that $w_{13} = \infty$ for instance indicates that the MS1 is not in the radio coverage of BS3. Given $\delta = 0.95$ (the threshold in the load balancing index), there are 80 different solutions $\sigma^{\star}$ that satisfy the load balancing condition (7.4). These assignments are found using Matlab. Here is one of these solutions:

$$\sigma_1^{\star} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

In this solution, the MS1 and MS2 connect to BS1; MS3 and MS4 connect to BS3; MS5 and MS6 connect to BS2. In this case, the load of the three BSs is $\rho = [0.9 \quad 0.4 \quad 0.9]$ and clearly $\xi_2 = 0$. Now, we assume that $\eta = \begin{bmatrix} 200 & 200 & 200 & 200 & 200 & 200 \end{bmatrix}$ is the required guaranteed bandwidth vector of the six users. For each solution, we compute the achievable bandwidth T_i for each MS (assuming that $g(\gamma) = 1$, $\eta_i^{min} = 0$ and $\zeta = 3$). According to the assignment $\sigma_1^{\star}$, we have $T = \begin{bmatrix} 222 & 222 & 222 & 222 & 500 & 500 \end{bmatrix}$ and the total user satisfaction of the six MSs is $U = 4.193$.

Taking into account the second optimization objective, the best assignment that maximizes the total user satisfaction is:

$$\sigma^{\dagger} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

This assignment results in a total user satisfaction level of $U = 4.701$ (the achievable throughput is T=[333.33 333.33 285.71 333.33 285.71 285.71]). This example shows how the optimal solution can be used to distribute and balance the load over a lightly loaded system to reach the best resource utilization. Though the second optimization does not require much computation time, the complexity of the first optimization remains high and thus hard to solve.

7.3.2 Proposed on-line load balancing algorithm

Our aim is to design a feasible and suboptimal solution for load balancing to minimize the resource rearrangement and the computation effort. When a user initiates a connection, its terminal selects a suitable access network among available ones using the network selection mechanism given in Chapter 2 and 3. The load value of each access node will be used in the network selection evaluation if the terminal has access to this information. The user will be able to not select the heavily loaded access node. Otherwise, the access node may refuse the user's connection request based on its admission control policy if it is heavily loaded. Despite the use of an admission control, the overload of an access node may still happen due to the transmission channel fluctuation or user's volume data rate changes. To handle the load balancing, on-going communications will be transferred from an access network to another by triggering a handover. The two main targets of our proposed algorithm are the admission control and the network-initiated handover.

7.3.2.1 Admission control

The admission control is employed to admit or reject a new originating communication in order to avoid overload situations. A connection request to a specific BS will be accepted if the BS's load, including the contribution of the incoming communication, is below an admission threshold δ_{AC}, that is $\rho \leq \delta_{AC}$. Otherwise, the new incoming communication will be redirected to the least loaded overlapped access network. If all BSs in the coverage area could not accommodate the new communication, the connection request is rejected. If the incoming communication is a handing-over one, the admission threshold is greater than the one used for a new originating communication. It is generally preferable to allocate resources to on-going communications rather than to new initiating ones. In our solution, we propose to always accept the handing-over users.

It is noteworthy that a number of previous publications [180, 181] [188] have considered the admission control as a means to achieve load balancing. However, the admission control is just a first step in the load balancing process as it only deals with incoming communications and it does not treat the load fluctuation of on-going ones. Moreover, trying to redirect an originating communication to a less loaded access system (redirect from one technology to another) may not be possible if the communication is initiated from a single-mode terminal. In this case, it may be better to force a multi-mode user to make a vertical handover to a coordinated access system and accommodate the originating single-mode user. That motivates the need to use handover enforcement to effectively distribute the load over the heterogeneous systems.

7.3.2.2 Handover enforcement

In addition to the admission control, it is essential to have a mechanism to detect and handle imminent overload situations. Such a mechanism is known as a handover enforcement since its main role is to select *suitable* users in a heavily loaded access network and force them to handover to *suitable* lightly loaded overlapped ones. The main output of the handover enforcement is to determine a set of pairs, *suitable* user and *suitable* target access network, for the handover execution.

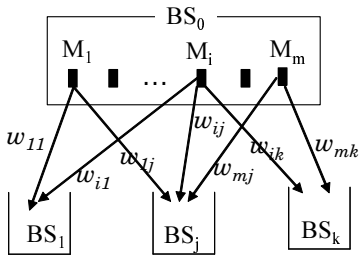

Figure 7.4: Illustration of load balancing algorithm

Let $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$ denote a set of mobile users currently connected to a heavily loaded BS_0 that needs to be unloaded. The load of BS_0 is $\rho_0 > \delta$. The set of neighboring BSs overlapped with BS_0 is denoted by $\mathcal{B} = \{BS_1, BS_2 \ldots, BS_k\}$ and the current load of each of the neighboring BSs is $\wp = \{\rho_1, ..., \rho_k\}$. The load balancing model is illustrated in Figure 7.4. While the load of BS_0 is still greater than δ and the load balance index ξ_2 can still be decreased, then identify a pair (I, J) of suitable user and suitable BS for load balancing handover. (I, J) is given by

$$(I, J) = \arg\min_{(i,j)} \xi_2(i, j) \tag{7.10}$$

$$\xi_2(i, j) = (\rho_0 - w_{i0} - \delta)^+ + (\rho_j + w_{ij} - \delta)^+ + \sum_{l \neq \{0,j\}} (\rho_l - \delta)^+ \tag{7.11}$$

where w_{ij} is the load contribution of user M_i at BS_j while M_i connects to BS_j. Also, $w_{ij} = \infty$ if M_i is not in the radio coverage of BS_j. The proposed algorithm to achieve this goal is as follows:

If $\rho_0 > \delta$ *then*
  Initiate k:=0 and cond:=true
  Calculate $\xi_2[k]$
  While $\xi_2[k] \neq 0$ *& cond=true do*
    Find (I,J) that minimizes $\xi_2(i, j)$
    k:=k+1
    If $\xi_2[k] \leq \xi_2[k-1]$ *then*
      cond:=false
    end
  end
end

Instead of balancing the resources of the overall system as described in the optimal algorithm, our proposed solution aims at redistributing locally the load of a heavily loaded BS around its neighboring overlapped BSs. In turn, the neighboring BS will redistribute its load to its own neighboring BSs and so on. By doing so, the load of the overall system will be then balanced. The handover enforcement will be triggered when the load of a specific BS is greater than δ. The algorithm execution is continued

until $\xi_2 = 0$ or we cannot find a handover to improve index ξ_2. Also, the decision to make M_i handover to BS_j only relies on the load comparison between BS_0 and BS_j. From (7.11), (i, j) is selected if

$$(\rho_0 - w_{i0} - \delta)^+ + (\rho_j + w_{ij} - \delta)^+ < (\rho_0 - \delta)^+ + (\rho_j - \delta)^+ \tag{7.12}$$

A ping-pong effect where M_i is decided to handover back to BS_0 could not happen due to constraint (7.12). Therefore, our algorithm is ensured to converge.

7.3.3 Performance evaluation

In this section, we first show the effectiveness of our new load balance index ξ_2 which is used as an objective function in our proposed load balancing scheme. Next, the performance of our proposed solution is compared with the optimal solution and a reference scheme. The chosen reference solution employs an *advanced* admission control [180] [181] [188], in which a new incoming communication will be redirected to the least loaded BS. This smallest load value includes the load of the new incoming communication.

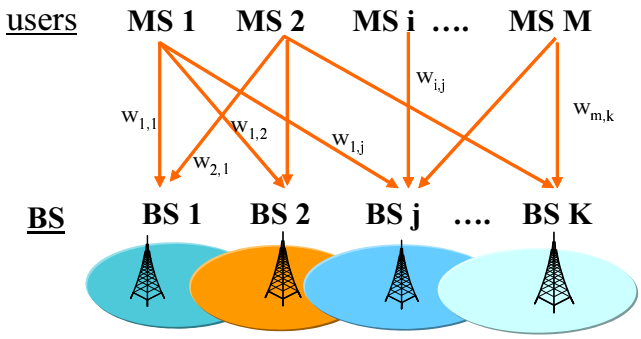

Figure 7.5: Simulation scenario

We consider a simulation scenario in which users activate and deactivate dynamically their communication sessions. Each communication is associated with a guaranteed bandwidth η which is randomly generated in the interval $\eta \in [100, 1200] Kbps$. Assume that a user has only one communication session at a time and the duration of each communication follows an exponential distribution with a selected averaged value of 5 minutes. A user has the possibility to connect to a random number of BSs (as illustrated in Figure 7.5). As we focus on the load balancing operation, the simulation of the physical and MAC layers is not necessary in order to observe the load balancing performance. Therefore, the radio link quality between a user and its reachable BSs (i.e., the modulation and coding rates) are also randomly selected at the beginning of each communication session. The modulation and coding rate R_{mc} varies from 0 (i.e., radio link is very poor for the connection or user is outside the BS's radio coverage) to 4 $bits/symbol$. The capacity of each BS is randomly selected in the interval $[1, 10]$ $Msymbol/s$.

The performance is evaluated by means of a non-satisfaction index (NSI) which is the ratio of the total number of users that are currently connected to the overloaded BSs to the total number of users in the system. The use of this metric is motivated by the fact that if a BS is overloaded, no matter how it schedules its users, it will not guarantee the required QoS for all its served users.

7.3.3.1 Validation of the load balancing index ξ_2

We employ indexes ξ_1 and ξ_2 as load-balancing objectives. Another strategy that consists in mini-

mizing the total load of all BSs is also examined. The performance of the three strategies is illustrated in Figure 7.6. In this simulation, the number of BSs in the system is fixed at 10. The value of threshold δ here is selected as $\delta = 0.8$. Note further that when we change the number of BSs or users in the system, the whole system configuration (e.g., R_{mc}, BS's capacity, η) is modified. The comparison between *NSI* of these different network configurations is not much relevant. So we do not discuss about the fluctuation of *NSI* between different network configurations. Note also that we keep the same initial network configuration to test the different load balancing indexes.

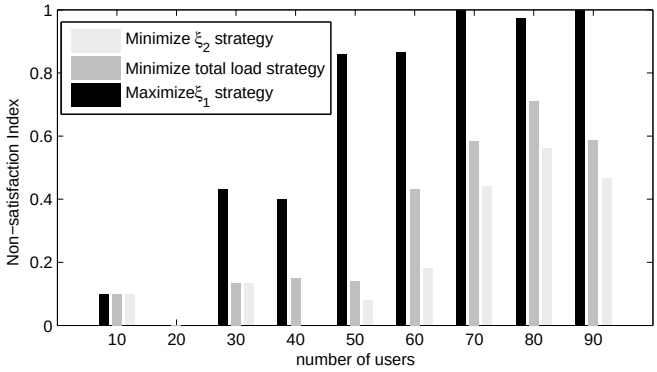

Figure 7.6: Non-satisfaction index vs. load balancing objective function strategies

From Figure 7.6, we observe that the ξ_2-based strategy gives the best performance compared to the two other strategies in any simulated network-load contexts. The positive values of *NSI* for a small number of users are explained by the fact that the users connected to the overloaded BS are outside any overlapped BS. The three strategies give approximately the same results for a small number of users. When the number of users increases (i.e., $M \geq 30$), the ξ_1 strategy exposes clearly its limitation. The ξ_1 strategy is not suitable since an equalization of all BSs' load does not lead to a good system performance. Also, minimizing the total load does not result in an efficient resource utilization either because minimizing the total load does not mean a minimization of the system overload level. One can see that the performance of the total load minimization strategy is still better than the ξ_1-based strategy. The results confirm the efficiency of the ξ_2-based load balancing strategy.

7.3.3.2 Performance of the proposed load balancing strategy

We compare the performance of our proposed scheme with the impractical optimal solution (using only the first optimization objective). As the optimal solution requires a great computation time, the number of users arriving at a time is limited to 15 and a small number of BSs is considered. However, each user requires a high η ($600 \leq \eta \leq 1200$) to introduce a high load in the system. According to Figure 7.7, our proposed algorithm performs very well compared to the optimal one. Indeed, the balance indexes ξ_2 given by our solution are very close to those of the optimal one. More interestingly, the *NSI* of the two solutions is lightly different (18% for the optimal one vs. 20% for our scheme in the case of 4 BSs) for heavily loaded scenarios (small number of BSs) and mostly the same for lightly loaded scenarios.

We compare now our proposed scheme with the reference solution [180, 181, 188] in which a new communication will be redirected to the least loaded BS computed including the load of the incoming communication. In practice, we start two separate simulations using the same initial load-balanced system, the same user arrival process and the same user's running application scenario. The number of BSs is set to 10 and the number of users is set to 20. The load variation of the system is due to the

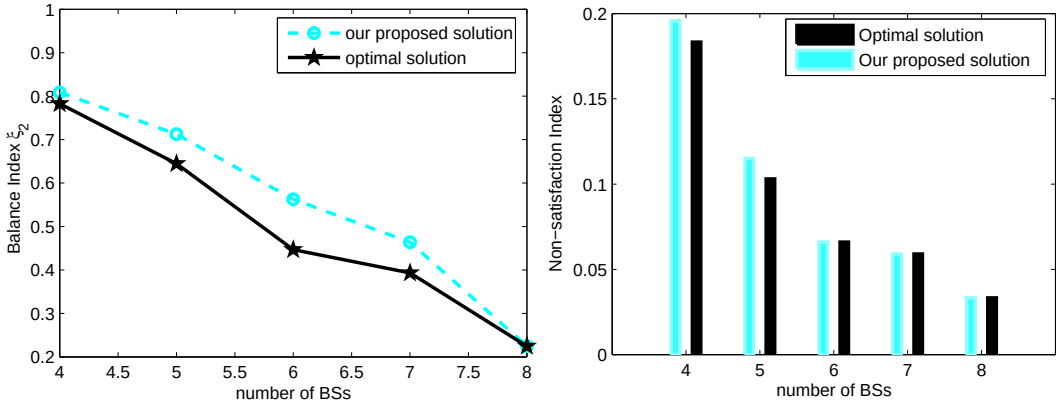

Figure 7.7: Performance comparison between our solution and the optimal one

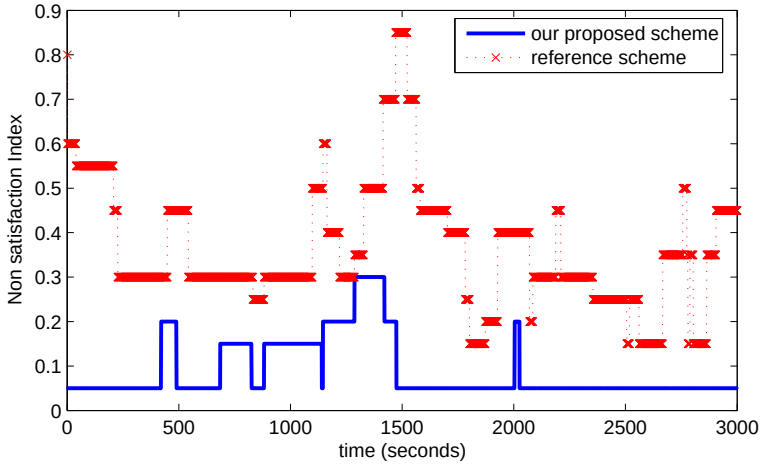

Figure 7.8: Performance comparison between our solution and the reference one (using advanced admission control)

communications initiation/termination. The NSI of the two systems, the one managed by our proposed load balancing scheme and the one managed by the reference scheme, is observed at every instant and is depicted in Figure 7.8. We observe that the ratio of non-satisfied users in the system managed by our proposed scheme is smaller than in the system managed by the reference one. In fact, our proposed scheme uses a simple admission control compared to the advanced admission control of the reference one. The key of our scheme is based on the handover enforcement process that handles imminent overload situations. The results show clearly the effectiveness of our solution which is furthermore feasible for implementation in both homogeneous and coordinated heterogeneous networks.

7.4 Summary

This chapter treated another aspect of the mobility management over heterogeneous networks where the role of the network control is vital: load balancing. We defined a new load metric which makes it possible to formulate the load balancing as a classic optimization problem. This novel load metric for wireless packet networks is based on the packet scheduling and the radio link quality information.

Thank to this new metric, the heterogeneity of different access technologies can be removed. It also facilitates the load balancing operations since it allows load variation anticipation. We introduced a new load balancing index to measure the overload degree of a system. This balancing index leads to minimize the overload degree of a system instead of equalizing the load among the access nodes within a system. We designed a load balancing scheme which consists of an admission control and a handover enforcement. The proposed iterative algorithm is one of the feasible suboptimal solutions to the problem. The solution can be used in on-line system because it requires less computation time and because it operates in a distributed way instead of a usual centralized way. However, this iterative search may result in a local optimum solution. To overcome this limitation, some advanced optimization techniques like Tabu search [196], ant colony optimization [197], particle swarm optimization [198], genetic algorithm, evolutionary optimization [199]...may be employed. The use of these techniques are left for future works.

∼∼ △♡△ ∼∼

Conclusions

As foreseen by many researchers and analysts, the next generation wireless mobile communications (4G) will be based on the heterogeneous underlying infrastructure integrating different wireless access technologies in a complementary manner. Looking at the portfolio of mobile and fixed operators today, a dominance of access bundles and flat rates, network convergence has become a means to solely reduce operational cost and maintain competitiveness in a market of flattened subscriber growth and decreasing per-user revenues. Future mobile users need to enjoy seamless mobility and ubiquitous access to services in an *always best connected* mode. In this context, the inter-system mobility management is an important and challenging technical issue to be solved. The underlying idea of this thesis was to optimize the inter-system mobility management and find out solutions to ensure seamless handovers across a wide range of networks and devices.

The contribution of the thesis was divided into two parts: the handover enhancements controlled by the user terminal side (user-controlled approach) and those controlled by the network (network-controlled approach).

The end-users are not anymore passive - they are starting to design their own services and sharing them in communities. They are increasingly demanding higher quality and more personalized services. The best way to offer customized services in the context of mobility management is to allow users to influence and even to control the access network choice, the handover decision and the handover preparation.

As the terminal can access to information on its inner capabilities, surrounding networks, active sessions and the user preferences, it is in a much better position to select the best access network than the network side. This real-time information can be gathered by the terminal but it is not available to the service platform or network entities. Transferring all gathered information to the network for the handover decision implies the frequent update from every user towards the network, which is not feasible. In the first part of this thesis, we presented a framework where the mobility management is considered as a third party service. The user terminal can select the best access network among the available ones and control a holistic vertical handover procedure. Based on the gathering information, the terminal determines the adaptive handover threshold in order to ensure seamless handover. Furthermore, the terminal can predict the handover and initiate some handover preparation techniques like pre-buffering to assure the seamless user experience during handovers. Last but not least, the terminal-controlled mobility management makes it possible to optimize the power consumption of multiple-radio interface devices.

In the gathering information phase, it is preferable that the mobile terminal has some provisioning information elements about its neighboring cells to be able to quickly discover, synchronize and measure their signal strength. The terminal can detect neighboring cells by naively scanning all the channels, which may take a long latency. Such a *blind* scanning can be possible if the terminal is equipped with multiple radio interfaces. But, in the case of single reconfigurable radio interface device, the long scanning latency implies a large overlap between adjacent cells, a high power consumption and probably

large handover interruption time. In addition to the access network discovery and the handover measurement, the transfer of user's mobility and security contexts from the serving network to the target network to reduce the handover latency requires control from the network. Similarly, the QoS guarantee, radio resource allocation and load balancing are subjects to be controlled by the network. In fact, only the network has the entire information about the traffic load in the access nodes (i.e., number of users currently connected to this access node) and its available radio resources.

In the two main parts of the thesis, we have pointed out the advantages and the capabilities to ensure seamless handovers of both user control and network control authorities. The vertical handover procedure can be completed under the full control of the network or the user terminal. However, to effectively manage the user mobility in heterogeneous networks, the control from both user terminal and network should be jointly coordinated. One possible question is that what happens if both the mobile terminal and the network want to control the handover procedure, for example, the handover decision. If the terminal decides to handover to a specific access node, the network can either help terminal to prepare the handover or not. In the latter case, the terminal prepares and executes the handover itself using a third party mobility service. A seamless handover can be possible if the terminal is equipped with multiple radio interfaces and it is multi-homed. Otherwise, a seamless handover seems to be difficult to achieve. Now, if the network forces the terminal to handover to an indicated access node, the terminal can either obey this recommendation or not. In the latter case, the terminal can consider this command as an access network selection triggering condition. The terminal starts to select another access node and then makes the handover to its selected access node using terminal-controlled handover scheme. In short, the coordination of the decisions should reflect the reasonable and dynamic compromise between user autonomy and network control. The network can help and guide the user terminal in the choice of the suitable access network and in the information gathering. The user terminal is responsible for selecting the best access network according to user preferences, managing the radio interface activation mode, handover preparation (for example, pre-buffering streaming data) and triggering the handover when appropriate. The network is responsible for authentication, roaming establishment if necessary, handover preparation (for example, mobility and security context transfer), QoS maintenance and load balancing-based handover management.

Contributions of this thesis

Summing up, the major contributions of this work include:

- A novel utility-based access network selection which includes:
  - A new sigmoidal function form to best model the utility of each access network characteristic,
  - A new multiplicative aggregate utility function form,
  - A context-aware (user situation and application) user preferences configuration.
- A terminal-controlled handover management framework which includes:
  - A very-loose coupling interworking architecture,
  - A user-centric access network selection scheme,
  - A power-saving multiple radio interface management,
  - An adaptive handover initiation threshold to ensure seamless mobility.
- A terminal-controlled seamless media streaming scheme which includes:

  - A practical handover prediction scheme based received signal strength measurements using the Grey-Model filter and CUSUM-based movement detection techniques,
  - An adaptive pre-buffering policy to maintain the media content in the buffer for seamless streaming.

- A UMTS-WiMAX interworking solution which includes:
  - A loose-coupling interworking architecture between UMTS and WiMAX systems,
  - A scheme for inter-system measurement using an SDR-enabled terminal,
  - A required minimum cell overlap for seamless handover in a UMTS/WiMAX system.

- An intermediary platform for interworking and roaming between different access networks in a multi-operator environment.

- A novel load balancing scheme which includes:
  - A novel load metric definition,
  - A new load balancing index definition,
  - A practical load balancing scheme in heterogeneous wireless packet networks.

Future Work

Mobility management in heterogeneous networks is a complex problem comprising of a large number of challenging issues. Regarding the aspects addressed in this thesis, there are still many possible research areas that the future work may take.

In the access network selection, recently researchers have considered the use of Multi-Attribute Decision Making (MADM) algorithms [69, 70, 200]. Though the MADM is much related to the utility theory, further analysis and comparison between the utility-based and MADM approaches will be needed to figure out the best one or a possible coordination of the two approaches. An automatic user situation profile update using the information from terminal status, the currently connected networks, location servers, running services, the sensors integrated in the terminal device and etc. is an interesting research direction not only for access network selection but also many other application fields.

Another research topic much related to the mobility management and access network selection is the radio resource management (RRM). The load balancing over heterogeneous networks was addressed in this work, however further improvements using some advanced optimization techniques may be envisioned. In an open access heterogeneous networks, two independent operators battle to get user connections (by allocating an appropriate resource amount to each user) to maximize their resource utilization and their revenue knowing that users have liberty to select the access network of highest utility level (according to preferences of each user). The combination between the access network selection and the resource allocation in this context is a difficult problem that can be modeled and solved by the game theory.

The terminal-controlled mobility management framework and the inter-system handover coupled with the roaming establishment between two access network domains having no direct agreement can be considered as a beginning of more extensive studies, not the final answer. Inspired by the localized network-based management solution, a new network-based *inter-domain* mobility management may be envisioned to enhance the handover performance.

In heterogeneous networks, the users will be always best connected (ABC) through the best access network using the best available device. The network operator has to satisfy ABC property for users (or

aid users to get ABC property) and has to maintain the best possible utility out of its investment. From the network operators' perspective, the operators need an environment which fulfils the requirements for "Always Best Managed" (ABM) infrastructures, networks, and services. An intelligent network selection and an autonomous mobility management are required to achieve ABC and ABM simultaneously. The autonomous mobility management including autonomous handover, autonomous load balancing and autonomous location management is an enabler for ABM networks. The autonomous mobility management in heterogeneous networks becomes one of future research directions towards seamless mobility.

In the last few years, the limited available spectrum and the inefficiency in the spectrum usage necessitate a new communication paradigm to exploit the existing wireless spectrum opportunistically [201]. This new networking paradigm is referred to as Dynamic Spectrum Access (DSA) or cognitive radio networks. The dynamic spectrum access techniques allow cognitive radio users to select the *best available channel* for the communication. The users can switch from one spectrum hole to another to maintain the connectivity, which is known as *spectrum handover*. The purpose of spectrum mobility management, including best channel selection and spectrum handover, is to make sure that such transitions are seamless and as soon as possible such that the applications running on cognitive radio users perceive minimum performance degradation during a spectrum handover. One can see that the mobility management should be take into consideration spectrum handover operations to maintain the seamless service delivery. One interesting future research direction is to study the coordination between the spectrum mobility management and traditional mobility management in cognitive radio networks.

This dissertation presents the comprehensive and pragmatical improvements on different facets of the inter-system handover. This research will facilitate the evolution of seamless mobility of the next generation network.

∼∼ △♡△ ∼∼

Bibliography

[1] 3GPP, "3GPP System Architecture Evolution: report on technical options and Conclusions," Tech. Rep. TR23.882 v1.15.0, 2008.

[2] J. Strassner, "Seamless service mobility: How autonomic networking and communications will shape the future of next generation services," in *Proc. of the 1st international workshop on Seamless Service Mobility (SSMO)*, Marrakech, 2007.

[3] W. Webb, *Wireless Communications: The Future*. John Wiley & Sons, 2007.

[4] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*. NJ, USA: Prentice Hall PTR, 2007.

[5] "Defining 4G: Understanding the ITU Process for the Next Generation of Wireless Technology," White Paper, 3G Americas, June 2007.

[6] E. Bohlin, S. Lindmark, J. Bjrkdahl, A. Weber, B. Wingert, and P. Ballon, "The Future of Mobile Communications in the EU: Assessing the Potential of 4G," *IPTS Technical Report Prepared for the European Commission-Joint Research Centre. Seville. EUR*, vol. 21192, 2004.

[7] S. Frattasi, H. Fathi, F. Fitzek, R. Prasad, and M. Katz, "Defining 4G technology from the user's perspective," *Network, IEEE*, vol. 20, no. 1, pp. 35–41, 2006.

[8] "What Exactly is 4G? Sorting Out the 4G Stew," Anritsu News, Anritsu, Vol. 26 No.123, Spring 2007.

[9] M. Jaseemuddin, "An architecture for integrating UMTS and 802.11 WLAN networks," in *Proc. of 8th ISCC*.

[10] N. Vulic, S. Groot, and I. Niemegeers, "A comparison of interworking architectures for WLAN integration at UMTS radio access level," in *ConWIN05*, 2005.

[11] S.-L. Tsao and C. Lin, "Design and evaluation of UMTS-WLAN interworking stategies," in *Proc. of VTC 2002-Fall*, 2002, pp. 777–781.

[12] C. Omidyar, *Mobile and Wireless Communications*. Kluwer Academic Publishers, 2003.

[13] 3GPP, "Feasibility study on 3GPP system to WLAN interworking," Tech. Rep. TR22.934 v6.2.0, 2003.

[14] ——, "3GPP system to Wireless Local Area Network (WLAN) interworking; System description," TS 23.234 v7.2.0, 2006.

[15] ——, "Radio access network; generic access to the A/Gb interface; stage 2," Tech. Spec. TS43.318 v6.6.0, 2006.

[16] L. Gras, Q.-T. Nguyen-Vuong, Y. Ghamri-Doudane, N. Agoulmine, M. Kassar, B. Kervella, and G. Pujolle, "Terminal Mobility in Mobile and Wireless Realm: tight, loose vs. very loose coupling," in *Proc. of the 1st international workshop on Seamless Service Mobility (SSMO)*, Marrakech, 2007.

[17] V. Varma, S. Ramesh, K. Wong, M. Barton, G. Hayward, and J. Friedhoffer, "Mobility management in integrated UMTS/WLAN networks," in *Proc. of ICC'03*, Anchorage, USA, May 2003.

[18] C. Perkins, "IP mobility support for IPv4, IETF RFC 3344," 2002.

[19] D. Johnson, C. Perkins, and J. Arkko, "Mobility support in IPv6, IETF RFC 3775," 2004.

[20] N. Montavont and T. Noel, "Handover management for mobile nodes in IPv6 networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 38–43, 2002.

[21] S. Dixit, "Wireless IP and Its Challenges for the Heterogeneous Environment," *Wireless Personal Communications*, vol. 22, no. 2, pp. 261–273, 2002.

[22] R. Moskowitz and P. Nikander, "Host Identity Protocol (HIP) Architecture, IETF RFC 4423," May 2006.

[23] P. Eronen, "IKEv2 Mobility and Multihoming Protocol (MOBIKE), IETF RFC 4555," June 2006.

[24] T. Kivinen and H. Tschofenig, "Design of the IKEv2 Mobility and Multihoming (MOBIKE) Protocol, IETF RFC 4621," August 2006.

[25] C. Kaufman, "Internet Key Exchange (IKEv2) Protocol, IETF RFC 4306," Dec. 2005.

[26] E. Fogelstroem, A. Jonsson, and C. Perkins, "Mobile IPv4 Regional Registration, IETF RFC 4857," 2007.

[27] H. Soliman, C. Castelluccia, K. E. Malki, and L. Bellier, "Hierarchical mobile IPv6 mobility management (HMIPv6), IETF RFC 4140," 2005.

[28] R. Koodli, "Fast handovers for Mobile IPv6, IETF RFC 0468," 2005.

[29] A. Misra, S. Das, A. Dutta, A. McAuley, and S. K. Das, "Idmp-based fast handoffs and paging in ip-based 4g mobile networks," in *IEEE Communication Magazine*.

[30] A. Campbell, J. Gomez, S. Kim, A. Valko, W. Chieh-Yih, and Z. Turanyi, "Design, implementation, and evaluation of cellular ip," in *IEEE Personal Communications*, vol. 7.

[31] R. Ramjee, K. Varadhan, L. Salgarelli, S. Thuel, S. Wang, and T. L. Porta, "Hawaii: A domain -based approach for supporting mobility in wide-area wireless network," in *IEEE IEEE/ACM Trans. Networking*, vol. 10.

[32] E. J. Kempf, "Goals for Network-Based Localized Mobility Management (NETLMM), IETF RFC 4831," April 2007.

[33] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, "Proxy Mobile IPv6, Internet Draft, draft-ietf-netlmm-proxymip6-10, NETLMM WG," Feb. 2008.

[34] A. Snoeren and H. Balakrishnan, "An end-to-end approach to host mobility," in *Proceedings of the 6th annual international conference on Mobile computing and networking*. ACM Press New York, NY, USA, 2000, pp. 155–166.

[35] S. Koh and Q. Xie, "Mobile SCTP (mSCTP) for Internet Mobility," IETF Internet Draft, draft-riegel-tuexen-mobile-sctp-09.txt, Nov. 2007, work in progress, Tech. Rep.

[36] R. Stewart, Q. Xie, M. Tuexen, S. Maruyama, and M. Kozuka, "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration," RFC 5061, Tech. Rep., September 2007.

[37] IETF, "SIP: Session Initiation Protocol," Tech. Rep., June 2002.

[38] A. Dutta, S. Madhani, W. Chen, O. Altintas, and H. Schulzrinne, "Fast-handoff schemes for application layer mobility management," in *Proc. of 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications PIMRC*, vol. 3, 2004.

[39] H. Yokota, A. Idoue, T. Hasegawa, and T. Kato, "Link layer assisted mobile IP fast handoff method over wireless LAN networks," in *Proceedings of the 8th annual international conference on Mobile computing and networking*. ACM Press New York, NY, USA, 2002, pp. 131–139.

[40] R. Hsieh, Z. Zhou, and A. Seneviratne, "S-MIP: a seamless handoff architecture for mobile IP," in *Proc. of 22rd Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM 2003*, vol. 3, 2003.

[41] C. Politis, K. Chew, and R. Tafazolli, "Multilayer mobility management for all-IP networks: pure SIP vs. hybrid SIP/mobile IP," in *The 57th IEEE Semiannual Vehicular Technology Conference*, vol. 4, 2003.

[42] Q. Wang, M. Abu-Rgheff, and A. Akram, "Design and evaluation of an integrated mobile IP and SIP framework for advanced handoff management," in *IEEE International Conference on Communications*, vol. 7, 2004.

[43] Q. Wang and M. Abu-Rgheff, "A multi-layer mobility management architecture using cross-layer signalling interactions," in *Proc. of the 5th European Personal Mobile Communications Conference*, 2003, pp. 237–241.

[44] H. J. Wang, R. H. Katz, and J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks," in *IEEE Workshop on Mobile Computing Systems and Applications '99*, Louisiana.

[45] L. Chen, T. Sun, B. Chen, V. Rajendran, and M. Gerla, "A smart decision model for vertical handoff," in *Proc. of 4th ANWIRE*, Athens, Greece, 2004.

[46] E. Adamopoulou, K. Demestichas, A. Koutsorodi, and M. Theologou, "Intelligent Access Network Selection in Heterogeneous Networks-Simulation Results," in *Proc. of 2nd International Symposium on Wireless Communication Systems*, 2005, pp. 279–283.

[47] O. Ormond, G. Muntean, and J. Murphy, "Utility-based intelligent network selection in beyond 3G systems," in *Proc. of ICC*, Turkey, 2006.

[48] ——, "Economic model for cost effective network selection strategy in service oriented heterogeneous wireless network environment," in *Proc. of NOMS'06*, Canada, 2006.

[49] J. Chen, K. Yu, Y. Ji, and P. Zhang, "Non-cooperative distributed network resource allocation in heterogeneous wireless data networks," in *Proc. of IST Mobile Summit 06*, Greece.

[50] H. Chan, P. Fan, and Z. Cao, "A utility-based network selection scheme for multiple services in heterogeneous networks," in *Proc. of Intl Conf. on Wireless Networks, Communications and Mobile Computing*.

[51] W. Shen and Q. Zeng, "Cost-Function-Based Network Selection Strategy in Integrated Wireless and Mobile Networks," *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops-Volume 02*, pp. 314–319, 2007.

[52] V. Gazis, N. Houssos, N. Alonistioti, and L. Merakos, "On the complexity of "Always Best Connected" in 4G mobile networks," in *Proc. of IEEE Vehicular Technology Conference (VTC)*, Oct. 2003.

[53] J. Kleijnen, "Scoring methods, multiple criteria and utility analysis," *ACM Sigmetrics Performance Evaluation Review*, pp. 45–56, 1980.

[54] N. N. Ahmed Hasswa and H. Hassanein, "Generic vertical handoff decision function for heterogeneous wireless networks," in *IFIP Conference on Wireless and Optical Communications*, Dubai, 2005, pp. 239–243.

[55] O. Ormond, G. Muntean, and J. Murphy, "Network selection decision in wireless heterogeneous networks," in *Proc. of IEEE 16th Intl Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, 2005.

[56] M. Xiao, N. Shroff, and E. Chong, "A utility-based power-control scheme in wireless cellular systems," *IEEE/ACM Transactions on Networking,*, vol. 11, no. 2, pp. 210–221, 2003.

[57] H. Lin, M. Chatterjee, S. Das, and K. Basu, "ARC: an integrated admission and rate control framework for competitive wireless CDMA data networks using noncooperative games," *IEEE Transactions on Mobile Computing*, vol. 4, no. 3, pp. 243–258, 2005.

[58] L. Badia and M. Zorzi, "On utility-based radio resource management with and without service guarantees," in *Proc. of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*. ACM Press New York, 2004, pp. 244–251.

[59] X. Duan, Z. Niu, and J. Zheng, "Utility optimization and fairness guarantees for multimedia traffic in the downlink of DS-CDMA systems," in *Proc. of IEEE Global Telecommunications Conference*, vol. 2, 2003.

[60] L. Badia, M. Lindström, J. Zander, and M. Zorzi, "An economic model for the radio resource management in multimedia wireless systems," *Computer Communications*, vol. 27, no. 11, pp. 1056–1064, 2004.

[61] P. Fishburn, *Utility theory for decision making*. Wiley New York, 1970.

[62] L. Badia, M. Lindstrom, J. Zander, and M. Zorzi, "Demand and pricing effects on the radio resource allocation of multimedia communication systems," in *IEEE Global Telecommunications Conference GLOBECOM'03*, vol. 7, 2003.

[63] L. Badia and M. Zorzi, "An analysis of multimedia services in next generation communication systems with QoS and revenue management," in *IEEE 59th Vehicular Technology Conference,*, vol. 4, 2004.

[64] S. Elayoubi, L. Salahaldin, and T. Chahed, "Admission control-based pricing in UMTS," in *Proc. of International Conference on Information and Communication Technologies: From Theory to Applications*, 2004, pp. 207–208.

[65] X. Bai, L. Bölöni, D. Marinescu, H. J. Siegel, R. Daley, and I.-J. Wang, "Are Utility, Price, and Satisfaction Based Resource Allocation Models Suitable for Large-Scale Distributed Systems?" in *Proceedings of the 3rd International Workshop on Grid Economics and Business Models (GECON 2006)*, Singapore, May 2006, pp. 113–122.

[66] L. Badia, C. Taddia, G. Mazzini, and M. Zorzi, "Multi-radio Resource Allocation Strategies for Heterogeneous Wireless Networks," in *Proceedings of WPMC*, vol. 5, 2005.

[67] L. Giupponi, R. Agusti, J. Perez-Romero, and O. Sallent, "WLC05-2: An Economic-Driven Joint Radio Resource Management with User Profile Differentiation in a Beyond 3G Cognitive Network," in *IEEE Global Telecommunications Conference GLOBECOM*.

[68] T. Saaty, *What is the analytic hierarchy process?* Springer-Verlag New York, USA, 1988.

[69] Q. Song and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," *IEEE Wireless Communications*, vol. 12, no. 3, pp. 42–48, 2005.

[70] ——, "A network selection mechanism for next generation networks," in *Proc. of IEEE ICC*, Korea.

[71] J. Dyer, P. Fishburn, R. Steuer, J. Wallenius, and S. Zionts, "Multiple Criteria Decision Making, Multiattribute Utility Theory: The Next Ten Years," *Management Science*, vol. 38, no. 5, pp. 645–654, 1992.

[72] R. Winkler, "Decision Modeling and Rational Choice: AHP and Utility Theory," *Management Science*, vol. 36, no. 3, pp. 247–248, 1990.

[73] V. N. J. and O. Morgenstern, "Theory of Games and Economics Behavior," *Princeton Univ. Press, Princeton, N.J.*, 1953.

[74] S. Berry, J. Levinsohn, and A. Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, vol. 63, no. 4, pp. 841–890, 1995.

[75] S. Shenker, "Fundamental design issues for the future internet," *IEEE Journal on Selected Areas in Communication*, vol. 13, no. 7, Sept. 1995.

[76] L. A. Dasilva, "Pricing for QoS-Enabled networks: a survey," *IEEE Communications Surveys*, 2000.

[77] L. Giupponi, R. Agusti, J. Pérez-Romero, and O. Sallent, "Towards Balancing User Satisfaction and Operator Revenue in Beyond 3G Cognitive Networks," in *Proc. of IST Mobile Summit 06*, Greece, 2006.

[78] S. Pal, S. Das, and M. Chatterjee, "User-Satisfaction based Differentiated Services for Wireless Data Networks," in *Proc. of IEEE International Conference on Communications (ICC)*, vol. 2, 2005, pp. 1174–1178.

[79] 3GPP, "Quality of Service (QoS) concept and architecture (Release 6)," Tech. Spec. TS23.107 v6.3.0, 2005.

[80] Y. Yang, F. Mahon, M. Williams, and T. Pfeifer, "Context-Aware Dynamic Personalised Service Re-composition in a Pervasive Service Environment," in *The 3rd IFIP International Conference on Ubiquitous Intelligence and Computing, China*. Springer, 2006, pp. 3–6.

[81] N. Enderle and X. Lagrange, "User satisfaction models and scheduling algorithms for packet-switched services in UMTS," in *The 57th IEEE Semiannual Vehicular Technology Conference,*, vol. 3, 2003.

[82] X. Zhang, E. Zhou, R. Zhu, S. Liu, and W. Wang, "Adaptive multiuser radio resource allocation for OFDMA systems," vol. 6, 2005.

[83] S. Choudhury and J. Gibson, "Joint PHY/MAC based link adaptation for wireless LANs with multipath fading," in *Proc. of IEEE Wireless Communication and Networking Conference (WCNC)*, Las Vegas, 2006, pp. 757–762.

[84] M. Siebert and al., "Enhanced measurement procedures for vertical handover in heterogeneous wireless systems," in *Proc. of 14th IEEE on Personal, Indoor and Mobile Radio Communications*, vol. 1, 2003.

[85] L. Harju and J. Nurmi, "A baseband receiver architecture for UMTS-WLAN interworking applications," in *Proc. of 9th ISCC*, vol. 2, Egypt, 2004, pp. 678 – 685.

[86] J. Liu, K. Kazaura, and M. Matsumoto, "A low latency inter-system handover scheme for multiple interfaces terminal," in *Proceedings of the IEEE 6th Circuits and Systems Symposium on Emerging Technologies: Frontiers of Mobile and Wireless Communication, 2004*, vol. 2.

[87] L. Murphy, J. Noonan, P. Perry, and J. Murphy, "An application-quality-based mobility management scheme," in *Proc. 9th IFIP/IEEE International Conference on Mobile and Wireless Communications Networks (MWCN 2007)*, Cork, Ireland, 2007.

[88] S. Mohanty, "VEPSD: a novel velocity estimation algorithm for next-generation wireless systems," in *IEEE Transactions on Wireless Communications*, vol. 4, Nov. 2005, pp. 2655 – 2660.

[89] M. Turkboylari and G. Stuber, "Eigen-matrix pencil method-based velocity estimation for mobilecellular radio systems," in *IEEE International Conference on Communications (ICC)*, vol. 2, New Orleans, USA, 2000.

[90] G. Azemi, B. Senadji, and B. Boashash, "Velocity estimation in cellular systems based on the time-frequencycharacteristics of the received signal," in *Proc. of Sixth International Symposium on Signal Processing and its Applications*, vol. 2, Malaysia, 2001.

[91] J. How, N. Pohlman, and C. Park, "GPS Estimation Algorithms for Precise Velocity, Slip and Race-Track Position Measurements, Technical report, Society of Automotive Engineers (02MSEC-93)," 2002.

[92] A. Gafflaut and P. Mahonen, "Probabilistic Estimation of Achievable Maximum Throughput from Wireless Interface," Apr. 19 2007, wO Patent WO/2007/044,255.

[93] M. P. Michael, "Energy awareness for mobile devices," in *Researh Seminar on Energy Awareness, University of Helsinki*, Helsinki.

[94] M. Methfessel, F.-M. Krause, and K. Helmrich, "Research report on power consumption," WINDECT project, Tech. Rep., Feb. 2005.

[95] T. Melia, D. Corujo, A. de la Oliva, A. Vidal, R. Aguiar, and I. Soto, "Impact of heterogeneous network controlled handovers on multi-mode mobile device design," in *Proc. of IEEE Wireless Communication and Networking Conference*, Hong Kong, March 2007.

[96] P. Bernardin, M. Yee, and T. Ellis, "Estimating the range to the cell edge from signal strength-measurements," in *IEEE 47th Vehicular Technology Conference*, vol. 1, 1997.

[97] ——, "Cell radius inaccuracy: a new measure of coverage reliability," *Vehicular Technology, IEEE Transactions on*, vol. 47, no. 4, pp. 1215–1226, 1998.

[98] E. Shil, P. Bahl, and M. J. Sinclair, "Wake on wireless: An event driven energy saving strategy for battery operated devices," in *Proc. ACM Mobicom'02*, Atlanta, 2002, pp. 160 – 171.

[99] W.-S. Feng, L.-Y. Wu, Y.-B. Lin, and W.-E. Chen, "WGSN: WLAN-based GPRS support node with push mechanism," in *The Computer Journal*, 2004.

[100] S.-L. Tsao and E. cheng Cheng, "Reducing idle mode power consumption of cellular/VoWLAN dual mode mobiles," in *Proc. IEEE Globecom '05*, St Louis, MO, Dec. 2005, pp. 2902–2906.

[101] A. Yoon-Young and J. Junghoon, "Multi-radio power management consideration," presented at IEEE 802.21 session No 23, Nov. 2007.

[102] A. Zahran, B. Liang, and A. Saleh, "Signal threshold adaptation for vertical handoff in heterogeneous wireless networks," *Mobile Networks and Applications*, vol. 11, no. 4, pp. 625–640, 2006.

[103] A. Devlic and G. Jezic, "Location-aware information services using user profile matching," in *Proceedings of the 8th International Conference on Telecommunications, ConTEL 2005*, vol. 1, Croatia.

[104] 3GPP, "Selection procedure for choice of radio transmission technologies of the umts," Tech. Rep. TR 30.03U v.3.2.0, 1998.

[105] T. S. Rappaport, *Wireless Communications: Principles and Practice*. 2nd Edition, Prentice Hall, 2002.

[106] R. Wakikawa, K. Uehara, T. Ernst, and K. Nagami. (2005) Multiple CoA registration. Internet draft. [Online]. Available: draft-wakikawa-mobileip-multiplecoa-04.txt

[107] B. Lim, C.-W. Ng, and K. Aso. (November 2007) Verification of Care-of-Addresses in Multiple Bindings Registration. Internet draft. [Online]. Available: draft-lim-mext-multiple-coa-verify-00

[108] G. Tsirtsis, V. Park, and H. Soliman. (May 2007) Flow Movement for Mobile IPv4. Internet draft. [Online]. Available: draft-tsirtsis-mip4-flowmove-01

[109] Y.-W. Lin and T.-H. Huang, "Sip-based handoff in 4g mobile networks," in *Proc. of IEEE Wireless Communication and Networking Conference*, Hong Kong, March 2007, pp. 2806–2811.

[110] "WiMAX technology: LOS and NLOS environment," White Paper, SR Telecom, 2004.

[111] L. Phifer. (2005) Understanding WLAN signal strength. [Online]. Available: http://searchmobilecomputing.techtarget.com

[112] S. De-Gregorio, M. Budagavi, and J. Chaoui, "Bringing streaming video to wireless handheld devices," White paper, Texas Instruments, May 2002.

[113] Y. Birk and Y. Wiener, "A bucket-interleaving multiplexer for efficient near-on-demand streaming to resource-constrained clients," in *Proc. of IEEE International Conference on Multimedia and Expo ICME'02*, vol. 1, 2002.

[114] L. Cai and Y. Lu, "Energy management using buffer memory for streaming data," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 2, pp. 141–152, 2005.

[115] A. Ferreira, "Optimizing Microsoft Windows Media Services 9 Series," *Technical Brief, Microsoft Windows Digital Media Division, March*, 2005.

[116] W. Tan and A. Zakhor, "Real-time Internet video using error resilient scalable compression and TCP-friendly transport protocol," *IEEE Transactions on Multimedia*, vol. 1, no. 2, pp. 172–186, 1999.

[117] N. Aboobaker, D. Chanady, M. Gerla, and M. Sanadidi, "Streaming Media Congestion Control using Bandwidth Estimation," *Proc. of MMNS'02*, 2002.

[118] A. Balk, D. Maggiorini, M. Gerla, and M. Sanadidi, "Adaptive MPEG-4 Video Streaming with Bandwidth Estimation," in *Proc. of 2nd international workshop on QoS in multiservice IP Networks*, 2003.

[119] G. De Los Reyes, A. Reibman, S. Chang, and J. Chuang, "Error-resilient transcoding for video over wireless channels," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1063–1074, 2000.

[120] A. Reibman, "Optimizing multiple description video coders in a packet loss environment," *Packet Video Workshop*, 2002.

[121] C. Blondia, N. Van den Wijngaert, G. Willems, and O. Casals, "Performance Analysis of Optimized Smooth Handoff in Mobile IP," in *Proc. of 5th ACM Int. Workshop on Modeling analysis and simulation of wireless and mobile systems*, Atlanta, USA, 2002.

[122] M. Jain and C. Dovrolis, "Pathload: A measurement tool for end-to-end available bandwidth," in *Proc. of Passive and Active Measurements Workshop*, 2002.

[123] V. Ribeiro, R. Riedi, R. Baraniuk, J. Navratil, and L. Cottrell, "pathChirp: Efficient Available Bandwidth Estimation for Network Paths," in *Proc. of Passive and Active Measurement Workshop*, 2003.

[124] C. Dovrolis, P. Ramanathan, and D. Moore, "What do packet dispersion techniques measure?" in *Proc. of IEEE INFOCOM*, vol. 2, 2001.

[125] J. Strauss, D. Katabi, and F. Kaashoek, "A measurement study of available bandwidth estimation tools," in *Proc. of the 2003 ACM SIGCOMM conference on Internet measurement*, USA, 2003, pp. 39–44.

[126] B. Liang and Z. Haas, "Predictive Distance-Based Mobility Management for Multidimensional PCS Network," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, 2003.

[127] J. Francois, G. Leduc, and S. Martin, "Learning movement patterns in mobile networks: a generic method," in *Proc. of European Wireless*, 2004, pp. 128–134.

[128] H. Karimi and X. Liu, "A Predictive Location Model for Location-based Services," in *Proc. of ACM Int. Workshop Advances in Geographic Information Systems (GIS)*, 2003.

[129] H. Ebersnan and O. Tonguz, "Handoff Ordering using Signal Prediction Priority Queuing in Personal Communications Systems," in *IEEE Trans. Veh. Technology*, vol. 48, no. 1, 1999, pp. 20–35.

[130] S. Sheu and C. Wu, "Using Grey Prediction Theory to Reduce Handoff Overhead in Cellular Communication Systems," in *Proc. of IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications*, 2000.

[131] S. Rezaei and B. Khalaj, "Grey Prediction Based Handoff Algorithm," *Transactions on Engineering, Computing and Technology*, vol. 2, 2004.

[132] P. Bellavista, A. Corradi, and C. Giannelli, "Adaptive Buffering based on Handoff Prediction for Wireless Internet Continuous Services," in *Proc. of Int. Conf. on High Performance Computing and Communications (HPCC'05)*, 2005.

[133] D. Lee, C. Lee, and J. Kim, "Seamless media streaming over mobile IP-enabled wireless LAN," in *Proc. of IEEE Consumer Communications and Networking Conference (CCNC)*, 2005, pp. 116–121.

[134] D. Lee, J. Kim, and P. Sinha, "Handoff-aware Adaptive Media Streaming in Mobile IP Networks," in *Proc. of Int. Conf. on Communications*, France, 2004.

[135] P. Bellavista, A. Corradi, and C. Giannelli, "Evaluating filtering strategies for decentralized handover prediction in the wireless internet," in *Proc. of ISCC*, Los Alamitos, CA, USA, 2006, pp. 167–174.

[136] J. Doble, *Introduction to radio propagation for fixed and mobile communications*. Artech House, Boston, 1996.

[137] J. Deng, "Introduction to grey theory," *The Journal of Grey System*, vol. 1, no. 1.

[138] F. Gustafsson, *Adaptive Filtering and Change Detection*. Wiley New York, 2000.

[139] M. Khalil, J. El Falou, W. Duchêne, and D. Hewson, "Automatic threshold determination for a local approach of change detection in long term signal recordings," *EURASIP journal on Advances in Signal Processing*, 2007.

[140] M. Basseville and I. Nikoforov, *Detection of Abrupt changes: Theory and Application*. Prentice Hall, NJ, 1993.

[141] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *IEEE Electronics Letters*, vol. 27, pp. 2145–2146, Nov. 1991.

[142] ETSI, "Selection procedure for the choice of radio transmission technologies of the UMTS," Tech. Rep. TR 101.112 v3.2.0, April 1998.

[143] J. S. Bendat and A. G. Piersol, *Random data: Analysis and measurement procedures*. John Wiley & Sons, 1971.

[144] W. Forum, "WiMAX Forum Network Architecture Stage 2, Release 1, Version 1.2," January, 2008.

[145] "Wireless LAN roaming brokers and the role they play between WISPs and providers," White Paper, WeRoam, Jan. 2005.

[146] "Public WLAN hotspot deployment and interworking," Intel Technology Journal, Intel, 2003.

[147] B. Anton, B. Bullock, and J. Short. (2003) Best current practices for wireless internet service provider roaming. WiFi Alliance Document.

[148] I. F. Akyildiz, S. Mohanty, and J. Xie, "A ubiquitous mobile communication architecture for next generation heterogeneous wireless systems," in *IEEE Radio Communications*, June 2005, pp. 29–36.

[149] O. Zlydareva and C. Sacchi, "SDR application for implementing an integrated UMTS/WiMAX phy-layer architecture," in *Proc. of 3rd International Mobile Multimedia Communications Conf.*, Greece, 2007.

[150] B. Bing and N. Jayant, "A cellphone for all standards," *IEEE Spectrum*, vol. 39, pp. 34–39, May 2002.

[151] J. Glossner, D. Iancu, M. Moudgill, M. Schulte, and S. Vassiliadis, "Trends in Low Power Handset Software Defined Radio," *Lecture Notes in Computer Science*, vol. 4599, p. 313, 2007.

[152] D. Farinacci, T. Li, S. Hanks, D. Meyer, and P. Traina, "Generic routing encapsulation (GRE), IETF RFC 2784," 2000.

[153] G. Mommety, "Key and sequence number extensions to GRE, IETF RFC 2890," 2000.

[154] 3GPP, "Radio resource control (RRC)," TS 25.331 v6.6.0, 2005.

[155] *Part 16: Air Interface for fixed and mobile broadband wireless access system*, IEEE Std. 802.16, 2004.

[156] *Part 16 - Amendment 2:Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, IEEE Std. 802.16e, 2005.

[157] H. Wang and A. Prasad, "Security context transfer in vertical handover," in *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, PIMRC 2003*, vol. 3, China.

[158] V.-K. Gondi, N. Agoulmine, and R. Durand, "Deliverable 8.2: Integrated architecture detailed document (software and integration)," SEIMONET project, Tech. Rep., 2007.

[159] V. K. Gondi, E. Lehtihet, and N. Agoulmine, "Ontology-based network management in seamless roaming architectures," in *Proc. of 3rd IEEE international workshop on Broadband Converged Networks (BcN), in conjunction with NOMS 2008*, Brazil, 2008.

[160] H.-H. Choi, O. Song, and D.-H. Cho, "A seamless handoff scheme for UMTS-WLAN interworking," in *Proc. of IEEE Globecom '04*, vol. 3, Texas, Dec. 2004, pp. 1559–1564.

[161] J. Laiho, A. Wacker, and T. Novosad, *Radio Network Planning and Optimisation for UMTS*. John Wiley & Sons, NY, USA, 2002.

[162] M. D. Yacoub, *Foundations of mobile radio engineering*. CRC Press, Inc., 2000.

[163] J. Kim, D.-H. Kim, P.-J. Song, and S. Kim, "Design of Optimum Parameters for Handover Initiation in WCDMA," in *Proc. of IEEE 54th VTC*, USA, Oct. 2001, pp. 2768–2772.

[164] 3GPP, "Requirements for support of radio resource management (FDD)," Tech. Spec. TS25.133 v7.3.0, 2006.

[165] A. Murase, I. Symington, and E. Green, "Handover criterion for macro and microcellular systems," in *41st IEEE VTC'91*, St. Louis, May 1991, pp. 524–530.

[166] N. Zhang and J. Holtzman, "Analysis of handoff algorithms using both absolute and relative measurements," in *IEEE Transactions on Vehicular Technology*, vol. 45, Feb. 1996, pp. 174–179.

[167] M. Zonoozo, P. Dassanayaken, and M. Faulkner, "Optimum hysteresis level, signal averaging time and handover delay," in *47th IEEE Vehicular Technology Conference (VTC'97)*, Phoenix, 1997, pp. 310–313.

[168] M. Emmelmann, "Influence of velocity on the handover delay associated with a radio-signal-measurement-based handover decision," in *IEEE 62nd Semiannual VTC*, 2005.

[169] J. Korhonen, *Introduction to 3G Mobile Communications*. Norwood, MA, USA: Artech House, Inc., 2003.

[170] 3GPP, "Multiplexing and channel coding (FDD) R7," Tech. Spec. TS25.212 v7.0.0, 2006.

[171] ——, "Physical layer- Measurements (FDD) R7," Tech. Spec. TS25.215 v7.0.0, 2006.

[172] H. Holma and A. Toskala, *WCDMA for UMTS - Radio Access for Third Generation Mobile Communication*. 2nd Edition, Wiley, 2002.

[173] P. Bernardin, M. Yee, and T. Ellis, "Cell radius inaccuracy: a new measure of coverage reliability," *IEEE Trans. on Vehicular Technology*, vol. 47, pp. 1215–1226, Nov. 1998.

[174] T. Korkmaz, "Verifying Physical Presence of neighbors against Replay-based attacks in wireless networks," *Internation Journal of Information Technology*, vol. 11, June 2005.

[175] G. Siqueira, G. Ramos, and R. Vieira, "Propagation measurements of 3.5GHz signal: Path loss and variability studies," in *Proc. of the SBMO/IEEE MTT-S Int'l Microwave and Optoelectronics Conference*.

[176] 3GPP, "Improvement of RRM across RNS and RNS/BSS (Release 5)," Tech. Rep. TR25.881 v5.0.0, December, 2001.

[177] N. Vulic, S. Groot, and I. Niemegeers, "Common Radio Resource Management for WLAN-UMTS integration at Radio Access Level," in *Proc. of 14th IST Mobile Summit*, Germany, June, 2005.

[178] J. Pérez-Romero and al., "Common Radio Resource Management: Functional Models and Implementation Requirements," in *Proc. of IEEE 16th PIMRC*, Germany, 2005, pp. 2067–2071.

[179] R. Ferrus, A. Gelonch, O. Sallent, and J. Perez-Romero, "Vertical Handover Support in Coordinated Heterogeneous Radio Access Networks," in *Proc. of 14th IST Mobile Summit*, Germany, 2005.

[180] R. Agusti, O. Sallent, J. Pérez-Romero, and L. Giupponi, "A Fuzzy-Neural Based Approach for Joint Radio Resource Management in a Beyond 3G Framework," *1st International Conf. on Quality of Service in Heterogeneous Wired/Wireless Networks*, vol. 4, pp. 216–224, 2004.

[181] L. Giupponi, J. Agusti, J. Perez-Romero, and O. Sallent, "Joint Radio Resource Management algorithm for multi-RAT networks," in *Proc. of IEEE Globecom*, St Louis, MO, 2005, pp. 3851–3855.

[182] B. B. Chen and M. C. Chan, "Resource Management in Heterogeneous Wireless Networks with Overlapping Coverage," in *Proc. of 1st International conference on Communication System Software and Middleware*, India, 2006, pp. 1–10.

[183] P. Stuckmann, Z. Altman, H. Dubreil, *et al.*, "The EUREKA Gandalf project: monitoring and self-tuning techniques for heterogeneous radio access networks," in *IEEE 61st VTC 2005-Spring*, vol. 4, 2005.

[184] W. Shen and Q. Zeng, "Resource Allocation Schemes in Integrated Heterogeneous Wireless and Mobile Networks," *Journal of Networks*, vol. 2, no. 5, 2007.

[185] L. Giupponi, R. Agusti, J. Perez-Romero, and O. Sallent, "Improved Revenue and Radio Resource Usage through Inter-Operator Joint Radio Resource Management," in *Proc. of IEEE ICC'07*, 2007, pp. 5793–5800.

[186] D. Niyato and E. Hossain, "A Noncooperative Game-Theoretic Framework for Radio Resource Management in 4G Heterogeneous Wireless Access Networks," *Transactions on Mobile Computing*, vol. 7, no. 3, pp. 332 – 345, 2008.

[187] Y. Zhang, K. Zhang, Y. Ji, and P. Zhang, "Adaptive Threshold Joint Load Control in an End-to-end Reconfigurable System," in *Proc. of 15th IST Mobile & Wireless Communication Summit*, Greec, 2006.

[188] G. Bianchi and I. Tinnirello, "Improving Load Balancing mechanisms in Wireless Packet Networks," in *Proc. of IEEE ICC*, 2002, pp. 891–895.

[189] U. Bernhard, E. Jugl, J. Mueckenheim, H. Pampel, and M. Soellner, "Intelligent Management of Radio Resources in UMTS access networks," *in Bell Labs Technical Journal*, vol. 7, pp. 109–126, 2003.

[190] H. Velayos, V. Aleo, and G. Karlsson, "Load Balancing in Overlapping Wireless LAN Cells," *in proc. of IEEE ICC'04*, pp. 3833–3836, 2004.

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

[192] 3GPP, "High Speed Downlink Packet Access (HSDPA)-Stage 2 (Realease 7)," Tech. Spec. TS25.308 v7.1.0, December, 2006.

[193] *Part 11: Wireless LAN Medium Access Control and Physical Layer specifications: Higher Speed Physical Layer Extension in the 2.4 GHz band,* IEEE Std. 802.11, 1999.

[194] D. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, vol. 17, no. 1, pp. 1–14, 1989.

[195] R. Ramjee, R. Nagarajan, and D. F. Towsley, "On optimal call admission control in cellular networks," in *INFOCOM*, 1996, pp. 43–50.

[196] F. Glover, E. Taillard, and E. Taillard, "A user's guide to tabu search," *Annals of Operations Research*, vol. 41, no. 1, pp. 1–28, 1993.

[197] M. Dorigo and G. Di Caro, "The ant colony optimization meta-heuristic," *Mcgraw-Hill'S Advanced Topics In Computer Science Series*, pp. 11–32, 1999.

[198] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Neural Networks, 1995. Proceedings., IEEE International Conference on*, vol. 4, 1995.

[199] T. Bäck, *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press, USA, 1996.

[200] F. Bari and V. Leung, "Application of ELECTRE to Network Selection in A Hetereogeneous Wireless Network Environment," in *IEEE Wireless Communications and Networking Conference, WCNC 2007.*, 2007, pp. 3810–3815.

[201] I. Akyildiz, W. Lee, M. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, 2006.